UNIVERSIDADE DA BEIRA INTERIOR
Engenharia Informática

Adaptabilidade não Supervisionada Independente da Língua, ao Perfil Linguístico do Utilizador

Ramos Eduardo Pedro

Dissertação para obtenção de Grau de Mestre em
Engenharia Informática
(2ºciclo de estudos)
Versão final após defesa

Orientador: Prof. Doutor Sebastião Pais
Co-Orientador : Prof. Doutor João Paulo Cordeiro

Covilhã, Novembro de 2018

Dedicatória

Dedico este trabalho a Deus, o meu verdadeiro Mestre.
Pela conquista durante essa jornada e em toda minha vida esteve, sempre presente me dando saúde, sabedoria e muita força de vontade, a ele toda a minha gratidão.

Agradecimentos

Um trabalho desta natureza, jamais poderia ter sido concretizado sem a ajuda de várias pessoas que, direta e/ou indiretamente permitiram a sua conclusão. Assim, começo por agradecer a Deus Todo-Poderoso, por me ter dado saúde, vontade e força para chegar até aqui.
Agradeço aos meus pais e à minha querida esposa, por me terem apoiado e incentivado nesta etapa da minha formação.
As minhas filhas, pela ausência física, mas pela presença sempre! À minha família e aos meus irmãos, por tudo o que fizeram para me ajudarem a chegar a esta fase do meu crescimento.
Aos meus amigos, por estarem sempre lá, quando precisei.
Aos meus colegas do Departamento de Informática, pelo constante apoio que me concederam ao longo desta minha caminhada.
Ao meu Orientador, Professor Doutor Sebastião Pais, pelo seu constante incentivo, motivação, apoio e disponibilidade para me acompanhar ao longo desta minha formação.
A todos os meus professores, pela disponibilidade em partilhar os seus conhecimentos e saberes, pela sua disponibilidade e pelo apoio constante para que eu concretizasse a parte letiva deste Mestrado.
À Universidade da Beira Interior, pelas ótimas condições de trabalho que me foram disponibilizadas, pelo apoio e incentivo na realização das atividades académicas e sociais.

Prefacio

Ao longo do Mestrado em Engenharia Informática da Universidade da Beira Interior (UBI), foi apresentado um leque alargado de temas interessantes para a realização de diversas dissertações. Entre os temas expostos, cheguei a ponto de escolher o tema " Adaptabilidade não Supervisionada Independente da Língua ao Perfil Linguístico do Utilizador", proposto pelo Professor Doutor Sebastião Pais. A razão da minha escolha prendeu-se com a atualidade e extrema importância, quer para os engenheiros Informáticos, quer para a sociedade em geral, que a implementação de métodos estatísticos adaptáveis ao perfil linguístico do utilizador. Além disso, essas tecnologias abrangem um extenso campo de atuação, para que, na minha opinião, cada uma delas constitui, por si só, matéria suficiente para ser abordada noutras dissertações ou teses.

No entanto, tenho como forte convicção e modificar a nossa própria forma de viver e de interagir com o mundo. Desta forma ficarei satisfeito se a presente dissertação contribuir para uma melhor compreensão da estrutura da Adaptabilidade não Supervisionada Independente da Língua ao Perfil do Utilizador, bem como, constituir uma base de informação a todos os atuais e futuros investigadores.

Resumo

Vivemos atualmente num período de grande entusiasmo, devido ao exponencial do surgimento de diversas ferramentas associadas às normas e tecnologias agregadas às adaptabilidades não supervisionadas, independentes da língua do utilizador. Apesar disso, o uso dessas ferramentas não está acessível a todos os utilizadores, sendo que se assume como um obstáculo. Assumindo que o sucesso de qualquer tecnologia se encontra dependente da sua aceitação por parte dos seus utilizadores, a presente investigação pretende contribuir para uma exposição de conceitos, normas e tecnologias associadas às sociedades adaptáveis, com métodos e técnicas que permitem uma maior familiarização das interfaces adaptáveis, não supervisionadas ao perfil dos utilizadores.

Este trabalho propõe um estudo sobre a análise de uma adaptabilidade não supervisionada independente da língua não supervisionada para um perfil linguístico com uso de métodos e técnicas de processamento de linguagem natural (PLN) usando a similaridade assimétrica com o uso de técnicas de extração de textos e, posteriormente, avaliar as similaridades encontradas. Foram utilizados métodos linguísticos e técnicas uniformes, baseados no perfil linguístico do utilizador, enquanto medida comparada na variedade de línguas com base num vasto leque de línguas potencialmente heterogéneas variáveis, onde se focaliza especificamente uma característica metodológica baseada na adaptabilidade não supervisionada independente da língua do perfil linguístico do utilizador.

Devido à grande quantidade de informação disponível atualmente no meio eletrónico, uma das tarefas consiste na classificação adaptável do perfil de um utilizador e na extração de termos relevantes e não relevantes que tem vindo a ganhar importância nas pesquisas realizadas nas áreas de extração de termos e na recuperação de informações.

A média apresenta melhores resultados do que as combinações baseadas na estratégia de classificação de melhores resultados. O método proposto parece ser útil para desambiguar uma grande percentagem de consultas temporais para os utilizadores.

Palavras-chave: adaptabilidade, interface, homem-máquina, inteligente, adaptáveis, utilizador, *stopwords*, perfil

Abstract

We are currently living in a period of great enthusiasm, due to the exponential emergence of several tools associated with the norms and technologies added to the unsupervised adaptations, independent of the user's language. Despite this, the use of these tools is not accessible to all users, and is assumed to be an obstacle. Assuming that the success of any technology depends on its acceptance by its users, the present research intends to contribute to an exposition of concepts, norms and technologies associated to adaptive societies, with methods and techniques that allow a greater familiarization of adaptive interfaces , not supervised to the profile of the users.

his work proposes a study on the analysis of an unsupervised adaptability independent of the unsupervised language for a linguistic profile using natural language processing methods and techniques (PLN) using asymmetric similarity with the use of text extraction techniques and , later, to evaluate the similarities found. Uniform linguistic and technical methods were used, based on the user's linguistic profile, as a comparative measure in the variety of languages based on a wide range of potentially heterogeneous variable languages, where a methodological feature based specifically on unsupervised adaptability independent of the language of the profile language.

Due to the large amount of information currently available in the electronic environment, one of the tasks consists in adapting a user's profile and in extracting relevant and non-relevant terms that have gained importance in research in the areas of term extraction and in information retrieval.

The average has better results than the combinations based on the strategy of classification of better results. The proposed method seems to be useful for disambiguating a large percentage of temporal queries for users.

Keywords: adaptability, interface, man-machine, Smart, adaptive, user, stopwords, profile.

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Acrónimos

AAM - *Asymmetric Association Measures*.

AV-*Added Value*.

AIS - *Adapted Asymmetric InfoSimba Similarity*.

AIS - Similaridade Assimétrica.

BB - *Braunbanket*.

CD - Documento Comparável.

CF - *Certainty Factor*.

CO - *Conviction*.

DF - Frequência do Documento.

GI - *Gini Index*.

IDF - Frequência do Documento Inversa.

IE - *Information Extraction*.

IHC - Interação Humana-Computador.

IS - InfoSimba.

MIDP -*Mobile Information Device Profile*.

MT - Tradução Automática.

PC - Computador Pessoal.

PC- *Conditional Probabiliy*.

PLN - Processamento de Linguagem Natural.

SUM - *Summarization*.

TFISF - *Term Frequency-Inverse Sentence Frequency*.

UI - Utilização de Interface.

YAKE - Extração automática de palavras-chave.

Capítulo 1

Introdução

Neste capítulo apresentam-se as motivações para pesquisa e a descrição do problema abordado, os objetivos do trabalho e a estruturação do documento. Neste capítulo o principal objetivo é a apresentação dos elementos essenciais associados à investigação, pelo que procura situar o trabalho num determinado contexto histórico e social, procurando regatar a relevância que alguns elementos que intervém na problemática tratada assumem, bem como as motivações que presidiram e moveram para a sua concretização.

1.1 Enquadramento da Dissertação

A evolução dos equipamentos adaptáveis a um perfil linguístico e a crescente disponibilização de serviços adaptáveis, ainda inibidos por restrições comportamentais, é uma das principais tecnologias que motivou a presente análise. Por conseguinte, o presente trabalho procura fazer uma análise da adaptabilidade não supervisionada e Independente da língua de um perfil linguístico ao utilizador e as técnicas e metodologias de desenvolvimento das interfaces que tornam mais fácil e eficiente o seu uso.De facto, historicamente *Hettling* [61] foi o primeiro estudioso a propor as ideias gerais conceptuais de como se poderia construir uma interface com usabilidade [101]. Neste estudo, os princípios fundamentais das interfaces foram levados em consideração, mas Galitz [54] não expõe claramente um conjunto de regras que devem ser analisadas pelos envolvidos no processo de interface de uma adaptabilidade não supervisionada. Apesar de o tema da adaptabilidade não supervisionada, independentemente da língua do perfil linguístico do utilizador se ter intensificado com a preocupação de se criarem interfaces adaptáveis para uma aplicação melhor de um perfil linguístico em geral, nota-se que a literatura científica não conta com muitos estudos que vinculam as principais contribuições sobre as interfaces com a usabilidade para dispositivos móveis, tomando-se como base os princípios gerais que estão consolidados nessas interfaces [28].

Nos dias atuais existem uma série de mudanças no plano social, económico, político e cultural. Uma das causas dessas mudanças está relacionada com o rápido desenvolvimento tecnológico, principalmente após a segunda Guerra Mundial. Nesse sentido, entende-se que a tecnologia é uma ferramenta que proporciona ao homem muitas melhorias no seu dia-a-dia, visto que o homem está em constante evolução. A invenção e o crescente avanço tecnológico estão a modificar a compreensão do mundo, causando dessa forma, uma necessidade significativa de readaptação do modo de vida do homem.

Segundo Lobato [89] existem basicamente duas maneiras de um produto se adaptar ao utilizador durante a sua interação: a adaptabilidade não supervisionada e independente da língua do perfil linguístico do utilizador. Embora ambas reflitam a propriedade do produto alterar-se de acordo com o utilizador, utilizam diferentes estratégias para tal.

O desenvolvimento de interfaces para aplicações móveis tem vários desafios como a diversidade de dispositivos móveis, o ambiente heterogéneo, limitações físicas do aparelho, entre outros aspetos [14].Neste contexto, há uma necessidade de estudar e implementar interfaces

inteligentes com o objetivo de adaptar o seu desempenho, as suas funcionalidades e o seu conteúdo, às necessidades e preferências dos utilizadores, assim como personalizar a interação homem-computador baseada no modelo do mesmo. A interação entre o utilizador e o sistema ocorre através de interfaces que devem ser amigáveis e atraentes, para que não haja perda de interesse do assunto abordado.
A inteligência das interfaces deve fazer com que os sistemas se adaptem aos utilizadores, tirar as suas dúvidas, permitindo um diálogo entre utilizador e sistema, e/ou apresentar informações integradas e compreensíveis, utilizando vários modos de comunicação. O desafio da interface adaptativa não é simplesmente fornecer informações aos utilizadores sempre e em qualquer lugar, mas fornecer informações apropriadas para os utilizadores quando eles precisam.
Os sistemas adaptáveis conscientes do contexto ganharam popularidade nos últimos anos [114] sendo que a quantidade de investigações no campo da consciencialização do contexto está cada vez mais a crescer, considerando-se o contexto como uma propriedade de dispositivos móveis e potenciais desta propriedade para adaptação de interfaces e sistemas de utilizador.
A grande revolução provocada pela computação e a necessidade de satisfação do utilizador têm aumentado a dificuldade e a complexidade do desenvolvimento da interface Homem-máquina, principalmente para computadores de mão. Os estudos referentes ao contexto de interfaces para aparelhos móveis parecem, portanto, desconexos de algumas conclusões teóricas consolidadas cientificamente, pelo que este não vínculo poderia ter gerado estudos que não identificassem problemas de usabilidade de interfaces da forma que poderiam.

1.2 Objetivos

As interfaces de utilizador a adaptáveis fornecem benefícios potenciais para abordar os problemas de usabilidade. A adaptabilidade não supervisionada é identificada como um aspeto importante a ser considerado no desenho dos sistemas de informação modernos. Assim, as técnicas de adaptação incluem a adaptação das informações a apresentar, sabendo que a forma de apresentar esta informação à adaptação não supervisionada depende da língua e da interação com esta informação de adaptação de interfaces, bem como da adaptação do perfil linguístico de um utilizador.
As tecnologias de comunicação e de informação nos seus formatos móveis têm reconhecido, até ao momento, relativamente pouca atenção por parte das academias como objeto de estudo, muito embora as estatísticas indiquem que são as tecnologias mais rápidas de disseminação na história da humanidade, sendo que já são mais de sete biliões de telefones celulares no mundo, para uma população mundial de sete biliões de habitantes [125].
Deste modo, o trabalho aqui proposto consiste essencialmente na exploração de novas metodologias para a conceptualização e desenvolvimento de interfaces inteligentes adaptativas ao perfil linguístico de um utilizador, para que estas possam melhorar a eficácia e a eficiência dos estudos e abordagens já existentes.
A conceptualização e o desenvolvimento experimental de novas metodologias assentam numa adaptabilidade em tempo real não supervisionada e independente da língua, focando-se nas funcionalidades e no conteúdo de uma qualquer aplicação.
Em suma, o objetivo da presente proposta é conceptualizar e fazer um desenvolvimento experimental de novas metodologias não supervisionadas e independentes da língua, para o desenvolvimento de aplicações adaptativas ao utilizador, sendo que subsistem alguns objetivos mais específicos que passam pela identificação das caraterísticas do utilizador para o perfil linguís-

tico adaptável para facilitar o utilizador.
Também a criação de um novo modelo de utilizador agregado, ao modelo adaptativo não supervisionado e independente da língua para um perfil de utilizadores, irá permitir uma melhoria dos problemas de usabilidade dos utilizadores nas aplicações tecnológicas. As interfaces inteligentes visam superar problemas, devido à complexidade crescente da interação homem-máquina.
O desenvolvimento de uma interface baseada nestes conceitos deve obedecer a dois fatores fundamentais: ser amigável e facilitar o utilizador. Para que a interface seja considerada amigável tem que apresentar a facilidade de utilização e de aprendizagem. O utilizador deve interagir com a interface que é estudada pela Interação Humana-Computador (IHC), uma das interfaces adaptáveis de alta qualidade [62] .
Contar os pontos fortes e fracos na adaptabilidade não supervisionada independente da língua em aplicações móveis na melhoria das interfaces móveis, é um outro objetivo, que permite propor melhorias na adaptabilidade não supervisionada independente da língua através do perfil linguístico nas interfaces móveis. Por último, pretende-se descrever, de forma simples, o nível de desenvolvimento atual das principais tecnologias associadas à adaptabilidade independente da língua para interface.

1.3 Principais contribuições

Este trabalho apresenta uma contribuição quando se trata da adaptabilidade não supervisionada independente da língua em aplicações computacionais. De modo a concretizar o trabalho de investigação e desenvolvimento relatado nesta dissertação e do qual vão resultar algumas contribuições cientificas. A presente secção descreve, na opinião do autor, essas contribuições.
As interfaces adaptativas, conhecidas também como inteligentes, devem ser mais tolerantes a erros e oferecer formatos mais agradáveis provendo uma interação mais natural aos utilizadores. A interface inteligente vai fazer com que o sistema se adaptarem aos utilizadores para que tirem as dúvidas das mesmas, para permitir um diálogo entre o utilizador e o sistema. Algumas das maneiras como adaptação do utilizador pode ser útil envolvem suporte para os esforços de utilizador para operar um sistema com sucesso e efetivamente.

Uma das principais contribuições é adaptação da interface é levar em consideração deficiência percetivas ou físicas especiais de utilizador individuais de modo permitir que eles usem um sistema de forma mais eficiente com facilidade de seu uso. O ambiente também deve ser flexível o suficiente para acomodar as diferenças existentes nos perfis dos utilizadores, procurando minimizar o seu gasto de energia para se adaptar às tarefas, independentemente da língua.

1.4 Organização da Dissertação

O presente trabalho encontra-se estruturado em cinco capítulos que serão descritos em seguida. Assim, o capítulo 1 apresenta as definições dos elementos chaves das interfaces e adaptabilidade, perfil linguístico, prorrogado pela revisão das implementações existentes no desenvolvimento de sistemas adaptáveis no contexto do perfil do utilizador.
O capítulo 2 explora os trabalhos relacionados, especialmente focaliza-se no contexto da adaptabilidade não supervisionada e independente da língua em sistemas adaptados para o utilizador no uso de interfaces. É o capítulo dedicado ao estado da arte que corresponde ao estudo e defi-

nição da revisão bibliográfica existente para resolver de uma forma conveniente os problemas, apresentando-se alguns trabalhos contextualizados na literatura existente.
O capítulo 3 descreve o modelo das metodologias proposto nesta dissertação e a sua implementação e é descrita, bem como a metodologia utilizada, suas métricas e objetivos, analisando-se as abordagens das adaptabilidades de interfaces de um perfil linguístico do utilizador.
No capítulo 4 são apresentadas as novas medidas usadas, combinadas com diferentes medidas e os testes realizados com novas experiências e resultados.
O sistema das diretrizes projetado na prática é apresentado no capítulo 5, como prova de trabalho e eficiência da solução proposta, apresentando-se as principais conclusões e orientações para trabalhos futuros da adaptabilidade não supervisionada e independente da língua de um perfil do utilizador.

Capítulo 2

Estado da Arte

2.1 Introdução

Este capítulo apresenta uma base teórica para os elementos-chave do desenvolvimento da interface adaptativa do utilizador, enumera os motivos mais comuns para o uso de interfaces e, particularmente, concentra-se na implementação de sistemas móveis com exemplos de uso. O desenvolvimento de interfaces adaptativas é uma área bastante recente e tem vindo a ter um crescimento grande devido à popularidade e massificação dos perfis na adaptabilidade de interfaces para a criação de perfis. A evolução destes dispositivos tem sido grande, o que permite que esta área seja impulsionada.

2.2 Adaptabilidade de Interfaces

Muito embora a adaptabilidade pode ser realizada de múltiplas formas, em geral apenas dois tipos de caraterísticas de utilizadores que são tomadas em consideração como informação de base para adaptabilidade: a experiência do utilizador e o seu conhecimento relativamente aos princípios de utilização e de funcionamento do sistema. As preferências do utilizador dizem respeito ao estilo do diálogo, ao tipo de apresentação, aos objetos interativos, entre outros [67].Uma maneira diferente de ajudar uma pessoa a usar os sistemas mais eficazmente é adaptar a interface do utilizador de modo a que esta se encaixe melhor à forma de trabalhar com o sistema. Os elementos da interface que foram adaptados desta forma incluem menus, ícones e processamento do sistema de sinais de dispositivos de entrada, tais como os teclados [66].

A adaptabilidade pode fornecer benefícios potenciais para abordar os problemas de usabilidade da adaptação da interface, identificando-se como um dos aspetos mais importantes a ser considerado no desenho de sistemas de informação modernos ao nível das técnicas de adaptação.
A interface de utilizador adaptável é uma interface que adapta os seus elementos às necessidades de utilizadores individuais, sendo que um dos exemplos de interfaces adaptativas em dispositivos móveis é o navegador que indica uma posição atual do utilizador para que permita procurar o caminho para o seu destino, na barra de pesquisa expansível. A interface de utilizador adaptável tem vindo a assumir um papel importante na pesquisa de interação homem-máquina. Por conseguinte, só recentemente é que a pesquisa nesta área está a ganhar popularidade e maior interesse, graças principalmente à evolução tecnológica e ao crescimento da aplicação à complexidade e capacidade dos sistemas informáticos [78]. Além disso, os computadores portáteis e os dispositivos móveis são mais acessíveis para as pessoas e tornam-se, cada vez mais, parte da vida quotidiana, afigurando-se num novo desafio para as interfaces de utilizadores [87].

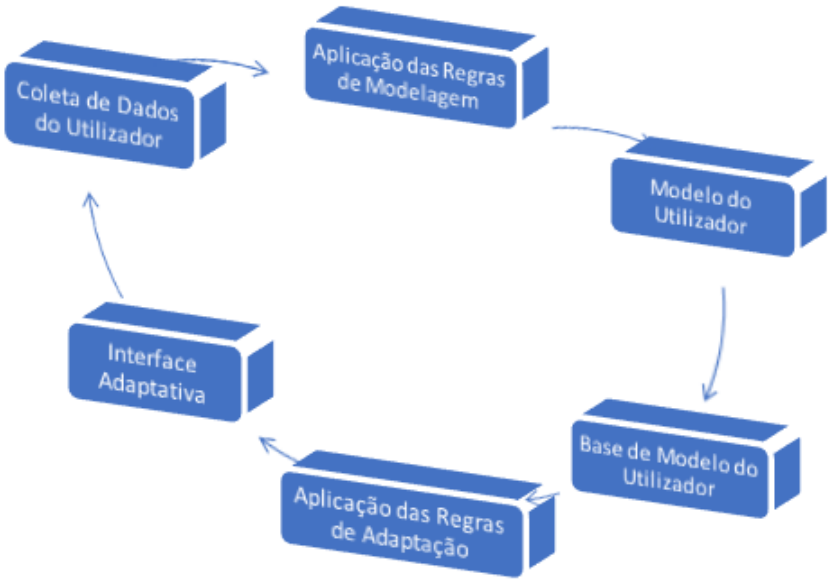


Figura 2.1: Figura nº1- adaptabilidade aprofundando (Fonte .Luciano Lobato)

2.2.1 Definição

A adaptabilidade consiste na propriedade que um produto tem em se adaptar ao utilizador, sem que este tenha que escolher as mudanças. O próprio sistema adapta-se de acordo com a "perceção" do utilizador, e não é o utilizador que solicita alterações. O sistema deve apresentar diferentes conteúdos para diferentes utilizadores, ou seja, adapta o conteúdo da interface de acordo com o modelo de utilizador. O sistema vai realizando alterações na sua arquitetura hipertextual (manipulando os links que estão disponíveis para o utilizador). Jameson descreveu o sistema adaptável aos utilizadores como um sistema interativo que adapta o seu comportamento a utilizadores individuais [4] com base em processos de aquisição e aplicação de modelos de utilizadores que envolvem algumas formas de aprendizagem, interferência ou tomada de decisão [65].
Consequentemente, a interface de utilizador adaptável é um tipo de sistema interativo, que adapta os seus elementos de interfaces e comportamento, sendo que tais sistemas são capazes de alterar as suas caraterísticas, configurações e elementos automaticamente, de acordo com as necessidades e objetivos de interação com o sistema. Assim, os sistemas adaptativos podem não só adaptar-se às necessidades do utilizador em determinado momento, mas também podem antecipar o futuro, etapas e requisitos [9].
O processo de adaptar um sistema de um perfil linguístico para as necessidades de um utilizador é chamado de adaptabilidade, e este aproveita o conhecimento adquirido do utilizador pela análise do seu comportamento e dos dados adquiridos. Em contrapartida, o processo de adaptabilidade é um processo iniciado e conduzido pelo utilizador, em que o mesmo fornece algum perfil através de um diálogo, no qual a personalização é iniciada e dirigida pelo sistema e, dessa forma, o sistema monitoriza o comportamento do utilizador a fim de realizar a adaptação automaticamente [85].
O utilizador pode configurar preferências nas suas interfaces e o sistema pode auto adaptar-se para melhor interagir com o utilizador, sem que o utilizador tenha que definir a ação; as interfaces adaptáveis podem também determinar o tipo de interface a apresentar ao utilizador, dependendo da análise do modelo do utilizador.

Nos sistemas adaptáveis, o utilizador pode fornecer algum tipo de perfil (através de um diálogo ou questionário), a partir da versão da aplicação do sistema [55] ajustando determinadas preferências da apresentação. Nos sistemas adaptativos, há uma adaptação automática, onde o sistema monitoriza o comportamento do utilizador e adapta a apresentação adequadamente [10].
A adaptabilidade não supervisionada refere-se, então, ao desenvolvimento do sistema não supervisionado aonde se permite uma utilização manual, mínima, aceitando a utilidade dos sistemas semi-supervisionados. Quando não existe um agente externo que indica a resposta desejada para os padrões de entrada na interface, este método é conhecido como forma de aprendizagem auto supervisionada ou de auto-organização pelo facto de não requerer uma saída desejada.

2.2.2 Interfaces para aplicações adaptativa

As interfaces adaptativas, conhecidas também como interfaces inteligentes, são interfaces capazes de se adaptarem a diferentes tipos de utilizador, fazendo com que os mesmos possam reorganizar os módulos apresentados na interface de forma mais agradável ao seu uso. Elas são artefactos que devem reconhecer os objetivos e metas dos utilizadores, bem como saber como atingi-los. Devem ser mais tolerantes a erros e oferecer formatos mais agradáveis, promovendo uma interação mais natural aos utilizadores, como também, empregar os recursos de inteligência artificial com o objetivo de facilitar o seu uso.
Utilizando-se interfaces adaptativas o sistema pode ser personalizado para estilos cognitivos individuais, necessidades de informações e tarefas personalizadas. As diferenças de cada utilizador que podem ser controladas, pelo projeto da interface passam pela personalidade, estilo cognitivo, estilo de aprendizagem e experiência [2].
O desenvolvimento de interfaces para aplicações envolve vários desafios, como a diversidade de dispositivos, o ambiente heterogéneo, as limitações físicas do aparelho entre outros aspetos [13].Neste contexto, torna-se imperativa a necessidade de estudo e implementação de interfaces inteligentes com o objetivo de adaptar o seu desempenho, as suas funcionalidades e o seu conteúdo, às necessidades e preferências dos utilizadores, assim como personalizar a interação homem-máquina baseada no modelo do mesmo.
A grande revolução provocada pela computação e a necessidade de satisfação do utilizador têm aumentado a dificuldade e a complexidade do desenvolvimento da interface homem-máquina, principalmente para computadores portáteis. O *design* móvel está em grande expansão, pelo que são lançados constantemente, no mercado, novos modelos [12].
Uma interface de utilizador adaptável pode ser definida como um artefacto de sistema que melhora a sua capacidade de interagir com um utilizador, construindo um modelo de utilizador baseado na experiência parcial com esse utilizador.
As interfaces de utilizador adaptáveis estão conscientes do seu contexto de uso e são capazes de fornecer uma resposta automática às mudanças neste contexto [64]. Tais respostas podem variar de um ajuste de *loyout*,simples para uma mudança na funcionalidade da interface do utilizador, de pois de observar literatura existente conseguimos diferenciar os seguintes tipos de solução de adaptação de interface do utilizador.
As interfaces de utilizadores adaptáveis permitem que as partes interessados sejam manualmente adaptáveis as características desejadas. Um exemplo simples de comportamento adaptável é um sistema móvel que deve suportar personalização manual, adicionando e removendo.

2.2.3 Interfaces de utilizador adaptável

Um dos módulos mais importantes em ambientes de aprendizagem nas aplicações é a interface de utilizador. A interface é a parte do sistema interativa responsável por traduzir ações do utilizador em ativações das funções do sistema, permitindo que os resultados possam ser observados, coordenando a interação entre o utilizador e o sistema [46].A qualidade da interface tem grande influência no sucesso de sistemas. A interface ajuda a navegabilidade, sendo este um fator de preocupação importante para o processo de *design*, já que muitos utilizadores apresentam problemas para navegar por determinadas interfaces computacionais [37].Com o propósito de melhorar a navegação pela interface por parte do utilizador, criaram os seguintes conjuntos que são recomendados [132]:

- Flexibilidade do acesso e uso das Informações.
- Minimização da carga de memória do utilizador.
- Compatibilidade entre a entrada de dados e o conteúdo exibido.
- Consistência dos dados sendo exibidos.
- Minimização do número de ações asserem realizados pelo utilizador.

A interface nada mais é do que a ligação/conexão/transferência de informações entre dois meios distintos. De acordo com Peter Brusilovsky, os sistemas adaptativos monitorizam o padrão de atividade do utilizador e, automaticamente, ajustam a interface ao conteúdo proveniente do sistema, para acomodar-se aos conhecimentos e preferências, pois os sistemas adaptáveis permitem ao utilizador controlar estes ajustes e promovem, frequentemente, guia ou ajuda especializada para utilizador [11].
[1] As interfaces adaptadas são concebidas para ambientes particulares, procurando ser mais adequadas possível a circunstâncias ambientais para a sua utilização. São desenvolvidas por evolução gradual, utilizando protótipos com os quais todos os futuros utilizadores são previamente postos em contacto. Da apreciação do comportamento destes, particularmente dos erros mais frequentes e fatores de desempenho, o protótipo é ajustado e, por iteração, é atingido o desenho final da interface a implementar, sendo que a adaptação é portanto estática [16].
As interfaces adaptáveis são possíveis de ajuste individual, em geral a cargo de cada utilizador. Nestes, vários níveis de estilos e objetos de diálogo são oferecidos, escolhendo cada utilizador o que, na sua opinião, mais se ajusta a si próprio, ou então, a interface é programável e partindo de um conjunto de base, permite ao utilizador construir o seu próprio ambiente, em geral usando comandos pré-programados ou "macros". Em qualquer um dos casos a adaptação é ainda a cargo do utilizador, sendo estática na primeira hipótese e semi-dinâmica na segunda.
As interfaces adaptativas são diferentes das anteriores porque nestas o utilizador não controla a adaptação. Esta é realizada de forma automática e dinâmica, isto é, durante a utilização efetiva do sistema, pela interface do utilizador. Sendo o objetivo fazer a interface aproximar-se o mais possível, em cada momento, das necessidades do utilizador, tal implica que a interface se deve basear num "conhecimento" que a cada momento deverá possuir sobre este, em particular sobre o tipo de utilização do sistema pelo mesmo feito. Assim, e contrariamente às anteriores, são "interfaces inteligentes", nas quais a adaptabilidade é determinada com base neste conhecimento e em regras que determinam a sua dinâmica, baseando-se particularmente no contexto

[1] Os aplicativos adaptativos devem possuir mecanismos que permitam a recolha de dados a respeito do estado de seu ambiente de execução, analisar esses dados visando identificar mudanças significativas e alterar dinamicamente seu comportamento para atingir seus objetivos.

atual e histórico da interação. Por serem "inteligentes" e dinâmicas são, conforme se referiu atrás, por vezes referidas como interfaces auto-adaptativas realçando tais caraterísticas*(" Self-Adaptive User Interfaces")*[2] A interface de utilizador adaptável é uma interface que adapta os seus elementos às necessidades individuais de utilizador. Um exemplo de interface adaptativa em dispositivo é o navegador que indica uma posição atual do utilizador.
As aplicações modernas têm que operar em diferentes dispositivos com diferentes recursos usados em diferentes contextos para atender a diferentes utilizadores, pelo que a ideia de uma interface adaptável ao utilizador destina-se a lidar automaticamente com distinções [8].Uma aplicação flexível ou adaptável pode interagir com utilizadores de forma personalizada e, além disso, resolver os problemas de diferenciação descritos acima.

2.3 Interfaces Inteligentes

Uma interface inteligente é considerada aquela que entende os objetivos e metas do utilizador e sabe atingi-los ou que facilita uma interação mais natural, com uma maior tolerância a erros e com formatos mais recompensadores e agradáveis. A interface é a arte e a ciência de fazer com que a máquina seja de uso fácil e intuitivo para as pessoas, em particular, pelo que para a construção de interfaces é preciso analisar o processo de interação entre a máquina e as pessoas que o utilizam, a partir dos paradigmas físicos, psicológicos, cognitivos e ergonómicos.
Mais concretamente, o que torna uma interface inteligente é esta poder adaptar-se às necessidades de diferentes utilizadores; poder aprender novos conceitos e técnicas; poder antecipar as necessidades do utilizador; poder tomar iniciativas e oferecer sugestões para o utilizador. Existem múltiplas áreas para a aplicação de interfaces inteligentes, sobretudo onde o conhecimento sobre como resolver parcialmente uma tarefa reside no sistema de computador. Até o início dos anos 90, a interface era um elemento secundário nas interfaces ou sistemas inteligentes [15] ,isto porque o seu desenvolvimento é bastante recente.
Segundo Bruillard a explosão de tecnologias usadas para a construção de interfaces resulta na necessidade de personalizar o seu uso, tornando o utilizador o elemento central para o projeto deste importante componente que permite o acesso aos sistemas de aprendizagem [33].Para se poder tornar numa interface inteligente é preciso que esta se possa adaptar às necessidades de diferentes utilizadores, aprender novos conceitos e técnicas, antecipar as necessidades do utilizador, apresentar iniciativas, oferecer sugestões para o utilizador e fornecer explicações de suas ações [93].
As principais áreas de aplicação das interfaces inteligentes são todas aquelas onde o conhecimento sobre a resolução parcial de uma tarefa reside no sistema de computador [6].As interfaces inteligentes são importantes quando o objetivo é apoiar os utilizadores com diversas necessidades, habilidades e preferências (incluindo pessoas com necessidades especiais), desde que facilitem uma efetiva, eficiente e natural interação do utilizador, facilitando a comunicação homem-homem [51].
As interfaces inteligentes de utilizadores são interfaces homem-máquina que têm por objetivo melhorar a eficiência, efetividade e naturalidade da interação homem-máquina representando discursos médios (por exemplo: o uso de gráficos, linguagem natural. Para uma interface ser

[2]Mesmo que os sistemas auto-adaptativos atuem de forma autónoma em muitos aspetos, eles têm que manter o utilizador no circuito, fornecer utilizador com *feedback*,sobre o estado do sistema é crucial para estabelecer e manter confiança dos utilizadores. Para esse efeito, um sistema auto-adaptativo precisa expor aspetos de seu controle para os utilizadores da Interface.

considerada inteligente ela deve possuir um ou mais das seguintes componentes [35].[3]

- Modelo do utilizador: é uma compilação de informações que descreve o utilizador e é usada para determinar como apresentar informação, o tipo de ajuda a conceder e como o utilizador vai interagir com a interface, sendo um dos componentes mais importantes das interfaces inteligentes;
- Ajuda Inteligente: apresenta ao utilizador a ajuda que ele precisa num período de tempo particular ou numa situação particular; o sistema reconhece o erro e propicia a causa deste erro.
- Adaptabilidade da interface: o utilizador pode configurar as preferências nas suas interfaces, o sistema pode auto adaptar-se para melhor interagir com o utilizador, sem que este tenha que definir a ação; interfaces adaptáveis podem também determinar o tipo de interfaces a apresentar para o utilizador, dependendo da análise do modelo do utilizador.
- Comunicação Multimodal: o uso de vários meios de comunicação multimodal [48].
- Reconhecimento dos Planos: é usado para deduzir o que o utilizador planeia fazer. Este reconhecimento torna o sistema inteligente; neste reconhecimento o modelo do utilizador e as suas ações são considerados.
- Apresentação Dinâmica: Diferentes pessoas devem ser capazes de ver dados de diferentes maneiras, a forma como o sistema decide mostrar os dados é determinado pelo exame do modelo do utilizador.

Ainda existem muitas dúvidas quanto à caraterização do que se pode considerar uma interface inteligente ou "interface bem-sucedida", que são geralmente designadas por amigáveis, o que certamente é muito pouco ilustrativo das suas caraterísticas. As capacidades de flexibilidade, acessibilidade, entre outras, são propriedades apontadas como indispensáveis para se atingir tal sucesso, pelo que muitas destas caraterísticas resultam da implementação de princípios formalizados, tais como a previsibilidade, a coerência, a conservabilidade, entre outros [60].
Com base na premissa anterior, o problema seguinte consistirá na determinação do tipo de conhecimento que a interface inteligente deverá possuir para que possa observar e participar no fenómeno de interação com um ser humano de modo compreensivo, isto é, intervindo autonomamente e com coerência.

2.3.1 Tipos de interfaces Inteligentes

A interface deve permitir que o utilizador possa traduzir as suas intenções à linguagem de entrada do sistema e possa interpretar a resposta do sistema, avaliando se ele está a conseguir cumprir com os seus objetivos. Esta problemática permite estabelecer uma representação mental das intenções com o sistema, salientado a importância da relação entre o conhecimento e a sua representação, para facilitar a utilização. No entanto, os tipos de comunicação tradicionais, como o diálogo, não podem ser descartados, já que desempenham um papel crucial. Por esta razão, diversas pesquisas estão a avaliar como construir interfaces que se adaptem às necessidades específicas dos utilizadores [34].A adaptabilidade pode considerar unicamente as

[3] Devem ser utilizadas quando existe uma grande distância semântica entre a linguagem dos utilizador e a linguagem de máquina, o que poderá suficientemente complicar as tarefas do utilizador.

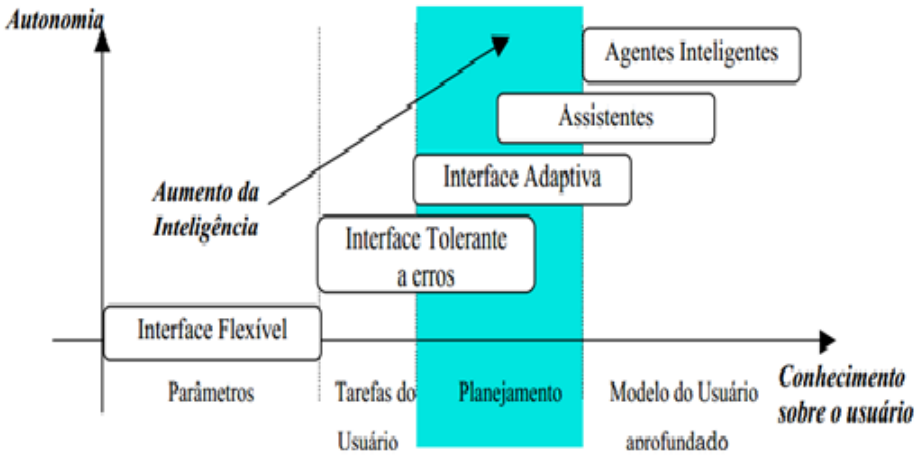


Figura 2.2: Figura nº2 Tipologia de Interfaces Inteligentes (modificado de KOLSKI, 1998)

preferências do utilizador ou considerar também as caraterísticas cognitivas, já que uma interface é definida por [76], sistema inteligente é o ultimo grau, sendo capaz de modelar o sistema humano-maquina como um sistema cooperativo [7].A inteligência das interfaces deve fazer com que os sistemas se adaptem aos utilizadores, tirem as suas dúvidas, permitam um diálogo entre o utilizador e o sistema, ou apresentar informações integradas e compreensíveis, utilizando vários modos de comunicação.
Todos os sistemas de hipertexto e de hipermédia refletem algumas caraterísticas do utilizador no seu modelo e aplicam de modo a adaptar vários aspetos visíveis do sistema, pelo que são sistemas de hipermédia adaptáveis. Ou seja, o sistema deve satisfazer três critérios:

1. Deve ser um sistema hipertexto ou hipermédia;
2. Deve ter um modelo de utilizador;
3. Deve ser capaz de adaptar a hipermédia usando este modelo, ou seja, o mesmo sistema pode parecer diferente aos utilizadores com diferentes modelos.

2.4 Interfaces adaptativas

A adaptação da interface foi identificada como um aspeto importante a ser considerado no desenho dos sistemas de informação, já que modernas técnicas de adaptação incluem a adaptação de informações a apresentar (adaptação da informação), como apresentar esta informação, adaptação da apresentação e como interagir com esta informação e adaptação de interface [68].As interfaces adaptativas apresentam-se promissoras na tentativa de superar os problemas atuais de complexidade na interação homem-máquina onde, cada vez mais, as aplicações se tornam mais complexas, levando o utilizador a tratar uma grande quantidade de informações simultaneamente. Para melhorar esta interação, é necessário que as interfaces sejam capazes de se adaptar às necessidades do utilizador.
Para que a interface seja considerada adaptativa, é necessário um modelo do utilizador, onde o sistema possa analisar as ações e perfis do utilizador, adaptando-se automaticamente ao mesmo. Utilizando-se interfaces adaptativas o sistema pode ser personalizado para estilos cognitivos individuais, necessidades de informações e tarefas personalizadas, sendo que vários estudos apontam para o facto de as diferenças de cada utilizador poderem ser controladas pelo projeto

da interface, nomeadamente pela personalidade, estilo cognitivo, estilo de aprendizagem e experiência [104].
A interface adaptativa é a interface homem-máquina que também é conhecida como interface utilizada nos *média*[4] para a troca de informações entre o utilizador e a máquina, que tem métodos de design tradicionais. A interface tradicional só pode ser adaptada a algumas pessoas, mas também não pode atender aos requisitos de uma pessoa em diferentes períodos [70].
A interface de utilizador adaptável pode-se ajustar a um utilizador ou a uma tarefa que surja e se desenvolva rapidamente, enquanto a exigência de desafio computacional tradicional requer três modelos:

- Modelo de sistema: descreve as caraterísticas do sistema que pode ser alterado, como o sistema para poder adaptar, a aquisição e a aplicação do modelo de utilizador que são as bases de uma interface de utilizadores adaptável para tornar o sistema que se adapta ao comportamento individual do utilizador.
- Modelo de interação: o modelo define como o sistema é modificado e o que ele pode-se adaptar, acima de tudo, o grau de adaptabilidade.
- Modelo do utilizador: é o processo que descreve os conhecimentos que podem ser utilizados para facilitar o uso das aplicações.

De acordo com Brusilovsky os sistemas adaptativos monitorizam o padrão de atividades dos utilizadores e, automaticamente, ajustam a interface ou conteúdo provido pelo sistema para acomodar-se ao utilizador, assim como às suas mudanças nas habilidades, conhecimentos e preferências [36].

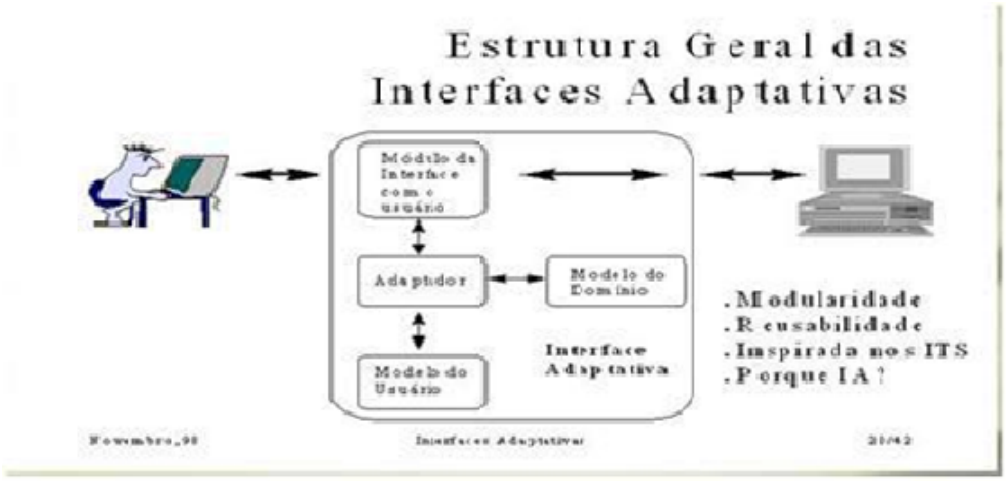


Figura 2.3: Figura Nº3 Estrutura Geral das Interfaces Adaptativas (Fonte: Elisa B.2015)

As interfaces adaptativas são diferentes das anteriores porque nestas, o utilizador não controla a adaptação, esta é realizada de forma automática e dinâmica, isto é, durante a utilização efetiva do sistema, pela interface, sendo o objetivo de fazer a interface aproximar-se o mais possível, em cada momento, das necessidades do utilizador, implicando que a interface se baseie num conhecimento que a cada momento deverá possuir sobre este, em particular sobre o tipo de utilização do sistema pelo mesmo. Assim, e contrariamente às anteriores, são interfaces

[4]Conjunto de meios de comunicação, meios de difusão de informação, meios de divulgação de informação, meios de veiculação de mensagens, suportes de campanhas publicitárias, suportes de propaganda.

inteligentes, nas quais a adaptabilidade é determinada com base neste conhecimento e que em regra determinam a sua dinâmica, baseando-se particularmente no contexto atual e histórico da interação, e por serem inteligentes e dinâmicas são tidas como interfaces auto adaptativas realçando tais caraterísticas.
Para que as interfaces adaptáveis sejam amigável é necessário que se construa uma interface com usabilidade, pois precisam de apresentar as caraterísticas e qualidades desejáveis relacionadas aos critérios de ergonomia de *software*. Uma interface é ergonómica quando as técnicas de construção de telas, de diálogo, de comunicação gráfica e visual, conduzem a comunicação homem-máquina a um estado de perfeito entendimento, qualidade e satisfação do utilizador no seu uso do sistema móvel [83].

