ZipfTool: Uma ferramenta bibliométrica para auxílio na pesquisa teórica

ZipfTool: A bibliometric tool for supporting in theoretical research

Diego Nunes Molinos [1]
Daniel Gomes Mesquita [2]
Debora Nayar Hoff [2 3]

Data de submissão: 14/10/2015, Data de aceite: 13/05/2016

Resumo: Devido ao volume de trabalhos publicados nos veículos de divulgação científica, ferramentas de análise de dados textuais tornam-se importantes para diversas áreas de conhecimento. Utilitários dessa natureza oferecem ao usuário funcionalidades tanto em âmbito quantitativo quanto qualitativo. Do ponto de vista quantitativo, possibilitam identificar a frequência de ocorrência de palavras no texto e diferenciar verbos, substantivos e artigos definidos. Já as análises qualitativas tratam do levantamento de palavras de maior conteúdo semântico. Este trabalho tem por finalidade apresentar o desenvolvimento de uma ferramenta de análise de dados que não somente possui primitivas de análise quantitativas mas também qualitativas. Quantitativamente a ferramenta fornece a frequência dos principais termos do texto, enquanto qualitativamente ela identifica as palavras de maior teor semântico. Utilizando de técnicas advindas da bibliometria, a ferramenta apresentada, chamada de ZipfTool implementa tanto a 1^a quanto a 2^a Leis de Zipf. Este trabalho também apresenta um estudo de caso na área de arquitetura de computadores e mostra uma redução do universo de artigos a serem analisados de 46785 para 1508, permitindo observar a importância da utilização da ferramenta ZipfTool principalmente no auxilio para observação de conceitos, termos e palavras.

***P*alavras-chave:** lei de Zipf, análise textual, bibliometria, conteúdo semântico, frequência de ocorrência, análise qualitativa, análise quantitativa, computação reconfigurável

[1]Universidade: Universidade Federal de Uberlândia, UFU - Uberlândia, Minas Gerais, Brasil. {diego.molinos@ufu.br}
[2]Universidade Federal do Pampa, UNIPAMPA - Santana do Livramento, Rio Grande do Sul, Brasil. {mesquita@unipampa.edu.br}
[3]{deborahoff@unipampa.edu.br}

Abstract: Due to the high number of scientific publications, textual data analysis tools are important for many knowledge fields. Such tool's features can be both in quantitative and qualitative levels. On one hand quantitative tools allows to identify the frequency of of words in the text and differentiate verbs, nouns and definite articles. On the other hand, qualitative tools analyzes identifies words of greater semantic content, identifies descriptors and keywords in the text. This study aims to present the development of a data analysis tool with both quantitative and qualitative approaches. Quantitatively the tool provides the frequency of the main terms of the text, while qualitatively it identifies the words of greater semantic content. Using techniques arising from bibliometrics, the presented tool, so called ZipfTool both implements the 1^a and 2^a Laws of Zipf. This article also presents a case study in the field of computer architecture that allows us to see the importance of using the ZipfTool, mainly with respect to a more accurate observation of concepts, terms, words and definitions.

***K*eywords:** Zipf's law, textual analysis, bibliometric, semantic content, occurrence frequencies, quantitative analysis, qualitative analysis, reconfigurable computing

1 Introdução

A busca pela construção do conhecimento remete para a importância de estabelecer-se uma discussão na direção do campo teórico. Dentro do campo teórico, conceitos e termos são apresentados como construções lógicas, os quais são estabelecidas de acordo com um sistema de referência [20]. O uso indiscriminado de termos e conceitos dentro das ciências, de modo geral, conduz ao empobrecimento do campo de estudo e dos próprios conceitos.

Pode-se, segundo [1] observar que a noção de $conceito$ tem-se confundido muito com a noção de $significado$, resultando em uma analogia errônea com o conceito de objeto, tendo a ideia de algo pronto, passível de ser apenas decorado e repetido. Sócrates mostrou que a definição conceitual se inicia com o raciocínio indutivo, expressando a essência ou a natureza de algo [22]. Segundo [8], o conceito é constituído de elementos que se articulam numa unidade estruturada. Dentre esses elementos há enunciados verdadeiros sobre o objeto que se deseja descrever. Esses enunciados são expressos através de signos que possam traduzir e fixar os enunciados que definem um objeto. Comumente esses signos são palavras. Disso abstrai-se a necessidade de identificação do peso semântico das palavras em um texto.

Ainda conforme [8], se no quotidiano a imprecisão na descrição de objetos pode não trazer grandes consequências, quando se trata de linguagens mais especializadas, essas consequências podem ser desagradáveis. Um exemplo de "linguagens mais especializadas"é a forma de expressão no meio científico, onde conceitos são utilizados para compreensão da

realidade e para replicação de experimentos. Neste espaço não há, ou pelo menos não deveria haver, margem para subjetividade.

Para [31], um dos maiores obstáculos ao desenvolvimento do conhecimento humano advém justamente da imprecisão dos termos utilizados na constituição dos saberes. Ainda segundo a autora, esta dificuldade gera confusões e inadequações de graves consequências.

Por outro lado, o processo de construção do conhecimento normalmente se dá dentro de um paradigma científico. Segundo [18] durante um período de tempo, a evolução científica acontece dentro de parâmetros aceitos pela comunidade de pesquisadores de um tema ou área. Entretanto o autor salienta que muitas das grandes revoluções científicas ocorrem em dissonância com os paradigmas vigentes, causando rupturas que podem levar à construção de novos paradigmas.

Em todo caso, seja para enquadrar-se no paradigma atual, seja para refutá-lo, o cientista precisa ler e compreender os trabalhos relacionados com o seu. Entretanto, considerando as facilidades das publicações digitais e da internet, a tarefa de seleção e leitura do que é relevante pode ser árdua[4]. Ainda que se faça uso de filtros disponibilizados pelas editoras de publicações eletrônicas ou ferramentas de busca, o número de artigos relacionados com determinadas palavras-chave pode ser elevado, dificultando o discernimento do pesquisador.

Uma forma de otimizar o tempo do pesquisador seria a utilização de técnicas que o permitissem analisar e catalogar, automaticamente, artigos baseados em termos específicos nos textos. Tais termos podem ser descritores ou palavras-chave relevantes para seu campo de pesquisa. Desta forma, artigos com conteúdo semântico menos relevante para o tema da pesquisa podem ser descartados sem a necessidade de uma leitura completa desses artigos, poupando tempo para textos mais significativos.

Técnicas advindas da bibliometria permitem analisar a frequência de ocorrência de palavras dentro de um texto, lançando mão de métodos matemáticos e estatísticos para investigar e quantificar os processos de comunicação e escrita. Neste contexto, destaca-se a lei de [33], que além calcular a frequência de ocorrência das palavras dentro do texto, permite a identificação de palavras-chaves e descritores, bem como as palavras de maior conteúdo semântico dentro de um determinado texto.

Este trabalho tem como objetivo apresentar uma ferramenta auxiliar para pesquisa científica, baseada na Lei de Zipf, bem como discutir um estudo de caso relacionado com pesquisa em arquitetura de computadores.

Para uma melhor compreensão deste trabalho, a seção 2 apresenta os fundamentos matemáticos da Lei de Zipf, além de mencionar o ponto de transição de Goffman. Já a seção

[4]Em 13 de abril de 2015 a IEEE Xplore Digital Library comemorava a chegada ao número de dois milhões de artigos publicados em HTML.

3 discute a importância da utilização de ferramentas de análise textual, bem como apresenta alguns aplicativos que implementam a Lei de Zipf nesse contexto. A ferramenta ZipfTool, proposta neste trabalho, é descrita na seção 4 . Um estudo de caso é discutido na seção 5.2, no qual a ferramenta foi utilizada para auxiliar na criação de um arcabouço conceitual para pesquisas na área da computação reconfiguráveis. Finalmente, a seção 6 discute os resultados obtidos e fornece um panorama dos trabalhos atuais e futuros relacionados com a ferramenta ZipfTool.

2 Lei de Zipf e o ponto de transição de Goffman

A bibliometria é a área do conhecimento que utiliza métodos matemáticos e estatísticos para investigar e quantificar os processos de comunicação e escrita [13]. Convergente a este conceito, [16] indica que a bibliometria compreende o exame dos aspectos quantitativos dos processos de produção, disseminação e uso da informação registrada, contendo medidas e modelos matemáticos que auxiliam os exercícios de prospecção e tomada de decisão. Desses modelos quantitativos, a Lei de [33] é uma técnica que identifica a frequência de ocorrência de palavras dentro de um texto longo. Como exercício, Zipf analisou a obra Ulisses, de James Joyce. Percebeu então uma correlação entre a frequência em que um termo aparecia e sua posição na lista de palavras ordenadas segundo sua frequência de ocorrência. Isso levou Zipf a concluir que havia uma regularidade na seleção e no uso das palavras. Também observou que a posição de um termo, multiplicada por sua frequência iguala-se a uma constante (≈ 26.500). Essa lei pode ser expressa como:

> "O produto da ordem de série (R) de uma palavra pela sua frequência de ocorrência (F) é aproximadamente constante (C).

A "ordem de série"é a representação temática da organização das palavras em ordem, de acordo com a quantidade de vezes que elas aparecem no texto. Isso significa que a palavra de maior número de ocorrências tem ordem 1, e que vem logo em seguida tem ordem 2, e assim por diante. Matematicamente, a Lei de Zipf pode ser descrita como na Equação 1.

$$R \times F = C \tag{1}$$

Entretanto foi observado que essa lei não se aplica para palavras de baixa frequência. O próprio Zipf propôs uma segunda Equação para tratar dessa anomalia. Essa Equação foi revisada e modificada por [2], que deu a forma da Equação 2.

$$\frac{I_1}{I_n} = \frac{n \times (n+1)}{2} \tag{2}$$

Na Equação 2, I_1 representa a quantidade de palavras que tem frequência 1, I_n representa a quantidade de palavras que tem frequência n, e 2 é a constante válida para a língua inglesa.

A constante válida empregada na equação 2 trata-se de um valor utilizado para análise em textos escritos na língua inglesa, existem trabalhos, [6] e [19] que fazem menção a utilização da Lei de Zipf em outros idiomas, porém, fazem uso de abordagens empíricas para adaptar este valor de constante.

De acordo com [19], a Lei de Zipf, em sua forma natural é valida para língua portuguesa, pois, nessa forma primitiva a lei trata do nível de ocorrência de palavras no texto e a indexação das mesmas. A modificação proposta por [2] a princípio não é aplicável para trabalhos escritos na língua portuguesa, porém, ao se alterar o denominador 2 da equação, pelo denominador 1.5, identificado por [19] para textos em língua portuguesa, é possível adequar a modificação proposta por [2] para esta língua.

As Equações 1 e 2 descrevem o comportamento das palavras situadas nas extremidades da lista de distribuição em um dado texto. Portanto pode-se inferir que há uma região com palavras cujas frequências de ocorrência são similares. Nessa região crítica há uma transição de comportamento de palavras de alta frequência para palavras de baixa frequência. [12] levantou a hipótese de que as palavras de maior conteúdo semântico (descritores, palavras-chave ou termos de indexação) de um determinado texto estariam nessa região.