2.5 Interfaces adaptáveis

Os sistemas adaptáveis permitem ao utilizador controlar os ajustes provenientes frequentemente do guia ou especializados para o utilizador. Uma interface de utilizador adaptável pode ser definida como um artefacto de *software* que melhora a sua capacidade de interagir com um utilizador, construindo um modelo de utilizador baseado numa experiência parcial com esse utilizador. Há uma premissa básica por detrás das interfaces que nos remete para o facto de os utilizadores serem diferentes e, portanto, possuem diferentes necessidades de um sistema interativo. O sistema deve-se adaptar ao utilizador em vez de forçar o utilizador a adaptar-se ao sistema; cada caraterística do utilizador e o comportamento passado são modelados na tentativa de se adaptar às suas necessidades e desejos. As interfaces podem ser classificadas de acordo com as variáveis de entrada que influenciam a adaptação e os tipos de efeitos de adaptação.
Hoje em dia, as pessoas podem transportar dispositivos para todos os lugares devido ao desenvolvimento e aplicação extensiva da comunicação móvel e das tecnologias da Internet. Neste caso, uma grande quantidade de requisitos concentrados nos interesses dos próprios utilizadores, as informações como uma das recomendações do interesse do utilizador ao adquirir a localização dos utilizadores, devem ser consideradas, bem como as questões pessoais para o espaço das informações usadas por mais e mais utilizadores [69].
A interface adaptativa tem sido amplamente estudada, mas há falta de autoaprendizagem de interação por parte do utilizador, pelo que iremos adotar um modelo de interface de utilizador que possa prever a intenção do utilizador com a sua experiência especial e, além disso, a proposta para o uso da consciência e da experiência auxilia na previsão do utilizador e sua intenção ao satisfazer o requisito.
Uma interface é considerada adaptável quando realiza as adaptações únicas no momento em que o utilizador requisita, ou seja, o utilizador adapta o sistema ao seu modo. Adaptar o utilizador ao sistema significa oferecer-lhe formação, documentação, tutores, facilidades de ajuda, entre outras, enquanto o sistema permanece fixo. Este enfoque apresenta desvantagens ao exigir uma maior dedicação do utilizador, de tempo para aprender a usar o sistema, tempo que não é utilizado numa atividade produtiva. Os sistemas adaptáveis permitem ao utilizador adaptar o seu próprio ambiente às suas preferências.
Quando se trata da adaptação do utilizador à interface necessita-se de um projeto para que os diferentes tipos de pessoas consigam aceder, de forma prevista, o conteúdo estabelecido, diminuindo a probabilidade de erros na navegação da interface do sistema, ou pode ser uma formação para que estes não sintam restrições quanto ao uso coletivo da forma como é apre-

sentada, antecipando possíveis erros e possibilitando a correção de maneira mais fácil.
As interfaces adaptáveis são interfaces possíveis de ajuste individual, em geral a cargo de cada utilizador, a vários níveis, estilos e objetos de diálogo que são oferecidos, escolhendo cada utilizador o que, na sua opinião, mais se ajusta a si próprio, ou então, a interface é programável a partir de um conjunto base, que permite ao utilizador construir o seu próprio ambiente em geral usando comandos pré-programados ou macros. Em qualquer dos casos a adaptação está a cargo do utilizador, sendo estática na primeira hipótese e semi-dinâmica na segunda.

2.6 Padrões adaptativos

O padrão de interface é um tipo de estrutura de interação que se torna comum num determinado ambiente a pontos de ser reconhecido como um padrão, dada a sua repetição, não sendo uma regra ou lei, mas sim apenas uma prática comum, segundo o arquiteto Christopher Alexander que apresentou a ideia de padrões nos anos 60 e explicou que eles formam uma espécie de linguagem do *design*.
O conceito de padrões adaptativos de interface foi criado pelo arquiteto Christopher Alexander no livro*Apattern Language*, mas somente com a publicação do Livro *Design Patterns*: *Elements of Reusable Object-Oriented Software de Erich Gamma, Richard Helm, Ralph Johnson e John* , os padrões de projetos se tornaram populares no âmbito do desenvolvimento de software. Diante da popularidade Bem Hui criou *Pattens* para solucionar problemas encontrados no desenvolvimento e interfaces em MIDP *(Mobile Information Device Profile)*. Segundo Hui, após trabalhar em vários projetos MIDP, ficou claro que muitos dos problemas recorrentes poderiam ser resolvidos com padrões adaptativos bem definidos, o que levou à seleção de vários padrões que se encaixassem no cenário do problema, adaptando-os, em seguida, para as condições das plataformas [18].

Mesmo que os sistemas operativos móveis tenham dado um grande salto evolutivo na melhoria dos interfaces de utilizador nos últimos anos, ainda existem muitos problemas de usabilidade desafiadores, muitas tarefas comuns ainda são mais fáceis e mais rápidas de atuar com o teclado e o *mouse* (rato) tradicionais num computador pessoal; sendo que o toque na tela como dispositivo de entrada é menos preciso do que o teclado e *mouse* (rato), e tem menor espaço para exibir os elementos da interface de utilizador. Estas tarefas são executadas de forma diferente num dispositivo móvel, ou seja, mais adaptado ao contexto em que são realizadas, no entanto, os sistemas operacionais ainda transportam muitos paradigmas herdados do *softwar*e de *desktop*, que nem sempre são adequados para dispositivos e, na sua maioria, estão cada vez mais a melhorar e introduzir conceitos novos bem-sucedidos.
Os dispositivos modernos estão equipados com sensores altamente sofisticados e esta caraterística particular pode ajudar o dispositivo a estar ciente sobre o estado dos sistemas e outros dados contextuais, podendo ser usados para tornar o comportamento do sistema mais flexível, sendo que um melhor conhecimento do ambiente circundante e a análise do uso do contexto podem trazer mais possibilidades para adaptação do sistema para atender os utilizadores de forma personalizada [77].
Dessa forma, resumidamente pode-se entender o padrão de projeto, como a solução recorrente para Um problema num contexto, mesmo que em projetos e áreas distintas, onde se observam termos-chave, torna obrigatória a compreensão inequívoca de cada um. De facto, um contexto diz respeito ao ambiente, e às circunstâncias dentro das quais existe algo, uma solução que se refere à resposta do problema que ajuda e que pode ser solucionado.

A seguir são apresentadas algumas das caraterísticas gerais dos padrões reunidas por Ficher, que integram as definições apresentadas de interfaces e da IHC:

- Formato de Apresentação: os padrões são descritos dividindo a sua apresentação num determinado número de elementos; estes são escolhidos de acordo com a ênfase e os detalhes que o autor deseja destacar no padrão, logo, percebe-se que não existe um formato único [116];
- Captura da Prática: um padrão apresenta uma solução para um problema existente na prática cuja solução pode ser aplicada, de modo eficiente, a diversos casos;
- Abstração: os padrões não podem ser abstratos demais, nem muito específicos; quando um padrão é muito abstrato, o utilizador do padrão necessita redescobrir como aplicar a solução, impedindo um melhor aproveitamento da solução do padrão; quando específicos demais também não são desejáveis, pois impediria uma maior aplicação do padrão;
- Princípio de Organização: Alexander observou que os padrões não estão isolados, que existe um relacionamento entre eles, estando organizados seguindo um dos dois princípios:

1. Catálogo de padrões; apresenta um conjunto de padrões;
2. Linguagem de Padrões, que apresenta um conjunto estruturado de padrões, o relacionamento entre eles e informações relevantes sobre o domínio e a aplicabilidade da linguagem, na linguagem, os padrões que a compõem que devem cobrir todos os aspetos importantes de um dado domínio.

Um padrão é escrito de forma compreensível, simples e coesa para facilitar a sua leitura e aplicação, pelo que possuem, basicamente, duas vantagens:

- Padrões de interface: são casos particulares de padrões de projeto e definem soluções para problemas comuns no projeto de interface de sistema;
- Captura da experiência: como os padrões apresentam soluções de sucesso que foram identificadas e desenvolveram ao longo do tempo, eles capturam a experiência, logo, os padrões podem ser utilizados para a transferência de conhecimento entre pessoas de nível de experiência diferentes. Os padrões possuem uma ampla aplicabilidade no desenvolvimento de sistemas interativos, essa aplicabilidade de padrões, tanto na interface como no IHC é ainda discutível.

Os padrões de programação ou idiomas são padrões de baixo nível, específicos para uma linguagem de programação, que descrevem como implementar aspetos particulares dos componentes ou a relação entre eles usando caraterísticas de uma determinada linguagem [19].

2.6.1 Desafios dos sistemas Adaptativos

Apesar das vantagens óbvias da implementação de aplicativos adaptativos de contexto em dispositivos, existem inúmeros desafios técnicos associados ao *design* de tais sistemas, sob o ponto de vista de *hardware e software*. Existem várias origens dos impactos, que afetam o *design* de aplicações adaptáveis ao contexto, nomeadamente a complexidade do projeto de sistemas adaptativos e problemas relacionados ao conhecimento do contexto.
A adaptabilidade aporta problemas com a correção, flexibilidade, sincronização e transparência do contexto, o trabalho do *design* da interface do utilizador tornando-se ainda mais complexo

devido às múltiplas situações de contexto, e para além disso, a mobilidade do dispositivo e as limitações que devem ser consideradas como um fator que afeta o *design* da interface do utilizador para aplicações.
Os problemas de privacidade (contexto-consciência, adaptação de interface) estão intimamente ligados, fazendo parte do contexto, o processo de manuseio e adaptabilidade de interfaces. Os sistemas devem proteger os dados recolhidos de acordo com as configurações e preferências de privacidade do utilizador, pelo que as informações como a atividade ou local atual devem ser publicadas com precisão e alinhadas às políticas de privacidades do utilizador

2.7 Usabilidade de Interface adaptável

A usabilidade pode ser definida como o estudo ou a aplicação de técnicas que proporcionam a facilidade de uso de um dado objeto, neste caso, um *site*. A usabilidade procura assegurar que qualquer pessoa consiga usar o *site* e que este funcione da forma esperada pela pessoa. Em resumo, a usabilidade tem como objetivos:

- Facilidade de uso
- Facilidade de aprendizado
- Facilidade de memorização de tarefas
- Produtividade na execução de tarefas
- Prevenção, visando a redução de erros
- Satisfação do indivíduo.

Assim, quando relacionada ao uso de interfaces digitais, a usabilidade refere-se ao potencial de efetivação das ações que os utilizadores desejam realizar (por exemplo: encontrar informações, ler textos, comprar produtos, jogar jogos).
O conceito inclui, também, o entendimento dos padrões de comportamento na procura e no uso de informações, no atendimento às necessidades dos utilizadores e grupos de utilizadores, na compreensão das motivações e nos processos de transformação subjetivos que se realizam através das informações e do uso. O aperfeiçoamento estrutural da usabilidade de *médias* digitais reflete-se diretamente na melhoria da qualidade da experiência do utilizador e no aperfeiçoamento dos seus processos de decisão, tanto em relação às ações que realiza quanto às informações que seleciona, refletindo-se na positiva perceção da marca associada à interface. Posto isto, a usabilidade está diretamente relacionada com o alcance dos objetivos da pessoa responsável pela sua publicação [50].
Dentro da IHC, o conceito de usabilidade tem vindo a ser reconstruido continuamente e tornou-se, cada vez mais rico e problemático, pois a usabilidade integra qualidades como a diversão, o bem-estar, a eficácia coletiva, a estética, a criatividade, o suporte para o desenvolvimento humano, entre outras. O entendimento atual da usabilidade é, portanto, diferente dos primeiros passos da IHC na década de 80. Na mudança de século, a ascensão dos serviços digitais (por exemplo: a web, o telemóvel ou a televisão interativa) acrescentou novas preocupações à IHC, dando origem a um outro conceito ainda mais significativo do que a usabilidade: a experiência do utilizador [43].
A experiência do utilizador vai além da eficiência, qualidade das tarefas e satisfação do utilizador, pois considera os aspetos cognitivos, afetivos, sociais e físicos da interação. Nesta

perspetiva, a experiência do utilizador contextualiza a usabilidade. Já não se espera que a usabilidade estabeleça o seu valor de forma isolada, mas que seja um dos contributos complementares para um projeto de qualidade, que não se concentre apenas nas caraterísticas e atributos dos sistemas [91] nomeadamente se são inerente-mente utilizáveis ou não, mas também, no que acontece quando os sistemas são utilizados. Tal permite contemplar aspetos como a diversão, o bem-estar, a eficácia, a estética, a criatividade e o suporte para o desenvolvimento humano, entre outros.
A ISO 9241-11 (1998) define a usabilidade como a medida na qual um produto pode ser usado por um utilizador específico com eficiência e satisfação num contexto específico de utilização. Para a usabilidade é um dos aspetos que pode influenciar a aceitação de um produto e aplica-se a todos os aspetos do sistema com os quais a pessoa pode interagir, incluindo os procedimentos de instalação [99] e manutenção, devendo ser sempre medida relativamente a determinados utilizadores que executam determinadas tarefas. Para que a usabilidade possa ser avaliada e medida, são necessários cinco atributos [100]:

1. Aprendizagem: o sistema deve ser de fácil aprendizagem para que o utilizador possa começar a utiliza-lo rapidamente.

2. Eficiência: o sistema deve ser eficiente no sentido de que uma vez o utilizador aprenda a utiliza-lo ele o faça com alta produtividade.

3. Memorização: o sistema deve ser de fácil lembrança, ou seja, ao passar um determinado período sem utilizar o sistema o utilizador pode utiliza-lo novamente sem ter que aprender tudo novamente.

4. Erros: a taxa de erros deve ser baixa, erros de extrema gravidade não devem ocorrer. Ao cometer algum erro, o utilizador deve ter a possibilidade de recuperar o sistema para o estado imediatamente anterior ao erro.

5. Satisfação: os usuários devem gostar do sistema, ele deve ser agradável de ser utilizado para que as pessoas se sintam satisfeitas com o seu uso.

Para a computação móvel segundo Bevan, a usabilidade de um dispositivo computacional depende de vários fatores, incluindo o utilizador, o ambiente e as caraterísticas do dispositivo, caraterísticas do utilizador (flexibilidade e destreza) [29]. O ambiente do utilizador afeta a escolha do dispositivo e as caraterísticas do dispositivo(os dispositivos têm caraterísticas próprias diferentes, que podem afetar a usabilidade total).
[5] A usabilidade, como já foi visto anteriormente, é um termo bastante amplo na computação e na usabilidade dos sistemas, referindo-se à facilidade para o utilizador final aprender a usar o sistema eficientemente, bem como tornar o seu uso agradável. Também, a frequência e a severidade dos erros do utilizador são considerados com partes constituintes da usabilidade, no entanto um utilizador pode achar um elemento da interface problemático por diversas razões: porque é um sistema difícil de aprender, porque é lento na execução das suas tarefas, causa erros de uso, ou pode ser simplesmente feio e desagradável. Por conseguinte, muito do trabalho da inspeção de usabilidade é classificar e contar o número de problemas de usabilidade, sendo que esta análise depende da exata definição do que é um problema de usabilidade e julgamentos

[5] Em suma, qualquer destes componentes servem perfeitamente para a definição de objetivos de usabilidade em qualquer tipo de projeto de sistemas interativos.

de como diferentes fenómenos constituem manifestações de um único problema. Frequentemente, é muito difícil fazer essas distinções mas, na maioria dos casos, o bom senso é suficiente para determinar o que é um problema de usabilidade, para uma definição geral de problemas de usabilidade, pode-se dizer que é qualquer aspeto de um *design* onde uma mudança pode melhorar uma ou mais medidas de usabilidade[3].

2.8 Personalização adaptável uso de Redes Sociais

O conceito de redes sociais antecede o surgimento das tecnologias atuais, podendo ser vistos como elementos que estão conectados de diversas formas, dependendo do tipo de relação adotado [133]. No entanto, com a massificação do uso da Internet o conceito de rede social tem-se misturado com os *sites* e aplicações móveis afins.
De acordo com Palazzo, as redes sociais possuem todos os recursos de uma rede social convencional, como o *Facebook*, perfil do utilizador, mensagens, mural, *chat*, listas de amigos, repositório, entre outros, para além disso, também permite a personalização da interface e da experiência do utilizador de acordo com as políticas específicas da comunidade [107]. A evolução tecnológica permitiu avanços na produção de interfaces cada vez mais interativas e voltadas para o utilizador, além disso, a popularização da Internet e dos dispositivos móveis tem possibilitado o aumento das atividades colaborativas na chamada web [63]. Sob a ótica da interface, há uma maior exigência dos *média* no fornecimento não apenas do acesso às informações, mas também a interação com elas contribuindo e modificando-as. Neste cenário, observaram que há diferentes possibilidades de interação em *sites* colaborativos e que nem todos os utilizadores contribuem ou lideram as discussões, o que os levou a uma classificação dos utilizadores [112].
A personalização, usando as redes sociais, pode ser mais efetiva porque os utilizadores dependem das escolhas dos seus pares como um atributo adicional [27], ou revisam preferências de atributos como resultado de influências de pares [98]. A importância das redes sociais é evidente: o *Fecebook* tem mais de 400 milhões de utilizadores móveis e os dois maiores dispositivos móveis serviços de redes sociais,*Foursquare e Instagram* , cada um tem cerca de 15 milhões de utilizadores [75]. Além disso,*Fecebook* tem parceria com o *Wall Street Journal e Washington Poste*, entre outros, para lançar aplicativos Social *Reader* para dispositivos móveis.[6]

2.8.1 Personalização Adaptável

Os dispositivos adaptáveis rapidamente se tornaram mais poderosos, versáteis e omnipresentes. Os telefones inteligentes e dispositivos similares fazem parte do dia-a-dia de qualquer utilizador, sendo que a tecnologia está tão avançada que permite a navegação na web e a utilização das redes sociais nesses dispositivos.
Os pesquisadores criaram um sistema de personalização adaptável, que combina os dados sobre o comportamento do utilizador em sistemas de informações sobre os interesses da rede social dos utilizadores para oferecer recomendações melhores e mais relevantes sobre artigos de notícias em dispositivos. Esta área é cada vez mais importante para a pesquisa, pois tendemos a ler as notícias cada vez mais em dispositivos dos utilizadores que impõem várias restrições, como o tamanho reduzido da tela, as conexões sem fio limitadas e a menor atenção dos leitores. É

[6]Computação Móvel desempenha um papel fundamental visto que os dispositivos móveis fornecem conectividade e pode permitir acesso, processamento e compartilhar as informações a qualquer tempo e em qualquer lugar.

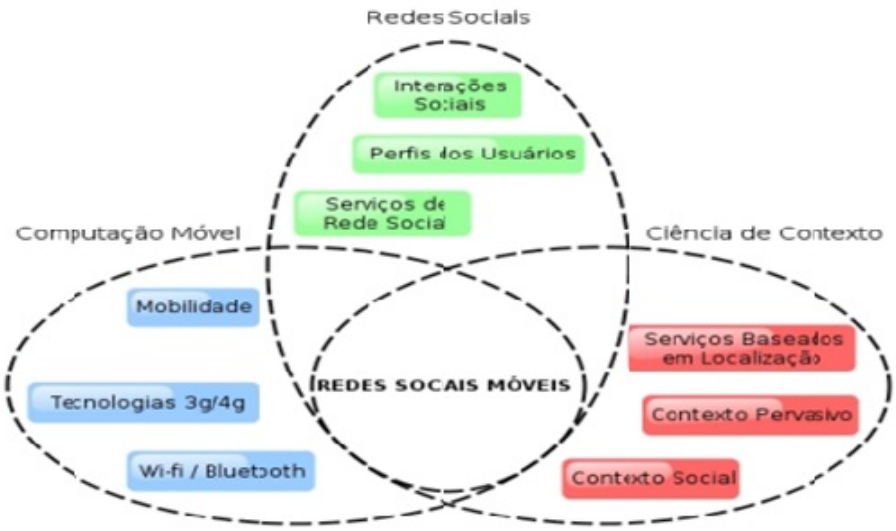


Figura 2.4: Figura nº4- Definição de Redes Sócias Móvel (Fonte:Livro de Mini cursos)

crucial reduzir o excesso de informações e exibir somente as notícias relevantes.
O sistema de personalização adaptável auxilia os utilizadores a descobrir de forma discreta notícias relevantes através da sua rede social, sem a necessidade de compartilhar o conteúdo ativamente e sem uma influência indevida de utilizadores pesados ou líderes de opinião. Este é um dos modelos mais importante para a compreensão da estrutura adaptativa, através de uma série de regras adaptativas, mencionados nas secções anteriores. A adaptação em conjunto com as informações dos modelos de utilizadores é responsável por oferecer ao utilizador uma estrutura personalizada aos seus objetivos e preferências [111]. Eles estão mais integrados na vida pessoal de um indivíduo e representam uma maneira mais natural de consumir serviços digitais (por exemplo, notícias móveis) em diferentes contextos. Mas, até agora, há apenas pesquisas limitadas que mostram como personalizar um ambiente móvel e quais são os benefícios de o fazer [41].
A localização (o termo em inglês*localization* é abreviado liOn, porque existem 10 letras do primeiro "1"até o ultimo "n") é o processo de adaptação de aplicações,serviços e produtos para um mercado internacional específico, para permitir a sua aceitabilidade numa cultura específica, incluindo a tradução da interface do utilizador, redimensionamento de caixas de diálogo, recursos de personalização, e resultados dos testes para garantir que o programa ainda funcione, ou seja, é o processo de fazer uma versão específica do produto para um mercado-alvo [120].
A proteção de informação de contexto de localização é particularmente importante nas Redes Sociais Móveis, devido à grande quantidade de aplicações que proveem recursos baseados na localização do dispositivo móvel. A privacidade de localização pode ser definida como o direito do utilizador de decidir como, quando e para qual o propósito das suas informações de localização podem ser reveladas a outros [21]. As seguintes categorias de privacidades de localização podem, então, ser identificadas [20]:

- Privacidade da Identidade: onde o objetivo é proteger a identificação do utilizador associadas a partir das informações de localização, informações de localização podem ser providas a alguma entidade, mas a identidade do utilizador deve ser preservada;

- Privacidade de Posição: onde o objetivo é adulterar a localização exata do utilizador a fim

de proteger sua real localização;

- Privacidade de caminho: onde o objetivo é não revelar localização anteriores ao qual o utilizador passou, ou seja o caminho atravessado pelo utilizador

A personalização adaptativa para atender o difícil teste de aceitação do utilizador, deve ser concebido com certas restrições, fornecer uma boa experiência inicial e aprender rapidamente para novos utilizadores, onde o primeiro uso deve fornecer uma experiência aceitável e não pessoal. Os benefícios da personalização adaptativa devem ser aparentes dentro dos primeiros usos se os utilizadores retornarem além disso, a transição de um não-pessoal para uma experiência personalizada deve ser boa, não passando por um estado de apresentação de itens aleatórios para o utilizador.

As redes sociais possuem funcionalidades para criar perfis que representam entidades, as quais se relacionam socialmente, trocando informações, entre indivíduos, organizações ou mesmo entre sistemas. A computação móvel possibilita os utilizadores estarem sempre *online*, devido ao suporte de mobilidade provido pelos dispositivos portáteis [74].

A análise das redes sociais surgiu como uma técnica chave na sociologia moderna e na antropologia social, no final do século XX. O termo passou a ser visto como um novo paradigma das ciências sociais, tendo vindo a ser aplicado e desenvolvido no âmbito de disciplinas tão diversas como a antropologia, a biologia, os estudos de comunicação, a economia, a geografia, as ciências de informação, a psicologia social, a sociolinguística e, sobretudo, no serviço social. Em 1954 J. A. Barnes começou a usar o termo sistematicamente ao público em geral, incorporando os conceitos tradicionalmente usados quer pela sociedade quer pelos cientistas sociais, grupos bem definidos, tribos familiares e categorias sociais [25].

Assim a rede social é uma estrutura social composta por pessoas ou organizações, conectadas

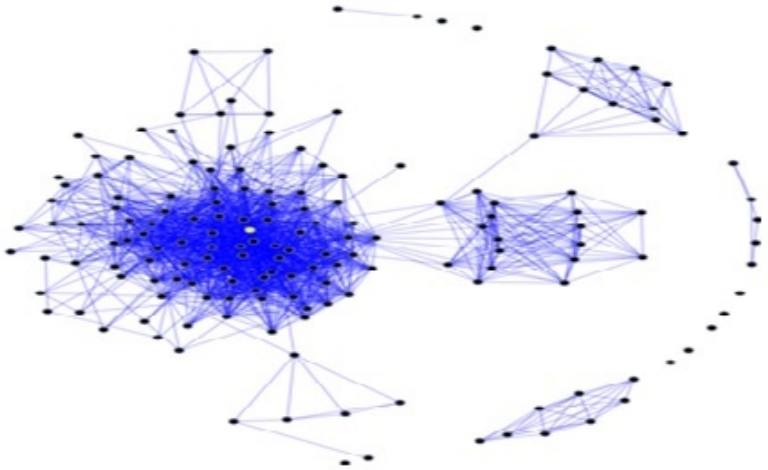


Figura 2.5: Figura nº5 exemplo de um diagrama de uma rede social. O nó com maior grau de centralidade de intermediação está representado em amarelo. (Fonte: Levina, e Timme, 2011)

por um ou vários tipos de relações, que compartilham valores e objetivos comuns. Uma das fundamentais caraterísticas na definição das redes sociais é a sua abertura, que possibilita relacionamentos horizontais e não hierárquicos entre os participantes. As redes não são, portanto, apenas uma outra forma de estrutura, mas quase uma não estrutura, no sentido de que parte da sua força está na habilidade de se fazer e desfazer rapidamente.[7]

[7] É a ordenação de elementos em ordem de importância, podendo significar também, mais especificamente: a distribuição ordenada de poderes

2.9 Perfil do utilizador

O perfil é um conjunto de caraterísticas ou competências necessárias ao desempenho de uma atividade; é um cadastro de dados pessoais, de contacto e preferenciais de um determinado utilizador. Parte destes dados podem ser públicos, sendo compartilhados com os demais utilizadores, ou privados dependendo do tipo de perfil, tipo de comunidade ou configurações de privacidade definida pelo utilizador.
Uma boa ergonomia e um *design* minimalista, deve caraterizar os dispositivos tornando-os mais fáceis de manusear, ser resistentes à degradação por ações do ambiente, como unidade e, além disso, nenhuma informação desnecessária deve ser exibida ao utilizador. Um sistema móvel deve ter uma boa interface, deve ter uma boa interação entre o utilizador e a tela, para que o mesmo possa fazer as suas tarefas da melhor forma possível. Permite cada um dos módulos para que tenham acesso, o perfil particular do utilizador, selecionando ou não cada uma das autorizações associados ao módulo:

- Pode alterar os formatos da base bibliografia do perfil.
- Pode alterar os formatos do Administrador.
- Pode alterar os registos.

É uma das técnicas que auxilia a gestão das permissões e acesso aos diversos na rede de telemóveis porque controla o acesso aos sistemas adaptáveis às aplicações com uma maior segurança e disposição racional dos recursos permitidos.
A evolução das preferências do utilizador e do seu conhecimento pode ser deduzida a partir dos acessos das interfaces, uma adaptação essencialmente baseada nas ações de navegação do utilizador. A adaptação pode ser feita modificando-se vários aspetos visíveis do sistema e formas de acesso preestabelecidas.
O utilizador apresenta diferentes caraterísticas de perfil que podem ser configuradas diretamente próprio, caraterísticas como a língua, o tipo de grafismo utilizado, predefinições do sistema, entre outros. No modelo do utilizador é onde se encontram as caraterísticas e preferências particulares que o sistema apresenta, sendo que essas informações podem ser modeladas pelo próprio utilizador de acordo com o nível em que ele se encontra.

2.10 Perfil Linguístico do Utilizador

O perfil linguístico tem vindo a ser desenvolvido pela ciência da informação, que tem investigado as propriedades e os comportamento da informação, as forças dos sistemas ou das aplicações através dos fluxos e dos meios de processamento de informações, tendo como objetivo a sua organização, no armazenamento, recuperação e disseminação, com estreita ligação com a linguística pela intermediação da análise de documentos com palavras, utilizando-se métodos e processos para descrever o conteúdo dos documentos. Muitas são as teorias que corroboram as pesquisas para uma interface com um perfil linguístico que possa permitir uma análise ou comparação de palavras na procura de subsídios para a aplicação.
[8] O bastante extenso, pode referir-se desde a linguagem dos animais até outras línguas mais recentes apontam para a tendência de uma terminologia e análise documentaria, subáreas mais

[8]Como o termos linguagem pode ter um uso não especializado bastante extenso , podendo referir-se desde até outras linguagens humanas

estudadas nos últimos 10 anos pelos teóricos da informação e semântica. Estes estudos visam procurar, na próxima década, o rigor para as práticas de construção de perfis linguísticos para documentos com semelhança de palavras numa interface. As regras entendidas por um orador refletem padrões ou regularidades específicas na forma como as palavras são formadas a partir de unidades menores no idioma que está a ser usado e como essas unidades menores interagem na fala e na escrita. Desta forma, a morfologia é o ramo da linguística que estuda padrões de formação de palavras dentro e entre idiomas e tenta formular regras que modelam o conhecimento dos falantes nas suas pronúncias em cada região ou do continente [5].

Os sistemas personalizados adaptam o seu comportamento a utilizadores individuais aprendendo as suas preferências durante a interação, para construir um perfil de utilizador, que pode ser posteriormente explorado no processo de procura. Esta abordagem tradicional é baseada em palavras-chaves e é principalmente conduzida por uma operação de correspondência de *string*. Esta ou alguma variante morfológica é encontrada tanto no perfil e no documento, uma correspondência ocorre e o documento é considerado relevante. Estes problemas exigem métodos alternativos para aprender perfis mais precisos que capturam conceitos, expressando interesses de utilizadores de documentos relevantes, no entanto os perfis semânticos devem conter referências a conceitos definidos em léxicos ou em ontologias, segundo Semeraro, no seu artigo, que define um perfil linguístico de utilizador como uma caraterística que é obtida por técnicas de aprendizagem integradas com uma desambiguação de sentido de palavra [119].
O perfil linguístico abrange a gramática (o tipo de formas verbais utilizados, as concordâncias nominais e verbais, entre outros); o léxico (os estrangeirismos e os empréstimos das línguas bantu e do Inglês, do árabe); a semântica, que depende e varia de cultura para cultura, e da fonética e fonologia que influencia a escrita, como é o caso do informante "A" que coloca /X/ em todas as palavras que têm o som [127]. Por exemplo, escreve *xtou para dizer "estou"; escreve *parabnx para dizer "parabéns"; *mexmo para dizer "mesmo". Esta adoção, ou melhor, esta adaptação da escrita apareceu com os informantes " A" e " D", e é um traço do perfil linguístico que ocorre nos dias de hoje, talvez motivado pelo nível de escolaridade dos informantes.
Evidentemente que as perspetivas das caraterísticas do perfil linguístico, são bem entendidas, textualmente e pragmaticamente, não se podendo entender o funcionamento da linguagem por modelos paradigmáticos. Além disso, o posicionamento que acolhe o funcionamento dos termos em todas as dimensões das suas realizações, acarreta profundas implicações sobre o tratamento dos termos com vista à termino grafia [38]. Os instrumentos termino gráficos elaborados com essa perspetiva têm condições de oferecer ao cliente informações. Com efeito, a contribuição da linguística para a terminologia aporta perspetivas promissoras para a realização de uma tarefa complexa que requer fundamentação científica apropriada para tratar adequadamente as formas de expressão especializada dos saberes humanos [130].
Segundo Smalls o perfil linguístico é um termo usado para descrever inferências que muitas vezes são feitas sobre o discurso de uma pessoa, concluindo que o orador pode ser homem ou mulher, se ele é nativo do país ou não, refletindo caraterísticas de fala aprendidas que comunicam muitas informações. Todavia este processo pode ser usado para fins discriminatórios da linguagem e das minorias raciais. O perfil linguístico é um termo recentemente criado para representar o equivalente auditivo do perfil racial, considerando que o perfil se baseia em pistas visuais que resultam na confirmação ou na especulação do plano, na pronúncia de palavras na escrita e nos atos pessoais na cultura, na raça e cor. Um estudo de Purnell, Isardi e Baugh indica que o perfil linguístico é baseado em dialetos na afiliação de grupos étnicos; com muito poucos discursos necessários para que o perfil seja descriminado entre dialetos, através de alguns correlatos fo-

néticos ou marcadores de dialetos que são recuperáveis de discurso [122].
O estudo determinou que alguns utilizadores conseguiram identificar corretamente o dialeto de um falante em mais de 70% do tempo, tendo estas descobertas sido significativas, uma vez que os utilizadores são capazes de discernir esta informação apenas a partir das palavras, ou de uma palavra que neutraliza, diferenças lexicais, sintáticas e fonológicas entre os dialetos. É importante notar que as diferenças linguísticas entre minorias dos utilizadores são anatómicas e determinadas ou um resultado da "genética racial", em vez disso, tais diferenças linguísticas são caraterísticas de fala aprendidas, com conhecimentos compartilhados sobre a sociedade em geral, nas suas posições dos interlocutores e normas apropriadas do discurso dadas, na situação do discurso. O perfil linguístico frequentemente é feito a partir das comunidades linguísticas e a fala com acentos indesejáveis também pode estar sujeita a perfis linguísticos.
Uma das primeiras tarefas que envolve um exame das caraterísticas das palavras para que permita verificar informações sobre a base regional, social e étnica dos falantes, envolve a análise comparativa contra uma amostra de referência conhecida por um utilizador particular. Embora a análise das caraterísticas vocais não possa determinar a identidade de um falante, ela pode fornecer uma grande quantidade de informações sobre o falante, embora com diferentes graus de precisão e confiança, pelo que a comparação de palavras aos falantes é preferida para a identificação e para que se realize as três metodologias gerais: impressão de palavras sinónimas, análise usando sistemas automáticos e análise linguística.

Segundo Timbane as caraterísticas lexicais e o vocabulário de um perfil linguístico são definidos em termos do conjunto de palavras e de expressões idiomáticas dadas no uso distinto dentro de uma variedade. Um exemplo concreto é o Português de Moçambique, que segundo Timbane, no seu artigo, refere que este tem caraterísticas lexicais próprias que se distingue das outras variedades usadas nas comunidades dos Países de Língua Oficial Portuguesa [126]. Por exemplo, as unidades lexicais lobolar ou anelar (ato de entregar dote), sograria (casa dos sogros), descamisar (ato de tirar a camisa), mukhero (negócios ilícitos fronteiriços), bichar (fazer fila), são palavras que ocorrem exclusivamente na variedade do Português de Moçambique, devido às caraterísticas peculiares do seu perfil linguístico. Segundo Leiria, o perfil linguístico de um utilizador permite aprender, mas, tudo isto demora tempo e exige esforço. É possível ler, interagir e escrever com um número relativamente reduzido de palavras muito frequentes, assim sendo, é conveniente prestar particular atenção e insistir no vocabulário mais frequente, usando cada um dos itens lexicais em contextos linguísticos mais frequentes, usando cada um dos itens lexicais em contexto linguísticos muito variados [81] (por exemplo, fazer um bolo, fazer uma festa, fazer anos, fazer frio, fazer a cama, fazer doer, entre outros). Como os vocabulários básicos constituem uma rede na qual se encaixam progressivamente nomes e verbos específicos de um dado domínio, isto é, vocabulários específicos, é sobre esta base lexical que se constrói a gramática de uma língua, pelo que quanto mais informação se tiver associada a cada uma das palavras, em particular às mais frequentes, maior será a nossa competência lexical e, melhor compreenderemos e falaremos uma língua.
A uniformidade linguística baseada no perfil é um dos métodos projetados para permitir comparar as variedades da linguagem com base numa ampla gama de linguagens potencialmente heterogéneas, com diferenças nas variáveis individuais que são resumidas em dissimilaridades globais; é uma série de variedades de linguagem que são subsequentemente agrupadas ou traçadas usando uma multivariedade técnica, como análises de*cluster* ou dimensionamento multidimensional. Segundo Geeraerts, na sua obra, em que o seu foco estava no léxico, a variação onomasiológica ocorreu quando termos diferentes são usados para se referir à mesma entidade

(ou à mesma propriedade, relação, ação, estado de coisas, entre outros). A variação onomasiológica da forma foi definida como variação em que o uso de diferentes termos não é devido a uma classificação conceptual diferente dos objetos relativos ao mesmo conceito. Assim, uma variação onomasiológica [9] conceptual é uma variação que não é formal, é uma situação em que a mesma entidade é referida uma vez por meio de um termo específico [124]. Um exemplo de variação onomasiológica formal é a situação em que a mesma entidade às vezes é chamada de "carro" e às vezes de "automóvel", ou seja, o termos alternativos para o mesmo conceito que devem ser usados como um método adaptativo ao perfil linguístico.
O perfil linguístico é visto ainda de forma morfológica, que pode ser um dos ramos da linguística que estuda padrões de formação de palavras dentro e entre idiomas e tenta formular regras que procuram modelar o conhecimento dos falantes das línguas. As modificações fonológicas e ortográficas entre umas palavras base e a sua origem podem ser parciais nas habilidades de alfabetização. Estudos mostram que a presença de modificação na fonologia e na ortografia torna as palavras morfologicamente complexas, mais difíceis de entender e que a ausência de modificação entre uma palavra base e a sua origem torna as palavras morfologicamente complexas que serão mais fáceis de entender; as palavras que são morfologicamente complexas são mais fáceis de compreender quando incluem uma palavra base.
Na sua forma mais simples e ingénua, esta maneira de analisar formas de palavras, chamada, item- e – arranjo, trata as palavras como se fossem feitas de morfemas, postas uma após a outra, concatenadas como se fossem as contas de uma corda. Abordagens mais recentes e sofisticadas, como a morfologia distribuída, procura manter a ideia do morfema, enquanto acomodam processos não concatenados, analógicos e outros que se mostraram problemáticos para teorias de itens e arranjos, e abordagens semelhantes de palavras para os perfis linguísticos. Esta representação da estrutura morfológica básica permite uma melhor caraterização das formas de base das palavras geradas por utilizador. Enquanto as palavras, juntamente com os clíticos, são geralmente aceites como as de menor unidades de sintaxe, na maioria dos idiomas, se não todas, muitas palavras podem ser relacionadas a outras palavras que por regras descrevem coletivamente a gramática para a linguagem. Em contraste, os Chineses clássicos têm pouca morfologia, utilizando morfemas quase não vinculados e dependendo da ordem das palavras para transmitir o significado, sendo a maioria das palavras no padrão chinês "mandarim", no entanto, são compostos e, a maioria das raízes está vinculada, é entendida como gramática e representam a morfologia da linguagem [32].

2.10.1 Perfil linguístico e suas características

Segundo Baubh o perfil linguístico é a prática de identificar as caraterísticas sociais de um indivíduo com base em pistas auditivas, em particular o dialeto e acento. A teoria foi desenvolvida pela primeira vez pelo professor para explicar as práticas discriminatórias no mercado imobiliário com base na redação auditiva da clientela prospetiva por administrador [1]. O perfil linguístico se estende para questões de processos legais, oportunidades de emprego e educação, pelo que a teoria é frequentemente descrita como o equivalente auditivo do perfil racial, sendo que a maior parte da pesquisa e evidência que apoia a teoria potencia distinções raciais

[9] A Onomasiologia é um ramo da lexicologia que estuda os significados partindo de um conceito existente na realidade, o significado; este é útil para os estudiosos de linguística que observam as mutações das palavras no tempo. É muito estudada juntamente com a semasiologia, que percorre o mesmo percurso em direção oposta.

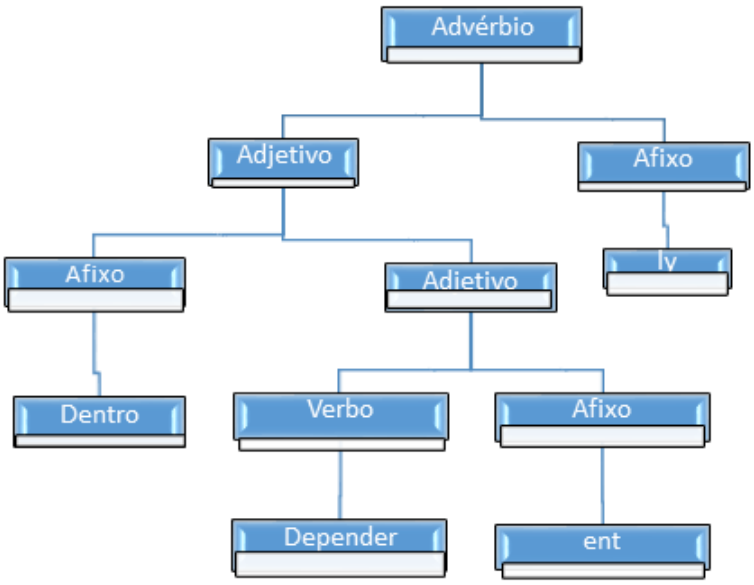


Figura 2.6: Figura nº6 Árvore de morfologia baseada em morfemas da palavra ”independentemente. Fonte:wikipedia.org.2018

e étnicas. Baugh, Purnell e Idsardi completaram um conjunto de quatro experiências com base na identificação de dialetos em Inglês americano, tendo chegado aos seguintes resultados [17].

- A discriminação baseada no dialeto e corre.
- É possível que os ouvintes naïves (ingénuo) identifiquem a etnia através da fala.
- É necessário muito discurso para fazer uma identificação precisa.

As caraterísticas baseadas numa série de abordagens linguísticas, para a análise de texto, e palavras são baseadas em palavras simples e frases simples (cabeça substantivos precedidos de pré-modificadores opcionais, mas sem recursivo incorporado). Cada uma dessas características morfológicas, sintáticas e semânticas têm varias variações, considerando, portanto, as seguintes possíveis combinações entre unidades de texto:

1. O que conhece, isto é; compartilha de uma única palavra entre unidades de texto, variações de recurso que restringe a correspondência aos casos em que as partes do discurso ou das palavras no texto também combinam ou relaxe nos casos em que apenas á duas palavras que são idênticas.
2. O que queres? *WordNet* [10] permite fornecer informações sensoriais, colocando palavras em conjuntos de sinónimos, estas combinações de palavras aparecem no mesmo sincronismo, as variações neste recurso restringem as palavras considerando uma classe especifica de parte da fala.
3. Classes semânticas comuns para verbos; Levin para verbos têm considerado útil para determinar o tipo de documento e a similaridade de texto , combinados entre dois verbos

[10]O WordNet é um banco de dados lexical para a língua inglesa que permite a agrupar palavras em inglês em conjuntos de sinónimos chamados sintaxes, fornece definições curtas e exemplos de uso e registar uma serie de relações entre esses conjuntos de sinónimos ou seus membros [57]. *WordNet* pode assim ser visto como uma combinação de dicionário de sinónimo

que compartilham a mesma classe [11]semântica [82].

As caraterísticas do perfil linguístico faz-nos compreender que na fala de uma língua, antes de mais, é indispensável saber palavras ou saber uma palavra é muito mais do que associar um significado a uma sequência sonora ou gráfica de cada indivíduo, dependendo da sua cultura etnia e cor. Cada palavra contém um conjunto de diferentes tipos de informação: fonológica, morfológica, sintática, semântica e pragmática. Assim, para saber uma palavra, é preciso saber, pelo menos, como se diz (e como se escreve) [23] conhecer a sua estrutura de base, as derivações mais comuns e a sua flexão; o seu comportamento numa frase ou num enunciado; o seu significado referencial, extensões metafóricas e a sua adequação pragmática; as suas relações com eventuais sinónimos e antónimos e a suas combinatórias mais frequentes. Mas, nem todos os itens lexicais são formados por uma só palavra, muitos são combinatórias de vários tipos, sequencias mais ou menos cristalizadas, que podemos aprender de cor, em conjunto. As chamadas expressões idiomáticas são, de entre estas, as combinatórias menos transparentes e, nos primeiros níveis de utilidade muito relativas.
As regras entendidas por um orador refletem padrões ou regularidades específicas na forma como as palavras são formadas a partir de unidades menores no idioma que é usado, como de menor integração na fala. A morfologia é o ramo da linguística que estuda padrões de formação de palavras dentro e entre idiomas e tenta formular regras que modelam o conhecimento dos falantes das línguas [72].
Uma das caraterísticas do perfil linguístico é a heterogeneidade linguística que é um dos exemplos dos alunos de português que não têm a língua portuguesa como língua materna e abrange um leque muito vasto de perfis linguísticos, com as diferentes caraterísticas pessoais e com diferentes graus de afastamento entre a língua materna dos utilizadores e o português, resultando [80]:

1. Diferentes graus de transferência de conhecimentos, de estruturas e de experiências comunicativas da língua materna do utilizador para o português.