Conforme proposto por [2], palavras que possuem baixa frequência tem seu número de ocorrência tendendo a 1. Então, substituindo I_n por 1 na Equação 2, obtém-se a Equação 3:

$$\frac{I_1}{1} = \frac{n \times (n+1)}{2} \tag{3}$$

Que pode ser rearranjada como:

$$n^2 + n - 2 \times I_1 = 0 \tag{4}$$

Resolvendo-se a Equação 4 através da popular fórmula de Bhaskara [5], levando-se em consideração apenas a raiz positiva, tem-se:

$$n = \frac{-1 + \sqrt{1 - 8 \times I_1}}{2} \tag{5}$$

[5] Ainda que não haja evidência que o brilhante matemático indiano do século XII tenha desenvolvido a resolução de equações de 2^o, essa é a nomenclatura ensinada nas escolas do Brasil.

Então, o n calculado na Equação 5 é denominado ponto de transição (T) de Goffman, que determina graficamente a localização da transição das palavras de alta frequência para as de baixa. Segundo sua hipóteses, existe uma região em torno desse ponto onde há maior probabilidade de encontrar-se as palavras com maior conteúdo semântico. A Figura 1 ilustra a área de transição em torno do ponto T de Goffman, bem como as regiões onde se encontram as palavras de alta frequência de ocorrência e as de baixa.

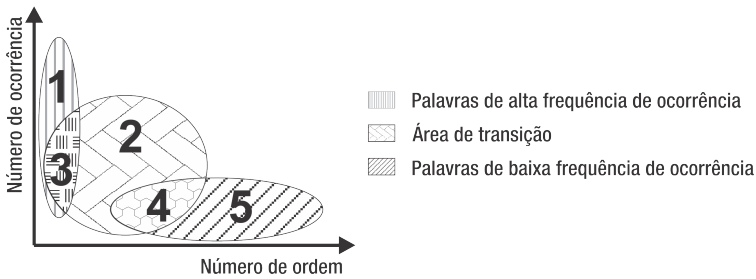

Figura 1. Zonas de ocorrência de palavras classificadas segundo a Lei de Zipf-Booth e interpretadas segundo Goffman

Ainda na Figura 1, a primeira zona de ocorrência, representada na Figura pelo número 1 é composta por palavras de maior número de ocorrências. Essas palavras normalmente são as raízes de sintaxe do idioma em que o texto é escrito (por exemplo, artigos definidos ou indefinidos). Já a segunda zona, descriminado na Figura pelo número 2 se caracteriza por conter uma quantidade maior de representantes de categorias morfológicas e informativas do que a primeira zona, como substantivos, adjetivos e verbos. A terceira zona, representada pelo numeral 3 é conhecida como ponto de transição, que de acordo com Goffman se encontram as palavras de maior teor semântico. Finalmente, a quarta e quinta zona contém instâncias que ocorrem uma única vez.

3 Referencial teórico e trabalhos relacionados

Esta seção ressalta a importância da análise textual e lista algumas ferramentas disponíveis para essa atividade.

3.1 Importância da análise textual

Conforme [24], a leitura e a produção textual são atividades habituais no quotidiano de milhares de pessoas, as quais estão diretamente relacionadas com o desenvolvimento intelectual e social. São consideradas de extrema importância para o aprendizado, no entanto,

são notáveis as dificuldades enfrentadas pelos leitores na tentativa de captar a ideia intrínseca do texto.

De acordo com [5], a análise de conteúdo textual trata-se de métodos qualitativos de extração de dados, sendo esses compreendidos por um conjunto de técnicas que convergem para a busca do entendimento de um determinado texto. Esse processo denomina-se análise semântica.

Ainda de acordo com [5], afirma-se que o método de análise de conteúdo é balizado por duas vertentes: A da linguística tradicional e a da interpretação do sentido das palavras (hermenêutica). A primeira vertente visa guiar o pesquisador a utilizar métodos de análise lógicos e estéticos, onde se tem a busca por aspectos formais típicos do texto. Já a segunda vertente prioriza métodos puramente semânticos, partindo da interpretação epistemológica e ontológica de palavras e frases de um texto.

A complexidade do processo de interpretação de um texto fica evidente, uma vez que exige a compreensão do conhecimento expresso na sua forma escrita, para além da ambiguidade natural de quase todos os idiomas. Ressalta-se aqui a importância das definições precisas dos conceitos no campo teórico, uma vez que favorecem a superação da ambiguidade. Outro aspecto, já citado, que contribui para a complexidade da tarefa diz respeito ao imenso volume de informações disponíveis atualmente. Buscar referencial teórico e trabalhos relacionados para embasar uma pesquisa científica têm se tornado uma tarefa hercúlea.

Considerados esses aspectos, é cada vez mais necessário o uso de ferramentas eficientes para auxiliar os cientistas na construção do conhecimento.

Ferramentas e algoritmos que possuem técnicas de análise de textos são largamente utilizados para extrair, organizar e observar o comportamento do conhecimento através dos textos, oferecendo apoio na identificação de termos relevantes, palavras chaves e descritores inseridos nos textos [17].

Na seção seguinte são listadas alguns aplicativos relacionados com este tema.

3.2 Ferramentas de análise textual

A análise textual, em grande parte dos casos, exige a manipulação de grandes quantidades de dados. Assim, o desenvolvimento de software específico para este fim contribui para a redução do esforço manual, gerando resultados de maneira mais rápida e organizada. Do ponto de vista prático e analítico a análise textual é definida como um método de extração de dados relevantes, utilizando bases de dados não estruturadas, ou semi-estruturadas [10]. Do ponto de vista do software de análise, os mesmo devem atender alguns requisitos funcionais que servem de apoio para a análise. Conforme [17] esses requisitos podem ser definidos como:

1. **Contagem de Termos:** Levantamento do número de termos utilizados no texto, tais como, artigos, preposições, verbos e sujeitos;
2. **Apresentação dos termos:** visualização de todos os termos catalogados;
3. **Apresentação dos termos relevantes:** identificação dos termos relevantes, tais como: descritores, palavras chaves e termos mais utilizados;
4. **Frequência dos termos:** visualização da frequência de ocorrência de cada termo dentro do texto;
5. **Relacionamento entre os termos:** visualização dos relacionamentos entre dois ou mais termos para identificação das ideias gerais do texto;
6. **Visualização gráfica dos termos e relacionamentos entre os termos:** visualização gráfica de todos os termos e relacionamento entre eles;

Através da análise desses requisitos pode-se delinear a diferença entre os conceitos aplicados juntamente com uma melhor compreensão dos termos utilizados, permitindo novas proposições e uma análise diferenciada sobre o texto. Abaixo segue uma breve descrição sobre algumas ferramentas avaliadas frente aos requisitos citados.

3.2.1 *TextAnalyzer*

O *TextAnalyzer* trata-se de uma ferramenta online gratuita que se encontra ativa desde abril de 2009 e foi desenvolvida pela iniciativa *Online-Utility.org* com o objetivo de auxiliar pesquisadores, escritores e alunos na análise de textos. Sob a luz da analise textual, a ferramenta permite várias análises do texto, tais como: identificação da frequência de palavras, identificação do número de palavras, identificação do número de sílabas. Além desses aspectos a ferramenta também consegue identificar frases correlacionando as palavras de maior frequência no texto. A ferramenta possui suporte para vários idiomas e não possui uma interface gráfica amigável, sendo bastante robusta para uma análise textual única, mas torna-se inviável para um grande número de textos.

3.2.2 *WordCounter*

O *WordCounter* é uma ferramenta de análise textual online, desenvolvida por [25] que tem como objetivo principal a mensuração das palavras que possuem maior frequência de ocorrência no texto. Assim como a grande maioria das ferramentas online, a mesma não dispõe de uma interface gráfica para auxilio visual das informações. Uma das aplicações possíveis para a o *WordCounter* é a mensuração de palavras que se encontram repetidas no texto, evitando possíveis repetição de palavras [17]. Assim como outras ferramentas online, oferece suporte para análise unitária e não para um grupo ou conjunto de textos.

3.2.3 *SOBEK*

O *SOBEK* é uma ferramenta desenvolvida com o intuito de servir de apoio para professores e pesquisadores na evolução do processo textual. A ferramenta apresenta aspectos bem atrativos do ponto de vista de usabilidade pois, permite acompanhar todo o processo de escrita de forma clara e inteligível [21]. Conforme [17] a ferramenta apresenta grande potencial para análise textual, principalmente em casos de bases de dados não estruturadas. A ferramenta trabalha de forma unitária, ou seja, analisando um texto por vez e possibilita ao usuário visualização dos principais termos e relacionamentos em forma de grafos, permitindo identificar a ideia original do texto através de conceitos analisados pré-definidos [21].

3.2.4 *TagCrowd*

O *TagCrowd* foi desenvolvido por [29], trata-se de uma ferramenta de análise textual simples, porém robusta, não apresenta muitas funcionalidades para o âmbito da análise de dados, oferecendo o básico da análise quantitativa de palavras [17]. Através de configurações parametrizadas, tais como: o número de palavras que deseja-se obter como resultado, idioma e a frequência de ocorrência, a ferramenta direciona a analise sobre o texto predefinido. Diferentemente de outras ferramentas online, o *TagCrowd* oferece uma interface gráfica mais amigável, ilustrando de forma diferenciada palavras que possuem maior frequência de ocorrência no texto. Assim, como as outras ferramentas online, não possui suporte para analisar vários textos de forma dinâmica, oferecendo apenas análise unitária.

3.2.5 *Uff - Lei de Zipf*

Desenvolvida pelo Instituto de Matemática da Universidade Federal Fluminense a ferramenta *Uff - Lei de Zipf* a ferramenta apresenta robustez no cálculo de frequência de ocorrência das palavras de um determinado texto, juntamente com resultados marginais convenientes para o estudo da Lei de Zipf. A mesma não possui uma interface visual amigável, porém oferece a possibilidade da visualização das informações através de gráficos o que possibilita a interpretação dos resultados de forma diferente. Para que o usuário parametrize os resultados a ferramenta oferece a possibilidade de inserção de filtros na análise. Assim como ferramentas online para esse propósito específico, recai-se sobre a mesma problemática, análise de textos unitária não permitindo análise de textos em lote.

3.2.6 *Iramuteq*

A ferramenta *IRAMUTEQ* trata-se de um software gratuito, desenvolvido por [26] e licenciado pela GNU GPL (v2). O software foi projetado tendo como base o software R (www.r-project.org) e a linguagem Python de programação (www.python.org)[4]. Dentre as principais análises realizadas pelo software, além das análises clássicas, ou seja, análise

lexicográfica e frequência de palavras, a ferramenta permite uma análise mais apurada, como, análise de similitude que utiliza teoria de grafos, possibilitando identificar as coocorrências entre as palavras e da conexão entre as mesmas [4] [26].

A Figura 2 faz uma ilustração das ferramentas apresentadas anteriormente face aos requisitos listados por [17].

Ferramenta	Funcionalidades						
	Online	Contagem de termos	Visualização dos termos	Termos relevantes	Frequência dos termos	Relacionamento dos termos	Visualização gráfica
TextAnalyser	X	X	X		X		
WordCounter	X	X	X		X		
SOBEK	X	X	X	X	X	X	X
TagCrowd	X			X	X		X
UffZipf	X	X	X	X	X	X	X
Iramuteq	X	X	X	X	X	X	X

Figura 2. Ferramentas analisadas face aos requisitos propostos por [17]

Faz-se necessário esclarecer que a Figura 2 foi adaptada dos requisitos funcionais propostos por [17], tendo como modificação mais significativa a coluna de termos relevantes. Para o autor supracitado, termos relevantes são termos que possuem alto nível de ocorrência no texto, discorda-se deste entendimento, pois acredita-se que palavras que possuem maior valor semântico são termos relevantes e nem sempre os mesmos se encontram entre as palavras de maior ocorrência do texto [13], dessa forma nenhuma das ferramentas atendem a esse critério de forma positiva.