2. Diferentes capacidades de discriminação e produção de sons do sistema fonológico do português.

3. Diferentes capacidades de segmentação de palavras e de deteção de valores das unidades de significado do português.

4. Diferentes desempenhos nas tarefas de associação som, grafia, motivados pelas características que distinguem o sistema de escrita alfabética, adotado pelo português, de outros sistemas.

Na obtenção dos radicais *(stems)*: em linguagem natural em diversas palavras que designam as variações indicando plural, flexões verbais ou variantes são sintaticamente similares entre si [56]. Um exemplo concreto são as palavras *delete, deletes, deleted e deleting* que tem a sua semântica relacionada, com um dos objetivos da obtenção dos radicais e a obtenção de um elemento único que o radical permite considerar como um único termo, portanto, com uma semântica única, estes elementos de texto, com este passo permite uma redução significativa no número de elementos que compõem o texto.
O conhecimento sobre as palavras de uma língua e os seus possíveis sentidos que pode organizar-se nas chamadas bases de conhecimentos léxicos onde, para o inglês se destaca a *(WordNet de*

[11]Semântica é a parte da gramática que estuda o sentido e a aplicação das palavras em um contexto. Estudo de palavras de uma língua

Princeton, entre as várias tarefas do processamento computacional da língua que podem recorrer a uma destas bases de conhecimento, destaca-se a similaridade semântica [52]. Nestes conjuntos de palavras vizinhas, podíamos incluir efetivamente todas as palavras diretamente relacionadas, ou poderíamos restringir apenas a alguns tipos de relação. Por exemplo, em algumas experiências utilizaram-se apenas sinónimos e hiperónimos para melhor obtenção do perfil linguístico [103]. Um perfil de uma aplicação tem a sua utilidade no que se refere à divulgação e à publicação da maneira com que os pesquisadores estão a utilizar os padrões de adaptação. Assim sendo, desenvolvem-se novos padrões e elementos que melhor descrevem as necessidades de aplicação. Segundo Befi-Lopes, a heterogeneidade é uma caraterística marcante no momento em que pode motivar algumas subdivisões para classificar as alterações linguísticas apresentadas, de maneira mais homogénea, sendo que uma delas sugere três divisões do que é distúrbio expressivo [26]; distúrbio expressivo e de perfil linguístico.

Capítulo 3

Metodologia para Identificação do Perfil Linguístico

A nossa abordagem refere-se à adaptabilidade de perfis linguísticos do utilizador para construção do perfil linguístico não supervisionado independente da língua do utilizador. Ultimamente, observa-se uma intensa atividade na investigação e no desenvolvimento de ferramentas eficientes na adaptabilidade dos perfis linguístico.

3.1 Metodologias e características de extração de termo

As informações importantes são obtidas normalmente pela criação de padrões e tendências através de meios de padrões estatísticos de aprendizagem, geralmente através de modelos de textos que envolvem o processo de estruturação do texto de entrada, que é frequentemente a análise, juntamente com a adição de algumas caraterísticas derivadas e com a retirada de outras, e com a subsequente inserção da derivação de padrões dentro da estrutura da avaliação e interpretação do resultado. Na análise automática de um vasto corpus textual os pesquisadores analisaram milhões de documentos em diversas línguas com pouca intervenção manual.
Existem vários tipos de metodologias para facilitar o cálculo de uma informação que foram criados para formar a taxonomia de modelos, que incluem alguns modelos: o Booleano, o espaço vetorial, o probabilístico, o fuso *(fuzzy)*,o da busca direta, o de aglomerados *(clusters)*, o lógico e o contextual ou conceptual, o método de seleção de recurso não supervisionado, e o método de seleção supervisionado, entre outros…
As principais fases para um utilizador identificar um texto ou frase são:

Figura 3.1: Fonte: elaborado pelo autor

uma adição de informação semântica, para que possa tornar possível o processamento computacional de textos, uma vez que é realizado um processo de tokenização, que é um dos passos que este trabalho vai conduzir, para que nos permita realizar tudo o que constitui conhecimento nos textos. O uso de técnicas de PLN em textos tem como objetivo identificar a real importância de cada termo em determinados contextos, onde a PLN se assume como sendo um dos mais utilizados para agregar valores semânticos. Este método consiste nos princípios da presente invenção e relaciona-se, geralmente, com a recuperação de informações e, mais particularmente, com as técnicas para localizar os termos-chave e frases. Normalmente, os termos importantes ou frases aparecem em consultas, porque os utilizadores formulam-nas, mesmo para termos-chave através de sistemas baseados, como se fossem destinados a um leitor humano.
As interfaces via texto estão rapidamente a tornar-se numa necessidade e, num futuro próximo, os sistemas interativos irão fornecer fácil acesso a milhares de informações e serviços

que vão afetar, de forma profunda, a vida quotidiana das pessoas. Nos dias de hoje, os sistemas de adaptabilidade de um perfil linguístico encontram-se limitados a pessoas com acesso aos equipamentos eletrónicos como telemóveis, tabletes e computadores, por isso, uma parte muito reduzida da população, mesmo nos países mais desenvolvidos, sendo necessário avanços nas tecnologias da linguagem humana para que o cidadão, com as condições médias (principalmente nos países subdesenvolvidos), possam aceder a estes sistemas, usando habilidades com maior facilidade de comunicação naturais e utilizando aparelhos domésticos, como os telefones. Na verdade, sem avanços fundamentais nas interfaces voltadas para o utilizador, uma larga fração da sociedade será impedida de participar da era da informação, resultando numa maior estratégia da sociedade, agravando ainda mais o panorama social atual. Um dos trabalhos de pesquisa futura é a implementação de uma interface via voz, na linguagem do utilizador, pois seria ideal uma vez que esta é mais natural, flexível, eficiente e económica a forma de comunicação humana.
Está a abordagem tem como exemplo são os tópicos e caracterizados a traves de modelos estatísticos de N-Gramas, com estes métodos temos o seguinte:

$$P(w1, w2, ...wn) = \prod_{i=1} T + 1P(w1/wi - 1, .., wi - n + 1) \quad \textbf{(3.1)}$$

onde w1 e wt define as fronteiras das frases,P(w1,w2,...wt) a ordem do processo de Markov e n é considerado.

3.2 Identificação do Perfil Linguístico

As palavras numa língua são a principal manifestação da inteligência humana, pois é através da língua que se expressam as palavras e a boa pronúncia das frases, bem como o enquadramento dos termos desde as necessidades básicas a conhecimentos técnicos, quando consideramos que uma linguagem natural é a aquela que é recorrente-mente utilizada no quotidiano para a comunicação humana. Em contrapartida com as linguagens artificiais, como as de programação, as linguagens naturais evoluíram de geração em geração e, embora tenham regras, estas são poucos estáticas, para os perfis linguísticos.

Nesta secção, abordamos um *Stopwords* que permite localizar as listas de termos vazios ou listas de termos–chave; ou seja, é basicamente uma lista de termos, preposições, determinantes, pronomes, verbos [128] entre outros, que são considerados de baixo valor semântico, e que por sua vez quando são identificados num documento, são eliminados, sem consideração aos termos de índice para a recolha de textos e para que possam ser analisados. Quando eliminados, estes termos evitam o ruído documental e supõem um rendimento considerável de recursos, uma vez que é um número relativamente pequeno de elementos que têm uma alta taxa de frequência nos documentos. Um sistema de recuperação de informações (por exemplo um motores de busca) que usa uma consulta baseada em termos-chave geralmente ignora palavras ou grupos de termos ou frases, que ocorrem muito comummente e, geralmente, não estão relacionados com a informação solicitada. Normalmente, os termos-chaves ou frases aparecem numa consulta, porque os utilizadores formulam as suas consultas, mesmo para termos-chaves, porque é um sistema baseado e como se fossem destinados a um leitor humano. Por exemplo, a palavra “a” na consulta “um hotel de Lisboa” é uma *stopword* e a frase “mostra- me” na consulta “mostra-

me Lisboa Hotels" é uma frase de parada, ambos "a" e "mostram-me" são sem sentido para a intenção dos utilizadores de encontrar informações sobre hotéis em Lisboa.
Não é só, consideramos as ocorrências de termos-chave, mas também leva-nos a encontrar as distâncias entre elas, sendo mais uma das maneiras adequadas de ver as distâncias entre *stopword* de um gráfico de termos-chaves e as distâncias entre eles capturados pelo seu peso das bordas [22].
As *stopword* baseiam-se num conhecimento ou palavras-ruído que transmitem muito pouca semântica sobre o significado numa frase, mas serve para adicionar detalhes. Um exemplo "Benfiquistas jogam no Jardim". Sem senhas ou termos-chaves: Benfiquistas jogar Jardim. Neste sentido, eles têm como função dos termos, mas as listas de parada, também incluem termos que estão fora das listas de termos de uma função, geralmente as *stopword* são identificados com base na sua prevalência no texto, ocupando uma percentagem significativa do texto que ronda pelo menos ou aproximadamente de 50%, encontrando-se muito em excesso da suas proporção no léxico.
Uma *stopword* pode ser chamada de palavra comum, ou palavra não-preditivas e não discriminatória, com baixo conteúdo de informação e uma baixa taxa de previsão e resultados de previsão, e que podem ser agrupadas em duas categorias [92]: a primeira inclui palavras-chave padrão, que estão disponíveis no público, parágrafos de domínio ou não-padrão que podem ser geradas dentro da recuperação de informações ou sistemas de categorização de textos. Na segunda categoria, as palavras-chave específicas de domínios são reconhecidas como um conjunto de palavras que não têm valor discriminante dentro de um domínio ou contexto específico, as palavras-chave específicas de domínio diferente de um domínio para outro. Um exemplo específico "aprendizagem", pode ser uma palavra-chave no domínio da educação, mas uma palavra-chave na ciência da computação ou na Engenharia Informática.

3.3 Modelos linguísticos Para identificação de perfil

O modelo linguístico é uma tarefa central para o natural processamento de idioma e compreensão, sendo que há modelos que podem colocar com precisão distribuições sobre frases, não só codificam as complexidades de linguagem como estrutura gramatical, mas também destilam uma quantidade razoável de informações sobre o conhecimento que um perfil pode conter. Na verdade, os modelos que são capazes de atribuir uma baixa probabilidade a frases que são gramaticalmente corretas, mas improváveis, podem ajudar outras tarefas na compreensão da linguagem fundamental, como resposta na identificação do perfil linguístico de um utilizador na sua tradução do texto. Os modelos linguísticos melhoram as métricas subjacentes da tarefa, como a taxa de erro dos termos para reconhecimento de fala ou pontuação, o que capacita melhores modelos linguísticos só para que o utilizador adapte melhor o seu perfil, onde uma frase no idioma é uma sequência de termos.
Assim, os modelos linguísticos são muitos úteis na sua ampla gama de aplicações, pelo que o mais óbvio, talvez seja o reconhecimento de um texto ou fala e a sua tradução, em aplicações, sendo muito útil ter uma distribuição onde p (x1……..xn), sobre quais frases são ou não prováveis numa língua [44].
Por exemplo, em reconhecimento de um texto ou de uma fala para que se identifique melhor o seu perfil linguístico, pode ser combinado com o modelo acústico que modela a pronúncia de termos diferentes: uma maneira de pensar sobre isso é que o modelo acústico gera um grande número de frases juntamente com probabilidades e, por isso, é usado o modelo linguístico para

reordenar estas possibilidades baseadas na probabilidade de serem uma frase associada ao idioma. Uma consulta é tratada como um processo de geração de dados sequenciais de termos numa consulta, calculando-se as probabilidades de geração desses termos, de acordo com cada modelo de documento; assim, a multiplicação dessas probabilidades é usada para classificar documentos que são recuperados e, quanto maior as probabilidades de geração, mais relevantes os documentos correspondentes para a consulta dada. Uma das fórmulas mas simples do modelo linguístico.

$$P(X1.....Xn) = \frac{(C(X1.....Xn)}{(N)} \quad \textbf{(3.2)}$$

N= é o número total de frases no corpo; (X1……Xn) é o número de vezes que o termo é visto no nosso corpo. Um modelo de documento não é estável no sentido de que existe um grande número de termos em falta, e pode haver distribuições anómalas de certos termos em falta, bem como distribuições de anomalias em certos termos conhecidos. Os modelos são estáveis uma vez que se obtém a partir de um grande número de documentos e, para além disso, pode-nos ajudar a diferenciar as contribuições dos diferentes termos em falta num documento. Existem duas formas gerais de combinar modelos de linguagem, a abordagem de soma ponderada (também chamada de interpolação) e a abordagem do produto ponderado [123].

Os modelos de linguagem fornecem um contexto para distinguir termos e frases que parecem semelhantes; as frases são pronunciadas da mesma forma, por exemplo, como no Inglês Americano, mas com significado muito diferente; essas ambiguidades são mais fáceis de resolver quando as evidências do modelo de linguagem são incorporadas com o modelo de pronúncia e o modelo acústico. O modelo de idioma está separado porque ele associa-se a cada documento ma sua coleção. Em geral, as interfaces modernas de reconhecimento de fala tendem a ser mais naturais e evitam o estilo de comando e controlo da geração anterior e, por essa razão, a maioria dos *designers* da interface prefere o reconhecimento da linguagem natural com modelo de linguagem estatística, ao invés de usar a gramática VXML antiga.
Posteriormente, são geradas palavras espontâneas, de acordo com o tamanho da frase original, como a sua posição na frase de destino depois da tradução; cada uma das palavras de origem é traduzida para um termo de destino, são reordenadas em função dos termos de origem que podem ser traduzidos para qualquer palavra da língua de destino, para que o utilizador possa fazer a sua tradução no sentido de melhor identificar o seu perfil linguístico.
Os modelos linguísticos têm uma relação significativa com os modelos tradicionais de TF-IDF, que são as frequências de prazo representadas diretamente em modelos TF-IDF, tendo havido muito trabalho recente para que pudesse ser reconhecida a importância da normalização do comprimento da frase.
A frequência de Documento (DF) é um esquema de classificação de termos, ou uma combinação de todos os métodos inspirados pela lei de *Zipf*, onde a quantidade de documentos nos dados é ainda alta. Nos termos comparados é uma das regras que falha, na maioria dos casos, incluindo aquele em que os documentos não estão uniforme-mente distribuídos nas categorias [90]. De acordo com a lei de *Zipf*, em que um corpus de natural de um texto de idioma, a frequência de uso de qualquer palavra é considerada inversamente proporcional à sua classificação de frequência, um texto tem uma frequência de documento baixa e alta (baixa DF e alta DF).
Um DF é uma medida que reflete a contribuição de um termo em que uma coleção de documentos, que possa ser usado na redução de *stopword* , como a DF, ignora os rótulos e informações de classe dos documentos e, uma medida de pontuação não supervisionada, que é utilizada no

agrupamento de texto. A frequência do documento inverso (IDF), que é uma das variantes do DF medida de classificação, que é calculada por diferentes formulações.

$$IDF(t^j) = log\frac{n - n_{tj} + 0.5}{n_{tj} + 0.5} \quad (3.3)$$

Nesta fórmula o n é o número de documentos no conjunto de dados de treinamento, e ntj é o número de documentos contendo o termo tj. O IDF é amplamente utilizado na remoção de altas palavras frequentes, que são consideradas como palavras-chave. Por exemplo, a regra DF≥n/2 falha quando há um domínio específico com inclinação de classe alta. A remoção de palavras comuns, sem significado relevante, pode ser uma preposição, um pronome, entre outros, pois depende do idioma. Por exemplo: Pouco se aprende com a vitória, mas muito com a derrota.

[“ pouco ”,”se”,” aprende,” “com,” “a”, “vitoria,” “mas,” ”muito” “com,” “a,” “ derrota ”].
[“pouco”, “aprende,” “vitoria,” “muito,” “derrota”].
É importante ressaltar que, para aplicar a técnica de remoção das *stopword* ,deve-se analisar o que se deseja manter do texto original, pois as palavras importantes para a aplicação podem ser consideradas *stopword* ou, dependendo do contexto, palavras que geralmente não compõem uma lista de *stopword* podem ser adicionadas. Segundo George Kingsley Zipf (1950) o termo de distribuição de frequência de classificação pode ser muito próximo de relacionamento:

$$F(r) = \frac{n}{r} \quad (3.4)$$

sendo que as nossas abordagens são inspirados pela lei de Zipf [86] com isso é possível determinar a melhor lista de palavras-chave para uma determinada coleção. A IDF usa a distribuição de frequência a termo, sendo que, na própria coleção, o valor do IDF de um determinado termo é dado por: :

$$IDFK = log\frac{(NDoc)}{(DK)} \quad (3.5)$$

Onde o NDoc é o número total de documentos no corpus e DK é o número de documentos que contém o termo K. Em outras palavras, raramente há uma maior probabilidade de ocorrência em documentos relevantes e deve ser considerado mais informativo e, portanto, mais importante nesses documentos. O DF normalizado, é a forma mais comum de ponderação de IDF e é usada por [115] pois normaliza em relação ao número de termos (NDoc-DK) e adiciona uma constante de 0,5 ao numerador e ao denominador para valores extremos moderados:

$$IDFKNorm = log\frac{(NDoc - DK + 0.5)}{(DK + 0.5)} \quad (3.6)$$

Onde NDoc é o número total de documentos na coleção e DK é o número de documentos contendo o termo K. As palavras com maior frequência de ocorrência deveriam ser consideradas pouco expressivas, porque este conjunto de palavras é [58] composto, normalmente, por artigos, preposições e conjunções; as palavras que muito raramente ocorrem deveriam ser consideradas pouco expressivas justamente em razão da baixa frequência. O TF-IDF é uma medida estatística que tem como intuito indicar a importância de um termo de um documento em relação a uma coleção de documentos. Uma das formas mais populares na formalização de ideias para matrizes de termos de documentos e palavras é a família TF-IDF (frequência de termo versus inverso

da frequência no documento) de funções de ponderação, sendo estas funções de ponderação da família TF-IDF [95] que podem produzir melhorias significativas em tarefas de recuperação de informação para um perfil de utilizador. As variações do esquema de ponderação TF-IDF são frequentemente utilizadas por motores de busca, como uma ferramenta central na pontuação e classificação da relevância de um documento, dada uma consulta do utilizador, o TF-IDF é utilizado com sucesso para a filtragem de palavras-chave em vários campos de assuntos, incluindo o resumo e classificação de textos, pelo uma das funções mais simples é calculada somando o TF-IDF para cada termo de consulta.

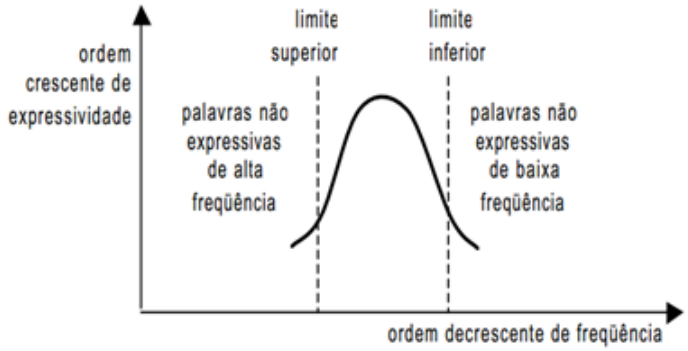


Figura 3.2: Figura Gráfico para relacionar expressividade e frequência de termos

O algoritmo TF-IDF é usado para pesar uma palavra-chave com base no número de vezes que aparece no documento. O TF (termo de frequência) de uma palavra-chave é calculado pelo número de vezes que aparece no documento. Assumindo que a palavra de ranking mais elevado (D) o corre pelo menos uma vez, temos D=AN/1, assim, a fração de palavras com frequência n é

$$\frac{In}{D} = \frac{1}{n(n+1)} \quad (3.7)$$

onde a fração de palavras que aparecem só uma vez é igual a ½. Os termos que ocorrem em muitos documentos são menos indicadores de um tópico específico, dfi= frequência documental do termo i = número de documentos que contem o termo I. Idfi= frequência do documento inversa do termo i, log (N/dfi) (Número total de documentos). O indicador de importância de um termo i num documento J, resulta da combinação tf-idf: Wij= tfij-idfi = tfij+log(N/dfi).Temos um texto de 100 palavras que contem o termo " Benfica" 12 vezes, o TF para palavra " Benfica" é tf cat= 12/100= 0,12.
Por exemplo, o termo " Benfica" apareceu 10 milhões de vezes no corpus. Vamos supor que existem 300.000 mil documentos que contêm um número tão grande de "Benfica", então o IDF é dado pelo número total de documentos que contém o termo "Benfica".
IDF(Benfica)= log(10.000.000/300.000)=1,52. Wcat=(tf-idf) cat=0,12*1,52=0,182.
Em termos experimentais, o TF*IDF tem demonstrado bons resultados na pesquisa de informação para um perfil de um utilizador. O IF-IDF é um método que foi criado inicialmente para resolver problemas de procura digitais e tem sido a técnica mais utilizada para a análise de similaridades de textos. O TF-IDF é um método que consiste em criar a representação de documentos em formato de vetores, para que a similaridade entre os seus conteúdos sejam medidas, sendo que não é apenas a tarefa de vectorização que é executada para esta representação, mas existe também a preparação do texto que basicamente se divide em etapas de *stopword* e no processo *stemming*,[94].

O processo de retirada de *stopword* consiste em retirar do texto, palavras menos relevantes no contexto, que não sejam diferenciais na análise, por serem palavras comuns a qualquer tipo de texto, por exemplo: artigos, preposições, advérbios, verbos e substantivos comuns, entre outros. Neste contexto, é importante salientar que existem listas de *stopword* disponíveis para uso, e que são frequentemente incrementadas, e dependentes da língua em que os textos são redigidos.

Exemplo de documentos para utilizador. Tratamento de termos que são muito usados numa coleção de documentos, fator TF. Quantidade de vezes que os termos A.B.G aparece no documento, fator IDF. O inverso da frequência do termo A,B,G, dentro da coleção de documentos, quanto menos usado for o termo, maior o IDF. Recuperação de documentos considerando o perfil do utilizador como consulta na base de documentos ou corpos.

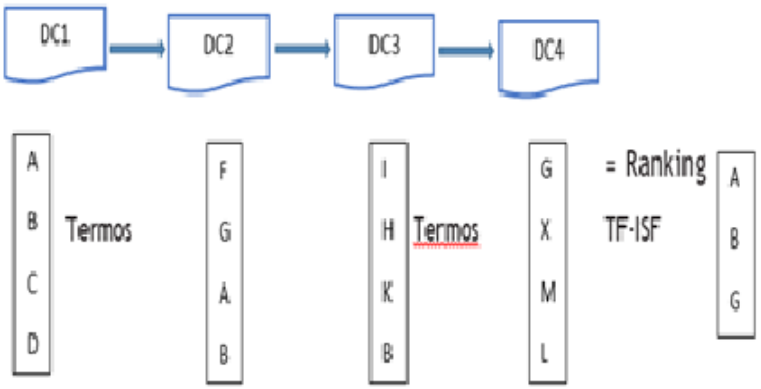


Figura 3.3: Recuperação de documentos considerando o perfil do utilizador (Fonte do autor)

3.3.1 Modelo de vetor

O modelo de espaço vetorial, ou simplesmente modelo vetorial, representa documentos e consultas como vetores de termos, estes termos são as ocorrências únicas nos documentos devolvidos como um resultado para uma consulta é montado através de um cálculo de similaridade, os termos das consultas e documentos são atribuídos um peso que especificam o tamanho e a direção de seu vetor de representação. Este modelo entende que os documentos podem ser expressos em termos de vetores que refletem a frequência de aparecimento de termos em documentos, os termos que formam a matriz que não está vazia ou seja, dotado com algum significado quando da sua recuperação de uma informação no outro lado eles seriam armazenados em forma que venha a reduzir termos de uma raiz comum, depois de um procedimento de isolamento de base que agruparia vários termos na mesma entrada.
Alguns autores venham a considerar que TF (frequência absoluta do aparecimento de um termo em um documento) venha ser como um fator que requerer uma correção, pois a importância de um termo baseado em sua distribuição torna excessiva [97]. Este vetor são representadas em todas palavras de conjunto, não somente as que se encontram presente no documento, os documentos em que o termos não contem recebem o grau de importância zero (0) e outros venham a ser calculadas através de uma formula de identificação de importância, os pesos próximos de um (1) indiquem termos extremamente importantes e pesos próximos de zero (0) caraterizem termos completamente irrelevantes em alguns casos a faixa pode variar entre -1 e 1. Os termos

de cada documentos ou texto são automaticamente atribuídos com base na frequência com que ocorrem na coleção inteira de documentos e na aparência de um termo num determinado documento.
As vantagens do modelo vetorial.
O modelo vetorial é muito versátil e eficiente quando se trata de gerar e classificações de precisão em grandes coleções, o que o torna ideal para determinar a correspondência parcial de documentos. Leva em conta as ponderações do TF-IDF para determinar a representatividade dos documentos na coleção.
Uma das desvantagens deste modelo e sendo um modelo estatístico-matemático, não leva em conta a estrutura sintática e semântica da PLN. Os pesos baseados no produto frequentes (k,s)xlog(N/nk), são chamados de abordagem tf-idf, onde cada elemento do vetor de termos é considerado uma coordenada dimensional. Assim, os documentos podem ser colocados num espaço quotidiano de n dimensões (onde n é o numero de termo) e aposição do documento em cada dimensão é dada pelo seu peso. Onde que cada dimensão corresponde um termo, e o valor do documento em cada dimensão que varia entre 0 (não presente) e 1 (totalmente relevante). e 1 (totalmente relevante).

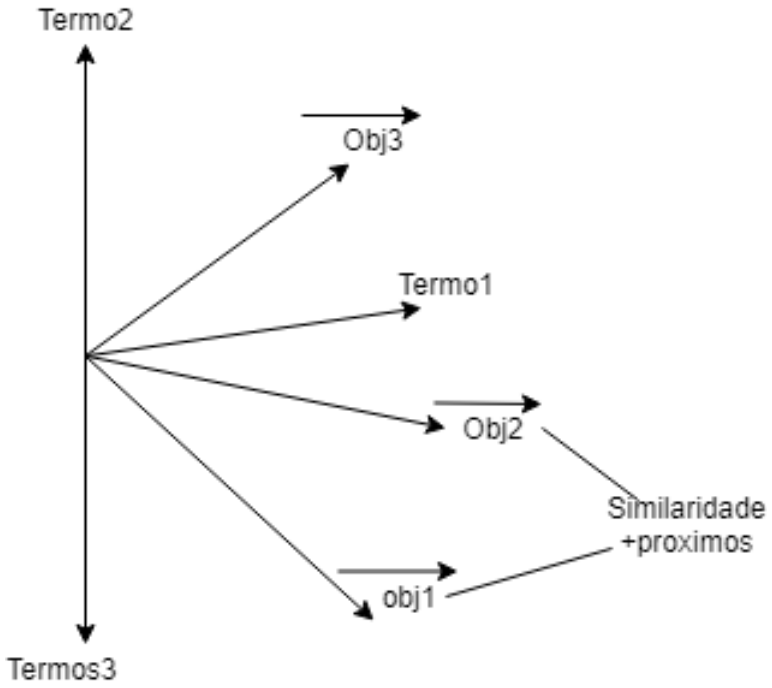


Figura 3.4: Modelo vetorial (Fonte do autor)

3.3.2 Modelo Booleano

É um dos modelos clássicos de recuperação de informação é mais adotado é baseada na lógica booleana e na clássica teoria dos conjuntos, na qual tanto os documentos a serem pesquisados quanto na consulta dos utilizadores são concebidos como um conjunto de palavras ou termos. Por exemplo o utilizador indica quais são as palavras que o documento deve ter para que seja retornado, assim todos documentos que possuem uma interseção com a mesma palavra que são retornados o mais alto é a aquele que contem todas as palavras específica na consulta do utilizador.Uma recuperação é baseada quando os documentos possuem ou não os [45] termos da consulta, exemplo: T= t1, t1,…..tj,……tm. Um dos princípios de Bayes: é o princípio de que, na

estimação de um parâmetro, poderia inicialmente assumir que cada valor Este modelo tem como vantagens, permitir processar rapidamente coleções muito grande, é sistemático e isso supõe uma grande velocidade de recuperação. Uma das vantagens é de efetuar uma recuperação de informação combinada, no sentido de que o sistema de informação apresenta a melhor resposta a uma necessidade de informação expressa por certas palavras- chaves, para o utilizador.
Uma das desvantagens do modelo booleano, em muitos casos, é as necessidades de informação complexas e isso envolve algumas dificuldades em expressar as consultas por meio de fórmulas lógicas que podem ate se tornar concatenadas. As vezes o utilizador pode impor uma semântica que não corresponde à lógica algébrica de Boole implicando um uso incorreto dos operador. Ou seja, não ordena os documentos em ordem de relevância, pois seria realizado em um modelo baseado em ponderações ou ponderações dos termos.

3.3.3 Modelo probabilístico

Este modelo foi desenvolvido por Robertson e Sparck Jones, e introduzido entre 1977 e 1979 é conhecido como modelo probabilístico ou independência binária, baseia-se na representação binária de documentos, como no modelo de recuperação booleano, e que indica a presença ou ausência de termos com 0 e 1, está diferença encontra-se no método estatístico e pressupostos em que o seu funcionamento é estabelecida, neste modelo, busca-se a saber a probabilidade de um texto x ser relevante a uma consulta y, no caso os termos específicos por esta apareçam, este método tem as seguintes afirmações de acordo as necessidades do utilizador.
De acordo com as pesquisas levantada pelos utilizadores, os textos da coleção são classificados em grupos;1) Conjunto de textos Relevantes e 2) Conjunto de textos irrelevantes.
Existe uma consulta ideal, que é a aquela que fornece um conjunto de resposta ideal ou o que é o mesmo conjunto de documentos relevantes para o utilizador. O objetivo do modelo probabilístico é para levar a consulta do utilizador para ser definido sucessivamente até que toda a resposta ideal, por reformulação sucessiva dos termos de sua consulta para poder empregar termos de ponderação, onde que os processos de passagem os termos da consulta estão a calcular a probabilidade de que o prazo em todos os documentos relevantes e a probabilidade de que está presente em todos os textos. Este modelo tem como uns dos objetivos de facilitar a localização de informação em textos escritos por ser humanos e para humanos para que venha facilitar o utilizador na adaptação.

$$(d, p) = \frac{P(r/d)}{P(R/d)} \quad (3.8)$$

em que:
r= Conjunto de termos relevantes.
R= Conjunto de termos não relevantes.
P(r/d)= Probabilidade de termos d ser relevante para a interrogação p.
P(R/d)= Probabilidade de termos d ser não relevante para p.
As vantagens deste método residem na independência dos termos da consulta, e atribuir pesos aos termos, permitindo recuperar os documentos que provavelmente que são relevantes, este método a sua forma de recuperação é por correspondência parcial, superando o método de correspondência exata do modelo booleano.
As desvantagens deste método ele mantém o modelo binário de recuperação de informação, não levando em conta todos os termos do documento como no modelo vetorial, por sua vez que atribui pesos para os termos, o que lhe permite recuperar textos suscetíveis de ser irrelevantes,

requer alta capacidade de computação, sendo complexo de implementar, este termo não leva em consideração a frequência de ocorrência de cada termos no texto. O calculo de Relevância de termos basicamente nem todos termos presentes em documentos possuem a mesma importância, os termos mais frequentes que são utilizados tem o significado mais importante, assim como os termos constantes em textos ou em outras estruturas, por sua vez que provavelmente foram colocados por serem considerados ou descritivas para a ideia de documentos no calculo de termos relevantes de um termo dentro de um documento com objetivo de obter um peso referente ao uso do termo dentro do texto. Uma vez que todos termos não tem a mesma importância no texto, onde o cálculo de relevância é realizado para destacar as que tenham mais significado.

3.3.4 Modelo Difuso (Fuzzy)

Este modelo permite a classificação de documentos, e admite o conceito de uma abordagem difusa para [131] a classificação, entende-se que o processo de classificação envolvem n categorias, considera a imprecisão em seu processo, este método tem como resultados de classificação permitira atribuir graus de pertencimento do documento para cada categoria.
Os documentos deste método são representados por vetores de palavras com seus respetivos graus de relevância as suas diferenças estão no conceito relacionado à relevância, a diferença é que presença pode ser medida e pode não ser exata, ou seja, pode haver incerteza, a um conjunto vazio, mais sim um conjunto cujos elementos possuem uma relevância importante que é muito baixa p próxima de zero. A teoria difusa permite trabalhar com os valores intermediários que indicam o quanto determinado objeto pertence ou não ao conjunto, pois esta foi construída com a finalidade de tratar incertezas e imprecisões. Este modelo é realizada a partir de um documento pré processado, que tem similaridade e certeza que são calculados de acordo com a definição de uma base de termos, que vem ser como entrada de modelo que o modelo fuzzy.

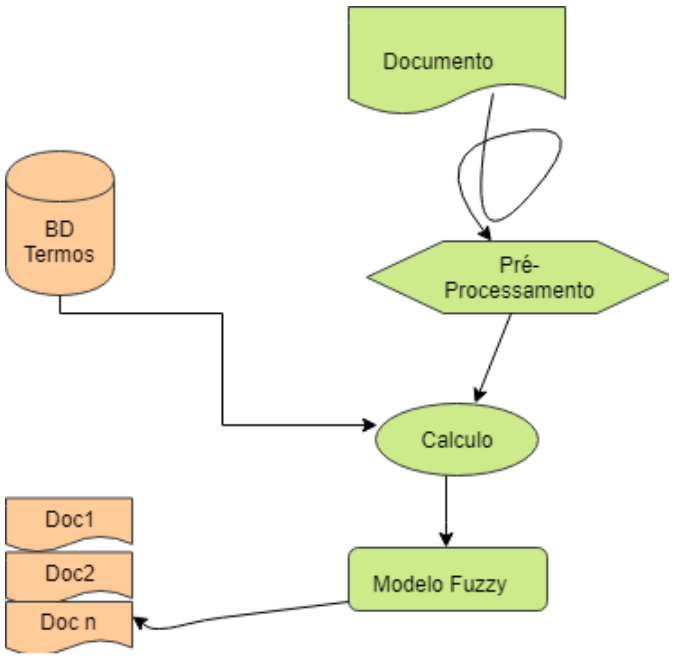


Figura 3.5: Modelo Organizacional de Documentos (Fonte do autor)

3.3.5 Modelo Busca Direta

Apesar da grande aplicabilidade dos algoritmos de busca padrão, a utilização de um numero finito de direções de busca implica que o seu desempenho pode não ser satisfatório no caso em que a função objetiva do problema e não é isto ocorre pois se a derivada da sua função em alguns pontos. O modelos de busca direta também são conhecido como Modelo de Busca de Padrões *(pattern search)*,[96] este modelo utiliza métodos de busca de *strings* para localizar documentos ou textos relevantes para o utilizador na localização da *strings* em documento e textos, as buscas são realizadas diretamente nos textos que devem ser originais, em tempo de execução, o seu resultado da busca é realizado em todas as ocorrências do padrão de consulta em um documento ou conjunto de documentos ou textos. O modelo de busca a sua utilização é aconselhável em caso onde a quantidade de documentos é pequena sendo muito utilizada em *softwares* de edição de documentos para que o utilizador possa localizar palavras ou expressões no texto que está escrito ou lendo.

3.3.6 Modelo Aglomerados *cluster*

Este modelo permite consistem em primeiro a escolher um cetroide aleatoriamente e adicionar alguns ruídos, estes métodos resultará em aglomerados de forma esférica na qual o armazenamento em *cluster* baseado em modelos pressupõe que os dados são gerados por um modelo e tentam recuperar o modelo original onde que a uma contribuição de documentos a *cluster* .
O funcionamento deste modelo consiste em identificar documentos de conteúdos similares que tratem de assuntos parecidos e armazenamento ou indexa-los em um mesmo grupo ou aglomerado *(cluster)*, a identificação de documentos similares em conteúdo dá-se pela quantidade de palavras similares e frequentes que eles contem, quando o utilizador especifica as suas consultas, o sistema identifica um documento relevante e retorna para o utilizador todos os documentos pertencentes ao mesmo grupo [71].

3.3.7 Modelo lógico

Este modelo permite que os linguistas descrever as suas regras de reconhecimento de padrões estruturais, usando um formalismo gramatical específico, onde os padrões estruturais são definidos por regras e pelas informações armazenadas nos dicionários de computadores.
Este método baseia-se em métodos e teorias provenientes da lógica matemática para modelar o processos de recuperação de documentos, basicamente este modelo funciona tornando-se necessário modelar os documentos através de lógica predicava o que exige um grande esforço no trabalho de modelagem incorporando semântica ao processo de recuperação, para que venha a julgar melhor a relevância para o seu utilizador.
Nesta pesquisa vamos nos limitar no método não supervisionados na qual este método que é a frequência do Documento (TF-IDF) é o número de Documentos na qual o termo ocorre em um conjunto de dados ou documentos, é um dos critérios mais simples para que permite uma seleção de termos e de textos que podem ser dimensionar para um grande conjunto de dados com a complexidade computacional e na categorização do texto na facilidade do utilizador [73].
A primeira forma de frequência de termo (TF) é dada a Hans Peter Luhn (1957) baseou-se na suposição de Luhn que :" o peso que um termo ocorre em um documento é simplesmente proporcional à frequência do termo" este termo mede a frequência com um termo que ocorre em um texto ou documento [102] .
O TF-IDF é um peso frequentemente utilizado na recuperação de informação e de mineração de

texto, esta frequência é uma das medidas estatísticas utilizadas para avaliar o quão importante é uma palavra para um documento em uma coleção ou texto, na qual aumenta a importância de forma proporcional o numero dez que uma palavra aparece no documento, mas é compensada pela frequência da palavra, venha a distinguir o fato da ocorrência,de algumas palavras a serem geralmente mais comuns que as outras no texto ou documento.

3.4 Pontuação

Falando da pontuação que está presente na escrita desde o século II a.C, quando, de acordo com Caglairi (1995), foi introduzido o primeiro sistema de pontuação por Aristófanes de Bizan- cio, a pesar disso, até os dias atuais, o uso dessas marcas apresenta grande flutuação, o que pode ser explicado por razões históricas [105]. A importância dos sinais de pontuação no texto escrito que mostram que a ausência ou a alteração de pontuação compromete a compreensão de textos e o reconhecimento de palavras, o que permite a firmar que a presença desses marcadores vai além de uma questão de estilo, onde que um texto escrito, então, os sinais de pontuação têm importância semântico-sintático-discursiva.

Na Idade média, duas orientações principais para o uso da pontuação já ocorriam, a pontuação como uma função semântica, respondendo à necessidade de clareza e de lógica, e a pontuação com uma função prosódia a tendo as pausas para respirar [79]. Conceber que a pontuação pode também desempenhar uma função prosódia não significa, que no entanto, encara-la como sendo essa a sua função primordial, para que podemos conceber essa função, como sendo a função essencial seria quando se afirma que existe reciprocidade total entre respiração e pontuação, dessa forma, cada utilizador com o seu perfil linguístico é completamente livre para pontuar, o que não condiz com o uso moderno do recurso, existe flutuações, restrições que não são determinadas pela prosódia, mas sim pela estruturação e característica do texto escrito pelo utilizador.
A pontuação é um ramo ou é um recurso da ortografia, que permite expressar exclusivamente na língua escrita um espectro de matizes rítmicas e melódicas características da língua falada, pelo uso de um conjunto sistematizado de sinais sintáticos. Os sinais de pontuação são símbolos que usamos na língua escrita e que tem por função organizar e estabelecer uma hierarquia entre as informações de texto, alguns deles também nos ajudam a identificar a intenção comunicativa do autor, ou seja, o que o autor do texto quis dizer em certos problemas.
Uma das funções da pontuação é a de organização sintática, ou seja, a pontuação a tua no sentido de unir e separar partes de um discurso realizado e textos para umas junções, disjunções, inclusões, exclusões, dependências e hierarquizações no âmbito da organização do texto escrito, auxiliando o utilizador ele percebe as relações entre as partes do texto, com essa forma, a pontuação atua como recurso coesivo no texto. Uma das funções é a de suplementar semântica, que na realidade, está subjacente as propriedades relacionadas aos efeitos da enunciação criam algumas dificuldades por implicarem em escolhas ocasionadas pelos efeitos de sentido, que se pretende provocar no texto, utilizador de texto pode em grande parte das ocasiões, selecionar um entre vários signos que podem ser utilizados, mudando em diversos momentos, a estrutura das sentenças, é nesse sentido que se concebe que existe uma grande flutuação ou liberdade no pontual.
Por um lado a flutuação e as decisões dependendo do domínio ou competência do utilizador em refletir sobre a estrutura do texto, sobre as relações entre as diversa partes que o compõem

e sobre os efeitos de sentido que se pretende provocar, por outro lado essas decisões e flutuações estão também nas características dos géneros textuais a serem produzidos, os textos são predominante-mente declarativos, onde que predominam o uso de virgulas e pontos finais, numa outra forma em que os textos, como aqueles em que ocorrem trechos conversacionais, onde utiliza-se frequentemente interrogações, exclamações e entre outros. Em função das diversidades abordados é necessário os utilizadores sejam com versatilidade em que utilizador, que precisa utilizar seus conhecimentos textuais para tomar decisões a cerca de uma estrutura do texto e dos caracteres de pontuação a utilizar.
Sabendo que um dos objetivos do ensino é facilitar a aprendizagem, os textos tem de estar adequados às capacidades de processamento dos utilizadores não se figurando uma tarefa demasiado difícil, logo o trabalho em torno das seleção textual terá de ser controlado. A escrita é um sistema de estruturação para o uso da língua falada, sempre contextualizado no entanto, a condição básica para o uso escrito da língua que é a apropriação dos sistema alfabético envolve, da parte dos utilizadores ou do publico alvo, que aprendizados muito específicos, independente do contexto de uso, relativos aos componentes do sistema fonológico da língua e as inter-relações como a pontuação no sistema adaptativo para um perfil linguístico.
Um sistema de pontuação na fala e na escrita permanece a mesmo independentemente do género textual e da esfera social em que ele se apresenta, numa piada ou num processo jurídico, económico, cientifico, e outros processos o sistema de pontuação utiliza as mesmas regras [47].
Os sinais de pontuação são marcações gráficas que servem para compor a coesão e a coerência textual, além de ressaltar especificidades semânticas e pragmáticas.
Os sinais de pontuação podem ser classificados em dois grupos: o primeiro grupo compreende os sinais que, fundamentalmente, se destinam a marcação as pausa:

1. A vírgula (,);
2. O ponto (.);
3. O ponto e vírgula (;);

O segundo grupo abarca os sinais cuja função essência é marcar a melodia, a entoação:

1. Os dois pontos (:);
2. O ponto de interrogação (?);
3. O ponto de exclamação (!);

Um exemplo no uso de sinal de pontuação usar o ponto de interrogação (?), o utilizador do texto quis indicar que a frase, " mas qual o sentido de tudo isso?" é uma interrogação (pergunta).Onde cada tipo de frase, que usamos ao final é um sinal de pontuação diferente, assim nas frases assertiva (aquelas que são usadas para uma informação ou fazer afirmações ou negações), emprega-se o ponto final (.) Exemplo: Os jogadores do 1°de Agosto passaram para fase de grupos dos clubes campeões de África na África do Sul no dia 17/03/2018. Nas frases interrogativas (aquelas que servem para fazer uma pergunta, um convite ou um pedido), usa-se o ponto de interrogação (?). Exemplo: como será que os do Benfica estão contente com a vitória? Nas frases exclamativas (aquelas que usamos para expressar admiração, sentimento, ordem ou chamada),o sinal empregado é o ponto de exclamação (!). Exemplo: Ei! Não faça isso!.
Fórmula de Cálculo de Pontuação.

$$RP(x) = \frac{N(x)}{TP} \quad (3.9)$$

RP(x)= Ranking de pontuação existente no texto.
N(x)= é o numero de pontuação existente no Documento .
(TP)= Total de pontuação existente no texto.
Exemplo:

$$RP(!) = \frac{6}{13} = 0,461 orankingdepontuaesqueseencontranodocumento \quad (3.10)$$

Um dos problemas mais marcantes das línguas oral é, sem dúvida, o facto de que as pausas não corresponderem, de modo nenhum, aos nossos hábitos de pontuação na escrita, é por isso que a muitos pesquisadores que simplesmente abdicam da pontuação que á convencional, optando por sistemas de representação mais abstratos, por outro há, no entanto, que preferem pontuar os textos de ortografia [113] na verdade os textos não pontuados tornam-se praticamente incompreensíveis. Portanto, os sinais de pontuação para os utilizadores servem para integrar os sintagmas, viabilizando a organização textual, sem estes sinais a compreensão de um texto pode ficar seriamente comprometida. Em consonância com as afirmações acima, a pontuação torna-se um elemento pertinente na construção de significados, tendo em vista a necessidade da colocação de sistemas icónicos que possibilitem o sentido desejado, mesmo com a expressiva multiplicidade de utilizadores nas suas leituras ou escrita do texto. Sob outros viesses, os sinais de pontuação estão ligados ao estilo de cada utilizador e, principalmente, aos efeitos de sentido que podem ajudar a expressar.
Ao explicarmos a funcionalidade da pontuação tomando como parâmetro o sujeito que interpreta a materialidade linguística, temos que considerar a incompletude dos sentidos, pois os enunciados não são estáticos, mas estão em constante interação. Desta forma, não podemos considerar os sinais de pontuação como um elemento dificulta-dor na aprendizagem de língua ou na adaptação de um perfil mas como instrumento que pode nos auxiliar na legibilidade do texto escrito.
Falando da pontuação a partir da análise na adaptação do perfil linguístico de um utilizador incluímos que a heterogeneidade discursiva é uma das formas que acarretam os efeitos de sentidos nos discursos, no qual a heterogeneidade mostrada tem papel relevante neste processo, sob esta perspetiva, os sinais de pontuação tem papel importante, pois evidenciam as vozes presentes na escrita e no discurso.
Nas pesquisas sociolinguísticas tem buscado traçar um perfil da mudança em progresso e um perfil da variação estável através da combinação dos resultados das variáveis raça, etnias, idade, classe social e nível de escolaridade, a partir da noção de um prestígio, no que concerne à faixa etária, a variação estável se carateriza por um padrão no qual as faixas intermediariam apresentariam a maior frequência de uso das formas de prestígio, já na mudança em progresso, a distribuição seria inclinada, com os mais jovens que deve apresentar a maior frequência de uso das formas inovadoras.