4 Ferramenta ZipfTool

A ferramenta ZipfTool foi desenvolvida com o principal objetivo de auxiliar pesquisadores no processo de análise de textos. A ferramenta realiza a análise do texto em sua totalidade, extraindo informações quantitativas e qualitativas do mesmo. Cabe-se salientar que ferramentas de análise de dados textuais não são e não devem ser taxadas como ferramentas de mineração de dados ou como técnica de mineração, pois, conforme [9], o processo de mineração de dados consiste em uma sequência predefinida de tarefas aplicadas sob uma base da dados comum, como exemplo de tarefas pode-se citar: Análise de Regras de Associação, Análise de Padrões Sequenciais, Classificação e Predição, Aglomeração e Análise de Outliers. Diante deste contexto pode-se inferir que ferramenta ZipfTool de análise textual partilha de conceitos como associações de dados, classificação, aglomeração e prognósticos, porém não faz uso de nenhuma técnica de mineração muito menos uso de nenhuma ferramenta projeta para tal tarefa, tais como: Intelligent Miner de [3] ou DBminer de [30].

Projetada no ambiente matemático MATLAB, a ZipfTool realiza a leitura de arquivos com extensão .TXT [6]. Quanto ao formato, trata-se do mesmo utilizado pelas ferramentas *online* já mencionadas. Uma conversão do formato .PDF para .TXT ficou de fora do escopo desta ferramenta, uma vez que existem aplicativos e scripts gratuitos que se ocupam eficientemente dessa tarefa [7].

4.1 Especificação

Conforme [28], especificação de um software está diretamente relacionado a definição de suas funcionalidades, em resumo, o levantamento dos requisitos da ferramenta, não ignorando as restrições e limitações que a mesma possui, pois de fato, esse saldo entre requisitos e restrições é o que caracteriza a especificação de uma ferramenta. Abaixo seguem os principais tópicos que retratam as especificações da ferramenta ZipfTool.

1. Ambiente para execução;
2. Formato de textos suportados;
3. Quantidade e tamanho de textos para análise;e
4. Regras de configuração;

Abaixo é apresentado cada tópico da especificação juntamente com seus requisitos funcionais.

4.1.1 Ambiente para execução: A ferramenta ZipfTool foi desenvolvida utilizando o software MATLAB (Versão 2012a). Dessa forma para que a ferramenta ZipfTool funcione, a plataforma citada deve estar instalada. Não há necessidade de instalação de nenhum pacote adicional na ferramenta para o funcionamento da mesma.

4.1.2 Formato de textos suportados: Devido a plataforma MATLAB não ser exclusivamente projetada para trabalhar com diferentes arquivos do tipo texto, o modelo padrão mais comum e primitivo de texto foi adotado, o **.TXT**, que é reconhecido e manipulado por qualquer sistema operacional e qualquer plataforma de desenvolvimento. Cabe salientar que a plataforma MATLAB possui diversas bibliotecas de software já pré-definidas e configuradas com inúmeras funções que auxiliam o carregamento, análise e armazenamento de arquivos em **.TXT**.

[6] Formato reconhecido por todos os sistemas operacionais e aplicativos.
[7] Como por exemplo o "Free PDF to Text Converter"do fabricante LotApps, que pode ser gratuitamente obtido através do "http://lotapps-free-pdf-to-text-converter.soft32.com"

4.1.3 Idiomas dos textos: A ferramenta ZipfTool foi projetada para analisar trabalhos escritos em língua Inglesa, porém, existe trabalhos baseados em abordagens empíricas que preveem a modificação das equações da Lei de Zipf para adaptar a mesma para trabalhos na língua portuguesa.

4.1.4 Quantidade e tamanho de textos suportados: A ferramenta ZipfTool não possui um limite pré estabelecido com relação ao tamanho dos arquivos para análise, vale lembrar que esse limite está diretamente relacionado as limitações de hardware do computador, como por exemplo, memória *RAM* disponível no momento da execução dos scripts e também o espaço disponível nas mídias de armazenamento em massa (*Hard Disk*) para armazenamento dos resultados. Cabe salientar que o software MATLAB em execução sobre o sistema operacional possui as características de um processo em execução, sendo assim, o mesmo partilha de um tamanho limitado de memória física do hardware para execução de suas tarefas, essa caracterizada como memória virtual, a qual é dedicada ao processo no momento de sua execução. Os scripts do MATLAB carregam todos os arquivos .TXT no momento da execução, sendo o número máximo de arquivos bem como o tamanho dos mesmos diretamente relacionados com estes aspectos apresentados. Diante do contexto, diversos testes foram realizados, os scripts conseguiram carregar arquivos de até 162 páginas ou arquivos fragmentados que somados não ultrapassem 162 páginas.

Diferentemente das ferramentas *online*, as quais possuem a capacidade de análise de um texto por vez, a ferramenta ZipfTool já foi projetada tendo como um dos seus requisitos funcionais a análise em lote para quando se deseja analisar diversos textos. A ferramenta possui a capacidade de organizar os resultados em múltiplos arquivos de saídas.

4.1.5 Regras de configuração: A ferramenta ZipfTool utiliza-se de *scripts* para carregamento, execução e salvamento das tarefas. Por tratar-se de *scripts* os mesmo necessitam de pré-configuração, cujas variáveis de configuração são apresentadas na Tabela 1.

4.2 Desenvolvimento

A ferramenta ZipfTool foi projetada visando preencher algumas lacunas as quais foram observadas em outras ferramentas de análise textual, tais como: análise em lote de arquivos de texto e a identificação de termos de maior conteúdo semântico no texto. As subseções a seguir relatam aspectos do desenvolvimento da ZipfTool.

4.2.1 Plataforma de desenvolvimento

A ferramenta ZipfTool foi inicialmente instituída tendo como base os requisitos funcionais básicos de análise de dados conforme apresentado na seção 3. Como já mencio-

Tabela 1. Descrição das variáveis do arquivo de configuração da ferramenta ZipfTool

Variável de Configuração	Descrição
$MyDir$	Define o caminho ou diretório onde se encontram os arquivos para análise.
$Directory$	Define o caminho ou diretório de saída, usado para armazenar os resultados.
$NumberofChar$	Define um número mínimo de caracteres que as palavras analisadas devem conter para não serem descartadas.
$Exception$	Define um conjunto de palavras, separadas por espaço em branco, que serão descartadas no momento da análise textual.

nado anteriormente, para o desenvolvimento da ferramenta ZipfTool foi utilizado o software MATLAB R2012a, o qual possui um propósito de ser uma plataforma de desenvolvimento interativo, possuindo alto desempenho para cálculos numéricos, assim permitindo desenvolver soluções com otimização de tempo de desenvolvimento em relação a outras plataformas de desenvolvimento dessa natureza. Apesar de não ser uma plataforma própria para implementação de algoritmos de análise textual o mesmo possui diversas bibliotecas e funções já pré-definidas que permitem implementações neste âmbito.

Abaixo são apresentadas algumas características da plataforma MATLAB,

1. **Sintaxe** - O MATLAB utiliza a linguagem M-Code ou simplesmente M. Através de uma forma interativa o usuário interage com a plataforma de duas formas, a primeira dela é utilizando o *Comand Window*, um espécie de prompt de comando, onde o usuário insere os comandos e a ferramenta executa o processamento tendo como base o histórico de comandos já inseridos, podendo também interagir através de scripts, onde o usuário pode definir o cadenciamento dos mesmos e executá-los em lote, [11].

2. **Scripts** - Trata-se de um contentor de comandos, onde cada vez que o script é executado, os comandos são processados de forma *top-down*[8], iniciando na primeira linha e terminando somente na última. A utilização de scripts é bastante conveniente, pois os mesmos permitem alterações e atualizações do código e a possibilidade do mesmo ser executado inúmeras vezes, [11].

3. **Visualização gráfica** - Por se tratar de uma plataforma de desenvolvimento voltada para cálculos numéricos, a plataforma possui diversas formas de ilustração gráfica, partindo do 2D até gráficos tridimensionais.

[8]Método de execução de linhas de código, onde a execução é ordenada linha após linha, iniciando de cima para baixo

Diante do apresentado cabe-se esclarecer que do ponto de vista do ambiente de desenvolvimento, a ferramenta ZipfTool trata-se de um conjunto de scripts MATLAB, ordenados e finitos para realização das tarefas conforme é apresentado neste trabalho.

4.2.2 Scripts

A ferramenta ZipfTool possui 4 scripts desenvolvidos em linguagem M-Code. Um script principal, similar a um programa principal, que é responsável por executar todos os outros scripts como procedimentos. Abaixo segue em detalhes as funcionalidades de cada script.

1. **Script Conf.m:** Trata-se do script responsável pelo carregamento das informações de controle do processo de análise. O script **Conf.m** é o primeiro script a ser executado na cadeia de script, o mesmo é responsável por colher informações do arquivo *conf.txt*, onde tem-se as informações sobre o diretório de entrada (conjunto de textos a serem analisados), diretório de saída (conjunto de resultados da análise) e o número de descarte (palavras que possuem os números de caracteres iguais aos números de descarte são descartadas automaticamente). O script **Conf.m** também faz a leitura do arquivo *exception.txt*, o qual é responsável por armazenar palavras que não serão computadas (descartadas), e que não inferem de forma positiva nos resultados, como exemplo, artigos definidos e indefinidos da gramática.

2. **Script RunMyFiles.m:** Este script tem como principal objetivo fazer uma varredura dentro do diretório estabelecido como universo de dados (definido no arquivo de configuração) e organizar os arquivos para análise, de forma que, quando a análise estiver sendo realizada, todos os textos estejam completamente catalogados.

3. **Script Load.m:** Este script é responsável por fazer o carregamento de todos os caracteres do texto e o armazenamento na memória temporária (cache) do processo de análise, ficando essas disponíveis até o próximo carregamento de dados.

4. **Script Zipf.m:** Este script é responsável por executar (instanciar) todos os scripts explicados anteriormente, aplicar a tratativa básica de padronização (acentuação, maiúsculo e minúsculo) bem como executar de todas as primitivas matemáticas que envolve a 1^a e 2^a Lei de Zipf, calcular de frequência de palavras, sequenciar as palavras conforme nível de ocorrência, calcular o ponto de transição do Goffman bem como identificar as palavras que fazer parte desta érea de transição.

4.2.3 Arquivo de configuração

A ferramenta ZipfTool possui um arquivo de configuração, chamado de *conf.txt*, que é responsável pelas configurações de partida, contendo informações essenciais para o funciona-

mento de toda processo de análise. O script *Conf.m* está programado para ler cada linha deste arquivo como sendo uma configuração, abaixo segue é apresentado a estrutura do arquivo supracitado:

1. A primeira linha do arquivo *Conf.txt* deve ser atualizada com as informações do diretório que contém os arquivos para análise, já no formato **TXT**. O caminho deve ser especificado em sua completude, incluindo partição, pastas e subpastas.

2. A segunda linha do arquivo *Conf.txt* deve ser atualizada com as informações do diretório onde os resultados das análises são armazenados. O caminho deve ser específicado de forma completa, incluindo partição, pastas e subpastas.

3. Terceira linha do arquivos *Conf.txt* deve ser atualizada com o número de caracteres de descarte, esse número é interessante pois, o script Zipf.m armazena o número de caracteres de cada palavras que é catalogada e analisada, assim, as que possuírem o número menor ou igual ao definido são automaticamente descartadas da análise.