3.5 Comprimento de frases e de palavras

O comprimento de frases é um conjunto de diretrizes venham a firmar que uma frase longa pode ser de tamanho médio é aquela que contem de 12 a 17 palavras, mas pelas outras fontes eles relatam que uma frase longa é aquela que contem a media de 25 a 30 palavras. Curiosamente, as frases em alguns textos submetidos, ou publicados que são escritas por não-nativos da língua inglesa são geralmente mais curtas do que as escritas por nativos da língua inglesa. Com essa

diferença pode ser atribuído à tendência de nativos da língua inglesa em usar sentenças mais longas e complexas, alem disso, como os nativos da língua inglesa sentem menos pressão com relação à qualidade da escrita, é por isso que eles acabam por ser menos diligentes.
Para tentarmos adaptar cada sentença para conter um certo número de palavras, concentra-se na quantidade de informações transmitidas em cada sentença que não seja muito longa ao chegar ao utilizador não se lembra de como a sentença tem o início. Uma das alternativas é pensar nas paradas entre as frases como sendo lugares na narrativa onde o leitor deve pausar e analisar as informações apresentadas, antes de continuar lendo o resto do texto.

As frases longas têm vantagem de fluir melhor, mas exigem mais concentração por parte do utilizador, e por isso funcionam melhor quando o interesse do utilizador é aguçado, por outro lado frases curtas a traem mais a atenção, e são ideias para manter seu utilizador corajoso na leitura [109]. Os métodos que extrai as seguintes informações relacionadas com palavras e frases, onde que o número de palavras é diferentes o número de frases e frequência de palavras, os parâmetros extraídos são utilizados para que possamos calcular as informações presentes no conjunto de parâmetros relativos a medias e frequências e estão relacionados com o comprimento do texto, em por outros estudos sobre a inteligibilidade apresentados, as frases longas contem normalmente, mais informação estes por sua vez dificultam a tarefa de compreensão.
Uma das maiores partes dos textos de legibilidade consideram a estrutura de comprimentos das frases e palavras na qual a sua percentagem de palavras menos vulgares do vocabulário, podendo enquadrar-se em modelos aditivos ou multiplicativos como apresentado nas seguintes expressões, onde os literais que são constantes: Legibilidade= (Comp.frase) + (comp.palavra) + (palavras difíceis). As determinações do comprimento das palavras deixam-nos alguns problemas, sendo que um deles prende-se como o uso do hífen, na língua Portuguesa. O hífen é usado para a translineação das palavras, mas também para algumas palavras compostas, como por exemplo " fim-de-semana".
As palavras mais comuns são as mais fáceis então algumas métricas medem a legibilidade do texto pela sua percentagem de palavras. Observou se também que, com frequências, as palavras que ocorrem são muitas vezes curtas, então a palavra comprimento foi usado para aproximar a legibilidade mais robusto do que usar uma frequência de palavras predefinida no perfil linguístico.
O comprimento das palavras é utilizada (isto é o número médio de carateres por palavras), e o comprimento das frases são utilizadas (que consiste no numero de palavras por frase) são estes elementos linguísticos que é muito variáveis de autor para autor, ou de acordo com o tipo de escrita, tomando como exemplo o comprimento das palavras são as palavras mais longas que estão normalmente, associadas a uma escrita mais formal ou a textos escritos [121] enquanto a utilização de palavras mais curtas se encontra relacionada com uma escrita mais informal ou nos textos orais, alguns autores venhas apresentar preferências diferentes, sobre tudo em termos de comprimento das frases, estes marcadores linguísticos possuem capacidade discriminatórias, constituindo-se, por isso, como marcadores fiáveis e válidos na análise de autoria, ainda o comprimento das frases é um marcador de discurso particularmente significativo, uma vez que se encontra, normalmente, para que efetivamente o utilizador se decidir onde colocar a pontuação e, assim, continuar ou terminar a frase.
Um comprimento de frases ou de palavras é uma linha de texto muito longa para os olhos do utilizador terão curto tempo de foco no texto, isto porque o comprimento da linha faz com que seja difícil avaliar para se adaptar do perfil para onde começa e termina, também pode vir atrapalhar na hora de mudar de uma linha quando podemos confundir com a próxima. Em

relação para uma palavra ou frase curta se uma linha é muito curta e venha quebrando o ritmo de leitura, uma frase ou palavra muito curta também tendem a stressar os leitores, fazendo-os começar na próxima linha antes de terminar a atual.
O método extrai informações relacionados com termos e frases onde que os números de termos ou frases diferentes e com frequências de termos, os primeiros parâmetros extraídos são utilizados para o calculo de informações que são presentes no conjunto de parâmetros relativos a média e frequências estes estão relacionados com comprimento das palavras, na pontuação e comprimento das frases, com estes parâmetros citados, as motivações linguísticas para a utilização dos parâmetros é o facto de um vocabulário que é utilizado com mais fácil se torna a sua compreensão.
Em 1948, flesch modificou a sua fórmula inicial e por sua vez separou-a em duas partes (Flesch,1948), a primeira parte, foi usada formula *Reading Ease*, deixou se utilizar afixos tendo em conta apenas duas variáveis, o número de silabas e frases para cada amostra de um número de palavras. Segunda parte da formula calcula o interesse na matéria pela contagem do numero de palavras (tais como pronomes e nomes) e frases pessoais (tais como frases incompletas), a formula *Flesch Reading*, [53]. Com está fórmula vai nos permitir calcular o comprimento de palavras e de Frases, temos a seguinte formula a baixo.
A fórmula matemática específica é:

$$RE = 206,835 - (1,015 \times ASL) - (84,6 \times ASW) \quad \textbf{(3.11)}$$

RE = facilidade de legibilidade.
ASL = Comprimento Médio da Frase (ou seja, o número de palavras dividido pelo número de sentenças).
ASW = Número médio de sílabas por palavra (ou seja, o número de sílabas dividido pelo número de palavras).
A saída, ou seja, RE é um número que varia de 0 a 100. Quanto maior o número, mais fácil será o texto para ler.
A pontuação tem limite inferior, isto é, pode tomar valores negativos, no caso de uma palavra com muitas silabas.
Ainda temos uma outra fórmula que nos permite calcular o comprimento de palavras e de frases. E a seguinte.
Exemplificando uma outra fórmula do nosso trabalho para o cálculo de comprimento de palavras e de Frases recorrendo a valores numéricas.
Comprimento da palavra 1(W1)=8, TF(W1/R)=5, TF(W1|c|)=10 Comprimento da palavra 2(W2)=3, TF(W2|R)=20, TF(W2|C|)=300 |C|=1000 termos, |R|= 100 termos. I(W|R)= Comprimento de palavras* log2 $\frac{P(W|R)}{P(W)}$ Sendo que o comprimento de palavra é P(W)= FRE$\frac{W|C}{|C|}, P(W|R) = \frac{f}{r} eqW|R|R|$
Onde I(W|R) é o valor da importância de cada palavra de qualquer vocabulário, e a formula P(W|R) é a probabilidade da palavra W que ocorre em uma parte (R),P(W) é a probabilidade da palavra a que parece no corpus e |C| é o tamanho do corpus com frequência (W|R) é a frequência da palavra W na região e |R| é o tamanho de da região.

3.6 Abordagem adaptada para a extração de termos

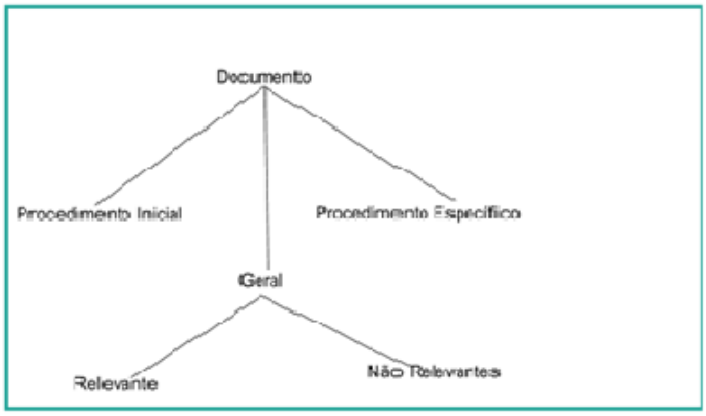
O método proposto neste trabalho não supervisionado e independente da Língua, o que implica que um dos principais requisitos é que tem a capacidade de extrair conhecimento tanto explícito como implícito de textos em linguagem natural e não estruturados. A depender da unidade textual básica a utilizar, os dois tipos de conhecimento podem ser extraídos. A aprendizagem não supervisionada pretende extrair informações sem auxílio humano, não existindo por isso uma necessidade de supervisão; diferente da Supervisionada.

Entender o significado (semântica) de textos não é uma tarefa trivial, representando um dos principais problemas em tarefas de PLN. É necessário quantificar e decidir quão semelhantes ou não, são o significado de dois textos. Identificar diferentes tipos de semelhanças entre palavras representa um desafio importante em PLN e algumas abordagens têm sido úteis no que diz respeito ao cálculo do grau da similaridade entre termos ou palavras. Na abordagem estatística, uma palavra é representada por um vetor de coocorrência de palavra em que cada entrada cor- responde a outra palavra no léxico.

É uma seleção de elementos importantes de um texto que podem ser termos, ou relações neste caso a extraindo termos em documentos no texto são palavras corretas em uma consulta, podem ser facilmente de saber quais documentos ou textos que devem ser lidos e que documentos devem ser postos ao lado, por enquanto a extração de documentos tem sido um campo de busca ativa pela comunidade cientifica, a extração de termos isoladas de n-gramas, tem sido basicamente ignorada devido á sua dificuldade. Os métodos atuais que foi proposto nesta secção é um dos métodos de última geração para extração de n-gramas foi proposto na extração de termos e palavras-chaves no documento que tem frequências e com um Ranking mais elevados, basicamente para que podemos encontrar os termos não relevantes= gerais é os relevantes= específicas.

A deteção de Textual é um desafio recentemente reconhecido no domínio da PlN e um dos mais exigente, neste facto, os sistemas participantes devem provar sua capacidade de entender como trabalhos de linguagem, vem afirmando em reconhecimento da vinculação tem semelhanças com o famoso teste de Turing para avaliar se as máquinas podem pensar, como o a cesso a diferentes fontes de conhecimento e a capacidade de desenhar inferências parece estar entre os principais ingredientes para um sistema inteligente com isso a PLN as suas tarefas tem uma forte ligações com vinculação, em particular, o desafio termo listou o seguinte. Dentro da *Summarization(SUM)*, um resumo deve ser vinculado pelo texto, em para frases (PP), que podem ser vistas como mútuas acarreta mente entre um texto X e uma hipótese de Y, em *Information Extraction* (IE), a informação extraída deve ser implicado pelo conteúdo do texto, enquanto na Tradução Automática (MT), a tradução automática deve, no máximo, ser implicado pela tradução humana padrão dourada, em documento comparável (CD).

Este método é representado numa estrutura de árvore.



fonte: do Autor .

Figura 3.6: Demonstração da árvore (Fonte do autor)

Depois de representar o método em grafo, demonstramos em detalhe o processo da estrutura proposta.
Com este método teremos:
-Pré-processamento do texto.
-Identificação de possíveis ocorrências relevantes.
-Identificação de termos relevantes.
-Identificação de termos não relevantes.
-Extração de termos relevantes e não relevantes no texto.

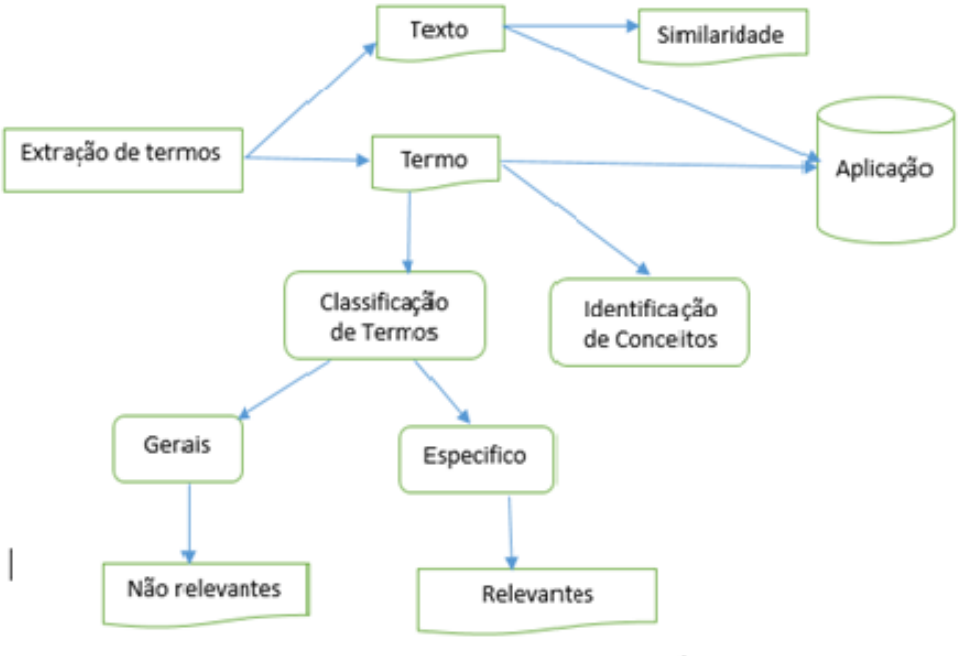


Figura 3.7: Fonte: Elaborado pelo autor (de talhe do método proposto)

Falar dos corpos compostos por vários documentos, um dos objetivos é para tentar entender quais os termos são relevantes e quais não são relevantes e qual é o *ranking* no usando métodos estatísticos, este tipo de classificação nem sempre é simples ou mesmo exato porque, embora

a noção de relevância seja um conceito fácil de entender, normalmente não há um consenso sobre o que se para a relevância da não relevância para que o utilizador possa intender dar valores a estes termos.
Por exemplo termo como “África” ou “Luanda” têm uma relevância muito significativa e termos como “ou” e “deste” não tem relevância alguma, mas o que dizer de que termos como “comer” “terminar” e o próximo”. Estes tipos de termos são problemático porque geralmente não há consenso sobre o seu valor semântico, portanto ainda a problemas sobre a relevância das palavras. No que diz respeito a contexto do nosso trabalho decidimos adotar um método de classificação de termos onde que os termos variam em gerais e específicos.
Os métodos incluindo em que nós chegamos de mencionar neste trabalho são capazes de identificar, até certo ponto, os termos relevantes em textos eles são, no entanto, incapazes de tomar decisões sobre a verdadeira relevância dos termos. Um dos problemas de alguns métodos só podem criar relevância ranking, dos quais só podemos identificar, por exemplo, que um determinado termos no top do texto deve ser mais relevante do que um termo na parte inferior, portanto, em algumas situações, pode ser necessário saber se um determinado termo é realmente relevante ou não relevante. Este tipo de método é absolutamente necessário para aplicações como no documento onde desejamos um conjunto de termos que realmente vem descrevendo um documento ou um conto de documentos. Um termos dentro de um documento é considerado relevante se ocorrer mais frequentemente do que outros termos em um com uma certa categoria, mas com ocasionalmente em outros lugares do documento. Lunh nas suas pesquisas também sugeriu que os termos com uma frequência muito alta de ocorrência são geralmente consideradas termos comuns e termos não frequentes poderiam ser consideradas raras, sendo termos não relevantes, embora esta abordagem pareça bastante intuitiva, não é necessariamente verdade [129].
A abordagem de extração de termos é uma das primeiras observações muito importante no que diz respeito à extração de termos, é o fato de que existem diferenças entre extração de termos simples e os termos com uma única palavra, e extração de termos compostos. Um termos composto é um conjuntos de duas ou mais palavras que possui um significado comum, e que por sua natureza são mais difíceis de detetar do que termos simples (com única palavra). De outra maneira devido à importância da qualidade na extração de termos, muitas pesquisas científicas dedicam-se a aperfeiçoar esse processo, e com a PLN, as abordagens para a extração de termos se dividem em abordagem estatística e linguísticas.

3.7 N-Gramas

Neste trabalho consideramos sequências de palavras do tipo Tri-grams que é o método proposto na qual podemos analisar os termos isolados, perde-se em alguns contexto da informação e chegamos a conclusão que desse problema passa pela implicação do Uni-Grams, que vai nos dar a sequencia dum termo num comprimento n, o n-grams é uma sequencia de caracteres consecutivos existindo em n tamanhos em que pode ser (n=1….3) onde que cada tamanho pode ter um significado de um nome Uni-Grams para 1 caráter Bi-grams para 2 carateres, tri-grams para 3 carateres, e assim sucessivamente. A pontuação final é o resultado da soma ponderada dos métodos dos diferentes valores da precisão, para n variando 1 a 3, para que não é aconselhável usar valores de n desde coincidências mais de três grams são muito incomuns [110].
Este método n-grams tem uma das grandes vantagens no nosso trabalho, permite uma captura raízes dos termos e palavras que são mas frequentes, independência em relação à linguagem

do documento, que é diferente de outras técnicas que usam dicionários específicos, para cada idioma. Este método n-grams permite tolerâncias com os erros ortográficos e as de formações. Considerando a importância dessas vantagens do método n-grams estes são usados em vários campos. A nossa abordagem neste trabalho é uma abordagem baseada nos n-grams, que tem a vantagem de ser independente em relação à língua dos textos e opera sem qualquer pré-processamento isso é independente da língua.
Segundo a nossa análise pode-se considerar que todos os termos no topo do ranking são relevantes, e quase os termos no fundo não são. Mas com isso, causa dois problemas, um dos primeiros problemas é para definir o espaço que separa os termos relevantes dos não relevantes, onde se estivermos muitos alto no texto provavelmente perderíamos termos relevantes para a falta de relevante.
Quando o *ranking* fosse muito baixo, seria o oposto, o outro problema é que embora o termos no ranking podem ser genericamente comparadas entre si, ou seja, podemos dizer que um determinado termo n no topo da classificação é mais relevante, que um certo termo y na parte inferior, onde que não podemos dizer que o termo n não é relevante, mesmo se estiver no final do texto. Isto porque quando temos n tem uma pontuação alta de relevância e por isso deve ser muito relevante em que todo o texto n pode ser muito relevante apenas em um contexto local, obtendo uma pontuação menor no final da lista, onde que sabemos existe método que extrai termos relevantes sobro a base, dos termos.
Defende-se que informação capturada por n-grams é, em grande medida, apenas um reflexo indireto das relações lexicais, sintáticas e semânticas na linguagem. E isso ocorre porque a produção de sequências consecutivas de palavras é o resultado de estruturas linguísticas mais complexas, e neste sentido os modelos n-grams demonstraram ter vantagens práticas por ser fácil formular modelos probabilísticos para eles, é possível extrai-los de um Corpus com baixo grau de dificuldade e, acima de tudo, provaram fornecer estimativas de probabilidade úteis para leituras alternativas da entrada.As similaridades de palavras obtidas por dados de n-grams podem refletir uma mistura de semelhanças sintáticas, semânticas e contextuais. Tais semelhanças são adequadas para melhorar um modelo de linguagem n-gramas, uma vez que os mesmos abarcam estes tipos de relacionamento.

3.8 Sentença

Sentença permite localizar o texto que fornece a classificação das sentenças, neste método pode conter varias substâncias na localização de termo no documento, para que primeiro termo seja classificada a pontuação mais alta. No texto as sentenças são identificadas por meio de regras simples baseadas na ocorrência de sinais de pontuação. Com esses método vem nos contribuir na pontuação às sentenças produzem um ranking, todas as letras das sentenças são transformadas em letras maiúsculas, onde que os termos do texto são substituídas pelas suas respetivas raízes.
A remoção de palavras-chave (que são termos muitos comuns e porque são irrelevantes para o processamento em questão) podem ser removidos no texto como uma pontuação de sentenças basicamente a sentença de pontuação pode ocorrer por alguns métodos mais nós vamos nos limitar no método TFISF *(Term Frequency–Inverse Sentence Frequency)*, o número de sentença que é selecionada para formar o texto, por sua vez depende da taxa de compressão específica pelo utilizador do sistema, a taxa de compressão é uma medida que determina o tamanho do texto. Uma sentença é dividida em sequências de termos contiguas em delimitadores de frase e para as

posições de termos. Neste método existem várias maneiras de calcular uma similaridade entre duas sentenças, na sua maioria das medidas de similaridade entre termos associadas a sentença ou seja modelo de espaço vetorial, no entanto na sua aplicação as medidas clássicas de semelhança entre dois termos apenas os índices do vetor de linha, para lidar com estes problemas foram propostas avaliação das métricas de similaridades [106].
Em vez de confiar nas correspondências exatas dos termos entre os textos, propomos que sentenças inferior a outra em termos de generalidade se duas restrições forem respeitadas: se e somente se ambos termos compartilham muitos termos relacionadas e (b) se a maioria dos termos de uma frase e mais geral do que os termos da outra sentença, até onde sabemos, somos os primeiros a propor um projeto não supervisionado dependente da língua, é uma metodologia independente de similaridade livre de língua no contexto da implicação textual por generalidade, embora a abordagem de que seja baseada em suposições similares, está nova proposta é exaustivamente avaliada em relação ao conjunto de dados estando diferentes medidas de associação assimétrica.
Os documentos extrativos visam encontrar os termos mais significativos em um determinado documento, para que uma pontuação de significado deve ser atribuída a cada sentença com maior significado naturalmente se tornam relevantes com uma taxa de compressão define o número de sentenças a serem comparadas na identificação de termos gerais e específicos no documento ou texto [88].
Exemplo de uma semelhança de termos:

1. Geraldo derrotou o guarda-rede mais uma vez.
2. O atacante do 1º de Agosto marcou novamente

Outro exemplo de uma semelhança de palavra:

1. As Correias do Norte e do Sul são países da Ásia que assinaram um tratado para acabar formalmente a Guerra da Coreia.
2. Os líderes das duas Coreis, Kim Jong-Un e Moon Jae-in, acordaram tomar medidas para a completa desnuclearização.

3.9 Similaridade de termos assimétricas

As medidas assimétricas tem sido amplamente utilizado no contexto da adaptabilidade, este método com precisão a similaridade entre os termos no sentido de que os termos próximos muitas vezes têm um coeficientes de baixa similaridade, em poucos termos a norma de t1 de um termo é o número de documentos indexados pela palavra devido à lei de Zipf, a distribuição das normas de termos t1 e da norma t2 é muito assimétrico neste facto altera a distancia entre ter- mos, em grande diferenças em suas normas. Para as visualizações segundo a similaridade, é importante ter uma boa estimativa de semelhança de termos, isto pode ser feito, utilizando diferentes métricas, de duas maneiras; a partir de uma representação que deve ser gerada no conteúdo dos documentos ou deve ser comparado termos sem adotar uma representação intermediaria [117].
As medidas de semelhança assimétricas são construídas a partir da análise de termos nos textos são baseadas na seguinte hipótese; dois termos são os mais próximos que são frequentemente usadas nos mesmos textos, assim, dois termos serão consideradas muito simples se aparecerem frequentemente " lado-a-lado" ou pode ser no mesmo documento e de forma complementar se

eles parecem raramente um sem o outro. Estas medidas são do tipo, probabilidade condicional (PCS), que tem o coeficientes de dados e Jaccard ou informações mútuas usam este princípio. Quando observamos que dois sinónimos raramente são usados juntos as medidas de associação permitem dois ou mais termos em relação sinonímia [42] enquanto, por definição, mesmo desta relação assimétrica, onde se usa as noções de "vetor de contexto" que caracterizam cada termos pelo seu conjunto, onde que foi proposto a medida InfoSimba que consiste em comparar dois para dois cada um dos elementos de dois vetores de contexto para deduzir a semelhança assimétrica entre dois termos.
De facto que é importante há existência de uma tendência para uma forte associação direta de um termo especifico para o termo mais geral mas associação inversa é a mais fraca, dentro de um escopo, onde a vários trabalhos recentes de termos gerais foi proposto o uso de medidas de similaridade assimétricas, por sua vez acreditamos que esta ideia tem o potencial de provocar melhorias significativas na aquisição de relações assimétricas de termos.
Na literatura a abordagem mais popular é a baseada em um modelo de esquema de frequências de palavras, que faz uso de um vetor de frequência de palavra para representar um documento. Função cosseno, Produto escalar e função de proporção, entre outras, são medidas de similaridade regulares de vetor. Essas medidas são medidas de similaridade simétrica.
Neste trabalho como mencionado, apresentamos um modelo de similaridade assimétrica, desenvolvendo o contexto em causa (extração de termos relevantes) uma medida assimétrica, *(Adapted Asymmetric InfoSimba Similarity)*, derivada da medida desenvolvida.
Um método de similaridade é uma função que calcula um coeficiente de semelhança entre vetores, utilizando uma medida de similaridade entre uma *query* e cada um dos documentos de um conjunto, que podem ser:

- Recuperar os documentos segundo a ordem de relevância presumida;
- Definir um valor limiar de modo a controlar o tamanho do subconjunto de documento recuperados.

3.9.1 Medidas de Associação Assimétrica

Nesta hipótese, as medidas de associação assimétricas são necessárias induzir associações de termos partir dos quais apresentar as assimetrias que serão usadas para medir o grau de atividade entre substantivos, que são respetivamente a frequência em função da probabilidade. Estás medida são baseadas em padrões que podem incorporar a assimetria que são definidas inicialmente para uma relação, basicamente esta abordagem aproveita ao máximo os padrões assimétricos, por instanciando para um mecanismo de busca um número de padrões preenchidos apenas com um candidato possível que pode garantir a extrações de termos quando existir padrão assimétricos, por entanto nós sabemos que as medidas que são baseadas em padrões sensíveis de termos e com a confusão do padrão, com estas técnicas vêm dependendo do idioma que são difíceis de argumentar para diferentes idiomas, para que permaneça dentro da metodologia independente da língua e não supervisionadas. A aplicação dos problemas de construção de uma taxonomia, linguística cognitiva é um dos termos geral específica.
Quando os termos ocorrem juntos com mais frequências do que o caso, isso pode ser uma evidência de que eles têm uma função especial que não é simplesmente explicada como resultado da sua combinação, esta propriedade é conhecida em linguística como não composicionalidade, e chegamos de pensar que num corpus como uma sequência de termos gerada aleatoriamente que é visto como uma sequência de termos de pendência de n-gramas é no nosso caso [108] es-

tas são as consequências que ocorrem frequentemente que usamos nas medidas de associação assimétrica, certa mente com estes estudos a propor o uso de probabilidade, para construção de taxonomia, apresentada na equação 3.10.

Conditional Probability

$$P(x/y) = \frac{P(x,y)}{P(y)} \quad (3.12)$$

Os termos de altas frequências não discriminam entre documentos relevantes, e não relevantes, a adição desses termos para a expansão da consulta é ineficaz, o nível de relevância de um termo que o específico que resulta na determinação por [30] ocorrência com o termo conceito geral, que pode ser extraído do corpus, para que os termos de conceito gerais numa consulta que são substituídos por um conjunto de termos conceituais específicos usados no corpus. Esta abordagem é adequada apenas para situações em que precisão é mais importante.
Assume-se que um termo T2 inclui um termo T1 se os documentos em que T1 ocorre forem um subconjunto dos documentos em que T2 ocorre limitado por P(T2/T1)≥0,8 e P(T1/T2)todas estas relações de integração, constrói-se a estrutura semântica de qualquer domínio, que corresponde a um gráfico acíclico dirigido, a relação de integração de subsunção é aliviada com a seguinte expressão, de comparação de termos semelhantes no documento, P(T2/T1)≥ P(T1/T2) e P(T2/T1)gt;T, onde T é um determinado limiar e todos os pares de termos encontrado para ter um relacionamento de integração que é passado através de um módulo de transitividade, que remove as relações de integração estranhas na maneira de que a transitividade é preferida em relação o caminho direto para que nos conduz assim a um gráfico acíclico que é direcionado não triangular.
As duas medidas propostas para modelar a noção das similaridade assimétrica, com as intenções de determinar em que medida essas duas de associação dirigida podem ser usados como modelo para associação psicológica dirigida na mente do utilizador, com estas medidas a probabilidade condicional é simples e a medida de classificação R$(. \parallel .)$ com na escala de person, X2, em particular o TI,I=1… n é a lista de todos termos que concorrem com o termo T ordenado com respeito para um determinado valor, X2, (T,TI),a R(TI$(. \parallel .)$T), é a classificação do termo TI nesta lista, com estas classificações dos termos os resultados são avaliados a favor de um grande número de normas de livre associação, com estas medidas é capaz de distinguir os pares simétricos e com os assimétricos e com certas medidas na previsão de graus assimétrico.
Por enquanto as pontuações finais dos *rankings* para o gráfico que é relevante diferentes significativamente em comparação às suas alternativas não relevantes com os números de iterações à convergência e a forma da convergência com uma curva é quase idêntico para poder permitir uma identificação de termos relevantes e não relevantes no gráfico.

Added Value

$$AV(x \parallel y) = P(x|y) - P(x) \quad (3.13)$$

Braun-Blanket

$$BB(x \parallel y) = \frac{f(x,y)}{f(x,y) + f(x^{-}, y)} \quad (3.14)$$

Certainty Factor

$$CF(x \parallel y) = \frac{P(x|y) - P(x)}{1 - P(x)} \quad (3.15)$$

Conviction

$$CO(x \parallel y) = \frac{P(x) \times P(y^-)}{P(x, y^-)} \quad (3.16)$$

Gini Index

$$GI(x \parallel y) = P(y) \times P(x|y)^2 + P(x^-|y)^2 - P(x)^2 P(y^-) \times (P(x|y^-)^2) + P(x^-|Y^-)^2) - P(x^-)^2 \quad (3.17)$$

J-measure.

$$JM(x \parallel y) = P(x, y) \times log\frac{P(x|y)}{P(x)} + P(x^-, y) \times log\frac{P(x^-|y)}{P(x^-)} \quad (3.18)$$

Laplace

$$LP(x \parallel y) = \frac{N \times P(x, y) + 1)}{N \times P(y)^+ 2} \quad (3.19)$$

3.10 Assimetrias entre Termos

Todas as oito (8) definições definidas na secção seguinte mostram as suas assimetria avaliando o valor máximo entre duas hipóteses, ou seja, por avaliar a sua tração de x em y, mas também a atração de y em cima de x com esta sequência o valor máximo decidir a direção da associacão geral específica. Existem várias maneiras de calcular assimetrias entre termos, a maioria das medidas assimétricas determinam a distância entre dois vetores associados em dois ou mais termos. Com aplicação das medidas clássicas de semelhança entre termos, apenas os índices do vetor das linhas Xi e Xj são levados em conta, o que pode levar a semelhanças calculadas equivoca mente, para se lidar com estes problemas, diferentes metodologias foram propostos, mas propicia na medida de assimetrias informativas, do InfoSimba, onde que cada medida de similaridades assimétricas.
No entanto, a pesquisa assimétrica de semelhança de termos com as coocorrências são notadas e exploradas, mas não existe em profunda estudo com relatos deste evento, onde que os esforços são direcionados para desenvolvimento de medidas de distribuição assimétricas, que é regularmente separado da divergência, com entropia cruzada descrita.

$$IS(Tx, Ty) = \frac{\sum_{i=1}^{N} \sum_{i=1}^{N} Tx_i * Ty_i * S(Tx_i, Ty_i)}{\begin{pmatrix} \sum_{i=1}^{N} \sum_{i=1}^{N} Tx_i * Tx_l * S(Tx_i, Tx_l) + \\ \sum_{i=1}^{N} \sum_{i=1}^{N} Ty_i * Ty_i * S(Ty_i, Ty_l) - \\ \sum_{i=1}^{N} \sum_{i=1}^{N} Tx_i * Ty_i * S(Tx_i, Ty_i) \end{pmatrix}}. \quad (3.20)$$

A similaridade assimétrica é uma família que consiste em termos que se sobrepõem em significado denotativo, significado conotativo, ou ambos, estes são mais conhecidos dessas relações é sinonímia em quais termos têm a mesma denotação semelhança dimensional que venha envolver acordo denotativo que não é suficiente para que podem vir ocorrer em pontos adjacentes numa dimensão comum. Com a similaridade InfoSimba (IS) que visa medir as relações entre todos os pares de palavras em vetores de contexto de duas, em vez de depender apenas da sua correspondência exata, para qualquer medida de similaridade assimétrica e cada Tij corresponde à palavra do atributo [40] encima com aposição no vetor Xi e P é o comprimento do Xi. Com as relação pragmática parte de caso que não envolvem semelhança de significado, em vez disso, estes dois conceitos são relacionados por uma associação que é pragmática.
Na existência de muitas medidas de similaridade assimétrica, elas evidenciam problemas que podem ser úteis, por um lado, as de associação assimétrica só podem avaliar a relação de generalidade e especificidade entre palavras que se sabe estarem numa relação semântica, como em [118] por exemplo, África, Angola. Noutra vertente a similaridade assimétrica de palavras de atributos no contexto avaliar o grau de familiarismo entre dois termos, par aproveitar esses problemas nos AIS, que é uma ideia subjacente é dizer que um termo x é semanticamente relacionado com palavra y é mais geral do X que vem compartilhar tantos termos relacionadas relevantes quando possível e cada termo x provavelmente será mais do que a maioria dos termos de y. Os AIS é definido na fórmula abaixo, onde que AIS é $(. \parallel .)$ que é qualquer medida de similaridade assimétrica, da mesma forma para o IS, na fórmula, onde que S (., .) significa que qualquer similaridade simétrica também definimos a sua versão simplificada.

$$AIS(T) = \frac{\sum_{i=1}^{N} AS(Tx_i \parallel Ty_l)}{\begin{pmatrix} \sum_{i=1}^{N} AS(Tx_i \parallel Tx_l) + \\ \sum_{i=1}^{N} AS(Ty_i \parallel Ty_l) - \\ \sum_{i=1}^{N} AS(Tx_i \parallel Ty_l) \end{pmatrix}}. \quad (3.21)$$

$$AISs(X_i \parallel X_J) = \sum_{L=1}^{N} X_{iK} \times X_{jl} \times AS(t_{ik}, \parallel t_{jl}) \Longleftrightarrow T_{ik} \# T_{jl} \quad (3.22)$$

Por baixo apresentamos um cálculo que mostra do AIS um simplificado com a medida do valor adicional para os seguintes termos, abaixo mencionados.

Xi: Fátima trabalha em África
Xj: Fátima trabalha em Angola
Em Problemas deste projeto surge que dois objetos"a" e"b" são mais parecidos com"c" e "d" se a importância acumulativa das propriedades compartilhadas por "a" e "b" é maior que as propriedades compartilhadas por"c" e "d". Considerando importância como um assunto altamente volátil, variando de acordo ao contexto e interesse [59]. Uma abordagem comum é tratar a similaridade como uma unificação de uma relação de equivalência, generalizando a reflexividade, a simetria e a transitividade.

3.11 Conclusão

O conjunto de recursos explorados a cima é usado em conjunto com uma estrutura de aprendizado de técnicas e métodos não supervisionados que vem a fornecer tarefas específica de

classificação do utilizador. Neste capitulo revisamos os métodos e técnicas usadas para construir perfis não supervisionados do utilizador e que incluem a extração de termos claves de interesse do utilizador, estes termos são apresentados em gráficos no quarto (4.º) capitulo. Está dinâmica do perfil linguístico do utilizador se refere a mudança que ocorre nos interesses do utilizador ao longo do tempo. Alguns pesquisadores forneceram muitos métodos para ser refletir nos interesses alterados ao longo do tempo de modo a construir perfil linguístico não supervisionado de um utilizador que podem ser úteis quando usado com aplicativos.

Capítulo 4

Fundamentação

Nossa abordagem baseai-se na utilidade por desta forma é proposto uma nova Medida denominada as similaridade assimétrica que é combinada de maneira significativa com diferentes Medidas de Associação Assimétricas visa proporcionar bons resultados no processo de seleção dos termos relevantes e não relevantes.

4.1 Desenvolvimento experimental

Com enorme procura e com uma grande divulgação de dados, o processamento de dado manual de documentos uma tarefa muito difícil com o surgimento da automatização nos últimos anos seguindo uma abordagem não supervisionada na qual viemos propor uma medida que as similaridades assimétricas onde seguimos uma metodologia não supervisionada com usos dos métodos uni-grams, bi-grams, tri-grams sendo um dos métodos conhecidos. Com estas abordagens acima referidas podem nos facilitar a escalar diferenças coleções, nos domínios de idiomas num curto intervalo de tempo com a proposta do método não supervisionado para extrairmos os termos chaves com as medidas propostas que se baseia em recursos estatísticos de texto para extrairmos termos chaves.

Nas maiorias das colocação lexical é determinada a combinações de sintagmaticamente relacionadas em termos que são associações de utilizadores também incluem muitos termos paradigmaticamente relacionadas, com estas medidas probabilísticas de associação também são aplicadas para a identificação de tais relações paradigmáticas, em que termos sinónimos e antonímia. Desta maneira, ao contrário das metodologias existentes, não precisamos definir ou ajustar os princípios, devido à definição assimétrica, a medida de similaridade assimétrica permite comparar os dois lados das vinculações. Por esta razão possível para a assimetria é o grau de generalidade dos termos, há uma tendência para um forte guia de associação de um termo específico como para um mais genérico. Depois destes dados Abaixo fornecemos um cálculo de amostra dos AIS simplificado, com medida de valor acrescentado, de cálculo de desvio padrão para o seguinte termo:
TEXTO 1"O presidente português foi recebido pelo presidente de Angola João Lourenço, em luanda."
Com estes termos conseguimos calcular o ranking com a média e desvio padrão o usando os métodos propostos conforme referimos nas fórmulas acima. As medidas comparadas os métodos como; Uni-Grams, Bi-Grams,Tri-Grams.
Na distribuição condicional de dados como uma contínua por sua função com a densidade de probabilidade conhecida como uma condicional, com as propriedades de uma distribuição, tal como o momento, são muitas vezes chamadas por nomes que corresponde, como a média condicional e variância. Em geral, está medida refere-se a uma distribuição condicional de um subconjunto de um conjunto de mais de dois termos ou variáveis está chegar sobre todas as incluídas em todos os subconjuntos que então estão distribuídos e é condicional das medidas

propostas.

Probabilidade Condicional

Para uma melhor explicação do que se trata de probabilidade condicional, considera-se um espaço amostral N finito não é vazio e um evento S de N se quisermos X desse espaço amostral N, essa nova é indicada por P(x‖ y) e dizemos que é a probabilidade condicional de X em relação a de Y.

Historicamente a probabilidade condicional é uma medida intimamente relacionada com a probabilidade, na qual as razões mais encontram para uma declaração, em comparação com as razões contra elas, é provável que o declarado pelo Jacob Bernoulli (1713/1987) o ponto de inflexão na emergência expressa por um número (entre 0 e 1) para um objeto [31]. Isso amplia o escopo da probabilidade muito além das situações assimétricas. Nestes contestamos e demonstramos os testes e os resultados encontrados na comparação dos termos usando os método e com as mediadas assimétricas, basicamente teremos os resultados em tabelas e gráficos.

4.1.1 Frequências de termos

Nas tabelas mencionadas onde foram selecionadas um conjunto de termos no texto em português e que determina qual deles tem maior relação com a sentença“ o”, “presidente” e “português”, mas apenas com esse procedimento não seria suficiente para completar a análise, pois, a muitos documentos provavelmente possuem os três termos, para que se poça melhor a distinção entre elas, podemos contar o número de vezes que as palavras ocorre no texto e somamos esse valor, o número de vezes que um termo que se repete no documento, isto é que demonstra as seguintes, tabelas 3.1, 3.3, e 3,5, é com estas nos permite encontrar quantas vezes em que o termo se encontra repetido no texto, com um objetivo é marcar a todos os termos que se encontra acima da média dos da palavra para equilibrar o grau de dispersão de dados pelo desvio de padrão.

Também é importante mencionar que não são estabelecidas condições em relação à frequência da frase que um termo-chave candidata deve ter [39]. Isso vem significar que podemos ter um termo-chave considerada significativa / insignificante com uma ocorrência ou com ocorrências múltiplas. Cada termos-chave candidata será então é atribuída uma média, de tal forma que quanto menor a pontuação, mais significativa será o termo-chave ou o relevante. A tabela 4.1

Tabela 4.1: Frequências de termos em Uni-Grams

Termos	Frequências
O	1/13=0,0769
presidente	2/13=0,1538
português	1/13=0,0769
foi	1/13=0,0769
Recebido	1/13=0,0769
pelo	1/13=0,0769
presidente	2/13=0,0769
de	1/13=0,0769
Angola	1/13=0,0769
João	1/13=0,0769
Lourenço	1/13=0,0769
em	1/13=0,0769
Luanda	1/13=0,0769

apresenta a distribuição de numero de frequências em que os termos são agrupados no texto, ou

pode ser o numero de ocorrência em que os termos estão divididas usando o método Uni-Grams.
AISs (T|| Ti)=Uni-Grams);
A=(P (O || presidente))+ (P(O || português)) +(P(O || foi)) +(P(O || recebido))+(P(O || pelo))+(P(O || de)) +(P(O || Angola))+(P(O || João)) + (P(O || Lourenço)) + (P(O|| em)) + (P(O || luanda))A=0,8459
B=(P(Presidente|| português)+(P(Presidente|| foi)+(P(Presidente|| recebido)+(P(Presidente|| pelo)+(P(Presidente || de)+(P(Presidente|| Angola)+(P(Presidente|| João)+(P(Presidente|| Lourenço)+(P(Presidente|| em)+(P(Presidente luanda)+(P(Presidente|| o) B= 1,6918
C=(P(português|| foi)+(P(português|| recebido)+(P(português || pelo)+(P(português|| presidente)+(P(português|| de)+(P(português|| Angola)+(P(português|| João)+(P(português|| Lourenço)+(P(português|| em)+(P(português|| luanda)+(P(português|| presidente)+(P(português|| o)= C=0,8459
D=(P(foi|| recebido)+(P(foi|| pelo)+(P(foi|| presidente)+(P(foi|| de)+(P(foi|| Angola)+(P(foi|| João)+ (P(foi|| Lourenço)+(P(foi|| em)+(P(foi|| luanda)+(P(foi|| português)+(P(foi|| presidente)+(P(foi|| o=)D=0,8459
E=(P(recebido|| pelo)+(P(recebido|| presidente)+(P(recebido|| de)+(P(recebido|| Angola)+(P(recebido|| João)+ (P(recebido|| Lourenço)+(P(recebido||l em)+(P(recebido|| Luanda)+(P(recebido||l foi)+(P(recebido||l português)+(P(recebido|| o)E=0,8459
F=(P(pelo|| presidente)+(P(pelo|| de)+(P(pelo||l Angola)+(P(pelo|| Lourenço)+(P(pelo|| em)+ (P(pelo||l Luanda)+ (P(pelo|| recebido)+(P(pelo|| foi)+(P(pelo|| português)+(P(pelo|| o) F=0,8459
G=(P(de|| Angola)+(P(de|| João)+(P(de|| Lourenço)+(P(de|| em)+(P(de|| Luanda)+ (P(de|| presidente)+(P(de|| pelo)+(P(de|| recebido)+(P(de|| foi)+(P(de|| português)+(P(de|| presidente)+(P(de|| o) G=0,8459
H=(P(Angola|| João)+(P(Angola|| Lourenço)+(P(Angola|| em)+(P(Angola|| Luanda)+(P(Angola||l de)+(P(Angola|| presidente)+ (P(Angola|| pelo)+(P(Angola|| recebido)+(P(Angola|| foi)+(P(Angola|| português)+(P(Angola||l presidente)+(P(Angola|| o) H=0,0,8459
I=(P(João|| Lourenço)+(P(João|| em)+(P(João|| Luanda)+(P(João|| Angola)+((João|| de)+(P(João|| presidente)+ (P(João|| pelo)+(P(João|| recebido)+(P(João|| foi)+(P(João|| português)+(P(João|| presidente)+(P(João o) I=0,8459
J=(P(Lourenço||l em)+(P(Lourenço|| luanda)+(P(Lourenço|| João)+(P(Lourenço|| Angola)+(P(Lourenço|| de)+(P(Lourenço || presidente)+(P(Lourenço|| pelo)+(P(Lourenço|| recebido)+(P(Lourenço|| foi)+(P(Lourenço|| português)+(P(Lourenço|| o)J=0,8459
L=(P(em|| luanda)+(P(em|| Lourenço)+(P(em|| João)+(P(em|| Angola)+(P(em|| de)+(P(em|| presidente)+(P(em|| pelo)+(P(em|| recebido)+(P(em|| foi)+(P(em|| português)+(P(em|| o)L=0,8459
M=(P(Luanda|| em)+(P(Luanda|| Lourenço)+(P(Luanda|| João)+(P(Luanda|| Angola)+(P(Luanda|| de)+(P(Luanda|| presidente)+(P(Luanda|| pelo)+(P(Luanda|| recebido)+(P(Luanda|| foi)+(P(Luanda|| português)+(P(Luanda|| o)M=0,8459

$$P(x/y) = \frac{P(x,y)}{P(y)} \quad (4.1)$$

P(O|| Presidente)-P(o)= P(o)$\times P(Presidente) - \frac{P(O)}{P}$
$= \frac{P(O)\times P(presidente) - P(O)}{P} = \frac{(O\times Presidente)}{P}$
$P(0,0769||$0,1538)=-(0,0059)$\frac{}{13=\frac{0,0118}{13}=0,0118-0,0059=\frac{0,0059}{13}=0,8459}$
P=(1,6918+0,8459+0,8459+0,8459+0,8459+0,8459+0,8459+0,8459+0,8459+0,8459+0,8459)=10,9967
Media=0,9997
Desvio padrão =0,9457
TR $\geq Media - Desviopadro.$
$TR \geq 0,9997 - 0,9457 = 0,054$
$TR \geq Media + Desviopadro.$
$TR \geq 0,9997 + 0,9457 = 1,9454$

Tabela 4.2: Frequências de termos e do desvio padrão em Uni-Grams

Ordem	Frequências	(Fre-Media)^2
B	1,6918	2,86218724
D	0,8459	0,71554681
E	0,8459	0,71554681
A	0,8459	0,71554681
J	0,8459	0,71554681
L	0,8459	0,71554681
M	0,8459	0,71554681
F	0,8459	0,71554681
C	0,8459	0,71554681
G	0,8459	0,71554681
H	0,8459	0,71554681
I	0,8459	0,71554681
SOMA	10,9967	0,97574565
Média	0,9997	
Desvio padrão	0,9457	

A tabela 4.2 apresenta os dados de uma maneira mais concisa e que nos permite extrair informações sobre o seu comportamento dentro do texto mostrar a frequência dos termos relevantes a soma, a média, e o desvio padrão usando o método uni-grams.

?? A tabela 4.3 apresenta a distribuição de numero de frequências em que os termos são agru-

Tabela 4.3: Frequências de termos em Bi-Grams

Termos	Frequências
O presidente	1/13=0,0769
presidente português	2/13=0,0769
foi recebido	1/13=0,0769
pelo presidente	1/13=0,0769
Presidente de	1/13=0,0769
de Angola	1/13=0,0769
Angola João	2/13=0,0769
João Lourenço	1/13=0,0769
Lourenço em	1/13=0,0769
Em luanda	1/13=0,0769

pados no texto, ou pode ser o numero de ocorrência em que os termos estão divididas usando o método Bi-Grams.
AlSs(T‖ Ti)= Bi-grams

A=(P(O presidente‖ presidente português))+(P(O presidente‖ foi recebido))+(P(O presidente ‖ pelo presidente))+(P(O presidente‖ de Angola))+(P(O presidente‖ João Lourenço))+(P(O presidente‖ em Luanda))A=0,4612
B=(P(presidente português‖ foi recebido))+(P(presidente português‖l pelo presidente))+(P(presidente português ‖ de Angola)) + (P(presidente português‖ João Lourenço)) + (P(presidente português ‖ em Luanda))B=0,3845
C=(P(português foi‖l recebido pelo)) +(P(português foi ‖ presidente de)) + (P(português foi‖ Angola João)) +(P(português foi‖ Lourenço em))+ (P(português foi‖ em luanda))+(P(português foi‖l presidente o))C=0,4612
D=(P(foi recebido ‖ pelo presidente)) + (P(foi recebido‖ de Angola)) +(P(foi recebido‖ João Lourenço))+(P(foi recebido‖ em Luanda)) +(P(foi recebido ‖ o presidente)) +(P(foi recebido ‖ presidente português))D=0,3845
E=(P(recebido pelo‖ presidente de))+(P(recebido pelo‖ Angola João))+(P(recebido pelo‖ Lourenço em)) +(P(recebido pelo‖ em luanda))+(AV(recebido pelo‖ foi português))+(P(recebido pelo‖ presidente))E=0,4612
F=(P(pelo presidente‖ de Angola))+(P(pelo presidente‖ João Lourenço)) +(P(pelo presidente‖ em luanda))+(P(pelo presidente‖ foi recebido))+(P(pelo presidente ‖ português presidente))F=0,3845
G=(P(Presidente de‖ Angola João)) +(P(Presidente de‖ Lourenço em)) +(P(Presidente de‖ em luanda)) +(P(Presidente de‖ pelo recebido))+(P(Presidente de ‖ foi português))+(P(Presidente de‖ presidente o))G=0,4612
H=(P(de Angola‖ João Lourenço)) +(P(de Angola‖ em luanda))+(P(de Angola‖ presidente pelo))+(P(de Angola‖ recebido foi)) +(P(de Angola‖ português presidente))H=0,3845
I=(P(de Angola‖ Lourenço em)) +(P(de Angola‖ em luanda))+(P(de Angola ‖ de presidente))+(P(de Angola‖ pelo recebido))+(P(de Angola‖ foi português))+(P(de Angola‖ presidente o))I=0,4612
J=(P(Angola João‖ Lourenço em)) +(P(Angola João‖ em Luanda))+(P(Angola João‖ de presidente)) +(P(Angola João‖ pelo recebido))+(P(Angola João‖ foi português))+(P(Angola João‖ presidente o)) J=0,4612
L=(P(João Lourenço‖ em Luanda))+(P(João Lourenço‖ Angola de))+(P(João Lourenço‖ presidente pelo))+(P(João Lourenço‖ recebido foi))+(P(João Lourenço‖ português presidente))L=0,3845

K=(P(Lourenço em || em luada))+(P(Lourenço em || João Angola))+(P(Lourenço em || de presidente))+(P(Lourenço em|| pelo recebido))+(P(Lourenço em || foi português))+(P(Lourenço em || presidente o))K=0,4612
M=(P(em luanda || Lourenço João))+(P(em luanda || Angola de))+(P(em luanda || presidente pelo))+(P(em luanda || recebido foi))+(P(em luanda || português presidente))M=0,3845
P=(0,4612+0,3845+0,4612+0,3845+0,4612+0,3845+0,4612+0,3845+0,4612+0,4612+0,3845+0,4612+0,3845)/13
Média =0,4258
Desvio Padrão=0,03823
TR $\geq Media - Desviopadro.$
$TR \geq 0,4258 - 0,0382 = 0,3876$
$TR \geq Media + Desviopadro.$
$TR \geq 0,4258 + 0,0382 = 0,464$

Tabela 4.4: Frequências de termos do desvio padrão em Bi-Grams

Ordem	Frequências	(Fre-Media)^2
A	0,4612	0,00125316
C	0,4612	0,00125316
E	0,4612	0,00125316
G	0,4612	0,00125316
I	0,4612	0,00125316
J	0,4612	0,00125316
K	0,4612	0,00125316
B	0,3845	0,00170569
D	0,3845	0,00170569
F	0,3845	0,00170569
H	0,3845	0,00170569
L	0,3845	0,00170569
M	0,3845	0,00170569
SOMA	5,5354	0,00146202
Média	0,4258	
Desvio padrão	0,03823637	

A tabela 4.4 apresenta os dados de uma maneira mais concisa e que nos permite extrair informações sobre seu comportamento dentro do texto nos permite mostrar a frequência dos termos relevantes a soma , a média, e o desvio padrão usando o método Bi-Grams.

Tabela 4.5: Frequências de termos em Tri-Grams

Termos	Frequências
O presidente português	1/13=0,0769
presidente português foi	1/13=0,0769
português foi recebido	1/13=0,0769
foi recebido pelo	1/13=0,0769
recebido pelo presidente	1/13=0,0769
pelo presidente de	1/13=0,0769
presidente de Angola	1/13=0,0769
de Angola João	1/13=0,0769
Angola João Lourenço	1/13=0,0769
João Lourenço em	1/13=0,0769
Lourenço em luanda	1/13=0,0769

A tabela 4.5 apresenta a distribuição de numero de frequências em que os termos são agrupados no texto, ou pode ser o numero de ocorrência em que os termos estão divididas usando o método Tri-Grams.

AISs(T || Ti)=Tri-grams
A=(P(O presidente português|| foi recebido pelo)) + (P(presidente português foi|| l presidente de Angola))+ (P(O presidente português|| João Lourenço em)) +(P(O presidente português|| Lourenço em Luanda)) A=0,3076
B=(P(presidente português foi|| recebido pelo presidente))+(P(presidente português foi|| de Angola João))+ (P(presidente português foi|| Lourenço em luanda))B=0,2307
C=(P(português foi recebido|| pelo presidente de))+(P(português foi recebido|| Angola João Lourenço))+(P(português foi recebido|| Lourenço em Luanda))C=0,2307
D=(P(foi recebido pelo || presidente de Angola))+(P(foi recebido pelo|| João Lourenço em))+(P(foi recebido pelo || Lourenço em Luanda))+(P(foi recebido pelo || português presidente o))D=0,3076
E=(P(recebido pelo presidente|| de Angola João))+ (P(recebido pelo presidente||l Lourenço em Luanda))+(P(recebido pelo presidente|| foi português presidente))E=0,2307
F=(P(pelo presidente de|| Angola João Lourenço))+(P(pelo presidente de|| Lourenço em Luanda))+(P(pelo presidente de|| recebido foi português))F=0,2307
G=(P(presidente de Angola|| João Lourenço em))+(P(presidente de Angola|| Lourenço em Luanda))+(P(presidente de Angola|| pelo recebido foi))+(P(presidente de Angola|| português presidente o))G=0,3076
H=(P(de Angola João|| Lourenço em Luanda))+(P(de Angola João|| presidente pelo recebido))+(P(de Angola João ||l foi português presidente))H=0,2307
I=(P(Angola João Lourenço|| de presidente pelo))+(P(Angola João Lourenço|| recebido foi português))I=0,1538
J=(P(João Lourenço em|| Angola de presidente))+(P(João Lourenço em|| Angola de presidente))+(P(João Lourenço em|| pelo recebido foi))+(P(João Lourenço em|| português presidente o))J=0,3076
K=(P(Lourenço em Luanda|| João Angola de))+(P(Lourenço em Luanda|| presidente pelo recebido))+(P(Lourenço em Luanda|| foi português presidente))K=0,2307

P=(0,3076+0,2307+0,2307+0,3076+0,2307+0,2307+0,3076+0,2307+0,1538+0,03076+0,02307/11
Media=0,25167
Desvio Padrão=0,20968
TR $\geq Media - Desviopadro.$

$TR \geq 0,25167 - 0,20968 = 0,04199$
$TR \geq Media + Desviopadro.$
$TR \geq 0,25167 + 0,20968 = 0,46135$

Tabela 4.6: Frequências de termos do desvio padrão em Tri-Grams com probabilidade condicional

Ordem	Frequências	(Fre-Media)^2
A	0,3076	0,06769415
D	0,3076	0,06769415
G	0,3076	0,06769415
J	0,3076	0,06769415
B	0,2307	0,033593519
C	0,2307	0,033593519
E	0,2307	0,033593519
F	0,2307	0,033593519
H	0,2307	0,033593519
K	0,2307	0,033593519
I	0,1538	0,011317842
SOMA	2,7684	
Média	0,25167	
Desvio padrão	0,2096	

A tabela 4.6 apresenta os dados de uma maneira mais concisa e que nos permite extrair informações sobre seu comportamento dentro do texto nos permite mostrar a frequência dos termos relevantes a soma , a média, e o desvio padrão usando o método Tri-Gramas.
Medida Laplace

$$LP(x \parallel y) = \frac{N \times P(x,y) + 1)}{N \times P(y)^{+}2} \quad (4.2)$$

$$LP(0,0769 \parallel 0,1538) = \frac{13 \times P(0,0769 + 0,1538) + 1}{13 \times P(0,1538) + 2} = \frac{13 \times P(0,2307) + 1}{13 \times P(0,1538) + 2} \frac{3,9991}{3,9994} = 0,9999 \quad (4.3)$$

P=(5,5544+10,9989+5,5538+5,5538+5,5544+4,9990+5,5538+5,5538+5,5538+5,5544+5,5544+5,5544)/12

Media=5,961575
Desvio padrão =1,5264
TR $\geq Media - Desviopadro.$
$TR \geq 5,961575 - 1,5264 = 4,4351$
$TR \geq Media + Desviopadro.$
$TR \geq 5,961575 + 1,5264 = 7,4879$

A tabela 4.7 apresenta os dados de uma maneira mais concisa e que nos permite extrair informações sobre seu comportamento dentro do texto nos permite mostrar a frequência dos termos

Tabela 4.7: Frequências de termos e do desvio padrão em Uni-Gramas com medida Laplace

Ordem	Frequências	(Fre-Media)^2
B	10,9989	25,37464316
A	5,5544	0,165791481
E	5,5544	0,165791481
J	5,5544	0,165791481
L	5,5544	0,165791481
M	5,5544	0,165791481
C	5,5538	0,166280451
D	5,5538	0,166280451
G	5,5538	0,166280451
H	5,5538	0,166280451
I	5,5538	0,166280451
F	4,999	0,926550631
SOMA	71,5389	2,330129454
Média	5,961575	
Desvio padrão	1,5264	

relevantes a soma , a média, e o desvio padrão usando o método Uni-Grams.
LP=(3,333+3,333+3,333+3,333+3,333+3,333+3,333+3,333+3,333+2,7775+2,7775+2,7775+2,7775+2,7775)/13
Média =3,1193
Desvio Padrão=2,8691
TR $\geq Media - Desviopadro.$
$TR \geq 3,1193 - 2,8691 = 0,2502$
$TR \geq Media + Desviopadro.$
$TR \geq 3,1193 + 2,8691 = 5,9884$

Tabela 4.8: Frequências de termos do desvio padrão em Bi-Grams usando a método de Laplac

Ordem	Frequências	(Fre-Media)^2
A	3,333	9,425154321
C	3,333	9,425154321
D	3,333	9,425154321
E	3,333	9,425154321
G	3,333	9,425154321
I	3,333	9,425154321
J	3,333	9,425154321
K	3,333	9,425154321
B	2,7775	6,322918553
F	2,7775	6,322918553
H	2,7775	6,322918553
L	2,7775	6,322918553
M	2,7775	6,322918553
SOMA	40,5515	8,231986718
Média	3,119346154	
Desvio padrão	2,8691439	

A tabela 4.8 apresenta os dados de uma maneira mais concisa e que nos permite extrair informações sobre seu comportamento dentro do texto nos permite mostrar a frequência dos termos

relevantes a soma , a média, e o desvio padrão usando o método Bi-Grams

Medida laplace com o método Tri-Grams

$$LP(0,0769 \parallel 0,0769) = \frac{13 \times P(0,0769 + 0,0769) + 1}{13 \times P(0,0769) + 2} \frac{13 \times P(1,1538)}{13 \times P(2,0769)} = \frac{14,9994}{26,9997} = 0,5555 \quad (4.4)$$

LP=(2,222+2,222+2,222+2,222+1,6665+1,6665+1,6665+1,6665+1,6665+1,6665+1,111)/13
Média =1,818
Desvio Padrão=1,135
TR $\geq Media - Desviopadro.$
$TR \geq 1,818 - 1,135 = 0,683$
$TR \geq Media + Desviopadro.$
$TR \geq 1,818 + 1,135 = 2,953$

Tabela 4.9: Frequências de termos do desvio padrão em Tri-Grams com a medida Laplac

Ordem	Frequências	(Fre-Media)^2
A	2,222	0,163216
D	2,222	0,163216
G	2,222	0,163216
J	2,222	0,163216
B	1,6665	0,02295225
C	1,6665	0,02295225
E	1,6665	0,02295225
F	1,6665	0,02295225
H	1,6665	0,02295225
K	1,6665	0,02295225
I	1,111	0,499849
SOMA	19,998	1,2904265
Média	1,818	
Desvio padrão	1,13596941	

A tabela 4.9 apresenta os dados de uma maneira mais concisa e que nos permite extrair informações sobre seu comportamento dentro do texto nos permite mostrar a frequência dos termos relevantes a soma , a média, e o desvio padrão usando o método Tri-Grams

Medida de Cobertor de Braun(BB) temos os teste utilizados a baixo com a granularidade Uni-Grams.

$$BB(x \parallel y) = \frac{f(x,y)}{f(x,y) + f(x^-,y)} \quad (4.5)$$

$$(BB)\frac{f(x) + f(y)}{f(x) + f(y) + 1} = \frac{0,0769 + 0,1538}{0,0769 + 0,1538 + 1} = \frac{0,2307}{1,3076} = 0,1764 \quad (4.6)$$

$$(BB)\frac{f(x) + f(y)}{f(x) + f(y) + 1} = \frac{0,0769 + 0,0769}{0,0769 + 0,0769 + 1} = \frac{0,1538}{1,1538} = 0,1332 \quad (4.7)$$

BB=(1,9404+1,6848+1,6848+1,6848+1,6848+1,6848+1,5084+1,5084+1,5084+1,5084+1,5084+1,3752)/12
Média =1,6068
Desvio Padrão=1,493
TR $\geq Media - Desviopadro$
$TR \geq 1,6068 - 1,493 = 0,1138$
$TR \geq Media + Desviopadro$
$TR \geq 1,6068 + 1,493 = 3,0998$

Tabela 4.10: Frequências de termos e de desvio padrão com a medida Braun-Blanket(BB)granularidade Uni-Grams

Ordem	Frequências	(Fre-Media)^2
B	1,9404	3,31167204
C	1,6848	2,44672164
D	1,6848	2,44672164
G	1,6848	2,44672164
H	1,6848	2,44672164
I	1,6848	2,44672164
A	1,5084	1,92598884
E	1,5084	1,92598884
J	1,5084	1,92598884
L	1,5084	1,92598884
M	1,5084	1,92598884
F	1,3752	1,57402116
SOMA	19,2816	2,2291038
Média	1,6068	
Desvio padrão	1,493	

A tabela 4.10 apresenta os dados de uma maneira mais concisa e que nos permite extrair informações sobre seu comportamento dentro do texto nos permite mostrar a frequência dos termos relevantes a soma , a média, e o desvio padrão usando a granularidade Uni-Gramas.
BB=(0,7992+0,7992+0,7992+0,7992+0,7992+0,7992+0,7992+0,7992+0,666+0,666+0,666+0,666+0,666)/13
Média =0,7479
Desvio Padrão=0,6879
TR $\geq Media - Desviopadro$
$TR \geq 0,7479 - 0,6879 = 0,06$
$TR \geq Media + Desviopadro$
$TR \geq 0,7479 + 0,6879 = 1,4358$

A tabela 4.11 apresenta os dados de uma maneira mais concisa e que nos permite extrair informações sobre seu comportamento dentro do texto nos permite mostrar a frequência dos termos relevantes a soma , a média, e o desvio padrão usando a granularidade Bi-Grams.

BB=(0,5328+0,5328+0,5328+0,5328+0,3996+0,3996+0,3996+0,3996+0,3996+0,3996+0,2664)/11
Média =0,4359
Desvio Padrão=0,0821
TR $\geq Media - Desviopadro$
$TR \geq 0,4359 - 0,0821 = 0,3538$
$TR \geq Media + Desviopadro$
$TR \geq 0,4359 + 0,0821 = 0,518$

A tabela 4.12 apresenta os dados de uma maneira mais concisa e que nos permite extrair informações sobre seu comportamento dentro do texto nos permite mostrar a frequência dos termos relevantes a soma , a média, e o desvio padrão usando o método Tri-Grams.

Tabela 4.11: Frequências de termos e de desvio padrão com a medida Braun-Blanket(BB) com granularidade Bi-Grams

Ordem	Frequências	(Fre-Media)^2
A	0,7992	0,541912031
C	0,7992	0,541912031
D	0,7992	0,541912031
E	0,7992	0,541912031
G	0,7992	0,541912031
I	0,7992	0,541912031
J	0,7992	0,541912031
K	0,7992	0,541912031
B	0,666	0,363544778
F	0,666	0,363544778
H	0,666	0,363544778
L	0,666	0,363544778
M	0,666	0,363544778
SOMA	9,7236	0,473309241
Média	0,7479	
Desvio padrão	0,6879	

Tabela 4.12: Frequências de termos do desvio padrão em Tri-Grams com a medida Braun-Blanket(BB)

Ordem	Frequências	(Fre-Media)^2
A	0,5328	0,009384325
D	0,5328	0,009384325
G	0,5328	0,009384325
J	0,5328	0,009384325
B	0,3996	0,001319671
C	0,3996	0,001319671
E	0,3996	0,001319671
F	0,3996	0,001319671
H	0,3996	0,001319671
K	0,3996	0,001319671
I	0,2664	0,006744984
SOMA	4,7952	0,006744984
Média	0,4359	
Desvio padrão	0,0821	

Medida de Condenação(CO) temos os teste utilizados a baixo com a granularidade Uni-Grams.

$$CO(x \parallel y) = \frac{P(x) \times P(\bar{y})}{P(x, \bar{y})} \quad (4.8)$$

$$CO(x \parallel y) = \frac{P(x) \times P(\bar{y})}{P(x, \bar{y})} = \frac{x \times P\bar{y}}{\bar{y}} = P(x) = (x); \quad (4.9)$$

$$CO = \frac{0,0769 \times 0,1538 - 1}{0,0769 \times 0,1538 - 1} = \frac{0,0769 + 0,1538 - 1}{0,1538 - 1} = 0,0769 \quad (4.10)$$

CO=(1,6918+0,9228+0,9228+0,9228+0,9228+0,9228+0,8459+0,8459+0,8459+0,8459+0,8459+0,769)/12
Média =0,9420
Desvio Padrão=0,7476
TR $\geq Media - Desviopadro$
$TR \geq 0,9420 - 0,7476 = 0,1944$
$TR \geq Media + Desviopadro$
$TR \geq 0,9420 + 0,7476 = 1,6896$

Tabela 4.13: Frequências de termos e de desvio padrão com a medida Condenação (CO) granularidade Uni-Grams

Ordem	Frequências	(Fre-Media)^2
B	1,6918	2,134033464
C	0,9228	0,478633094
D	0,9228	0,478633094
G	0,9228	0,478633094
H	0,9228	0,478633094
I	0,9228	0,478633094
A	0,8459	0,378142767
E	0,8459	0,378142767
J	0,8459	0,378142767
L	0,8459	0,378142767
M	0,8459	0,378142767
F	0,769	0,28947966
SOMA	11,3043	0,558949369
Média	0,942025	
Desvio padrão	0,7476	

A tabela 4.13 apresenta os dados de uma maneira mais concisa e que nos permite extrair informações sobre seu comportamento dentro do texto nos permite mostrar a frequência dos termos relevantes a soma , a média, e o desvio padrão usando a granularidade Uni-Grams.

CO=(0,4614+0,4614+0,4614+0,4614+0,4614+0,4614+0,4614+0,3845+,3845+,3845+,3845+,3845+,3845)/13
Média =0,4259
Desvio Padrão=0,3895
TR $\geq Media - Desviopadro$
$TR \geq 0,4259 - 0,3895 = 0,0364$
$TR \geq Media + Desviopadro$
$TR \geq 0,4259 + 0,3895 = 0,8154$

A tabela 4.14 apresenta os dados de uma maneira mais concisa e que nos permite extrair informações sobre seu comportamento dentro do texto nos permite mostrar a frequência dos termos relevantes a soma , a média, e o desvio padrão usando a granularidade Bi-Grams.

Tabela 4.14: Frequências de termos e de desvio padrão com a medida Condenação (CO) granularidade Bi-Grams

Ordem	Frequências	(Fre-Media)^2
A	0,4614	0,179079209
C	0,4614	0,179079209
E	0,4614	0,179079209
G	0,4614	0,179079209
I	0,4614	0,179079209
J	0,4614	0,179079209
K	0,4614	0,179079209
B	0,3845	0,119908117
D	0,3845	0,119908117
F	0,3845	0,119908117
H	0,3845	0,119908117
L	0,3845	0,119908117
M	0,3845	0,119908117
SOMA	5,5368	0,151769474
Média	0,4259	
Desvio padrão	0,3895	

CO=(0,3076+0,3076+0,3076+0,3076+0,2307+0,2307+0,2307+0,2307+0,2307+0,2307+0,1538)/11
Média =0,2516
Desvio Padrão=0,0474
TR $\geq Media - Desviopadro$
$TR \geq 0,2516 - 0,0474 = 0,2042$
$TR \geq Media + Desviopadro$
$TR \geq 0,2516 + 0,0474 = 0,299$

Tabela 4.15: Frequências de termos do desvio padrão em Tri-Grams com Condenação(CO)

Ordem	Frequências	(Fre-Media)^2
A	0,3076	0,00312786
D	0,3076	0,00312786
G	0,3076	0,00312786
J	0,3076	0,00312786
B	0,2307	0,000439855
C	0,2307	0,000439855
E	0,2307	0,000439855
F	0,2307	0,000439855
H	0,2307	0,000439855
K	0,2307	0,000439855
I	0,1538	0,009579071
SOMA	2,7684	0,002248149
Média	0,2516	
Desvio padrão	0,0474	

A tabela 4.15 apresenta os dados de uma maneira mais concisa e que nos permite extrair informações sobre seu comportamento dentro do texto nos permite mostrar a frequência dos termos relevantes a soma , a média, e o desvio padrão usando granularidade Tri-Grams.

O fator de Certeza é abordagem dos métodos da similaridade para o raciocínio no uso de fatores de certeza teve como pioneiro o sistema MYCIN qual tenta recomendar para os utilizadores o grau de confirmação o grau de originalidade do texto definido como um fator de certeza(FC) [84].
O fator de Certeza (CF) é um modelo que tem sido aplicado por diferentes pesquisadores no mapeamento de suscetibilidade a um escorregamento, a abordagem de CF é uma das possíveis funções das propostas para lidar com o problema de uma combinação diferente em camadas de dados, a heterogeneidade incerteza de entrada, estes fatores de certeza são dados pela seguinte equação abaixo mencionada [49].

$$CF(x \parallel y) = \frac{P(x \parallel y) - P(x)}{1 - P(x)} \quad (4.11)$$

Com está formula demonstramos os resultados obtidos nesta medida para uma melhor explicação
.

$$FC(0,0769 \parallel 0,1538) = FC\frac{(0,0059)}{1-(0,0059)} = \frac{0,0118}{1-0,0059} = \frac{0,0118-0,0059}{1-(0,0059)} = \frac{0,0059}{0,0059} = 1 \quad (4.12)$$

$$FC(0,0769 \parallel 0,0059) = FC\frac{(0,0059)}{1-(0,0059)} = \frac{4,5371}{1-(0,0059)} = \frac{4,5371-0,0059}{1-(0,0059)} = \frac{0,0059}{0,0059} = 1 \quad (4.13)$$

(4.14)

Esta medida após o teste feito permitiu concluir que uma medida que nos dá um valor negativo e com este valor não pode continuar, como as seguintes medidas abaixo mencionadas.
Esta medida de probabilidade condicional vem possuir tanto uma interpretação subjetiva como também com as frequências, com está interpretação de frequências, é um processo de experimentação é repetido um número grande de vezes para proporção de repetições nos eventos.

$$AV(x \parallel y) = P(x|y) - P(x) \quad (4.15)$$

$$AV(x \parallel y) = 0,0769 \times 0,0769P(0,0769) = 1 \quad (4.16)$$

$$JM(x \parallel y) = P(x,y) \times log\frac{P(x|y)}{P(x)} + P(x^-,y) \times log\frac{P(x^-|y)}{P(x^-)} \quad (4.17)$$

Depois dos testes realizados estas medidas usadas não dão resultado satisfatório para a comparação e controlo do *Ranking* e da semelhança de termos no texto.
Com estas medidas de identificação na organização de linguagem de um texto e na comparação na distância de termos e ordenação de frequências de Uni-Grams, Bi-Grams e Tri-Grams com base do idioma onde que as listas de frequências ordenadas podem ser calculadas, para cada medida no texto poder haver uma correspondência em perfil de idiomas de qualquer utilizador atual na qual é comparado com as granularidades uni-grams, Bi-Grams e Tri- Grams.
A nossa quantidade de interesse chama-se desvio padrão (σ) que vem ser o desvio padrão das medidas com a relação para de terminamos a média, o procedimento adotado foi a partir das nossas observações de ordem estatística, obtermos a melhor estimativa para o desvio padrão. O desvio padrão significa que ele indica o erro que teríamos caso fizéssemos uma única observação, equivalente, a um erro de um dado conjunto de X determinações com uma dada percentagem

de probabilidade.
Com a medida de similaridade assimétrica uma vez que a variância a soma de quadrados, a unidade em que se exprime não é a mesma que os dados, basicamente podem obter uma de variabilidade ou dispersão com os mesmos para que podemos tomar a raiz quadrada obtém o desvio padrão.
O desvio padrão é uma medida que só pode assumir valores não negativos e quanto maior for, será a dispersão dos dados. É definido como o afastamento em relação a uma média próxima da aritmética [24].Algumas propriedades do desvio padrão, que resultam imediatamente da definição, são:
-O desvio padrão é sempre positivo e será tanto maior quanto mais variabilidade houver entre os dados.
-Se n = 0, então não existe variabilidade, isto é, os dados são todos iguais.
Depois de o uso de medida condicional na qual conseguimos resultados satisfatórios e decidimos comparar com as medidas, dos fatores assimétricos, onde chegamos de calcular o *Ranking* , a média, Desvio padrão e usando os métodos Uni-Grams, Bi-Grams, Tri-Grams,que vamos demonstrar por tabelas com gráficos a comparação , que permite apresentar os termos relevantes e não relevantes no texto.

4.2 Avaliação

Para nossa avaliação a eficácia do nosso método com a sua generalidade de testar em 7 medidas de probabilidades estatísticas com um conjunto de dados diferentes com caracterizados por métodos a qualquer idioma e usando uma medida que é similaridade assimétrica (AIS) segundo os nossos testes uma das medidas de uma abordagem não supervisionada, com os nossos experimentos de encontrar os termos-chaves com tamanhos determinados.
Para que a nossa avaliação seja justa chegamos de comparar o nosso método contra o YAKE que um método usado neste trabalho *(A Text Feature Based Automatic Keyword Extraction Method for Single Documents)* , ao contrário do nosso método o YAKE é um método não supervisionado, com ostros três metodos como, *RAKE, TextRank e SingleRank* está abordagem bem conhecidas como não supervisionadas. Estas coleções diferentes para ilustrar a generalidade da nossa abordagem, os nossos resultados experimentais sugerem que a extrair termos — chaves de documento usando os nossos métodos resulta numa eficácia superior quando chegamos a ponto de comparar com as abordagens semelhantes.

Uma das vantagens da nossa medida retorna mais palavras em relação a outra avaliação. Para a nossa principal avaliação que começamos com uma tradicional, por cada texto, combinamos com termos-chaves no que diz a verdade com usos dos métodos estatísticos, calculamos a probabilidade com a precisão e pontuação valíamos a eficácia dos termo que diferentes coleções para o estudo de dados idiomas ajudando a esta medida. A nossa medida é uma melhoria significativa de um método ao longo de outros todos os resultados refletem a similaridade assimétrica (AIS) uma medida que alcança melhor estatisticamente com um significativo com este todo o conjunto de dados os melhores são alcançados, conforme referido não há diferença entre qualquer método, é importante ressaltamos que em quanto o YAKE depende e beneficia de técnicas de PLN, como também os AIS, simplesmente tomam como entradas, com estas medidas podem ser entendidas como uma das vantagens e são disponíveis para algumas línguas menores para os quais há falta de interesse de ferramentas, basicamente os AIS devolvem melhor resultado na

coleção de números de termos-chaves entre todo o texto, em quantos que em alguns cálculos são muito baixos em comparação com os outros com estes leva de outras áreas de pesquisa nos domínios de extração de termos-chaves.

De acordo o que foi já mencionado neste trabalho a cima de que os termos relevantes para as frases que compõem o par é comparado com algumas medidas em que chegamos de realizar menos cálculos, devido dos resultados obtidos, como; *Added Value* (Equação 3.11) *Centainty Factor*(Equação 3.13) e J-Measure (Equação 3.15) estas três medidas deu-nos uns valores negativos nos testes é por isso que nós não chegamos de implementar.
Com esta abordagem, a nossa metodologia nos mostra um comportamento aceitável nos cálculos de médias e dos desvio padrão, com as seguintes metodologias estes são as melhores medidas que constatamos nos nossos testes, condicional Probabilista (Equação 3.10) com as seguintes médias em que cada granularidade em Uni-Grams=0,237, Bi-Grams=0,425, Tri-Grams=0,251.
Com a medida (Laplace) (Equação 3.17) com os seguintes métodos, Uni-Grams=5,961,Bi-Grams=3,119, Tri-Grams=1,818
Braun-Blanket(Equação 3.12) com os seguinte métodos Uni-Grams=1,606, Bi-Grams=0,747, Tri-Grams=0,435.
Conviction(Equação 3.14) com os seguintes métodos como Uni-Grams=0.942, Bi-Grams=0.425, Tri-Grams=0.251.
Gini Index(Equação 3.15) Uni-Grams =22,1533, Bi-Grams=19,7691, Tri-Grams=19,221 estas sãos as medidas com um melhor resultado nos nossos testes realizados.

Comparação dos termos com Probabilidade Condicional

Tabela 4.16: Termos com maior e menores frequências no texto

Granularidade dos Conjuntos (Uni-Grams)		
Probabilidade Condicional(PC)		
M-DP(0.054)	M(0,9997)	M+DP(1,9454)
A E D F	A E D F G H I J L M	B
SOMA	10.9967	
Média	0.9997	
Desvio padrão	0,9457	

Com base na tabela 4.16 probabilidade condicional usando a medida assimétrica presentar as conclusões dos termos com maior e menor frequência em cada método. No entanto, que mostra as pequenas diferenças podem ser visto embora difiram em cada medida, onde mostramos a lista ordenada que é uma dada pela nossa metodologia é feita desta maneira com base a tabela acima com a granularidade Uni-Grams.
O gráfico 4.1 de probabilidade associados à distribuição normal, apresentam o valor com média

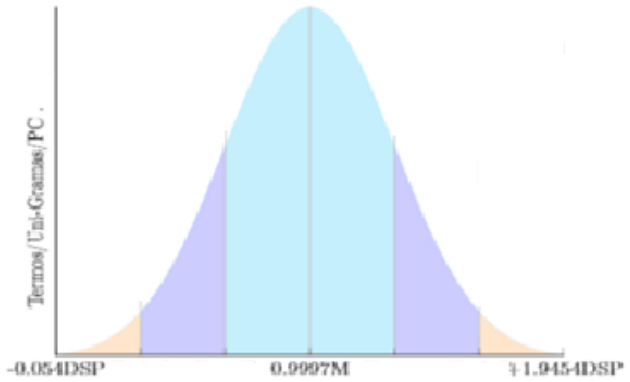


Figura 4.1: Fonte:Elaborado pelo autor (Gráfico de comparação de termos(PC))

e com os desvio padrão que são medidas em relação à , está corresponde a uma com os valores na qual são caracterizados por uma curva, e com uma linha que dividi a média desvio padrão, usando o Uni-Gramas onde que as cores demonstram o conjunto de termos por cores que se encontram no gráfico.

Tabela 4.17: Termos com maior e menores frequências no texto

Granularidade dos Conjuntos (Bi-Grams)		
Probabilidade Condicional(PC)		
M-DP(0.3876)	M(0,4258)	M+DP(0,464)
B	A	A
D	C	C
F	E	
H	G	
L	I	
M	J	
	K	
SOMA	5,5354	
Média	0,4258	
Desvio padrão	0,03823637	

Com base na tabela 4.17 probabilidade condicional usando a medida assimétrica presentar as conclusões dos termos com maior e menor frequência em cada método. No entanto, que mostra as pequenas diferenças podem ser visto embora difiram em cada medida, onde mostramos a lista ordenada que é uma dada pela nossa metodologia é feita desta maneira com base a tabela acima como o método Bi-Grams.

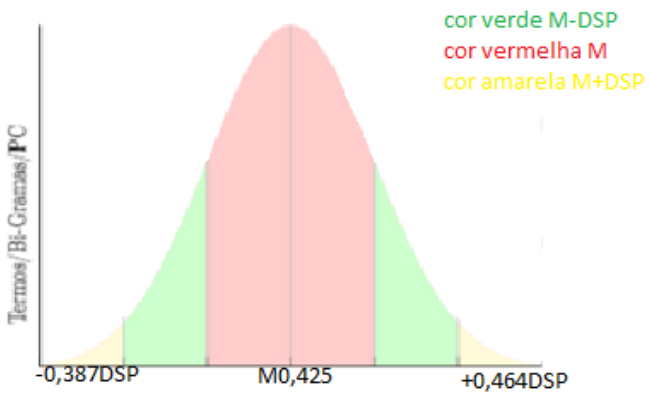


Figura 4.2: Fonte:Elaborado pelo autor (Gráfico de comparação de termos(PC)usando Bi-Gramas)

O gráfico 4.2 de probabilidade associadas a distribuição normal, apresentam o valor com média e com o desvio padrão que são medidas em relação à, está corresponde a uma com os valores na qual são caracterizados por uma curva, e com uma linha que dividia média e desvio padrão, usando o Bi-Grams onde que as cores demonstram o conjunto de termos por que se encontram no gráfico.

Tabela 4.18: Termos com maior e menores frequências no texto

Granularidade dos Conjuntos (Tri-Grams)		
Probabilidade Condicional(PC)		
M-DP(0,0419)	M(0.2516)	M+DP(0.4613)
I	B C E F H K	A D G J
SOMA	2,7684	
Média	0,25167	
Desvio padrão	0,2096	

A tabela 4.18 com probabilidade condicional usando a medida assimétrica, está a presenta as conclusões dos termos com maior e menor frequência em cada método. No entanto, que mostra as pequenas diferenças podem ser visto embora difiram em cada medida, onde mostramos a lista ordenada que é uma dada pela nossa metodologia é feita desta maneira com base a tabela acima com a granularidade Tri-Grams.

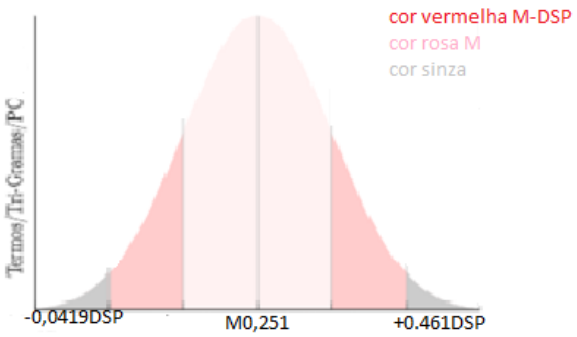


Figura 4.3: Fonte:Elaborado pelo autor (Gráfico de comparação de termos(PC)usando Tri-Gramas)

O gráficos 4.3 de probabilidade associadas a distribuição normal, apresentam o valor com média e com os desvios padrão que são medidas em relação a média, está corresponde a uma distribuição normal com os valores na qual são caracterizados por uma curva, e com uma linha que dividi a media e os desvios padrão, usando o Tri-Gramas onde que as cores demonstram o conjunto de termos por cores que se encontram no gráfico.

Comparação dos termos com o Laplace

Tabela 4.19: Termos com maior e menores frequências no texto

Granularidade dos Conjuntos (Uni-Grams)		
Laplace (PL)		
M-DP(4,4351)	M(5,9615)	M+DP(7,4879)
F	A E J L M C D G H I	B
SOMA	71,5389	
Média Desvio padrão	5,9615 1.5264	

A tabela 4.19 com o Laplace usando a medida assimétrica está a presenta as conclusões dos termos com maior e menor frequência em cada método. No entanto mostra as pequenas diferenças pode ser visto embora difiram em cada medida, onde mostramos a lista ordenada que é uma lista dada pela nossa metodologia é feita desta maneira com base a tabela acima com o método Uni-Grams.

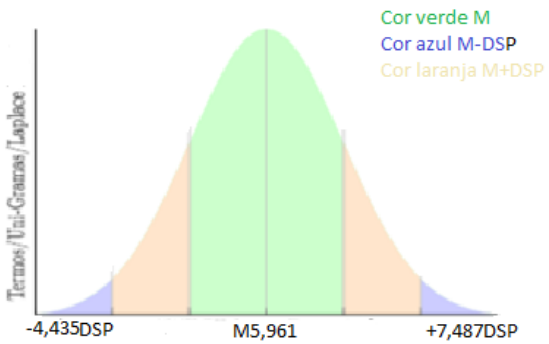


Figura 4.4: Fonte:Elaborado pelo autor (Gráfico de comparação de termos(Laplace)usando Uni-Gramas)

O gráfico 4.4 de probabilidade associadas a distribuição normal, apresentam o valor com média e com os desvios padrão que são medidas em relação a média, está corresponde a uma distribuição normal com os valores na qual são caracterizados por uma curva, e com uma linha que dividi a media e os desvios padrão, usando o Uni-Grams onde que as cores demonstram o conjunto de termos por cores que se encontram no gráfico.

Tabela 4.20: Termos com maior e menores frequências no texto

Granularidade dos Conjuntos (Bi-Grams)		
Laplace (PL)		
M-DP(0,2502)	M(3.1193)	M+DP(5.9884)
B F H L M	A C D E G I J K	
SOMA	40.5515	
Média Desvio padrão	3.1193 2,8691	

A tabela 4.20 com o Laplace usando a medida assimétrica está a presenta as conclusões dos termos com maior e menor frequência em cada método. No entanto que mostra as pequenas diferenças pode ser visto embora difiram em cada medida, onde mostramos a lista ordenada que é uma lista dada pela nossa metodologia é feita desta maneira com base a tabela acima com o método Bi-Grams.

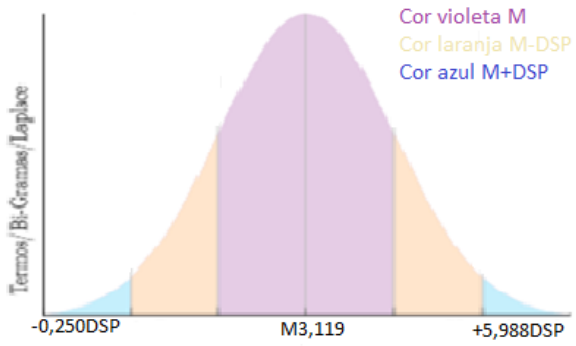


Figura 4.5: Fonte:Elaborado pelo autor (Gráfico de comparação de termos(Laplace)usando Bi-Gramas)

O gráfico 4.5 de probabilidade associadas a distribuição normal, apresentam o valor com média e com os desvios padrão que são medidas em relação a média, está corresponde a uma distribuição normal com os valores na qual são caracterizados por uma curva, e com uma linha que dividi a media e os desvios padrão, usando o Bi-Grams onde que as cores demonstram o conjunto de termos por cores que se encontram no gráfico.

Tabela 4.21: Termos com maior e menores frequências no texto

Granularidade dos Conjuntos (Tri-Grams)		
Laplace (PL)		
M-DP(0,683)	M(1,818)	M+DP(2,953)
I	B C E F H K	A D G J
SOMA	19.998	
Média Desvio padrão	1.818 1.135	

A tabela 4.21 com o Laplace usando a medida assimétrica está a presenta as conclusões dos termos com maior e menor frequência em cada método. No entanto que mostra as pequenas diferenças pode ser visto embora difiram em cada medida, onde mostramos a lista ordenada que é uma lista dada pela nossa metodologia é feita desta maneira com base a tabela acima com o método Tri-Grams.