4.3 Aplicações da ferramenta

A ferramenta ZipfTool é um mecanismo de análise textual, cabe salientar, que a ferramenta possui implementado as primitivas matemáticas da 1^a e a 2^a Leis de Zipf, permitindo assim, uma análise dos termos de maior valor semântico no texto, além de cálculo da frequência de ocorrência das palavras. A Figura 3, apresenta possíveis aplicações da ferramentas dentro do âmbito da análise textual.

APLICAÇÃO	DESCRIÇÃO	ÁREA DE ATUAÇÃO
Identificação da frequência de ocorrência de palavras no texto	A ferramenta permite identificar a frequência de ocorrência de palavras em um determinado texto. Aplicável em processos, onde se deseja identificar palavras repetidas no texto, palavras de maior ocorrência bem como as de menor ocorrência no texto.	Auxílio nas escritas de trabalhos científicos, artigos e revistas. Análises textuais de uma forma geral.
Identificação de palavras de maior teor semântico	A ferramenta permite identificar palavras de maior teor semântico no texto, as quais são classificadas como descritores, termos relevantes e palavras chave. Aplicável em processos onde necessita-se verificar a natureza do texto, bem como o índice de utilização de termos dentro do texto.	Análises conceituais em artigos, revistas, livros, etc. Auxílio na avaliação terminológica dos termos.
Análise de conjunto de textos	A ferramenta permite a análise sobre um conjunto de textos de forma dinâmica, sem a necessidade de carregamento unitário. Aplicável em processos onde o sucesso está condicionado ao número de trabalhos analisados.	Avaliações conceituais, onde o número de definições comprometem a consistência dos resultados. Avaliação de inúmeros trabalhos em geral onde o aspecto quantitativo é importante.

Figura 3. Possíveis aplicações para ferramenta ZipfTool

5 Resultados e Análise

Esta seção é responsável por apresentar os resultados quantitativos no que tange às funcionalidades da ferramenta ZipfTool em relação a outras ferramentas de análise textual. Além disso, traz um estudo de caso da ferramenta aplicada no auxílio em uma pesquisa na área de computação reconfigurável, onde fica claro a contribuição qualitativa da ferramenta no auxílio de pesquisas em âmbito teórico. Cabe salientar que em nenhum momento foi efetuado testes qualitativos para avaliar a qualidade dos resultados gerados das ferramentas citadas na seção 3, não é objetivo deste trabalho apresentar resultados nesse âmbito.

5.1 Comparação entre a ZipfTool e outras ferramentas

De acordo com o apresentado na seção 3, [17] enfatiza alguns requisitos funcionais os quais as ferramentas de análise de textos devem incorporar.

A Figura 4 ilustra um comparativo entre os requisitos propostos por [17] face as funcionalidades da ferramenta ZipfTool.

Ferramenta	Funcionalidades									
	Online	Contagem de termos	Visualização dos termos	Termos relevantes	Frequência dos termos	Relacionamento dos termos	Visualização gráfica	Análise em Lote	Termos de maior conteúdo semântico	Parametrização da Análise
TextAnalyser	X	X	X		X					
WordCounter	X	X	X		X					
SOBEK	X	X	X	X	X	X	X			
TagCrowd	X			X	X		X			X
UffZipf	X	X	X	X	X	X	X			X
Iramuteq	X	X	X	X	X	X	X			X
ZipfTool		X	X	X	X		X	X	X	X

Figura 4. Análise quantitativa das ferramentas de análise de dados

Pode ser observado que a ferramenta ZipfTool atende a praticamente todos os requisitos apresentados por [17], não abrangendo os requisitos de *Ferramenta OnLine* e *Relacionamento dos termos*. Nota-se que alguns requisitos funcionais os quais não são contemplados no trabalho de [17] são adicionados, pois, entende-se que os mesmo são importantes para análise dos textos, são eles:

1. **Análise em Lote:** Permite a ferramenta analisar diversos arquivos de textos não necessitando o carregamento unitário. Funcionalidade importante para análise de grandes volumes de dados de entrada.
2. **Termos de maior conteúdo semântico:** Discorda-se de [17] no que se refere a termos relevantes, para o mesmo são palavras que possuem maior frequência de ocorrência

dentro do texto. De acordo com [13], palavras de maior conteúdo semânticos são termos relevantes que aparecem no texto e definem a ideia principal do autor, normalmente, palavras chaves, descritores e indexadores, os quais podem auxiliar a definir aspectos morfológicos do texto.

3. **Parametrização das variáveis de análise:** Ilustrando uma análise textual em forma gráfica, gera-se algo similar a Figura 3.2.6, onde pode ser observado que a região onde se localizam as palavras de maior frequência de ocorrência no texto são compostas palavras, em aspectos morfológicos conhecidas como: artigos, conjunções e interjeições; Sendo em uma região mais nobre, denominada de região de transição, encontra-se palavras, em aspectos morfológicos conhecidas como: adjetivos, verbos, preposições, pronomes e sujeitos, sendo esses termos o mais prováveis a serem descritores, palavras chave ou termos de maior conteúdo semântico do texto. Formas de parametrização de análise que possibilitem ignorar um grupo de palavras que possuem um determinado número de letras ou até mesmo um grupo seleto de palavras se torna bem interessante do ponto de vista analítico.

Como pode ser observado através da análise da Figura 4, a ferramenta Zipftool apresenta como principal diferencial a possibilidade de análise de arquivos do tipo lote, ou seja, diversos arquivos sem a necessidade de carregamento um a um, porém também tem como ponto negativo a mesma não estar disponível online e não efetuar relacionamento dos termos de análise.

5.2 Estudo de Caso

O estudo de caso se justifica do ponto de vista científico como uma maneira metódica de descrever o exemplo de um determinado conhecimento [32]. O estudo de caso descrito nessa seção é compreendido por um trabalho de pesquisa realizado na Universidade Federal de Uberlândia, realizado no contexto de um mestrado acadêmico. Seu objetivo foi uma rigorosa e detalhada análise conceitual sobre um tema específico no campo da computação. Abaixo são apresentados maiores detalhes sobre essa dissertação.