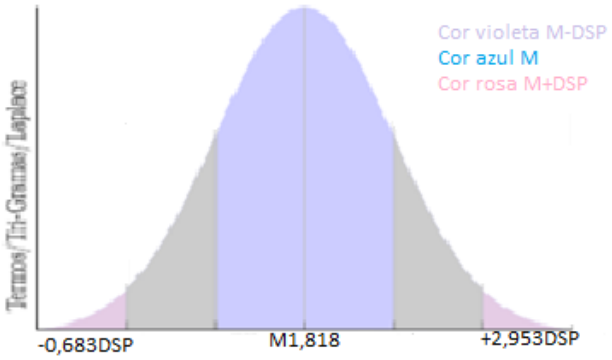


Figura 4.6: Fonte:Elaborado pelo autor (Gráfico de comparação de termos(Laplace)usando Tri-Gramas)

O gráfico 4.6 de probabilidade associadas a distribuição normal, apresentam o valor com média e com os desvios padrão que são medidas em relação a média, está corresponde a uma distribuição normal com os valores na qual são caracterizados por uma curva, e com uma linha que dividi a media e os desvios padrão, usando o Tri-Grams onde que as cores demonstram o conjunto de termos por cores que se encontram no gráfico.

Comparação dos termos com Braun-Blanket(BB)

Tabela 4.22: Termos com maior e menores frequências no texto

Granularidade dos Conjuntos (Uni-Grams)		
Braun-Blanket (BB)		
M-DP(0,1138)	M(1,6068)	M+DP(3,0998)
A E J L M F	B C D G H I	
SOMA	19,2816	
Média Desvio padrão	1.6068 1.493	

No entanto a tabela 4.22 Braun-Blanket usando a medida assimétrica com que mostra as pequenas diferenças pode ser visto embora diferem em cada medida, onde mostramos a lista ordenada que é uma lista dada pela nossa metodologia é feita desta maneira com base a tabela acima com a granularidade Uni-Grams.

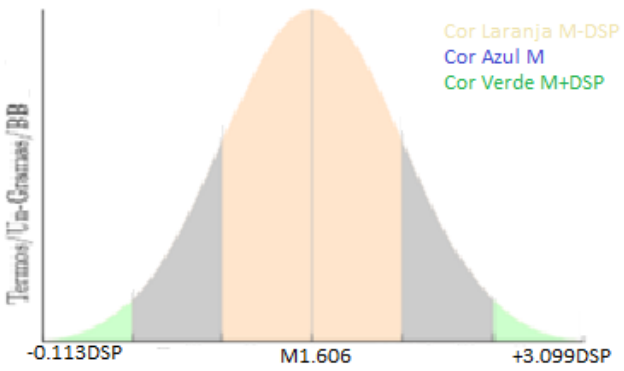


Figura 4.7: Fonte:Elaborado pelo autor (Gráfico de comparação de termos(BB)usando Uni-Gramas)

O gráfico 4.7 de probabilidade associadas a distribuição normal, apresentam o valor com média e com os desvios padrão que são medidas em relação a média, está corresponde a uma distribuição normal com os valores na qual são caracterizados por uma curva, e com uma linha que dividi a media e os desvios padrão, usando o Uni-Gramas onde que as cores demonstram o conjunto de termos por cores que se encontram no gráfico.

Tabela 4.23: Termos com maior e menores frequências no texto

Granularidade dos Conjuntos (Bi-Grams)		
Braun-Blanket (BB)		
M-DP(0.06)	M(0.7479)	M+DP(1.4358)
B F H L M	A C D E G I J K	
SOMA	9.7236	
Média Desvio padrão	0.7479 0.6879	

No entanto a tabela 4.23 Braun-Blanket usando a medida assimétrica com que mostra as pequenas diferenças pode ser visto embora difiram em cada medida, onde mostramos a lista ordenada que é uma lista dada pela nossa metodologia é feita desta maneira com base a tabela acima com o método Bi-Grams.

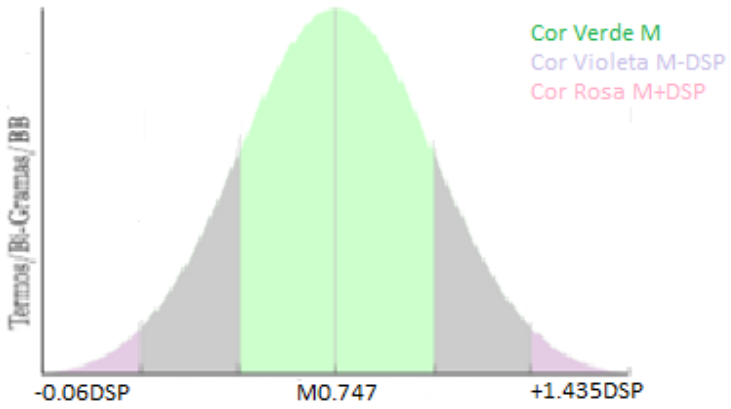


Figura 4.8: Fonte:Elaborado pelo autor (Gráfico de comparação de termos(BB)usando Bi-Gramas)

O gráfico 4.8 de probabilidade associadas a distribuição normal, apresentam o valor com média e com os desvios padrão que são medidas em relação a média, está corresponde a uma distribuição normal com os valores na qual são caracterizados por uma curva, e com uma linha que dividi a media e os desvios padrão, usando o Bi-Grams onde que as cores demonstram o conjunto de termos por cores que se encontram no gráfico.

Tabela 4.24: Termos com maior e menores frequências no texto

Granularidade dos Conjuntos (Tri-Grams)		
Braun-Blanket (BB)		
M-DP(0.3538)	M(0.4359)	M+DP(0.518)
B C E F H K I	A D	G J
SOMA	4.7952	
Média Desvio padrão	0.4359 0.0821	

No entanto a tabela 4.24 Braun-Blanket o usando a medida assimétrica com que mostra as pequenas diferenças pode ser visto embora difiram em cada medida, onde mostramos a lista ordenada que é uma lista dada pela nossa metodologia é feita desta maneira com base a tabela acima com o método Tri-Grams.

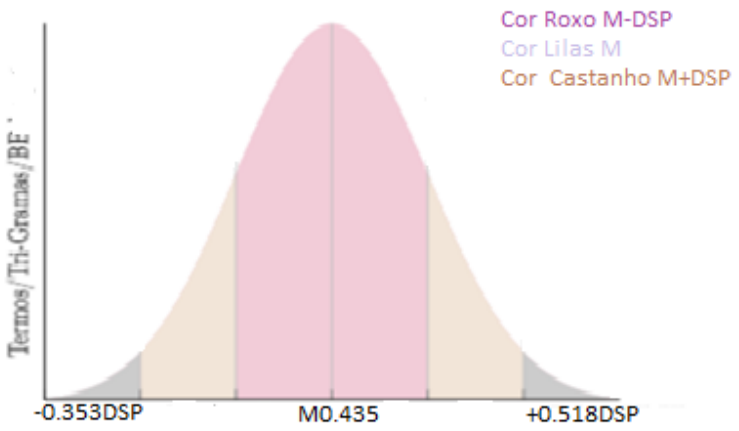


Figura 4.9: Fonte:Elaborado pelo autor (Gráfico de comparação de termos(BB)usando Tri-Gramas)

O gráfico 4.9 de probabilidade associadas a distribuição normal, apresentam o valor com média e com os desvios padrão que são medidas em relação a média, está corresponde a uma distribuição normal com os valores na qual são caracterizados por uma curva, e com uma linha que dividi a media e os desvios padrão, usando o Tri-Grams onde que as cores demonstram o conjunto de termos por cores que se encontram no gráfico.

Comparação dos termos com Conviction

Tabela 4.25: Termos com maior e menores frequências no texto

Granularidade dos Conjuntos (Uni-Grams)		
Conviction (CO)		
M-DP(0.194)	M(0.942)	M+DP(1.689)
A E J L M F	C D G H I	B
SOMA	11.304	
Média Desvio padrão	0.942 0.747	

Nesta tabela 4.25 Conviction usando a medida assimétrica com que mostra as pequenas diferenças pode ser visto embora difiram em cada medida, onde mostramos a lista ordenada que é uma lista dada pela nossa metodologia é feita desta maneira com base a tabela acima com o método Uni-Grams.

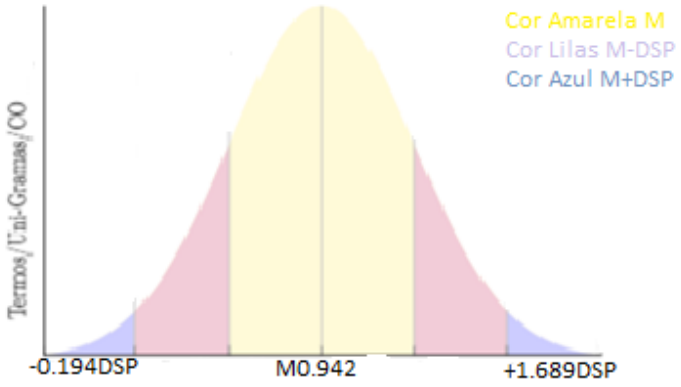


Figura 4.10: Fonte:Elaborado pelo autor (Gráfico de comparação de termos(CO)usando Uni-Gramas)

O gráfico 4.10 de probabilidade associadas a distribuição normal, apresentam o valor com média e com os desvios padrão que são medidas em relação a média, está corresponde a uma distribuição normal com os valores na qual são caracterizados por uma curva, e com uma linha que dividi a media e os desvios padrão, usando o Uni-Gramas onde que as cores demonstram o conjunto de termos por cores que se encontram no gráfico.

Tabela 4.26: Termos com maior e menores frequências no texto

Granularidade dos Conjuntos (Bi-Grams)		
Conviction (CO)		
M-DP(0.0364)	M(0.4259)	M+DP(0.8154)
B D F H L M	A C E G I J K	
SOMA	5.5368	
Média Desvio padrão	0.4259 0.3895	

Nesta tabela 4.26 Conviction usando a medida assimétrica com que mostra as pequenas diferenças pode ser visto embora difiram em cada medida, onde mostramos a lista ordenada que é uma lista dada pela nossa metodologia é feita desta maneira com base a tabela acima com o método Bi-Grams.

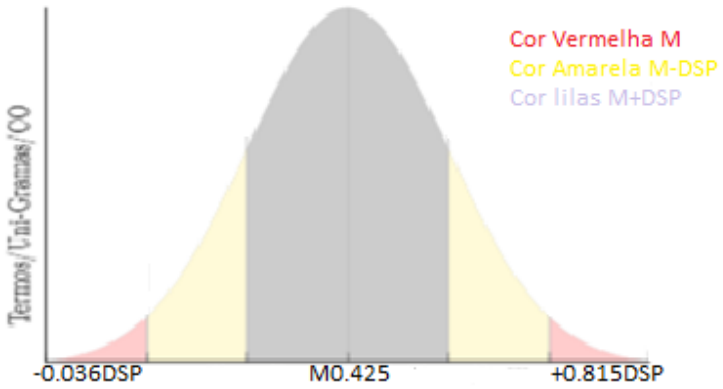


Figura 4.11: Fonte:Elaborado pelo autor (Gráfico de comparação de termos(CO)usando Bi-Gramas)

O gráfico 4.11 de probabilidade associadas a distribuição normal, apresentam o valor com média e com os desvios padrão que são medidas em relação a média, está corresponde a uma distribuição normal com os valores na qual são caracterizados por uma curva, e com uma linha que dividi a media e os desvios padrão, usando o Bi-Grams onde que as cores demonstram o conjunto de termos por cores que se encontram no gráfico.

Tabela 4.27: Termos com maior e menores frequências no texto

Granularidade dos Conjuntos (Tri-Grams)		
Conviction (CO)		
M-DP(0.2042)	M(0.2516)	M+DP(0.299)
B C E F H I	A D G J	
SOMA	2.768	
Média Desvio padrão	0.2516 0.0474	

A tabela 4.27 Conviction usando a medida assimétrica com que mostra as pequenas diferenças pode ser visto embora difiram em cada medida, onde mostramos a lista ordenada que é uma lista dada pela nossa metodologia é feita desta maneira com base a tabela acima com o método Tri-Grams.

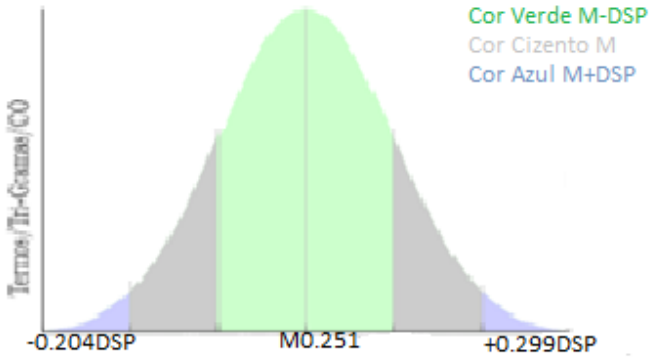


Figura 4.12: Fonte:Elaborado pelo autor (Gráfico de comparação de termos(CO)usando Tri-Grams)

O gráfico 4.12 de probabilidade associadas a distribuição normal, apresentam o valor com média e com os desvios padrão que são medidas em relação a média, está corresponde a uma distribuição normal com os valores na qual são caracterizados por uma curva, e com uma linha que dividi a media e os desvios padrão, usando o Tri-Grams onde que as cores demonstram o conjunto de termos por cores que se encontram no gráfico.

Comparação dos termos com Gini-Index(GI)

Tabela 4.28: Termos com maior e menores frequências no texto

Granularidade dos Conjuntos (Uni-Grams)		
Gini Index (GI)		
M-DP(0.9423)	M(22.1533)	M+DP(43.3643)
B A C D E F		
SOMA	243.6863	
Média Desvio padrão	22.1533 21.211	

A tabela 4.28 Gini Index usando a medida assimétrica com que mostra as pequenas diferenças pode ser visto embora difiram em cada medida, onde mostramos a lista ordenada que é uma lista dada pela nossa metodologia é feita desta maneira com base a tabela acima com o método Uni-Grams.

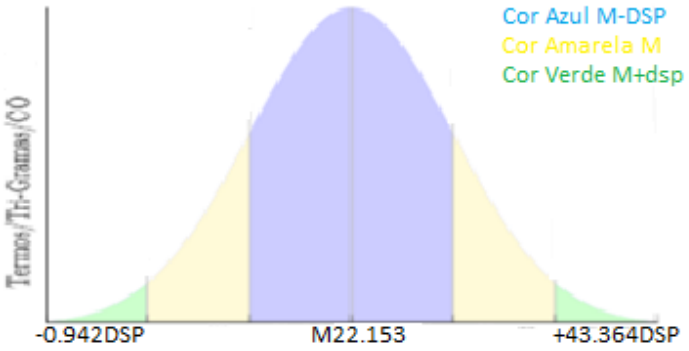


Figura 4.13: Fonte:Elaborado pelo autor (Gráfico de comparação de termos(CO)usando Uni-Gramas)

O gráfico 4.13 de probabilidade associadas a distribuição normal, apresentam o valor com média e com os desvios padrão que são medidas em relação a média, está corresponde a uma distribuição normal com os valores na qual são caracterizados por uma curva, e com uma linha que dividi a media e os desvios padrão, usando o Uni-Grams onde que as cores demonstram o conjunto de termos por cores que se encontram no gráfico.

Tabela 4.29: Termos com maior e menores frequências no texto

Granularidade dos Conjuntos (Bi-Grams)		
Gini Index (GI)		
M-DP(12.7797)	M(19.7691)	M+DP(26.7585)
	A B C D E F G H	
SOMA	243.6863	
Média Desvio padrão	22.1533 6.9894	

A tabela 4.29 Gini Index usando a medida assimétrica com que mostra as pequenas diferenças pode ser visto embora difiram em cada medida, onde mostramos a lista ordenada que é uma lista dada pela nossa metodologia é feita desta maneira com base a tabela acima com o método Bi-Grams.

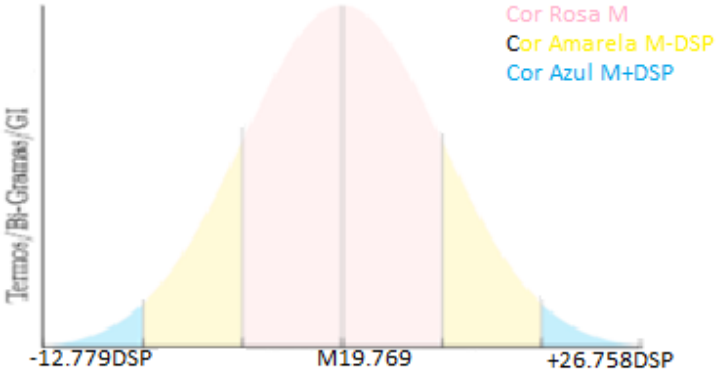


Figura 4.14: Fonte:Elaborado pelo autor (Gráfico de comparação de termos(CO)usando Bi-Gramas)

O gráfico 4.14 de probabilidade associadas a distribuição normal, apresentam o valor com média e com os desvios padrão que são medidas em relação a média, está corresponde a uma distribuição normal com os valores na qual são caracterizados por uma curva, e com uma linha que dividi a media e os desvios padrão, usando o Bi-Grams onde que as cores demonstram o conjunto de termos por cores que se encontram no gráfico.

Tabela 4.30: Termos com maior e menores frequências no texto

Granularidade dos Conjuntos (Tri-Grams)		
Gini Index (GI)		
M-DP(14.4158)	M(21.6236)	M+DP(28.8314)
A B C D	E F G H I	
SOMA	172.989	
Média Desvio padrão	21.6236 28.8314	

A tabela 4.30 Gini Index usando a medida assimétrica com que mostra as pequenas diferenças pode ser visto embora difiram em cada medida, onde mostramos a lista ordenada que é uma lista dada pela nossa metodologia é feita desta maneira com base a tabela acima com o método Tri-Grams.

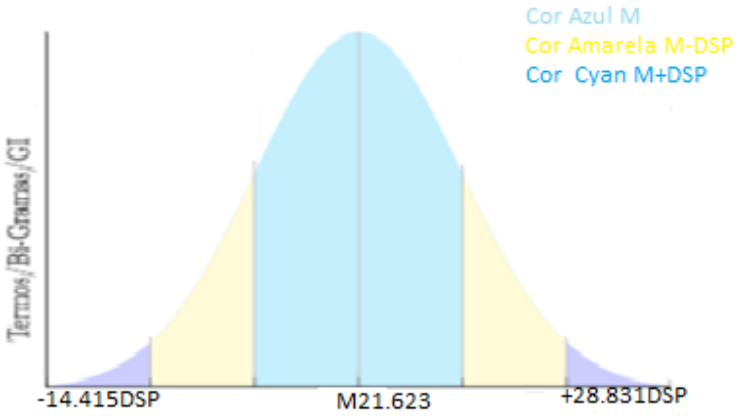


Figura 4.15: Fonte:Elaborado pelo autor (Gráfico de comparação de termos(CO)usando Tri-Grams)

O gráfico 4.15 de probabilidade associadas a distribuição normal, apresentam o valor com média e com os desvios padrão que são medidas em relação a média, está corresponde a uma distribuição normal com os valores na qual são caracterizados por uma curva, e com uma linha que dividi a media e os desvios padrão, usando o Tri-Grams onde que as cores demonstram o conjunto de termos por cores que se encontram no gráfico.

Tabela 4.31: Resumo dos Resultados do primeiro testo

Método	Média	M-DSP	M+DSP
Probabilidade Condicional (PC)	0,237	0,056	0.417
Medida Laplace	5,961	4,435	7,487
Braun-Blanket (BB)	1,606	0,113	3,099
Conviction (CO)	0,942	0,194	1,680
Gini Index (GI)	22,153	0.942	43,364

A tabela 4.31 representa todo conjunto dos resultados de todas medidas usados para este texto. Depois da avaliação ou de testes feitos em textos em Língua Portuguesa pensamos fazer uma outra avaliação com texto em Língua Francesa onde temos o seguinte texto com dados abaixo fornecemos um cálculo de amostra do AIS simplificado, com medida de valor acrescentado, de calculo de desvio padrão para os seguintes termos:

TEXTO 2- "João Lourenço Le président angolais est en visite officielle en France pour la première fois depuis son élection en septembre dernier. Le chef de l'Etat angolais lors de cette visite en France a été reçu lundi à l'Elysée par son homologue Emmanuel Macron. Le chef de l'Etat français s'est dit "très proche" du renforcement des relations avec l'Angola. Des questions régionales ont été évoquées en France et en RDC, Paris et Luanda appellent au respect de l'accord."

Neste texto escrito em língua Francesa com estes termos foram cuidadosamente extraídos, uma vez os resultados de todo processo dependem da qualidade dos palavras encontrados no texto, onde se encontra a representatividade dos termos em questão como uma compreensibilidade,

todos termos de em documentos são importantes para se descrever o seu conteúdo, que se encontra em língua francesa, com está forma, quando se tem sistema manual as expressões são determinados manualmente por um conhecedor da matéria, ou um especialista que domina a coleção. Tanto como o sistema automáticos ou semi-automáticos, estes são utilizadas técnicas especificas para este fim, com estas ajudam a extração de termos relevantes no texto ou no documento, podem vir aumentar ou reduzir as formas em que aparecem com a utilização das uni-grams ,Bi-Grams e Tri-Grams usando as cinco medidas de as similaridades assimétricas.

Tabela 4.32: Resultados dos termos com maior termos e com o maior Ranking com o método Uni-Grams

Granularidade dos Conjuntos Uni-Gramas		
Probabilidade Condicional (PC)		
M-DSP(0,3347687)	M(0,980769231)	M+DSP(1,626769762)
en	en	en
de	de	de
le	le	le
france	france	france
angolais	angolais	
visite	visite	
son	son	
chef	chef	
été	été	
des	des	
et	et	
AISs	51	
DESVIO PADRÃO	0,646000531	

A tabela 4.32 faz Comparação e demonstração dos resultados dos termos em media como as as técnicas Uni-Gramas, esta tabela tem 52 termos mas nós resumimos só os termos mas relevantes e que tem o maior ranking no texto, usando a probabilidade condicional.

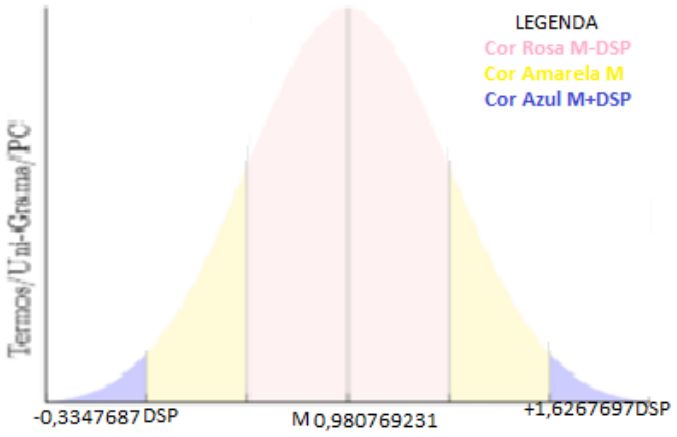


Figura 4.16: Fonte:Elaborado pelo autor (Gráfico de texto em Francês (PC)usando Uni-Gramas)

Este gráfico 4.16 mostra o Ranking dos 52 termos em língua Francesa com a medida assimétrica usando o método Uni-Gramas.

Tabela 4.33: Este resultado mostra o Ranking com a medida (BB) com o método Uni-Gramas

Granularidade dos Conjuntos Uni-Gramas		
Braun-Blanket (BB)		
M-DSP(1,263290478)	M(1,876142752)	M+DSP(2,488995026)
en	en	en
de	de	de
le	le	le
france	france	france
angolais	angolais	
visite	visite	
son	son	
chef	chef	
été	été	
AISs	98	
DESVIO PADRÃO	0,612852274	

A tabela 4.33 mostra o Ranking dos 52 termos em língua Francesa com a medida assimétrica usando o método (BB) Uni-Gramas.

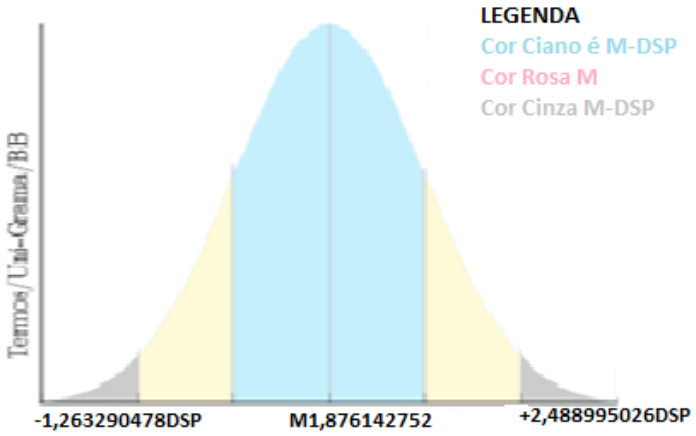


Figura 4.17: Fonte:Elaborado pelo autor (Gráfico de texto em Francês (BB)usando Uni-Gramas)

O gráfico 4.17 mostra o ranking dos termos por distribuição com a medida assimétrica usando a técnica Uni-gramas.

A tabela 4.34 mostra o Ranking dos 52 termos em língua Francesa com a medida assimétrica usando o método (CO) com a técnica Uni-Gramas.

Tabela 4.34: Este resultado mostra o Ranking com a medida (CO) com o método Uni-Gramas

Granularidade dos Conjuntos Uni-Gramas		
Conviction (CO)		
M-DSP(0,350166895	M(0,936158766)	M+DSP(1,522150637)
en	en	en
de	de	de
le	le	le
france	france	france
angolais	angolais	
visite	visite	
son	son	
chef	chef	
été	été	
des	des	
et	et	
AISs	49	
DESVIO PADRÃO	0,585991871	

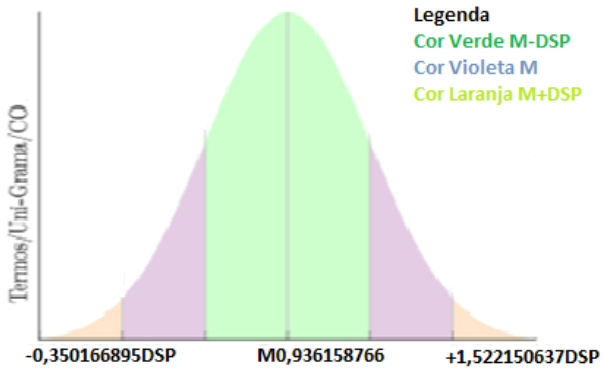


Figura 4.18: Fonte:Elaborado pelo autor (Gráfico de texto em Francês (CO)usando Uni-Gramas)

O gráfico 4.18 mostra o ranking dos termos por distribuição com a medida assimétrica usando o Uni-gramas.

Comparação dos termos com Gini Index e com o texto em Francês (GI)

A tabela 4.35 é uma das amostras de medida Gini Index com a técnica Uni-Gramas com os resultados a presentados onde que o M+DSP não tem termos que fazem parte neste conjunto, estes termos representados de mostram o rankig na qual não encontramos um termo mais relevante. Este gráfico 4.19 mostra o ranking dos termos por distribuição com a medida assimétrica usando o Uni-gramas, com (GI).

A tabela 4.36 é uma das amostras da medida Gini Index com o método Uni-Gramas com os resultados a presentados onde que o M-DSP, M+DSP e a Médias que fazem parte dos resultados deste conjunto.

Tabela 4.35: Este resultado mostra o Ranking com a medida (GI) com o método Uni-Gramas

Granularidade dos Conjuntos Uni-Gramas		
Gini Index (GI)		
M-DSP(0,966646285)	M(0,980262121)	M+DSP(0,993877957)
joão	joão	
lourenço	lourenço	
président	président	
est	est	
officielle	officielle	
pour	pour	
la	la	
première	première	
fois	fois	
depuis	depuis	
élection	élection	
septembre	septembre	
dernier	dernier	
lors	lors	
cette	cette	
a	a	
reçu	reçu	
lundi	lundi	
à	à	
par	par	
homologue	homologue	
emmanuel	emmanuel	
macron	macron	
français	français	
dit	dit	
très	très	
proche	proche	
du	du	
renforcement	renforcement	
relations	relations	
avec	avec	
questions	questions	
régionales	régionales	
ont	ont	
évoquées	évoquées	
rdc	rdc	
paris	paris	
luanda	luanda	
appellent	appellent	
au	au	
respect | respect		
AISs	51	
DESVIO PADRÃO	0,993877957	

O gráfico 4.20 mostra o ranking dos termos por distribuição com a medida assimétrica usando o Uni-gramas, com (Laplace).

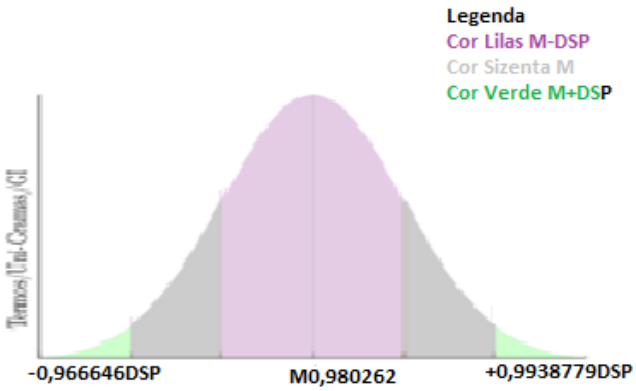


Figura 4.19: Fonte:Elaborado pelo autor (Gráfico de texto em Francês (GI)usando Uni-Gramas)

Tabela 4.36: Este resultado mostra o Ranking com a medida (Laplace) com o método Uni-Gramas

Granularidade dos Conjuntos Uni-Gramas		
Laplace (Laplace)		
M-DSP(22,97267257)	M(37,41217949)	M+DSP(51,85168641)
en	en	en
de	de	de
le	le	le
france	france	france
angolais	angolais	
visite	visite	
son	son	
chef	chef	
été	été	
des	des	
et	et	
AISs	1945	
DESVIO PADRÃO	14,43950692	

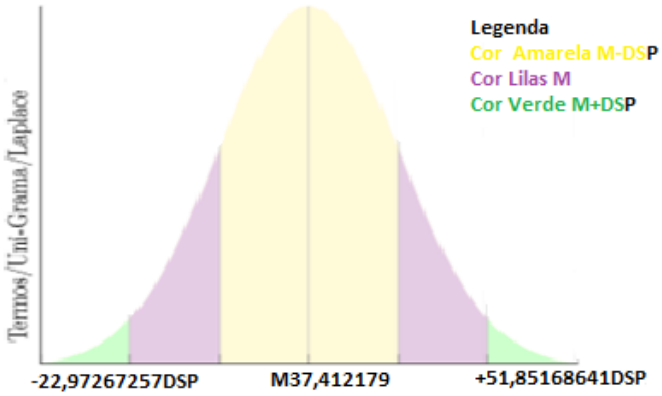


Figura 4.20: Fonte:Elaborado pelo autor (Gráfico de texto em Francês (Laplace)usando Uni-Gramas)

Estes gráficos acima a presentado tem uma distribuição obtida com medições reias que são obtidos com cálculos e formulas de as similaridades assimétricas, em relação com as frequências, a curva normal do gráfico correspondentes as distribuições que é igual a todos lados, onde que a linha é igual a outra como na direita e na esquerda quando temos assimetrias negativas e assimetrias positivas ou termos relevantes e não relevantes que é uma distribuição enforma de sino.
Os gráficos de probabilidade associadas a distribuição normal, apresentam o valor com média e com os desvios padrão que são medidas em relação a média, está corresponde a uma distribuição normal com os valores na qual são caracterizados por uma curva, e com uma linha que dividi a media e os desvio padrão, estes resultados foram calculadas em cinco medidas e três técnicas, usadas em Processamento de Linguagem Natural (PLN).

As tabelas abaixo mencionadas não tem os resumos em gráficos porque o procedimento é o mesmo e achamos melhor não repetirmos os mesmo métodos, neste contesto só teremos resultados em tabelas.

Amostra dos resultados dos termos em Língua Francesa usando a técnica Bi-Grama com 66 termos (PC)

Tabela 4.37: Resultados de Ranking com as cinco medida com o método Bi-Gramas (PC)

Granularidade dos Conjuntos Bi-Gramas		
Probabilidade Condicional(PC)		
M-DSP(0,699525779)	M(0,984848485)	M+DSP(1,270171191)
en france	en france	en france
le chef	le chef	le chef
chef de	chef de	chef de
João lourenço		
lourenço le		
e président		
président angolais		
angolais est		

A tabela 4.37 é o resultado apresenta-nos a comparação ou o Ranking de todos os termos que apresentam o maior Raking são os termos de M-DSP(Média menos do Desvio padrão) e o dois resultados tem os termos iguais, esta tabela nos mostra os resultados dos termos mais relevantes usando a técnica Bi-Grams com 66 termos com medida de (PC) Probabilidade Condicional.

Amostra dos resultados dos termos em Língua Francesa usando a técnica Bi-Grams com 66 termos (BB)

A tabela 4.38 mostra o resultado do Ranking de cada termos no seu conjunto para que demonstramos os termos com maior ranking ou com mais relevância no texto em língua francesa com 66 termos que é composto este texto.

Tabela 4.38: Resultados de Ranking com a medida com o método Bi-Gramas (BB)

Granularidade dos Conjuntos Bi-Gramas		
Braun-Blanket (BB)		
M-DSP(1,647145731)	M(1,909829087)	M+DSP(2,172512443)
en france	en france	en france
le chef	le chef	le chef
chef de	chef de	chef de
João lourenço		
lourenço le		
e président		
président angolais		
angolais est		

Amostra dos resultados dos termos em Língua Francesa usando a técnica Bi-Grams com 66 termos (CO)

Tabela 4.39: Resultados de Ranking com a medida com o método Conviction (CO)

Granularidade dos Conjuntos Bi-Gramas		
Conviction (CO)		
M-DSP(0,697690249)	M(0,954166963)	M+DSP(1,210643677)
en france	en france	en france
le chef	le chef	le chef
chef de	chef de	chef de
João lourenço		
lourenço le		
e président		
président angolais		
angolais est		
est en		
en visite		
visite officielle		
officielle en		
france pour		
pour la		
la première		
première fois		
fois depuis		
depuis son		
son élection		
élection en		

A tabela 4.39 amostra deste resultado do Ranking de cada termos no seu conjunto para que demonstramos os termos com maior ranking ou com mais relevância no texto em língua francesa com 66 termos que é composto este texto, em que tem 66 termos mais nós reduzir os termos para os termos mais relevantes .

Amostra dos resultados dos termos em Língua Francesa usando a técnica Bi-Grams com 66 termos (GI)

A tabela 4.40 mostra o resultado do Ranking de cada termos no seu conjunto para que demonstramos os termos com maior ranking ou com mais relevância no texto em língua francesa onde que M+DSP não tem termos, neste contexto não temos termo mais relevante nesta medida.

Tabela 4.40: Resultados de Ranking com a medida com o método Bi-Gramas (GI)

Granularidade dos Conjuntos Bi-Gramas		
Gini Index (GI)		
M-DSP(0,00703153)	M(0,984605796)	M+DSP(1,962180062)
João lourenço	João lourenço	
lourenço le	lourenço le	
e président	e président	
président angolais	président angolais	
angolais est	angolais est	
est en	est en	
en visite	en visite	
visite officielle	visite officielle	
officielle en	officielle en	
france pour	france pour	
pour la	pour la	
la première	la première	
première fois	première fois	
fois depuis	fois depuis	
depuis son	depuis son	
son élection	son élection	
élection en	élection en	
en septembre	en septembre	
septembre dernier	septembre dernier	
dernier le	dernier le	
de angolais	de angolais	
golais lors	golais lors	
lors de	lors de	
de cette	de cette	

Amostra dos resultados dos termos em Língua Francesa usando a técnica Bi-Grama com 66 termos (Laplace)

A tabela 4.41 mostra o resultado do Ranking de cada termos no seu conjunto para que demonstramos os termos com maior ranking ou com mais relevância no texto em língua francesa com a medida (Laplace), com a técnica Bi-Grams e com 66 termos resumido e apresentamos os resultados que se encontram nesta tabela.

Tabela 4.41: Resultados de Ranking com a medida com o método Bi-Grams (Laplace)

Granularidade dos Conjuntos Bi-Gramas Laplace (Laplace)		
M-DSP(37,65598011)	M(44,0439399)	M+DSP(50,43189868)
en france	en france	en france
le chef	le chef	le chef
chef de	chef de	chef de
João lourenço		
lourenço le		
e président		
président angolais		
angolais est		
est en		
en visite		
visite officielle		
officielle en		
france pour		
pour la		
la première		
première fois		
fois depuis		
depuis son		
son élection		

As amostras dos resultados dos termos em Língua Francesa usando a técnica Tri-Grams com 68 termos (PC)

Tabela 4.42: Resultados de Ranking com a medida com o método Tri-Gramas (PC) com 68 termos mais nós resumimos

Granularidade dos Conjuntos Tri-Grams Probabilidade Condicional (PC)		
M-DSP(0,868410351)	M(0,985294118)	M+DSP(1,102177885)
le chef de	le chef de	
joão lourenço le		
lourenço le président		
le président angolais		
président angolais est		
angolais est en		
est en visite		
en visite officielle		
visite officielle en		
officielle en france		
en france pour		
france pour la		
pour la première		
la première fois		
première fois depuis		
fois depuis son		
depuis son élection		
son élection en		
élection en septembre		
en septembre dernier		
septembre dernier le		

A tabela 4.42 nos apresenta uma mostra em que não tem o termos mais relevantes é uma das amostradas em que calculamos as frequências usando a medida (PC) Probabilidade Condicional e com a técnica Tri-Grams com 68 termos.

As amostras dos resultados dos termos em Língua Francesa usando a técnica Tri-Grams com 68 termos (BB)

Tabela 4.43: Resultados de Ranking com a medida com o método Tri-Gramas (BB)

Granularidade dos Conjuntos Tri-Grams		
	Braun-Blanket (BB)	
M-DSP(1,801902879)	M(1,913938101)	M+DSP(2,025973323)
le chef de	le chef de	le chef de
joão lourenço le		
lourenço le président		
le président angolais		
président angolais est		
angolais est en		
est en visite		
en visite officielle		
visite officielle en		
officielle en france		
en france pour		
france pour la		
pour la première		
la première fois		
première fois depuis		
fois depuis son		
depuis son élection		
son élection en		

A tabela 4.43 nos apresenta uma mostra em que não tem o termos mais relevantes é uma das amostradas em que calculamos as frequências usando a medida (BB) Braun-Blanket e com a técnica Tri-Grams com 68 termos com estas amostramos encontramos um termos que é mais relevante neste texto.

As amostras dos resultados dos termos em Língua Francesa usando a técnica Tri-Grama com 68 termos (CO)

Tabela 4.44: Resultados de Ranking com a medida com o método Tri-Gramas (CO)

Granularidade dos Conjuntos Tri-Grams		
	Conviction (CO)	
M-DSP(0,621622034)	M(0,956338663)	M+DSP(1,291055292)
le chef de	le chef de	le chef de
joão lourenço le		
lourenço le président		
le président angolais		
président angolais est		
angolais est en		
est en visite		
en visite officielle		
visite officielle en		

A tabela 4.44 com esta medida encontramos uma palavra relevante em que conseguimos usar a

medida Conection(CO) usando as técnicas Tri-Grams.

As amostras dos resultados dos termos em Língua Francesa usando a técnica Tri-Grams com 68 termos (GI)

Tabela 4.45: Resultados de Ranking com a medida com o método Tri-Gramas (GI)

Granularidade dos Conjuntos Tri-Grams		
Gini Index (GI)		
M-DSP(0,605719238)	M(0,985078169)	M+DSP(1,3644371)
joão lourenço le	joão lourenço le	
lourenço le président	lourenço le président	
le président angolais	le président angolais	
président angolais est	président angolais est	
angolais est en	angolais est en	
est en visite	est en visite	
en visite officielle	en visite officielle	
visite officielle en	visite officielle en	
officielle en france	officielle en france	
en france pour	en france pour	
france pour la	france pour la	
pour la première	pour la première	
la première fois	la première fois	
première fois depuis	première fois depuis	
fois depuis son	fois depuis son	
depuis son élection	depuis son élection	
son élection en	son élection en	
élection en septembre	élection en septembre	
en septembre dernier	en septembre dernier	
septembre dernier le	septembre dernier le	
dernier le chef	dernier le chef	
chef de angolais	chef de angolais	
de angolais lors	de angolais lors	
angolais lors de	angolais lors de	

A tabela 4.45 amostras apresentadas nesta tabela não encontramos termos relevantes onde que nós usamos as medida Gini Index e com a técnica Tri-Grams, usando 68 termos.

As amostras dos resultados dos termos em Língua Francesa usando a técnica Tri-Grams com 68 termos (Laplace)

Tabela 4.46: Resultados de Ranking com a medida com o método Tri-Grams (Laplace)

Granularidade dos Conjuntos Tri-Grams Laplace (Laplace)		
M-DSP(42,12249357)	M(44,83088235)	M+DSP(47,53927114)
le chef de	le chef de	le chef de
joão lourenço le		
lourenço le président		
le président angolais		
président angolais est		
angolais est en		
est en visite		
en visite officielle		
visite officielle en		
officielle en france		
en france pour		
france pour la		
pour la première		

Com os teste realizados nestas medidas está tabela 4.46 mostra os resultados dos termos em só temos 1 termo relevante onde temos 68 termos usando a medida Laplace e com a técnica tri-Grams.
A tabela 4.47 a baixo mostra o resumo dos resultados no texto escrito em Língua Francesa, e usando as três técnicas e as Cinco medidas.

Tabela 4.47: esta tabela faz o resumo de todos resultados das medidas mencionadas na tabela

Medidas	Uni-Grams			Bi-Gramas		
	Media	M-DSP	M+DSP	Media	M-DSP	M+DSP
PC	0,980769231	0,3347687	1,6267697	0,984848485	0,699525779	1,270171191
BB	1,876142752	1,263290478	2,488995026	1,909829087	1,647145731	2,172512443
CO	0,936158766	0,350166895	1,522150637	0,954166963	0,697690249	1,210643677
GI	0,980262121	0,96664285	0,993877957	0,984605796	0,00703153	1,962180062
LAPLACE	37,41217949	22,97267257	51,85168641	44,0439399	37,65598011	50,43189868

Tabela 4.48: esta tabela faz o resumo de todos resultados das medidas mencionadas na tabela(CONTINUAÇÃO)

Medidas	Tri-Grams		
	Media	M-DSP	M+DSP
PC	0,985294118	0,868410351	1,102177885
BB	1,913938101	1,801902879	2,025973323
CO	0,956338663	0,621622034	1,291055292
GI	0,985078169	0,605719238	1,3644371
LAPLACE	44,83088235	42,12249357	47,53927114

Neste contexto quando da nossa avaliação realizada em outros textos em línguas diferentes como língua Portuguesa e Língua Francesas também preferimos a avaliar na língua Espanhola, com as cinco (5) medidas e três (3) granularidades como Uni-Grams, Bi-Grams, Tri-Grams, com estas medidas e técnicas vão nos ajudar a calcular a de mostrar o AIS na simplificação com as medidas usadas e vai nos permitir mostrar o ranking e os termos mais relevantes neste texto, abaixo mencionado:

TEXTO 3- "La oferta de Irán, Portugal, que participó en esta oferta juego capitán Cristiano Ronaldo y Pouraliganji llevó al árbitro a utilizar el var, previa consulta al var, la oferta y las imágenes de Cristiano Ronaldo de Portugal elegido para mostrar la oferta y Cato amarilla a Cristiano Ronaldo capitán de Portugal, y Pouraliganji mostró una tarjeta amarilla, este encaje permitido Cristiano Ronaldo le llevó a jugar con Uruaguai,Portugues."

Depois de algumas avaliações ou testes utilizados em outras línguas como ; língua Portuguesa e Francesa, nós achamos melhor realizar outros testes em língua Espanhola porque estamos a tratar de medidas não supervisionadas independente da Língua.

As amostras dos resultados dos termos em Língua Espanhola usando a técnica Uni-Grama com 41 termos (PC)

Tabela 4.49: Resultados de Ranking com a medida e com a técnica Uni-Gramas (Probabilidade Condicional) em Língua Espanhola com 41 termos

Granularidade dos conjuntos usando a técnica Uni-Gramas Probabilidade Condicional (PC) em Língua Espanhola		
M-DSP(0,357846481)	M(0,97560975609756)	M+DSP(1,593373031)
oferta	oferta	oferta
irán	irán	irán
ronaldo	ronaldo	ronaldo
y	y	y
pouraliganji	pouraliganji	pouraliganji
la	la	
que	que	
tilizar	tilizar	
cristiano	cristiano	
al	al	
árbitro	árbitro	
previa	previa	
mostró	mostró	
llevó	llevó	
Portugal		
participó		
en		
esta		
juego		
capitán		
AISs	39.99996	
Desvio Padrão	0.6177632	

A tabela 4.49 mostra Resultados de Ranking com a medida e com a técnica Uni-Grams (Probabi-

lidade Condicional) em Língua Espanhola com 41 termos.

As amostras dos resultados dos termos em Língua Espanhola usando a técnica Uni-Grama com 41 termos (BB)

Tabela 4.50: Resultados de Ranking com a medida e com a técnica Uni-Gramas (Braun-Blanke) em Língua Espanhola

Granularidade dos conjuntos usando a Técnica Uni-Gramas Braun-Blanket (BB) em Língua Espanhola		
M-DSP(1,268513612)	M(1,845730179)	M+DSP(2,422946746)
oferta	oferta	oferta
irán	irán	irán
ronaldo	ronaldo	ronaldo
y	y	y
pouraliganji	pouraliganji	pouraliganji
la	la	la
que	que	que
tilizar	tilizar	tilizar
cristiano	cristiano	
llevó	llevó	
al	al	
árbitro	árbitro	
previa	previa	
mostró	mostró	
portugal		
participó		
en		
AISs	75.6749373	
DSP	0.5772165673	

A tabela 4.50 Resultados de Ranking com a medida e com a técnica Uni-Grams (Braun-Blanke) em Língua Espanhola.

As amostras de resultados de termos em Língua Espanhola usando a técnica Uni-Grama com 41 termos (CONVICTION)

Tabela 4.51: Resultados de Ranking com a medida e com a técnica Uni-Gramas (Conviction) em Língua Espanhola

Granularidade dos conjuntos usando a Técnica Uni-Gramas Conviction (CO) em Língua Espanhola		
M-DSP(0,357265776)	M(0,920155245)	M+DSP(1,483044714)
oferta	oferta	oferta
irán	irán	irán
ronaldo	ronaldo	ronaldo
y	y	y
pouraliganji	pouraliganji	pouraliganji
la	la	la
que	que	que
tilizar	tilizar	tilizar
cristiano	cristiano	
llevó	llevó	
al	al	
árbitro	árbitro	
previa	previa	
mostró	mostró	
portugal		
participó		
AISs	37.7263650569	
DSP	0.5628894691559	

A tabela 4.51 mostra o resultados de Ranking com a medida e com a técnica Uni-Gramas (Conviction) em Língua Espanhola.

As amostras dos resultados dos termos em Língua Espanhola usando a técnica Uni-Grams com 41 termos (Gini Index)

Tabela 4.52: Resultados de Ranking com a medida e com a técnica Uni-Gramas (Gini Index) em Lingua Espanhola

Granularidade dos conjuntos usando a Técnica Uni-Grams Gini Index (GI) em Língua Espanhola		
M-DSP(0,958357184)	M(0,974813275)	M+DSP(0,991269367)
portugal	portugal	
participó	participó	
en	en	
esta	esta	
juego	juego	
capitán	capitán	
a	a	
el	el	
var	var	
consulta	consulta	
las	las	
imágenes	imágenes	
elegido	elegido	
para	para	
mostrar	mostrar	
cato	cato	
amarilla	amarilla	
una	una	
tarjeta	tarjeta	
este	este	
encaje	encaje	
permitido	permitido	
l	l	
e	e	
jugar	jugar	
con	con	
uruaguai	uruaguai	
AISs	39.967344290657	
Dsp	0.01645609115861	

A tabela 4.52 mostra resultados de Ranking com a medida e com a técnica Uni-Grams (Gini Index) em Lingua Espanhola. Está medida mostra os resultados obtidos nos cálculos utilizados nesta medida usando a técnica Bi-Gramas, em temos 41 termos e com estes medida não temos termos relevantes porque a medida M+DSP não encontramos um ranking neste método.

As amostras dos resultados dos termos em Língua Espanhola usando a técnica Uni-Grama com 41 termos (Laplace)

Tabela 4.53: Resultados de Ranking com a medida e com a técnica Uni-Grams (Laplace) em Língua Espanhola

Granularidade dos conjuntos usando a Técnica Uni-Grams Laplace (Laplace) em Língua Espanhola		
M-DSP(18,5456131)	M(31,01626016)	M+DSP(43,48687901)
oferta	oferta	oferta
de	de	de
cristiano	cristiano	cristiano
ronaldo	ronaldo	ronaldo
y	y	y
la	la	la
portugal	portugal	portugal
a	a	a
capitán	capitán	
pouraliganji	pouraliganji	
llevó	llevó	
al	al	
var	var	
amarilla	amarilla	
irán		
que		
AISs	1271.666666666	
DSP	12.470618850847	

A tabela 4.53 mostra o resultado mostra-nos os termos mais relevantes consoante os cálculos realizados usando a medida de Laplace e com granularidade Uni-Grams.

As amostras de resultados de termos em Língua Espanhola usando a técnica Bi-Grams com 59 termos (Probabilidade Condicional) está tabela mostra os resultados dos termos e o ranking dos termos mais relevantes deste testo usando as medidas assimétricas.

A tabela 4.54 mostra o resultados de Ranking com a medida e com a técnica Bi-Grams (Probabilidade Condicional) em Lingua Espanhola.

As amostras de resultados de termos em Língua Espanhola usando a técnica Bi-Grams com 59 termos (Braun-Blanket) está tabela mostra os resultados dos termos e o ranking dos termos mais relevantes deste testo usando as medidas assimétricas.

A tabela 4.55 mostra o resultados de Ranking com a medida e com a técnica Bi-Gramas (Braun-Blanket) em Lingua Espanhola.

Tabela 4.54: Resultados de Ranking com a medida e com a técnica Bi-Gramas (Probabilidade Condicional) em Lingua Espanhola

Granularidade dos conjuntos usando a Técnica Bi-Gramas Probabilidade Condicional(PC) em Língua Espanhola		
M-DSP(0,547797104)	M(0,983050847)	M+DSP(1,418304591)
cristiano ronaldo	cristiano ronaldo	cristiano ronaldo
la oferta	la oferta	la oferta
y pouraliganji	y pouraliganji	y pouraliganji
oferta y	oferta y	oferta y
de portugal	de portugal	de portugal
oferta de		
de irán		
irán portugal		
portugal que		
que participó		
participó en		
en esta		
AISs	58.0000068	
DSP	0.4352537437404	

Tabela 4.55: Resultados de Ranking com a medida e com a técnica Bi-Gramas (Braun-Blanket) em Lingua Espanhola

Granularidade dos conjuntos usando a Técnica Bi-Gramas Braun-Blanket (BB) em Língua Espanhola		
M-DSP(1,48126472)	M(1,896355473)	M+DSP(2,311446226)
cristiano ronaldo	cristiano ronaldo	cristiano ronaldo
la oferta	la oferta	la oferta
y pouraliganji	y pouraliganji	y pouraliganji
oferta y	oferta y	oferta y
de portugal	de portugal	de portugal
oferta de		
de irán		
que participó		
participó en		
en esta		
esta oferta		
oferta juego		
juego capitán		
capitán cristiano		
AISs	111.8849728907	
DSP	0.41509075289018	

As amostras de resultados de termos em Língua Espanhol usando a técnica Bi-Grams com 59 termos *(CONVICTION)* está tabela mostra os resultados dos termos e o ranking dos termos mais relevantes deste testo usando as medidas assimétricas A tabela 4.56 mostra o resultados de Ranking com a medida e com a técnica Bi-Grams (Conviction) em Lingua Espanhola.

Tabela 4.56: Resultados de Ranking com a medida e com a técnica Bi-Grams (Conviction) em Lingua Espanhola

Granularidade dos conjuntos usando a Técnica Bi-Grams Conviction (CO) em Língua Espanhola		
M-DSP(0,545841787)	M(0,947057702)	M+DSP(1,348273618)
cristiano ronaldo	cristiano ronaldo	cristiano ronaldo
la oferta	la oferta	la oferta
y pouraliganji	y pouraliganji	y pouraliganji
oferta y	oferta y	oferta y
de portugal	de portugal	de portugal
oferta de		
de irán		
irán portugal		
portugal que		
que participó		
participó en		
en esta		
esta oferta		
oferta juego		
juego capitán		
capitán cristiano		
ronaldo y		
pouraliganji llevó		
llevó al		
al árbitro		
árbitro a		
a utilizar		
utilizar el		
el var		
var previa		
previa consulta		
consulta al		
al var		
var la		
y las		
las imágenes		
imágenes de		
de cristiano		
ronaldo de		
portugal elegido		
elegido para		
para mostrar		
AISs	55.87640443463319	
DSP	0.4012159152850336	

As amostras de resultados de termos em Língua Espanhola usando a técnica Bi-Grams com 59 termos (GINI INDEX) está tabela mostra os resultados dos termos e o ranking dos termos mais relevantes deste testo usando as medidas assimétricas

A tabela 4.57 mostra o resultados de Ranking com a medida e com a técnica Bi-Grams (Gini Index) em Lingua Espanhol. Nesta medida de pois dos cálculos realizados não encontramos termos mais relevantes ou no método M+DSP(Media mais o desvio padrão) onde que temos 59 termos foi usado a medida Gini Index e com a técnica Bi-Grams.

Tabela 4.57: Resultados de Ranking com a medida e com a técnica Bi-Grams (Gini Index) em Lingua Espanhol

Granularidade dos conjuntos usando a Técnica Bi-Grams Gini Index (GI) em Língua Espanhola		
M-DSP(0,974746618)	M(0,982716782)	M+DSP(0,990686946)
oferta de	oferta de	
de irán	de irán	
irán portugal	irán portugal	
portugal que	portugal que	
que participó	que participó	
participó en	participó en	
en esta	en esta	
esta oferta	esta oferta	
oferta juego	oferta juego	
juego capitán	juego capitán	
capitán cristiano	capitán cristiano	
ronaldo y	ronaldo y	
pouraliganji llevó	pouraliganji llevó	
llevó al	llevó al	
al árbitro	al árbitro	
árbitro a	árbitro a	
a utilizar	a utilizar	
utilizar el	utilizar el	
el var	el var	
var previa	var previa	
previa consulta	previa consulta	
consulta al	consulta al	
al var	al var	
var la	var la	
y las	y las	
las imágenes	las imágenes	
imágenes de	imágenes de	
de cristiano	de cristiano	
ronaldo de	ronaldo de	
portugal elegido	portugal elegido	
elegido para	elegido para	
para mostrar	para mostrar	
AISs	57.98029012877228	
DSP	0.00797016400667541	

As amostras de resultados de termos em Língua Espanhola usando a técnica Bi-Grams com 59 termos (Laplace) está tabela mostra os resultados dos termos e o ranking dos termos mais relevantes deste testo usando as medidas assimétricas

A tabela 4.58 mostra os resultados de Ranking com a medida e com a técnica Bi-Grams (Laplace) em Lingua Espanhola.

Tabela 4.58: Resultados de Ranking com a medida e com a técnica Bi-Grams (Laplace) em Lingua Espanhola

Granularidade dos conjuntos usando a Técnica Bi-Grams Laplace (Laplace) em Língua Espanhola		
M-DSP(30,5694629)	M(40,14943503)	M+DSP(49,72940716)
cristiano ronaldo	cristiano ronaldo	cristiano ronaldo
la oferta	la oferta	la oferta
y pouraliganji	y pouraliganji	y pouraliganji
oferta y	oferta y	oferta y
de portugal	de portugal	de portugal
oferta de		
de irán		
irán portugal		
portugal que		
que participó		
participó en		
en esta		
esta oferta		
oferta juego		
juego capitán		
capitán cristiano		
ronaldo y		
pouraliganji llevó		
llevó al		
al árbitro		
árbitro a		
a utilizar		
AISs	2368.8166666666652	
DSP	9.579972127873136	

As amostras de resultados de termos em Língua Espanhola usando a técnica Tri-Grams com 65 termos (Probabilidade Condicional) está tabela mostra os resultados dos termos e o ranking dos termos mais relevantes deste testo usando as medidas assimétricas

A tabela 4.59 tem resultados de Ranking com a medida e com a técnica Tri-Grams (Probabilidade Condicional) em Lingua Espanhola.

Tabela 4.59: Resultados de Ranking com a medida e com a técnica Tri-Grams (Probabilidade Condicional) em Lingua Espanhola

Granularidade dos Conjuntos Usando Técnicas Tri-Grams Probabilidade Condicional(PC)		
M-DSP(0,865268065)	M(0,984615385)	M+DSP(1,103962704)
la oferta y	la oferta y	
la oferta de		
oferta de irán		
de irán portugal		
irán portugal que		
portugal que participó		
que participó en		
participó en esta		
en esta oferta		
esta oferta juego		
oferta juego capitán		
juego capitán cristiano		
capitán cristiano ronaldo		
cristiano ronaldo y		
ronaldo y pouraliganji		
y pouraliganji llevó		
pouraliganji llevó al		
llevó al árbitro		
al árbitro a		
árbitro a utilizar		
AISs	63.99999999999995	
DSP	0.11934731934731921	

As amostras de resultados de termos em Língua Espanhola usando a técnica Tri-Grams com 65 termos (Braun-Blanket) está tabela mostra os resultados dos termos e o ranking dos termos mais relevantes deste testo usando as medidas assimétricas

A tabela 4.60 mostra o resultados de Ranking com a medida e com a técnica Tri-Grams (Braun-Blanket) em Lingua Espanhola.

Tabela 4.60: Resultados de Ranking com a medida e com a técnica Tri-Grams (Braun-Blanket) em Lingua Espanhola

Granularidade dos Conjuntos Usando Técnicas Tri-Grams		
	Braun-Blanket (BB)	
M-DSP(1,795888817)	M(1,910071105)	M+DSP(2,024253394)
la oferta y	la oferta y	la oferta y
la oferta de		
oferta de irán		
de irán portugal		
irán portugal que		
portugal que participó		
que participó en		
participó en esta		
en esta oferta		
esta oferta juego		
oferta juego capitán		
juego capitán cristiano		
capitán cristiano ronaldo		
cristiano ronaldo y		
ronaldo y pouraliganji		
y pouraliganji llevó		
pouraliganji llevó al		
llevó al árbitro		
al árbitro a		
árbitro a utilizar		
a utilizar el		
utilizar el var		
el var previa		
var previa consulta		
previa consulta al		
AISs	124.15462184873913	
DSP	0.11418228829993544	

As amostras de resultados de termos em Língua Espanhol usando a técnica Tri-Grama com 65 termos (CONVICTIONt) está tabela mostra os resultados dos termos e o ranking dos termos mais relevantes deste testo usando as medidas assimétricas

A tabela 4.61 mostra os resultados de Ranking com a medida e com a técnica Tri-Grams (Conviction) em Lingua Espanhola.

Tabela 4.61: Resultados de Ranking com a medida e com a técnica Tri-Grams (Conviction) em Lingua Espanhola

Granularidade dos Conjuntos Usando Técnicas Tri-Grams Conviction (CO)		
M-DSP(0,84201576)	M(0,954346674)	M+DSP(1,066677588)
la oferta y	la oferta y	la oferta y
la oferta de		
oferta de irán		
de irán portugal		
irán portugal que		
portugal que participó		
que participó en		
participó en esta		
en esta oferta		
esta oferta juego		
oferta juego capitán		
juego capitán cristiano		
capitán cristiano ronaldo		
cristiano ronaldo y		
ronaldo y pouraliganji		
y pouraliganji llevó		
pouraliganji llevó al		
llevó al árbitro		
al árbitro a		
árbitro a utilizar		
a utilizar el		
utilizar el var		
el var previa		
AISs	62.03253381898995	
DSP	0.11233091363918513	

As amostras de resultados de termos em Língua Espanhola usando a técnica Tri-Grama com 65 termos (Gini Index) está tabela mostra os resultados dos termos e o ranking dos termos mais relevantes deste testo usando as medidas assimétricas

A tabela 4.62 mostra o resultados de Ranking com a medida e com a técnica Tri-Grams (Gini Index) em Lingua Espanhola.

Tabela 4.62: Resultados de Ranking com a medida e com a técnica Tri-Grams (Gini Index) em Lingua Espanhola

Granularidade dos Conjuntos usando Tecnicas Tri-Grams Gini Index (GI)		
M-DSP(0,982432505)	M(0,984379074)	M+DSP(0,986325643)
la oferta de	la oferta de	
oferta de irán	oferta de irán	
de irán portugal	de irán portugal	
irán portugal que	irán portugal que	
portugal que participó	portugal que participó	
que participó en	que participó en	
participó en esta	participó en esta	
en esta oferta	en esta oferta	
esta oferta juego	esta oferta juego	
oferta juego capitán	oferta juego capitán	
juego capitán cristiano	juego capitán cristiano	
capitán cristiano ronaldo	capitán cristiano ronaldo	
cristiano ronaldo y	cristiano ronaldo y	
ronaldo y pouraliganji	ronaldo y pouraliganji	
y pouraliganji llevó	y pouraliganji llevó	
pouraliganji llevó al	pouraliganji llevó al	
llevó al árbitro	llevó al árbitro	
al árbitro a	al árbitro a	
árbitro a utilizar	árbitro a utilizar	
a utilizar el	a utilizar el	
utilizar el var	utilizar el var	
el var previa	el var previa	
var previa consulta	var previa consulta	
previa consulta al	previa consulta al	
consulta al var	consulta al var	
al var la	al var la	
AISs	63.98463978	
DSP	0.0019465688	

As amostras de resultados de termos em Língua Espanhola usando a técnica Tri-Grama com 65 termos (Laplace) está tabela mostra os resultados dos termos e o ranking dos termos mais relevantes deste testo usando as medidas assimétricas

A tabela 4.63 mostra os termos mais relevantes usamos medida de laplace com técnica Tri-Grams com 65 termos.
Este resumo desta tabela 4.64 mostra os resultados dos testes utilizados em Língua Espanhola usando as técnicas Uni-Grams, Bi-Grams, Tri-Grams,e os cinco métodos.

Tabela 4.63: Resultados de Ranking com a medida e com a técnica Tri-Grams (Laplace) em Lingua Espanhola

Granularidade dos Conjuntos usando Técnicas Tri-Grams Laplace (Laplace)		
M-DSP(40,18461538)	M(42,83076923)	M+DSP(45,47692308)
la oferta y	la oferta y	la oferta y
la oferta de		
oferta de irán		
de irán portugal		
irán portugal que		
portugal que participó		
que participó en		
participó en esta		
en esta oferta		
esta oferta juego		
oferta juego capitán		
juego capitán cristiano		
capitán cristiano ronaldo		
cristiano ronaldo y		
ronaldo y pouraliganji		
y pouraliganji llevó		
pouraliganji llevó al		
llevó al árbitro		
al árbitro a		
árbitro a utilizar		
a utilizar el		
utilizar el var		
el var previa		
AISs	2783.99999999999	
DSP	2.64615384615384	

Tabela 4.64: Está tabela mostra o resumo de tudo que foi tratada sobre as medidas e técnicas que foram usadas

Média	Uni-Grams			Bi-Grams			Tri-Grams		
	Média	M-DP	M+DP	Média	M-DP	M+DP	Média	M-DP	M+DP
PC	0,975609	0,3578464	1,593373	0,9830508	0,5477971	1,4183045	0,984615	0,865268	1,103962
BB	1,8457301	1,268513	2,422946	1,8963554	1,4812647	2,3114462	1,910071	1,795888	2,024253
CO	0,920155	0,3572657	1,4830447	0,9470577	0,5458417	1,3482736	0,954346	0,842015	1,066677
GI	0,9748132	0,9583571	0,9912693	0,9827167	0,9747466	0,9906869	0,984379	0,982432	0,986325
LAPLA	31,016260	18,54561	43,486879	40,149435	30,569462	49,729407	42,830769	40,184615	45,476923

4.3 Métodos utilizados na extração manual de termos.

Para extrairmos os termos relevantes foram selecionados somente textos em língua portuguesa, Espanhola, Francesa tanto para as fontes escritas quanto para as orais. Isso porque o nosso intuito foi organizar a terminologia de revestimento na adaptabilidade independente da língua, usando os métodos de similaridade assimétricas e com técnicas que facilitam na extração dos termos para isso, é fundamental considerar o universo conceptual e termino-lógico do utilizador e eventual consulta. Encontra-se acima, a relação de todos as matérias a partir do qual constituímos para a extração manual, dos termos. Uma das etapas fundamentais desta nossa pesquisa neste capitulo é a coleta de termos nos textos especializados, na qual foi um facto que é muito relevante por enfrentamos muitíssima dificuldade durante, em revistas especializadas

para a elaboração dos dicionários, com estas dificuldades dizem respeito sobre tudo termos que são especialista, como também as ocorrências de léxicos complexas, na tentativa de oito métodos e cinco técnicas e com metodologia assimétrica no texto em quanto que em alguns trabalhos consultados as frequências são baseados na web e nós as frequências são baseados em documentos, chegamos de fazer comparações com vários trabalhos, mas não chegamos de encontrar trabalho que fizer o trabalho com a similaridade assimétrica com 8 métodos e três técnicas como uni-Grams, Bi-Grams, Tri-Grams.

A medida de perplexidade *(em inglês perplexity scores)* os modelos de língua, em que está medida mede o sucesso com que uma distribuição de probabilidade ou de um modelo estatístico prevê os resultados, na qual podendo ser utilizada para comparar diferentes modelos de estatística, onde que com um baixo valor de perplexidade por vez indica uma maior semelhança entre os de teste e do método, está técnica ajuda avaliar os encontrados. O modelo de língua usado que é o estatístico que permite calcular a probabilidade de uma sequência de palavras que ocorrem em modelos n-gramas, que são utilizados com n=1,2, e 3 (Uni-grams, bi-grams e tri-grams, respetivamente) para a nossa avaliação foi utilizada três textos notícias de jornais com diversas versões.

4.4 Considerações Finais

Neste trabalho recomendamos o uso das medidas acima mencionadas que foram úteis nos nossos testes permitem o teste de diferentes combinações assimétricas, segundo a ordem da probabilidade ou estatisticamente fácil a comparação de termos pares, com estas técnicas empregadas não são totalmente novas, mostraram melhores resultados quando aplicamos a específicos. A média mostra melhor resultados que todos com combinações baseadas numa estratégia de classificação, acreditamos que o método proposto é útil para desambiguar uma grande percentagem nas consultas temporais para os utilizadores. Neste trabalho, foram desenvolvidos métodos que vem a extrair termos-chaves de um único texto uma das principais vantagens é a sua simplicidade e o seu alto desempenho, na qual quanto mais documentos eletrónico torna-se disponível é por isso a creditamos que estas medidas serão úteis em muitas aplicações em domínios independentes em extração de termos-chaves.

Chegamos de alcançar melhores resultados com o uso das cinco (5) medidas de extração de termos-chaves não supervisionados de pendentemente da língua usando de última geração, com grandes números de textos em , como uni-grams, bi-grams, tri-grams, em conjuntos, com dados diferentes e como; Português, Inglês, Espanhol, Francês, que foram a plicados a um grande conjunto de termos de coocorrências onde obtemos resultados que foram avaliados com as normas obtidas novas medidas são capazes de descobrir entre pares de termos assimétricos, com base em probabilidade condicional funciona bem os pares assimétricos e faz previsões razoáveis para medidas assimétricas.

Com este trabalho a que presentado pode ser estendido em muitas maneiras, mais com estes resultados de avaliação devem ser encorajadores, mas mostrando que há considerável espaço para melhorias, com as medidas de associação assimétrica, com uso desta medida em que em seguidamente são baseados em classificação, que atualmente rakings de acordo com as estatísticas realizadas nos nossos cálculos acima demonstrados, com AIS, que é baseado em qualquer medida.

Neste trabalho representamos os resultados em gráficos de curva normal para que nos mostra

a média e os desvio padrão e chegamos mostrar em tabelas, e as contagens das frequências de um termos no determinado texto que foi um tal experimento não supervisionado que ainda não foi tentado os nossos resultados obtidos no uso de cinco medidas de associação assimétricas que foram baseados em frequências de termos de um texto.

Capítulo 5

Conclusões e Perspetivas de trabalho Futuro

Este capítulo refere-se à conclusão da existência do esforço de trabalho realizado, que foi abordado em relação à adaptabilidade não supervisionado Independente da língua ao perfil linguístico do utilizador, onde que até agora, os pesquisadores têm se concentrado em estratégias de medidas adaptáveis tradicionais sem considerar a diversidade de elementos diferente média social, com o enriquecimento dos perfis dos utilizadores recebeu-se uma maior atenção adequado, com as necessidades dinâmicas, com usos dos inteligentes que exigindo mais atenção na obtenção de melhores resultados.

5.1 Conclusões

Este trabalho faz análise da adaptabilidade não supervisionada e Independente da língua de um perfil linguístico do utilizador, usando as granularidades e metodologias de desenvolvimentos das interfaces que tornem mais fácil e eficiente o seu uso. A inteligência deve fazer os sistemas se adaptarem aos utilizadores, tirar as dúvidas dos mesmos, permitir um diálogo entre o utilizador e o sistema ou apresentar informações integradas e compreensíveis utilizando vários modos de comunicação.
Adaptabilidade tem sido amplamente estudada, mas a falta de auto aprendizagem de interação do utilizador, aqui adotamos e iremos a dotar o uso de modelo para prever a intenção com a sua experiência especial, além disso, a proposta da consciência auxilia na previsão de satisfazer o requisito.
Apesar de que o tema da adaptabilidade não supervisionada independente da língua em perfil linguístico do utilizador ter-se intensificado com a preocupação de se criarem técnica adaptável melhor para uma aplicação, em geral, nota-se que a literatura científica não conta com muitos estudos principais (contribuições) sobre a usabilidade adaptáveis, tomando-se como base os princípios que consolidam as interfaces de dispositivos, em geral.
As tecnologias de comunicação e de informação nos seus formatos têm recebido, ate o momento, relativamente pouca atenção por parte das academias como objeto de estudo, muito embora as estatísticas indicam que são tecnologia mais rápida disseminação na história da humanidade, onde que já sãos elevados números para uma população mundial de sete biliões de habitantes.
A conceptualização e desenvolvimento experimental de novas metodologias assenta numa adaptabilidade em tempo real não supervisionada e independente da língua, focando-se nas funcionalidades e conteúdo de uma qualquer aplicação.
Em suma, o objetivo da presente proposta é conceptualizar e fazer desenvolvimento experimental de novas metodologias não supervisionadas e independente da língua, para aplicações adaptativas ao perfil do utilizador, e para os objetivos específicos que põem o nosso trabalho acima. Identificar as caraterísticas do utilizador para o perfil linguístico adaptável para facilitar. Este trabalho aqui apresentado pode ser estendido em muitas maneiras, mais com estes resultados de avaliação devem ser encorajadores, mas mostrando que há considerável espaço para melhorias, com a medida de associação assimétrica, com uso desta em que seguidamente

são baseados em classificação, que atualmente usa rankings de acordo com as estatísticas realizadas nos nossos cálculos acima demonstrados, com AIS.
Isso porque o nosso intuito foi organizar a terminologia de revestimento na adaptabilidade independente da língua, usando os métodos de similaridade assimétricas e com técnicas que facilitam na extração dos termos para isso, é fundamental considerar o universo conceptual e terminológico do utilizador e eventual consulta. Encontra-se acima, a relação de todos matérias a partir do qual constituímos para a extração manual, dos termos. Uma das etapas fundamentais desta nossa pesquisa neste capitulo é a coleta de termos nos textos especializados, na qual foi um facto que é muito relevante por enfrentamos muita dificuldade durante a pesquisa em revistas especializadas para a elaboração do dicionário, com estas dificuldades dizem respeito sobre termos que são utilizados nas línguas geral por um não especialista, como também as ocorrências de léxicos complexas, na tentativa de oito métodos cinco deles forneceram resultados satisfatórios e com uso das metodologia assimétrica no texto em quanto que em alguns trabalhos consultados as frequências são baseados em notícias, chegamos de fazer comparações com vários trabalhos, mas não encontrar trabalho que usam as mais medidas que é Similaridade assimétrica com 8 métodos e três granularidades como uni-Grams, Bi-Grams, Tri-Grams usando as Línguas Portuguesa, Espanhola e Francesa. No término deste trabalho pode-se constar o quanto são vastos os assuntos que tratam de PLN, na adaptabilidade dos perfis linguísticos não supervisionada do utilizador, como nas redes sociais, no público, em geral, e as suas respetivas técnicas de busca de conhecimentos.
Ao termino deste trabalho pode-se constar o quanto são vastos os assuntos que tratam de processamento de linguagem natural (Pln), na adaptabilidade dos perfil linguísticos não supervisionada do utilizador, como nas redes sociais, no publico em geral, e suas respetivas técnicas de busca de conhecimentos.
Neste trabalho recomendamos o uso das medidas acima mencionadas que foram úteis e permitem testar diferentes combinações assimétricas, segundo a ordem de probabilidades ou de fácil comparação estatística de termos pares. Estas metodologias não são totalmente novas e mostraram melhores resultados quando aplicadas a estes problemas específicos. Neste vamos nos concentrar no próximo passo no futuro na adaptabilidade não supervisionada para perfil do utilizador.

5.2 Sugestões Para trabalhos Futuros

Como um possível trabalho futuro, pode se citar a inclusão de novos paramentos humanos a datáveis relativos à forma com o qual utilizador vai interagir com os dispositivos computacionais de forma geral, e com dispositivos computacionais de forma específica que estão em constante evolução e possuem a dificuldade, já dependem da adaptabilidade, e dos resultados da psicologia, Sociológica, da mesma, segundo os trabalhos relacionados chegou-se a ponto de fazer comparações que adaptado perfil linguístico dos utilizadores, e criação de interfaces adaptáveis em componentes informáticos e Moveis é uma das nossas tarefas futuras, com estas propostas futuras vão mudar com o passar dos anos, que vão tornando-se necessária a revisão das diretrizes expostas neste trabalho.
Outra atividade futura seria a realização de um interface adaptável a um perfil linguístico independente da língua não supervisionado para um utilizador numa interface Móvel.
Este trabalho levou-nos em consideração aspetos intrínsecos à camadas de visualizações de *softwares*,e leituras de muitos artigos, onde existem, porém, pesquisas que visam a criação de

perfis adaptáveis a um perfil linguístico que abranjam outras questões adaptáveis e integração com tecnologias que suportam acessibilidade. Outra proposta de continuação deste estudo é criação de sistemas adaptável ao perfil do utilizador fim de avaliar os modelos e as técnicas.

Bibliografia

[1] Baugh, John . perfil linguístico, em linguagem negra: idioma, sociedade e política na África e nas américas. pages 155–160, 2003. 24

[2] Benyon D . Sistemas adaptativos: uma solução para problemas de usabilidade. modelagem de usuário e interação adaptada ao usuário. pages 65–87, 1993. 7

[3] Brian S, apud Peece et, al. Avaliação de interfaces de utilizadores. pages 166–168, 2003. 18

[4] Browne, D., Norman, M. and Riches, D . Why build adaptive systems? adaptive user interfaces. pages 16–37, 1990. 6

[5] DataGramaZero . Relações históricas entre biblioteconomia, documentação e ciência da informação. 2009. 22

[6] Ito G,C. Mauricio,F. Ana,N. Utilização de interfaces adaptativas para a computação móvel. 2005. 9

[7] LE STRUGEON, E., GRISLIN M., MILLOT P. Toward the application of multiagente techniques to the design of human-machine systems organizations. vi ifac/ifip/ifors/iea symposium of analysis, design and evaluation of manmachine systems. 1995. 11

[8] L´opez-Jaquero and Francisco Montero. Model-based design of adaptive user interfaces through connectors. pages 257–258, 2003. 9

[9] Opperman, R . Adaptive user support: Ergonomic design of manually and automatically adaptable software. 1994. 6

[10] P. A. Akiki, A. K. Bandara, and Y. Yu . Simplifying enterprise application user interfaces through engineering role-based adaptive behavior. in proceedings of the 5th acm sigchi symposium on engineering interactive computing systems. pages 12–13, 2013. 7

[11] Peter Brusilovsky, Elmar Schwarz . User as student: Towards an adaptive interface for advanced web-based applications. 1997. 8

[12] Wesson, J. L., Singh, A. et van Tonder, B. Can adaptive interfaces improve the usability of mobile applications. page 187–198, 2010. 7

[13] Williams, E. Predictive, adaptive mobile user interfaces: State of the art and open problems. page 30:1–35:3, 2014a. 7

[14] Williams, E. Predictive, adaptive mobile user interfaces: State of the art and open problems.in proceedings of the 2014 acm southeast regional conference. 3 ed:pages 35:1–35:3, 2014b. 1

[15] WINKELS R., BREUKER J . What's in an its? functional decomposition. in new directions for intelligent tutoring systems. pages 57–71, 1992. 9

[16] Akiki, P.A., Bandara, A.K. Yu, Y . Adaptive modeldriven user interface development systems. page 64:1–64:33, 2015. 8

[17] Alim, Samy H . Consciência da linguagem crítica nos estados unidos: revisão de questões e revisão de pedagogias em uma sociedade ressecada". pesquisador educacional. pages 31–34, 2005. 25

[18] ANDRÉ,et,al . PadrÕes de projeto para o desenvolviimento de aplicativos java para dispositivos mÓveis. 2014a. 14

[19] ANDRÉ,et,al. Aplicabilidade de padrões de engenharia de software e de ihc no desenvolvimento de sistemas interativos. pages 118–119, 2004b. 15

[20] Anthony, D., Henderson, T., and Kotz, D . Privacy in location aware computing environments. ieee pervasive computing. page 64–72, 2007. 19

[21] Ardagna, C. A., Cremonini, M., Damiani, E., di Vimercati, S.D. C., and Samarati, P . Privacy-enhanced location services information. in digital privacy.theory, technologies and practices. page 307–326, 2007. 19

[22] Rajkumar Arun, Venkatasubramaniyan Suresh, and CE Veni Madhavan. Stopword graphs and authorship attribution in text corpora. In Semantic Computing, 2009. ICSC'09. IEEE International Conference on, pages 192–196. IEEE, 2009. 31

[23] N. C. MADDEN S. KRASHEN BAILEY. Is there a 'natural sequence' in adult second language learning? language learning. 1974. 26

[24] Pedro Alberto Barbetta, Marcelo Menezes Reis, and Antonio Cezar Bornia. Estatística: para cursos de engenharia e informática, volume 3. Atlas São Paulo, 2004. 72

[25] Barnes, John . Classe e comitês em uma paróquia da ilha norueguesa". relações humanas. pages 39–58, 1954. 20

[26] Debora Maria Befi-Lopes, Ana Manhani CÃ¡ceres, and Lucila Esteves. Perfil linguÃstico de crianÃcom alteraÃ§ÃespecÃfica de linguagem. Revista da Sociedade Brasileira de Fonoaudiologia, 17:274 – 278, 00 2012. 27

[27] Bell, D. R., Song, S. Y. Neighborhood effects and trial on the internet: Evidence from online grocery retailing. quantitative marketing and economics. page 361–400, 2007. 18

[28] Ben Shneiderman, Pattie Maes. Designing the user interface: Strategies for effective human-computer interaction. 2009. 1

[29] Bevan, N., Claridge, N. Petrie, H . Tenuta: Simplified guidance for usability and accessibility. in proceedings of hci internationa. 2005. 17

[30] Jagdev Bhogal, Andrew MacFarlane, and Peter Smith. A review of ontology based query expansion. Information processing & management, 43(4):866–886, 2007. 51

[31] Manfred Borovcnik. Multiple perspectives on the concept of conditional probability. Avances de Investigación en Educación Matemática, (2), 2012. 56

[32] Dunstan Brown. "morphological typology", in jae jung song, the oxford handbook of linguistic typology. 2012. 24

[33] BRUILLARD E. Les machines a enseigner. 1997. 9

[34] BRUILLARD E. Les machines a enseigner. page 365–382, 2013. 10

[35] Brusilovsky, P. Adaptive hypermedia. user modeling and user-adapted interaction. page 87–110, 2001. 10

[36] BRUSILOVSKY, P.; SCHWARZ, E. User as student: Towards an adaptive interface for advanced web-based applications, in: User modeling. pages 177–188, 1997a. 12

[37] BRUSILOVSKY, P.; SCHWARZ, E. User as student: Towards an adaptive interface for advanced web-based applications. pages 177–188, 1997b. 8

[38] M.T CABRÉ. La terminología: teoría, metodología, aplicaciones. barcelona: Antartida/empúries. 1993. 22

[39] Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Mário Jorge, Célia Nunes, and Adam Jatowt. A text feature based automatic keyword extraction method for single documents. In European Conference on Information Retrieval, pages 684–691. Springer, 2018. 56

[40] Roger Chaffin and Douglas J Herrmann. The similarity and diversity of semantic relations. Memory & Cognition, 12(2):134–141, 1984. 53

[41] Chung, T. S., Rust, R. T., Wedel, M . My mobile music: Na adaptive personalization system for digital audio players. marketing science. 1st ed:52–68, 2000. 19

[42] Guillaume Cleuziou and Gaël Dias. Apprentissage de mesures de similarité sémantiques: étude d'une variante de la mesure infosimba. In Proceedings of the 1st Joint Meeting of the Société Francophone de Classification and the Classification and Data Analysis Group of the Italian Society of Statistics (SFC-CLADAG 2008), pages 233–236, 2008. 50

[43] Cockton, G . Usability evaluation. in soegaard, m. friis,encyclopedia of human-computer interaction.the interaction design foundation. 2012. 16

[44] Shay B Cohen, Karl Stratos, Michael Collins, Dean P Foster, and Lyle Ungar. Spectral learning of latent-variable pcfgs. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1, pages 223–231. Association for Computational Linguistics, 2012. 31

[45] Juan Antonio Martínez Comeche. Los modelos clásicos de recuperación de información y su vigencia. Memoria del Tercer Seminario Hispano-Mexicano de investigación en biblioteconomía y documentación 29 al 31 de marzo de 2006, page 187. 36

[46] DE ROSSI, F.; PIZZUTILO, S. Formal description and evaluation. user-adapted interfaces. international journal- humam-computer studies. 49:95–120, 1998. 8

[47] Ana Tereza Ribeiro de Vasconcelos, Darcy F De Almeida, Mariangela Hungria, Claudia Teixeira Guimarães, Regina Vasconcellos Antônio, Francisca Cunha Almeida, Luiz GP De Almeida, Rosana De Almeida, José Antonio Alves-Gomes, Elizabeth Mazoni Andrade, et al. The complete genome sequence of chromobacterium violaceum reveals remarkable and exploitable bacterial adaptability. Proceedings of the national academy of sciences of the United States of America, pages 11660–11665, 2003. 41

[48] Delfa M. H. Zuasnábar. Um ambiente de aprendizagem via www baseado em interfaces inteligentes para o ensino de engenharia.instituto tecnológico de aeronáutica – ita praça mal. 2003. 10

[49] Krishna Chandra Devkota, Amar Deep Regmi, Hamid Reza Pourghasemi, Kohki Yoshida, Biswajeet Pradhan, In Chang Ryu, Megh Raj Dhital, and Omar F. Althuwaynee. Landslide susceptibility mapping using certainty factor, index of entropy and logistic regression models in gis and their comparison at mugling–narayanghat road section in nepal himalaya. Natural Hazards, 65(1):135–165, Jan 2013. 71

[50] Duarte.Avellar . Interface e usabilidade. 2015. 16

[51] ENCARNAÇÃO, L. M . Concept and realization of intelligent user support in interactive graphics applications. 1997. 9

[52] Christiane Fellbaum. Wordnet: An electronic lexical database (language, speech, and communication). the mit pres. 1998. 27

[53] Edward Fry. A readability formula that saves time. Journal of reading, 11(7):513–578, 1968. 44

[54] GALITZ, Wilbert. The essential guide to user interface design: An introduction to gui design principles and techniques. 2 ed, 2003. 1

[55] Gasparini . nterface adaptativa no ambiente adaptweb: navegação e apresentação adaptativa baseada no modelo do usuário. Junho 2003. 7

[56] Chu Chia Gean and Celso Antônio Alves Kaestner. Classificação automática de textos usando subespaços aleatórios e conjunto de classificadores. 26

[57] Christiane Fellbaum Derek Gross George A. Miller, Richard Beckwith and Katherine J. Miller. Introduction to wordnet: An on-line lexical database. international journal of lexicography. pages 31–34, 1990. 25

[58] Marco Gonzalez and Vera Lúcia Strube Lima. Recuperação de informação e processamento da linguagem natural. In XXIII Congresso da Sociedade Brasileira de Computaçao, volume 3, pages 347–395, 2003. 33

[59] Nelson Goodman. Seven strictures on similarity. 1972. 53

[60] H. Thimbleby, and A. Cox. A performance review of number entry interfaces. proceedings of the 14th ifip tc 13 international conference on human-computer interaction. page 365–382, 2013. 10

[61] Hettling, Manfred; Reinke,Andreas; Conrads, Norbert. In breslau zu hause? jude in einer mitteleuropaeischen metrropol der neuzeit hamburg. pages 264–265, 2003. 1

[62] Hewett . This publication is a report of the acm special interest group on computer-human interaction (sigchi)curriculum development group. The Association for Computing Machinery, 1992. 3

[63] Ibope . Número de usuários de redes sociais ultrapassa 46 milhões de brasileiros, 2017. 18

[64] J. M. C. Fonseca . Model-based ui xg final report, 2010. 7

[65] Jameson, A . Numerical uncertainty management in user and student modeling: An overview of systems and issues.user modeling and user-adapted interaction. page 193–251, 2009. 6

[66] Jameson, A . Adaptive interfaces and agents. 2012. 5

[67] Jameson, A., Großmann-Hutter, B., March, L., Rummer, R.,Bohnenberger, T., Wittig, F. When actions have consequences: Empirically based decision making for intelligent user interfaces. Knowledge-Based Systems, page 75–92, 2001. 5

[68] JANET et,al . Interfaces adaptativas podem melhorar a usabilidade de aplicações móveis. 2014. 11

[69] Janet L. As interfaces adaptativas podem melhorar a usabilidade de aplicações móveis. 2009. 13

[70] JIANGFATAN. Interface inteligente contexto e adaptativo para mobil. 2015. 12

[71] Rong Jin, Alex G Hauptmann, and Cheng Xiang Zhai. Language model for information retrieval. In Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, pages 42–48. ACM, 2002. 39

[72] Peter Roach James Hartmann e Jane Setter. Jones, Daniel. ., english pronouncing dictionary , cambridge: Cambridge university press. 2003. 26

[73] S Kamalakannan, S Karthikeyan, and K Sathyamoorthy. Implementation of error correction technique based on decimal matrix code. International Journal of Advanced Research Trends in Engineering and Technology (IJARTET), 2(4):1–6, 2015. 39

[74] Kayastha, N., Niyato, D., Wang, P., and Hossain, E . Applications, architectures, and protocol design issues for mobile social networks: A survey. proceedings of the ieee. page 2130–2158, 2011. 20

[75] Keath, J . Instagram becomes the largest mobile social network, 2011. 18

[76] KOLSKI C., LE STRUGEON E . A review of intelligent human-machine interfaces in the light of the arch model.internationa journal of human-computer interactio. pages 193-231, 1998. 11

[77] KORNILOVA. Adaptive user interface patterns for mobile aplications. 25, 2014. 14

[78] L. Zouhaier, Y. B. Hlaoui, and L. Jemni Ben Ayed, . Building adaptive accessible context-aware for user interface tailored to disable users. in Proceedings of 37th IEEE Annual Computer Sofware and Applications Conference Workshops, page 157–162, 2013. 5

[79] Telma Leal Ferraz and Gilda Guimarães Lisbôa. Por que é tão difícil ensinar a pontuar? Revista Portuguesa de Educação, 15(1), 2002. 40

[80] Isabel Leiria, A Martins, J Cordas, M Mouta, and R Henriques. Orientações programáticas de português língua não materna (plnm) ensino secundário. http://www. dge. mec. pt/sites/default/files/Basico/Documentos/orientprogramatplnmvers aofinalabril08. pdf. Acesso em, 20(09):2016, 2008a. 26

[81] Isabel Leiria, A Martins, J Cordas, M Mouta, and R Henriques. Orientações programáticas de português língua não materna (plnm) ensino secundário. http://www. dge. mec. pt/sites/default/files/Basico/Documentos/orientprogramatplnmvers aofinalabril08. pdf. Acesso em, 20(09):2016, 2008b. 23

[82] Beth Levin. English verb classes and alternations: A preliminary investigation. university of chicago press, chicago. 1993. 26

[83] Lieberman, H . Your wish is my command: Programming by example. 2001. 13

[84] Kathya Silvia Collazos Linares. Aspectos teóricos do Datamining: descoberta de conhecimento em medicina. PhD thesis, Universidade Federal de Santa Catarina, Centro Tecnológico. Programa de Pós-Graduação em Engenharia Elétrica., 2003. 71

[85] Xin LIU, Katia Vega, Jing Qian, Joseph Paradiso, and Pattie Maes. Fluxa: Body movements as a social display. In Proceedings of the 29th Annual Symposium on User Interface Software and Technology, UIST '16 Adjunct, pages 155–157, New York, NY, USA, 2016. ACM. 6

[86] Rachel Tsz-Wai Lo, Ben He, and Iadh Ounis. Automatically building a stopword list for an information retrieval system. In Journal on Digital Information Management: Special Issue on the 5th Dutch-Belgian Information Retrieval Workshop (DIR), volume 5, pages 17–24, 2005. 33

[87] L´opez-Jaquero and Francisco Montero . Model-based design of adaptive user interfaces through connectors, 2003. 5

[88] David Machado, Tiago Barbosa, Sebastião Pais, Bruno Martins, and Gaël Dias. Universal mobile information retrieval. In International Conference on Universal Access in Human-Computer Interaction, pages 345–354. Springer, 2009. 49

[89] Machado,E. Formas de interação homen-maquina, 2013. 1

[90] Masoud Makrehchi and Mohamed S Kamel. Automatic extraction of domain-specific stopwords from labeled documents. In European Conference on Information Retrieval, pages 222–233. Springer, 2008. 32

[91] Martins, A., Queirós, A., Cerqueira, M., Alvarelhão J., Teixeira, A. Rocha, N . Assessment of ambient assisted living services in a living lab approach: a methodology based on icf. In 2nd, 2011. 17

[92] Makrehchi Masoud and Mohamed S Kamel. Automatic extraction of domain-specific stopwords from labeled documents. Advances in information retrieval, pages 222–233. 31

[93] MILLARAY,et,al . Metodologia para a construção de interfaces adaptáveis em sistemas tutores inteligentes. 2002. 9

[94] Lucas Borges Monteiro. Ligação de entidades: uma nova abordagem para ligação de conceitos concretos com entidades wiki utilizando modelos de espaço vetorial. 2016a. 34

[95] Lucas Borges Monteiro. Ligação de entidades: uma nova abordagem para ligação de conceitos concretos com entidades wiki utilizando modelos de espaço vetorial. 2016b. 34

[96] Edison Andrade Martins Morais and Ana Paula L Ambrósio. Mineração de textos. Relatório Técnico–Instituto de Informática (UFG), 2007a. 39

[97] Edison Andrade Martins Morais and Ana Paula L Ambrósio. Mineração de textos. Relatório Técnico–Instituto de Informática (UFG), 2007b. 35

[98] Narayan, V., Rao, V. R., Saunders, C . How peer influence affects attribute preferences: A bayesian updating mechanism. marketing science. page 368–384, 2011. 18

[99] Nielsen, J . Usability engeneering. boston, academic press. 1993. 17

[100] Nielsen, J . sability 101: Introduction to usability, 2003. 17

[101] Niesen,J.and Molich, R. Heurististic evaluation of user interfaces.in procedings of the sigchi conference on human factors in computing systems chi. pages 249–256, March 2013. 1

[102] Bruno Magalhaes Nogueira. Avaliação de métodos não-supervisionados de seleção de atributos para Mineração de Textos. PhD thesis, Universidade de São Paulo, 2009. 39

[103] Ana Oliveira Alves, Ricardo Rodrigues, and Hugo Gonçalo Oliveira. Asapp: alinhamento semântico automático de palavras aplicado ao português. Linguamática, 8(2):43–58, 2016. 27

[104] P. A. Akiki, A. K. Bandara, and Y. Yu . Integrating adaptive user interface capabilities in enterprise applications. in proceedings of the 36th international conference on software engineering. 2014. 12

[105] Vera Pacheco. Percepção dos sinais de pontuação enquanto marcadores prosódicos (punctuation signals perception while prosody markers). Estudos da Língua (gem), 3(1):205, 2006. 40

[106] Sebastiao Pais, Gaël Dias, Katarzyna Wegrzyn-Wolska, Robert Mahl, and Pierre Jouvelot. Textual entailment by generality. Procedia-Social and Behavioral Sciences, 27:258–266, 2011. 49

[107] PAlAZZo, l. A. m.; UlbrIChT, V. r.; VAnZIn, T. ; Flores, A.r.b; lIndner, l. Redes sociais temáticas apoiando avea - i. in: busarello, r.i.; bieging. Mídia e Educação: novos olhares para a aprendizagem sem fronteiras, 1 ed:118–130, 2013. 18

[108] Pavel Pecina and Pavel Schlesinger. Combining association measures for collocation extraction., 01 2006. 50

[109] Alexandre Pereira and Carlos Poupa. Como escrever uma tese, monografia ou livro científico usando o word. Lisboa: Edições Sílabo, 2004. 43

[110] Diana Pérez and Enrique Alfonseca. Using bleu-like algorithms for the automatic recognition of entailment. In Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment, pages 191–204. Springer, 2006. 47

[111] PREECE, J . Designing usability and supporting socialbilty. 1st ed, 2000. 19

[112] PreeCe, J.; shneIdermAn, b . The reader-to-leader framework: Motivating technology-mediated social participation, ais transactions on human-computer interaction. 1:13–25, 2009. 18

[113] Maria Celeste Ramilo and Tiago Freitas. Transcrição ortográfica de textos orais: problemas e perspectivas. Actas do encontro comemorativo dos, 25:55–68, 2002. 42

[114] Robek, Donna D, Amy M.Hageman, and Charles F.Kelliher . Analyzing the role of social norms in tax compliance behavior. Journal of Business Ethics, 115:451–468, 2013. 2

[115] Stephen E Robertson and K Sparck Jones. Relevance weighting of search terms. Journal of the Association for Information Science and Technology, 27(3):129–146, 1976. 33

[116] S. FINCHER . Hci pattern-form gallery. [online]. 15

[117] Frizzi Alejandra San Roman Salazar. Um estudo sobre o papel de medidas de similaridade em visualização de coleções de documentos. PhD thesis, Universidade de São Paulo, 2012. 49

[118] Mark Sanderson and Bruce Croft. Deriving concept hierarchies from text. In Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '99, pages 206–213, New York, NY, USA, 1999. ACM. 53

[119] Giovanni Semeraro, Marco Degemmis, Pasquale Lops, and Pierpaolo Basile. Combining learning and word sense disambiguation for intelligent user profiling. 22

[120] Shen, Siu-Tsen; Woolley,M; PRIOR,S . Towards culture-centred design.interacting with computer. v.18:820–852, 2006. 19

[121] Rui Sousa Silva. Riqueza lexical como critério de detecção de autoria. Textos Seleccionados. XXIV Encontro Nacional da Associação Portuguesa de Linguística, pages 575–587, 2009. 43

[122] Dawn L Smalls. Perfil lingüístico e lei. Stan. L. Pol'y Rev., 15. 23

[123] Fei Song and W. Bruce Croft. A general language model for information retrieval. In Proceedings of the Eighth International Conference on Information and Knowledge Management, CIKM '99, pages 316–321, New York, NY, USA, 1999. ACM. 32

[124] Dirk Speelman, Stefan Grondelaers, and Dirk Geeraerts. Profile-based linguistic uniformity as a generic method for comparing language varieties. Computers and the Humanities, 37(3):317–337, Aug 2003. 24

[125] Teleco. As tecnologias de comunicação e de informação em seus formatos móveis tem recebido, ate o momento, relativamente pouca atenção por parte da academias, 2012. 2

[126] Alexandre António TIMBANE. Identificação de perfis linguísticos no facebook durante a investigação policial. Linguagem: Estudos e Pesquisas, 20(1), 2016a. 23

[127] Alexandre António TIMBANE. Identificação de perfis linguísticos no facebook durante a investigação policial. Linguagem: Estudos e Pesquisas, 20(1), 2016b. 22

[128] Simon Tong, Uri Lerner, Amit Singhal, Paul Haahr, and Steven Baker. Locating meaningful stopwords or stop-phrases in keyword-based retrieval systems, August 5 2008. US Patent 7,409,383. 30

[129] Joao Ventura and Joaquim Ferreira da Silva. Ranking and extraction of relevant single words in text. In Brain, Vision and AI. InTech, 2008. 47

[130] Herbert Andreas Welker. O uso de dicionários. Horizontes de Linguistica Aplicada. Brasília: ano, 6, 2006. 22

[131] Beatriz Wilges et al. Um modelo para organização de documentos no contexto da memória organizacional. 2014. 38

[132] Liang Zeng. Designing the user interface: Strategies for effective human-computer interaction (5th edition) by b. shneiderman and c. plaisant. International Journal of Human–Computer Interaction, 25(7):707–708, 2009. 8

[133] ZhAn, m. Social netork analysis: History, concepts, and research. in: Furht. handbook of social networks technologies and applications. pages 18–21, 2010. 18

Apêndice A

Anexos

A.1 Resultados das probabilidade dos termos e técnicas Uni-Gramas

A tabela A.1 apresenta a distribuição das probabilidades em que os termos são agrupados no texto em frequências e com as medidas.
A tabela A.2 mostra os resultados de todas medidas em Bi-Gramas em 56 termos em língua Francesa. Está tabela apresenta um estudo comparativo das medidas de similaridade assimétricas de texto em língua francesa usando o método Bi-Gramas como elas afetam a qualidade da tabela baseadas no posicionamento em cada medida com o seu Ranking e com resultados diferentes para que seja destacado o melhor desempenho apresentado por estas cinco medidas com 66 termos .
A tabela A.3 mostra os resultados de todas medidas em Tri-Gramas em 68 termos em língua Francesa. Está tabela apresenta um estudo comparativo das medidas de similaridade assimétricas de texto em língua francesa usando o método Tri-Gramas como elas afetam a qualidade da tabela baseadas no posicionamento em cada medida com o seu Ranking e com resultados diferentes para que seja destacado o melhor desempenho apresentado por estas cinco medidas com 68 termos .
A tabela A.4 mostra o resultados das probabilidade e frequências dos termos com cinco medidas e com a técnica Uni-Gramas com texto em Lingua Espanhola com 41 termos.
A tabela A.5 mostra resultados das probabilidade e frequências dos termos com cinco medidas e com a técnica Bi-Gramas com texto em Lingua Espanhola.
A tabela A.6 mostra o resultados das probabilidade e frequências dos termos em cinco medidas e em técnica Tri-Gramas com texto em Lingua Espanhola.

Tabela A.1: Resultados das probabilidade e frequências dos termos com cinco medidas e com a técnica Uni-Gramas com texto em Lingua Francesa

Nº	TYPES	FREQ	Probili	PC	BB	CO	GI	LAPLACE
1º	joão	1	0,014084507	0,718309859	1,627415003	0,694449955	0,985719914	31,55
2º	lourenço	1	0,014084507	0,718309859	1,627415003	0,694449955	0,985719914	31,55
3º	le	3	0,042253521	2,154929577	2,988205417	2,027167712	0,956036557	63,63333333
4º	président	1	0,014084507	0,718309859	1,627415003	0,694449955	0,985719914	31,55
5º	angolais	2	0,028169014	1,436619718	2,307136546	1,369905877	0,971059844	47,575
6º	est	1	0,014084507	0,718309859	1,627415003	0,694449955	0,985719914	31,55
7º	en	6	0,084507042	4,309859155	5,038522028	3,897140063	0,90895502	111,8833333
8º	visite	2	0,028169014	1,436619718	2,307136546	1,369905877	0,971059844	47,575
9º	officielle	1	0,014084507	0,718309859	1,627415003	0,694449955	0,985719914	31,55
10º	france	3	0,042253521	2,154929577	2,988205417	2,027167712	0,956036557	63,63333333
11º	pour	1	0,014084507	0,718309859	1,627415003	0,694449955	0,985719914	31,55
12º	la	1	0,014084507	0,718309859	1,627415003	0,694449955	0,985719914	31,55
13º	première	1	0,014084507	0,718309859	1,627415003	0,694449955	0,985719914	31,55
14º	fois	1	0,014084507	0,718309859	1,627415003	0,694449955	0,985719914	31,55
15º	depuis	1	0,014084507	0,718309859	1,627415003	0,694449955	0,985719914	31,55
16º	son	2	0,028169014	1,436619718	2,307136546	1,369905877	0,971059844	47,575
17º	élection	1	0,014084507	0,718309859	1,627415003	0,694449955	0,985719914	31,55
18º	septembre	1	0,014084507	0,718309859	1,627415003	0,694449955	0,985719914	31,55
19º	dernier	1	0,014084507	0,718309859	1,627415003	0,694449955	0,985719914	31,55
20º	chef	2	0,028169014	1,436619718	2,307136546	1,369905877	0,971059844	47,575
21º	de	4	0,056338028	2,873239437	3,670519288	2,666991026	0,940666814	79,70833333
22º	lors	1	0,014084507	0,718309859	1,627415003	0,694449955	0,985719914	31,55
23º	cette	1	0,014084507	0,718309859	1,627415003	0,694449955	0,985719914	31,55
24º	a	1	0,014084507	0,718309859	1,627415003	0,694449955	0,985719914	31,55
25º	été	2	0,028169014	1,436619718	2,307136546	1,369905877	0,971059844	47,575
26º	reçu	1	0,014084507	0,718309859	1,627415003	0,694449955	0,985719914	31,55
27º	lundi	1	0,014084507	0,718309859	1,627415003	0,694449955	0,985719914	31,55
28º	à	1	0,014084507	0,718309859	1,627415003	0,694449955	0,985719914	31,55
30º	par	1	0,014084507	0,718309859	1,627415003	0,694449955	0,985719914	31,55
31º	homologue	1	0,014084507	0,718309859	1,627415003	0,694449955	0,985719914	31,55
32º	emmanuel	1	0,014084507	0,718309859	1,627415003	0,694449955	0,985719914	31,55
33º	macron	1	0,014084507	0,718309859	1,627415003	0,694449955	0,985719914	31,55
34º	français	1	0,014084507	0,718309859	1,627415003	0,694449955	0,985719914	31,55
35º	dit	1	0,014084507	0,718309859	1,627415003	0,694449955	0,985719914	31,55
36º	très	1	0,014084507	0,718309859	1,627415003	0,694449955	0,985719914	31,55
37º	proche	1	0,014084507	0,718309859	1,627415003	0,694449955	0,985719914	31,55
38º	du	1	0,014084507	0,718309859	1,627415003	0,694449955	0,985719914	31,55
39º	renforcement	1	0,014084507	0,718309859	1,627415003	0,694449955	0,985719914	31,55
40º	des	2	0,028169014	1,436619718	2,307136546	1,369905877	0,971059844	47,575
41º	relations	1	0,014084507	0,718309859	1,627415003	0,694449955	0,985719914	31,55
42º	avec	1	0,014084507	0,718309859	1,627415003	0,694449955	0,985719914	31,55
43º	questions	1	0,014084507	0,718309859	1,627415003	0,694449955	0,985719914	31,55
44º	régionales	1	0,014084507	0,718309859	1,627415003	0,694449955	0,985719914	31,55
45º	ont	1	0,014084507	0,718309859	1,627415003	0,694449955	0,985719914	31,55
46º	évoquées	1	0,014084507	0,718309859	1,627415003	0,694449955	0,985719914	31,55
47º	et	2	0,028169014	1,436619718	2,307136546	1,369905877	0,971059844	47,575
48º	rdc	1	0,014084507	0,718309859	1,627415003	0,694449955	0,985719914	31,55
49º	paris	1	0,014084507	0,718309859	1,627415003	0,694449955	0,985719914	31,55
50º	luanda	1	0,014084507	0,718309859	1,627415003	0,694449955	0,985719914	31,55
51º	appellent	1	0,014084507	0,718309859	1,627415003	0,694449955	0,985719914	31,55
52º	au	1	0,014084507	0,718309859	1,627415003	0,694449955	0,985719914	31,55
53º	respect	1	0,014084507	0,718309859	1,627415003	0,694449955	0,985719914	31,55

Tabela A.2: Resumos das a mostras e dos resultados Ranking com as cinco medida com o método Bi-Gramas

Nº	TYPES	FREQ	PROB	PC	BB	CO	GI	LAPLACE
1º	joão lourenço	1	0,014285714	0,928571429	1,855934882	0,901414164	0,98551312	42,73333333
2º	lourenço le	1	0,014285714	0,928571429	1,855934882	0,901414164	0,98551312	42,73333333
3º	e président	1	0,014285714	0,928571429	1,855934882	0,901414164	0,98551312	42,73333333
4º	président angolais	1	0,014285714	0,928571429	1,855934882	0,901414164	0,98551312	42,73333333
5º	angolais est	1	0,014285714	0,928571429	1,855934882	0,901414164	0,98551312	42,73333333
6º	est en	1	0,014285714	0,928571429	1,855934882	0,901414164	0,98551312	42,73333333
7º	en visite	1	0,014285714	0,928571429	1,855934882	0,901414164	0,98551312	42,73333333
8º	visite officielle	1	0,014285714	0,928571429	1,855934882	0,901414164	0,98551312	42,73333333
9º	officielle en	1	0,014285714	0,928571429	1,855934882	0,901414164	0,98551312	42,73333333
10º	en france	3	0,042857143	2,785714286	3,635135135	2,630282768	0,95538484	86
11º	france pour	1	0,014285714	0,928571429	1,855934882	0,901414164	0,98551312	42,73333333
12º	pour la	1	0,014285714	0,928571429	1,855934882	0,901414164	0,98551312	42,73333333
13º	la première	1	0,014285714	0,928571429	1,855934882	0,901414164	0,98551312	42,73333333
14º	première fois	1	0,014285714	0,928571429	1,855934882	0,901414164	0,98551312	42,73333333
15º	fois depuis	1	0,014285714	0,928571429	1,855934882	0,901414164	0,98551312	42,73333333
16º	depuis son	1	0,014285714	0,928571429	1,855934882	0,901414164	0,98551312	42,73333333
17º	son élection	1	0,014285714	0,928571429	1,855934882	0,901414164	0,98551312	42,73333333
18º	élection en	1	0,014285714	0,928571429	1,855934882	0,901414164	0,98551312	42,73333333
19º	en septembre	1	0,014285714	0,928571429	1,855934882	0,901414164	0,98551312	42,73333333
20º	septembre dernier	1	0,014285714	0,928571429	1,855934882	0,901414164	0,98551312	42,73333333
21º	dernier le	1	0,014285714	0,928571429	1,855934882	0,901414164	0,98551312	42,73333333
22º	le chef	2	0,028571429	1,857142857	2,744843528	1,777822208	0,970635569	64,35
23º	chef de	2	0,028571429	1,857142857	2,744843528	1,777822208	0,970635569	64,35
24º	de angolais	1	0,014285714	0,928571429	1,855934882	0,901414164	0,98551312	42,73333333
25º	golais lors	1	0,014285714	0,928571429	1,855934882	0,901414164	0,98551312	42,73333333
26º	lors de	1	0,014285714	0,928571429	1,855934882	0,901414164	0,98551312	42,73333333
27º	de cette	1	0,014285714	0,928571429	1,855934882	0,901414164	0,98551312	42,73333333
28º	cette visite	1	0,014285714	0,928571429	1,855934882	0,901414164	0,98551312	42,73333333
29º	visite en	1	0,014285714	0,928571429	1,855934882	0,901414164	0,98551312	42,73333333
30º	france a	1	0,014285714	0,928571429	1,855934882	0,901414164	0,98551312	42,73333333
31º	a été	1	0,014285714	0,928571429	1,855934882	0,901414164	0,98551312	42,73333333
32º	été reçu	1	0,014285714	0,928571429	1,855934882	0,901414164	0,98551312	42,73333333
33º	reçu lundi	1	0,014285714	0,928571429	1,855934882	0,901414164	0,98551312	42,73333333
34º	lundi à	1	0,014285714	0,928571429	1,855934882	0,901414164	0,98551312	42,73333333
35º	à par	1	0,014285714	0,928571429	1,855934882	0,901414164	0,98551312	42,73333333
36º	par son	1	0,014285714	0,928571429	1,855934882	0,901414164	0,98551312	42,73333333
37º	son homologue	1	0,014285714	0,928571429	1,855934882	0,901414164	0,98551312	42,73333333
38º	homologue emmanuel	1	0,014285714	0,928571429	1,855934882	0,901414164	0,98551312	42,73333333
39º	emmanuel macron	1	0,014285714	0,928571429	1,855934882	0,901414164	0,98551312	42,73333333
40º	macron le	1	0,014285714	0,928571429	1,855934882	0,901414164	0,98551312	42,73333333
41º	de français	1	0,014285714	0,928571429	1,855934882	0,901414164	0,98551312	42,73333333
42º	français dit	1	0,014285714	0,928571429	1,855934882	0,901414164	0,98551312	42,73333333
43º	dit très	1	0,014285714	0,928571429	1,855934882	0,901414164	0,98551312	42,73333333
44º	très proche	1	0,014285714	0,928571429	1,855934882	0,901414164	0,98551312	42,73333333
45º	proche du	1	0,014285714	0,928571429	1,855934882	0,901414164	0,98551312	42,73333333
46º	du renforcement	1	0,014285714	0,928571429	1,855934882	0,901414164	0,98551312	42,73333333
47º	renforcement des	1	0,014285714	0,928571429	1,855934882	0,901414164	0,98551312	42,73333333
48º	des relations	1	0,014285714	0,928571429	1,855934882	0,901414164	0,98551312	42,73333333
49º	relations avec	1	0,014285714	0,928571429	1,855934882	0,901414164	0,98551312	42,73333333

Tabela A.3: Resumos das a mostras e dos resultados Ranking com as cinco medida com o método Tri-Gramas é um texto em Francês

Nº	TYPES	FREQ	PROB	PC	BB	CO	GI	LAPL
1º	joão lourenço le	1	0,014492754	0,971014493	1,90025082	0,942863102	0,9853003	44,5
2º	lourenço le président	1	0,014492754	0,971014493	1,90025082	0,942863102	0,9853003	44,5
3º	le président angolais	1	0,014492754	0,971014493	1,90025082	0,942863102	0,9853003	44,5
4º	président angolais est	1	0,014492754	0,971014493	1,90025082	0,942863102	0,9853003	44,5
5º	angolais est en	1	0,014492754	0,971014493	1,90025082	0,942863102	0,9853003	44,5
6º	est en visite	1	0,014492754	0,971014493	1,90025082	0,942863102	0,9853003	44,5
7º	en visite officielle	1	0,014492754	0,971014493	1,90025082	0,942863102	0,9853003	44,5
8º	visite officielle en	1	0,014492754	0,971014493	1,90025082	0,942863102	0,9853003	44,5
9º	officielle en france	1	0,014492754	0,971014493	1,90025082	0,942863102	0,9853003	44,5
10º	en france pour	1	0,014492754	0,971014493	1,90025082	0,942863102	0,9853003	44,5
11º	france pour la	1	0,014492754	0,971014493	1,90025082	0,942863102	0,9853003	44,5
12º	pour la première	1	0,014492754	0,971014493	1,90025082	0,942863102	0,9853003	44,5
13º	la première fois	1	0,014492754	0,971014493	1,90025082	0,942863102	0,9853003	44,5
14º	première fois depuis	1	0,014492754	0,971014493	1,90025082	0,942863102	0,9853003	44,5
15º	fois depuis son	1	0,014492754	0,971014493	1,90025082	0,942863102	0,9853003	44,5
16º	depuis son élection	1	0,014492754	0,971014493	1,90025082	0,942863102	0,9853003	44,5
17º	son élection en	1	0,014492754	0,971014493	1,90025082	0,942863102	0,9853003	44,5
18º	élection en septembre	1	0,014492754	0,971014493	1,90025082	0,942863102	0,9853003	44,5
19º	en septembre dernier	1	0,014492754	0,971014493	1,90025082	0,942863102	0,9853003	44,5
20º	septembre dernier le	1	0,014492754	0,971014493	1,90025082	0,942863102	0,9853003	44,5
21º	dernier le chef	1	0,014492754	0,971014493	1,90025082	0,942863102	0,9853003	44,5
22º	le chef de	2	0,028985507	1,942028986	2,83098592	1,859201248	0,9701987	67
23º	chef de angolais	1	0,014492754	0,971014493	1,90025082	0,942863102	0,9853003	44,5
24º	de angolais lors	1	0,014492754	0,971014493	1,90025082	0,942863102	0,9853003	44,5
25º	angolais lors de	1	0,014492754	0,971014493	1,90025082	0,942863102	0,9853003	44,5
26º	lors de cette	1	0,014492754	0,971014493	1,90025082	0,942863102	0,9853003	44,5
27º	de cette visite	1	0,014492754	0,971014493	1,90025082	0,942863102	0,9853003	44,5
28º	cette visite en	1	0,014492754	0,971014493	1,90025082	0,942863102	0,9853003	44,5
29º	visite en france	1	0,014492754	0,971014493	1,90025082	0,942863102	0,9853003	44,5
30º	en france a	1	0,014492754	0,971014493	1,90025082	0,942863102	0,9853003	44,5
31º	france a été	1	0,014492754	0,971014493	1,90025082	0,942863102	0,9853003	44,5
32º	a été reçu	1	0,014492754	0,971014493	1,90025082	0,942863102	0,9853003	44,5
33º	été reçu lundi	1	0,014492754	0,971014493	1,90025082	0,942863102	0,9853003	44,5
34º	reçu lundi à	1	0,014492754	0,971014493	1,90025082	0,942863102	0,9853003	44,5
35º	lundi à par	1	0,014492754	0,971014493	1,90025082	0,942863102	0,9853003	44,5
36º	à par son	1	0,014492754	0,971014493	1,90025082	0,942863102	0,9853003	44,5
37º	par son homologue	1	0,014492754	0,971014493	1,90025082	0,942863102	0,9853003	44,5
38º	son homologue emmanuel	1	0,014492754	0,971014493	1,90025082	0,942863102	0,9853003	44,5
39º	homologue emmanuel macron	1	0,014492754	0,971014493	1,90025082	0,942863102	0,9853003	44,5
40º	emmanuel macron le	1	0,014492754	0,971014493	1,90025082	0,942863102	0,9853003	44,5
41º	macron le chef	1	0,014492754	0,971014493	1,90025082	0,942863102	0,9853003	44,5
42º	chef de français	1	0,014492754	0,971014493	1,90025082	0,942863102	0,9853003	44,5
43º	de français dit	1	0,014492754	0,971014493	1,90025082	0,942863102	0,9853003	44,5
44º	français dit très	1	0,014492754	0,971014493	1,90025082	0,942863102	0,9853003	44,5
45º	dit très proche	1	0,014492754	0,971014493	1,90025082	0,942863102	0,9853003	44,5
46º	très proche du	1	0,014492754	0,971014493	1,90025082	0,942863102	0,9853003	44,5
47º	proche du renforcement	1	0,014492754	0,971014493	1,90025082	0,942863102	0,9853003	44,5
48º	du renforcement des	1	0,014492754	0,971014493	1,90025082	0,942863102	0,9853003	44,5
49º	renforcement des relations	1	0,014492754	0,971014493	1,90025082	0,942863102	0,9853003	44,5
50º	des relations avec	1	0,014492754	0,971014493	1,90025082	0,942863102	0,9853003	44,5
51º	relations avec des	1	0,014492754	0,971014493	1,90025082	0,942863102	0,9853003	44,5
52º	avec des questions	1	0,014492754	0,971014493	1,90025082	0,942863102	0,9853003	44,5

Tabela A.4: Resultados das probabilidade e frequências dos termos com cinco medidas e com a técnica Uni-Gramas com texto em Lingua Espanhola com 41 termos

Nº	TYPES	FREQ	PROB	PC	BB	CO	GI	LAPLACE
1º	la	3	0,044117647	1,764705882	2,582212287	1,648047395	0,954021855	46,93333333
2º	oferta	4	0,058823529	2,352941176	3,133407844	2,167035263	0,937919805	58,83333333
3º	irán	4	0,058823529	2,352941176	3,133407844	2,167035263	0,937919805	58,83333333
4º	portugal	1	0,014705882	0,588235294	1,483966673	0,565223677	0,985081035	23,2
5º	que	3	0,044117647	1,764705882	2,582212287	1,648047395	0,954021855	46,93333333
6º	participó	1	0,014705882	0,588235294	1,483966673	0,565223677	0,985081035	23,2
7º	en	1	0,014705882	0,588235294	1,483966673	0,565223677	0,985081035	23,2
8º	esta	1	0,014705882	0,588235294	1,483966673	0,565223677	0,985081035	23,2
9º	juego	1	0,014705882	0,588235294	1,483966673	0,565223677	0,985081035	23,2
10º	capitán	1	0,014705882	0,588235294	1,483966673	0,565223677	0,985081035	23,2
11º	cristiano	2	0,029411765	1,176470588	2,03236018	1,114334546	0,969748626	35,05
12º	ronaldo	4	0,058823529	2,352941176	3,133407844	2,167035263	0,937919805	58,83333333
13º	y	4	0,058823529	2,352941176	3,133407844	2,167035263	0,937919805	58,83333333
14º	pouraliganji	4	0,058823529	2,352941176	3,133407844	2,167035263	0,937919805	58,83333333
15º	llevó	2	0,029411765	0,176470588	2,03236018	1,114334546	0,969748626	35,05
16º	al	2	0,029411765	1,176470588	2,03236018	1,114334546	0,969748626	35,05
17º	árbitro	2	0,029411765	1,176470588	2,03236018	1,114334546	0,969748626	35,05
18º	a	1	0,014705882	0,588235294	1,483966673	0,565223677	0,985081035	23,2
19º	tilizar	3	0,044117647	1,764705882	2,582212287	1,648047395	0,954021855	46,93333333
20º	el	1	0,014705882	0,588235294	1,483966673	0,565223677	0,985081035	23,2
21º	var	1	0,014705882	0,588235294	1,483966673	0,565223677	0,985081035	23,2
22º	previa	2	0,029411765	1,176470588	2,03236018	1,114334546	0,969748626	35,05
23º	consulta	1	0,014705882	0,588235294	1,483966673	0,565223677	0,985081035	23,2
24º	las	1	0,014705882	0,588235294	1,483966673	0,565223677	0,985081035	23,2
25º	imágenes	1	0,014705882	0,588235294	1,483966673	0,565223677	0,985081035	23,2
26º	elegido	1	0,014705882	0,588235294	1,483966673	0,565223677	0,985081035	23,2
27º	para	1	0,014705882	0,588235294	1,483966673	0,565223677	0,985081035	23,2
28º	mostrar	1	0,014705882	0,588235294	1,483966673	0,565223677	0,985081035	23,2
29º	cato	1	0,014705882	0,588235294	1,483966673	0,565223677	0,985081035	23,2
30º	amarilla	1	0,014705882	0,588235294	1,483966673	0,565223677	0,985081035	23,2
31º	mostró	2	0,029411765	1,176470588	2,03236018	1,114334546	0,969748626	35,05
32º	una	1	0,014705882	0,588235294	1,483966673	0,565223677	0,985081035	23,2
33º	tarjeta	1	0,014705882	0,588235294	1,483966673	0,565223677	0,985081035	23,2
34º	este	1	0,014705882	0,588235294	1,483966673	0,565223677	0,985081035	23,2
35º	encaje	1	0,014705882	0,588235294	1,483966673	0,565223677	0,985081035	23,2
36º	permitido	1	0,014705882	0,588235294	1,483966673	0,565223677	0,985081035	23,2
37º	l	1	0,014705882	0,588235294	1,483966673	0,565223677	0,985081035	23,2
38º	e	1	0,014705882	0,588235294	1,483966673	0,565223677	0,985081035	23,2
39º	jugar	1	0,014705882	0,588235294	1,483966673	0,565223677	0,985081035	23,2
40º	con	1	0,014705882	0,588235294	1,483966673	0,565223677	0,985081035	23,2
41º	uruaguai	1	0,014705882	0,588235294	1,483966673	0,565223677	0,985081035	23,2

Tabela A.5: Resultados das probabilidade e frequências dos termos com cinco medidas e com a técnica Bi-Gramas com texto em Lingua Espanhola

Nº	TYPES	FREQ	PROB	PC	BB	CO	GI	LAPLACE
1°	la oferta	3	0,044776119	2,597014925	3,435035722	2,443023101	0,953308751	75,66666667
2°	oferta de	1	0,014925373	0,865671642	1,784453635	0,838241815	0,984855185	37,56666667
3°	de irán	1	0,014925373	0,865671642	1,784453635	0,838241815	0,984855185	37,56666667
4°	irán portugal	1	0,014925373	0,865671642	1,784453635	0,838241815	0,984855185	37,56666667
5°	portugal que	1	0,014925373	0,865671642	1,784453635	0,838241815	0,984855185	37,56666667
6°	que participó	1	0,014925373	0,865671642	1,784453635	0,838241815	0,984855185	37,56666667
7°	participó en	1	0,014925373	0,865671642	1,784453635	0,838241815	0,984855185	37,56666667
8°	en esta	1	0,014925373	0,865671642	1,784453635	0,838241815	0,984855185	37,56666667
9°	esta oferta	1	0,014925373	0,865671642	1,784453635	0,838241815	0,984855185	37,56666667
10°	oferta juego	1	0,014925373	0,865671642	1,784453635	0,838241815	0,984855185	37,56666667
11°	juego capitán	1	0,014925373	0,865671642	1,784453635	0,838241815	0,984855185	37,56666667
12°	capitán cristiano	1	0,014925373	0,865671642	1,784453635	0,838241815	0,984855185	37,56666667
13°	cristiano ronaldo	4	0,059701493	3,462686567	4,262455016	3,211649635	0,936947031	94,75
14°	ronaldo y	1	0,014925373	0,865671642	1,784453635	0,838241815	0,984855185	37,56666667
15°	y pouraliganji	2	0,029850746	1,731343284	2,608995294	1,652224557	0,969284786	56,6
16°	pouraliganji llevó	1	0,014925373	0,865671642	1,784453635	0,838241815	0,984855185	37,56666667
17°	llevó al	1	0,014925373	0,865671642	1,784453635	0,838241815	0,984855185	37,56666667
18°	al árbitro	1	0,014925373	0,865671642	1,784453635	0,838241815	0,984855185	37,56666667
19°	árbitro a	1	0,014925373	0,865671642	1,784453635	0,838241815	0,984855185	37,56666667
20°	a utilizar	1	0,014925373	0,865671642	1,784453635	0,838241815	0,984855185	37,56666667
21°	utilizar el	1	0,014925373	0,865671642	1,784453635	0,838241815	0,984855185	37,56666667
22°	el var	1	0,014925373	0,865671642	1,784453635	0,838241815	0,984855185	37,56666667
23°	var previa	1	0,014925373	0,865671642	1,784453635	0,838241815	0,984855185	37,56666667
24°	previa consulta	1	0,014925373	0,865671642	1,784453635	0,838241815	0,984855185	37,56666667
25°	consulta al	1	0,014925373	0,865671642	1,784453635	0,838241815	0,984855185	37,56666667
26°	al var	1	0,014925373	0,865671642	1,784453635	0,838241815	0,984855185	37,56666667
27°	var la	1	0,014925373	0,865671642	1,784453635	0,838241815	0,984855185	37,56666667
28°	oferta y	2	0,029850746	1,731343284	2,608995294	1,652224557	0,969284786	56,6
29°	y las	1	0,014925373	0,865671642	1,784453635	0,838241815	0,984855185	37,56666667
30°	las imágenes	1	0,014925373	0,865671642	1,784453635	0,838241815	0,984855185	37,56666667
31°	imágenes de	1	0,014925373	0,865671642	1,784453635	0,838241815	0,984855185	37,56666667
32°	de cristiano	1	0,014925373	0,865671642	1,784453635	0,838241815	0,984855185	37,56666667
33°	ronaldo de	1	0,014925373	0,865671642	1,784453635	0,838241815	0,984855185	37,56666667
34°	de portugal	2	0,029850746	1,731343284	2,608995294	1,652224557	0,969284786	56,6
35°	portugal elegido	1	0,014925373	0,865671642	1,784453635	0,838241815	0,984855185	37,56666667
36°	elegido para	1	0,014925373	0,865671642	1,784453635	0,838241815	0,984855185	37,56666667
37°	para mostrar	1	0,014925373	0,865671642	1,784453635	0,838241815	0,984855185	37,56666667
38°	mostrar la	1	0,014925373	0,865671642	1,784453635	0,838241815	0,984855185	37,56666667
39°	y cato	1	0,014925373	0,865671642	1,784453635	0,838241815	0,984855185	37,56666667
40°	cato amarilla	1	0,014925373	0,865671642	1,784453635	0,838241815	0,984855185	37,56666667
41°	amarilla a	1	0,014925373	0,865671642	1,784453635	0,838241815	0,984855185	37,56666667
42°	a cristiano	1	0,014925373	0,865671642	1,784453635	0,838241815	0,984855185	37,56666667
43°	ronaldo capitán	1	0,014925373	0,865671642	1,784453635	0,838241815	0,984855185	37,56666667
44°	capitán de	1	0,014925373	0,865671642	1,784453635	0,838241815	0,984855185	37,56666667
45°	portugal y	1	0,014925373	0,865671642	1,784453635	0,838241815	0,984855185	37,56666667
46°	pouraliganji mostró	1	0,014925373	0,865671642	1,784453635	0,838241815	0,984855185	37,56666667
47°	mostró una	1	0,014925373	0,865671642	1,784453635	0,838241815	0,984855185	37,56666667
48°	una tarjeta	1	0,014925373	0,865671642	1,784453635	0,838241815	0,984855185	37,56666667
49°	tarjeta amarilla	1	0,014925373	0,865671642	1,784453635	0,838241815	0,984855185	37,56666667
50°	amarilla este	1	0,014925373	0,865671642	1,784453635	0,838241815	0,984855185	37,56666667
51°	este encaje	1	0,014925373	0,865671642	1,784453635	0,838241815	0,984855185	37,56666667
52°	encaje permitido	1	0,014925373	0,865671642	1,784453635	0,838241815	0,984855185	37,56666667
53°	permitido cristiano	1	0,014925373	0,865671642	1,784453635	0,838241815	0,984855185	37,56666667
54°	ronaldo le	1	0,014925373	0,865671642	1,784453635	0,838241815	0,984855185	37,56666667
55°	le llevó	1	0,014925373	0,865671642	1,784453635	0,838241815	0,984855185	37,56666667
56°	llevó a	1	0,014925373	0,865671642	1,784453635	0,838241815	0,984855185	37,56666667
57°	a jugar	1	0,014925373	0,865671642	1,784453635	0,838241815	0,984855185	37,56666667
58°	jugar con	1	0,014925373	0,865671642	1,784453635	0,838241815	0,984855185	37,56666667
59°	con uruaguai	1	0,014925373	0,865671642	1,784453635	0,838241815	0,984855185	37,56666667

Tabela A.6: Resultados das probabilidade e frequências dos termos em cinco medidas e em técnica Tri-Gramas com texto em Lingua Espanhol

Nº	TYPES	F	PROB	PC	BB	CO	GI	LAP
1º	la oferta de	1	0,015151	0,96969697	1,895798319	0,94030531	0,9846223	42,5
2º	oferta de irán	1	0,015151	0,96969697	1,895798319	0,94030531	0,9846223	42,5
3º	de irán portugal	1	0,015151	0,96969697	1,895798319	0,94030531	0,9846223	42,5
4º	irán portugal que	1	0,015151	0,96969697	1,895798319	0,94030531	0,9846223	42,5
5º	portugal que participó	1	0,015151	0,96969697	1,895798319	0,94030531	0,9846223	42,5
6º	que participó en	1	0,015151	0,96969697	1,895798319	0,94030531	0,9846223	42,5
7º	participó en esta	1	0,015151	0,96969697	1,895798319	0,94030531	0,9846223	42,5
8º	en esta oferta	1	0,015151	0,96969697	1,895798319	0,94030531	0,9846223	42,5
9º	esta oferta juego	1	0,015151	0,96969697	1,895798319	0,94030531	0,9846223	42,5
10º	oferta juego capitán	1	0,015151	0,96969697	1,895798319	0,94030531	0,9846223	42,5
11º	juego capitán cristiano	1	0,015151	0,96969697	1,895798319	0,94030531	0,9846223	42,5
12º	capitán cristiano ronaldo	1	0,015151	0,96969697	1,895798319	0,94030531	0,9846223	42,5
13º	cristiano ronaldo y	1	0,015151	0,96969697	1,895798319	0,94030531	0,9846223	42,5
14º	ronaldo y pouraliganji	1	0,015151	0,96969697	1,895798319	0,94030531	0,9846223	42,5
15º	y pouraliganji llevó	1	0,015151	0,96969697	1,895798319	0,94030531	0,9846223	42,5
16º	pouraliganji llevó al	1	0,015151	0,96969697	1,895798319	0,94030531	0,9846223	42,5
17º	llevó al árbitro	1	0,015151	0,96969697	1,895798319	0,94030531	0,9846223	42,5
18º	al árbitro a	1	0,015151	0,96969697	1,895798319	0,94030531	0,9846223	42,5
19º	árbitro a utilizar	1	0,015151	0,96969697	1,895798319	0,94030531	0,9846223	42,5
20º	a utilizar el	1	0,015151	0,96969697	1,895798319	0,94030531	0,9846223	42,5
21º	utilizar el var	1	0,015151	0,96969697	1,895798319	0,94030531	0,9846223	42,5
22º	el var previa	1	0,015151	0,96969697	1,895798319	0,94030531	0,9846223	42,5
23º	var previa consulta	1	0,015151	0,96969697	1,895798319	0,94030531	0,9846223	42,5
24º	previa consulta al	1	0,015151	0,96969697	1,895798319	0,94030531	0,9846223	42,5
25º	consulta al var	1	0,015151	0,96969697	1,895798319	0,94030531	0,9846223	42,5
26º	al var la	1	0,015151	0,96969697	1,895798319	0,94030531	0,9846223	42,5
27º	var la oferta	1	0,015151	0,96969697	1,895798319	0,94030531	0,9846223	42,5
28º	la oferta y	2	0,015151	1,939393939	2,823529412	1,852993983	0,9846223	64
29º	oferta y las	1	0,015151	0,96969697	1,895798319	0,94030531	0,9846223	42,5
30º	y las imágenes	1	0,015151	0,96969697	1,895798319	0,94030531	0,9846223	42,5
31º	las imágenes de	1	0,015151	0,96969697	1,895798319	0,94030531	0,9846223	42,5
32º	imágenes de cristiano	1	0,015151	0,96969697	1,895798319	0,94030531	0,9846223	42,5
33º	de cristiano ronaldo	1	0,015151	0,96969697	1,895798319	0,94030531	0,9846223	42,5
34º	cristiano ronaldo de	1	0,015151	0,96969697	1,895798319	0,94030531	0,9846223	42,5
35º	ronaldo de portugal	1	0,015151	0,96969697	1,895798319	0,94030531	0,9846223	42,5
36º	de portugal elegido	1	0,015151	0,96969697	1,895798319	0,94030531	0,9846223	42,5
37º	portugal elegido para	1	0,015151	0,96969697	1,895798319	0,94030531	0,9846223	42,5
38º	elegido para mostrar	1	0,015151	0,96969697	1,895798319	0,94030531	0,9846223	42,5
39º	para mostrar la	1	0,015151	0,96969697	1,895798319	0,94030531	0,9846223	42,5
40º	mostrar la oferta	1	0,015151	0,96969697	1,895798319	0,94030531	0,9846223	42,5
41º	oferta y cato	1	0,015151	0,96969697	1,895798319	0,94030531	0,9846223	42,5
42º	y cato amarilla	1	0,015151	0,96969697	1,895798319	0,94030531	0,9846223	42,5
43º	cato amarilla a	1	0,015151	0,96969697	1,895798319	0,94030531	0,9846223	42,5
44º	amarilla a cristiano	1	0,015151	0,96969697	1,895798319	0,94030531	0,9846223	42,5
45º	a cristiano ronaldo	1	0,015151	0,96969697	1,895798319	0,94030531	0,9846223	42,5
46º	cristiano ronaldo capitán	1	0,015151	0,96969697	1,895798319	0,94030531	0,9846223	42,5
47º	ronaldo capitán de	1	0,015151	0,96969697	1,895798319	0,94030531	0,9846223	42,5
48º	capitán de portugal	1	0,015151	0,96969697	1,895798319	0,94030531	0,9846223	42,5
49º	de portugal y	1	0,015151	0,96969697	1,895798319	0,94030531	0,9846223	42,5
50º	portugal y pouraliganji	1	0,015151	0,96969697	1,895798319	0,94030531	0,9846223	42,5
51º	y pouraliganji mostró	1	0,015151	0,96969697	1,895798319	0,94030531	0,9846223	42,5
52º	pouraliganji mostró una	1	0,015151	0,96969697	1,895798319	0,94030531	0,9846223	42,5
53º	mostró una tarjeta	1	0,015151	0,96969697	1,895798319	0,94030531	0,9846223	42,5