5.2.1 Arcabouço Conceitual para Computação Reconfigurável

A pesquisa compreendida pelo título supracitado foi desenvolvida a partir de uma análise feita em artigos publicados na área da computação reconfigurável. Os autores do trabalho constataram que havia uma inconsistência conceitual entre os termos utilizados dentro deste campo de estudo. Após realizarem uma vasta leitura em trabalhos publicados, observaram que alguns termos ora eram tratados como sinônimos e ora tratados como coisas bem divergentes.

Uma discussão conceitual dentro de qualquer campo de estudo, sempre apresenta importância significativa para área, já que "conceitos"são considerados instrumentos fundamentais para compreensão da área. O uso indiscriminado dos conceitos dentro da uma érea científica, de modo geral, conduz ao empobrecimento da mesma [20].

Em um trabalho prévio nosso, [23], para avaliar a qualidade e a forma com que os conceitos estavam sendo empregados, precisou-se desenvolver um processo de análise das principais definições conceituais encontrados em artigos na área da computação reconfigurável. A primeira dificuldade destacada por [23] foi a quantidade de trabalhos publicados que compõem o universo da Computação Reconfigurável, incluindo livros, jornais, revistas e artigos de congressos e outros eventos científicos. A partir de uma análise amostrada, foi observado que nem todos esses trabalhos publicados possuíam definições claras sobre os termos que envolvem a computação reconfigurável, sendo que os artigos lidos pareciam não valorizar a definição conceitual dos termos.

A Tabela 2 mostra a quantidade de trabalhos selecionados através da biblioteca *online* [15]. A localização destes trabalho foi feita usando-se os termos de busca apresentados na Tabela 5, sem o uso de qualquer tipo de filtro adicional.

Tabela 2. Número de publicações selecionados pela ferramenta online

Termo utilizado	Quantidade de Trabalhos selecionados
Reconfigurable Computing	5771
Reconfigurable Hardware	6550
Reconfigurable Architecture	9659
FPGA	24805

A partir dos dados apresentados na Tabela 2, pode-se dizer que o universo a ser analisado no campo sob observação é bastante grande. Mesmo que fossem detectadas redundâncias nos artigos selecionados em cada termos, trata-se de 40 mil trabalhos a serem analisados. Isso posto, fica evidente a necessidade de uma ferramenta que permita a identificação dos trabalhos mais relevantes.

5.2.2 Uso da ZipfTool como ferramenta auxiliar na construção do Arcabouço Conceitual para Computação Reconfigurável

A utilização da ferramenta ZipfTool no estudo de caso apresentado possui duas vertentes, uma qualitativa e outra quantitativa. Sob a luz do aspecto qualitativo, a ferramenta possui a capacidade de identificar os termos de maior conteúdo semântico do texto, como forma de evidenciar os trabalhos convergentes ao objetivo da busca, isso tende a poupar um grande esforço inicial manual de seleção e análise dos trabalhos, o que tende a otimizar o

tempo de pesquisa. No aspecto quantitativo, a ferramenta tende a induzir que sejam selecionados somente trabalhos que possuem alta relevância para com os objetivos de pesquisa, gerando assim, um universo amostral reduzido e de alto valor semântico.

Resultados Qualitativos

Tendo como hipótese inicial de que, muitos dos trabalhos que faziam parte do universo inicial de análise, mesmo após o procedimento de amostragem, não seriam de grande auxilio para a pesquisa. Isso porque, o primeiro levantamento de dados foi realizado utilizando a ferramenta *Online* [15], sendo que a mesma seleciona os trabalhos de sua base de dados correlacionando os termos de busca e os títulos dos trabalhos presentes em sua base de dados, não exercendo verificação de conteúdo sobre os trabalhos.

Os termos de busca instituídos inicialmente apenas faziam referência a termos que são constantemente citados na área de estudo de computação reconfiguravél, podendo ou não, terem relações com os objetivos da pesquisa. Para tanto, foi necessário identificar termos de busca que estimulassem a seleção natural de trabalhos que convergissem com os objetivos da pesquisa. Para tal tarefa, inicialmente foram identificados trabalhos consagrados na área da computação reconfigurável como tendo um ótimo embasamento conceitual e por terem sido construídos por pesquisadores expressivos e reconhecidos no campo, a julgar pelo número citações que os mesmos possuem. Esta estratificação focada elimina certas confusões conceituais e permite um direcionamento mais polido com relação aos potenciais termos de busca, eliminando termos menos impactantes. Sobre os artigos [27], [7] e [14] foi aplicada a ferramenta ZipfTool que selecionou os termos de maior conteúdo semântico dos trabalhos citados. O resultado é o apresentado na Figura 5.

Podem ser observados na Figura 5, número de palavras analisadas, frequência de ocorrência e identificação de palavras de maior teor semântico relacionadas a cada trabalho analisado.

Como resultado desta etapa de análise qualitativa, observa-se que as palavras *reconfigurable* e *architecture* aparecem em duas das três análises como palavras de alto teor semântico juntamente com a palavra *FPGA*. Diante do exposto, estas palavras se tornam os novos termos de busca da pesquisa, pois conforme a análise realizada, as mesmas possuem a capacidade de selecionar trabalhos com conteúdo impactante para a pesquisa.

Cabe-se ressaltar que, esta análise qualitativa com o objetivo de elencar termos de busca que incitem a ferramenta a selecionar trabalhos impactantes para a pesquisa proporcionou aos autores visualizar que, alguns veículos de publicações possuíam um maior número de trabalhos publicados relacionados com os novos termos de busca e outros veículos não. Este processo descrito auxiliou também de forma qualitativa os autores a selecionarem veículos de publicações com maior número de trabalhos impactantes para a pesquisa, tendo como resultado um número de amostras consideralvemente menor que o inicial, sendo composto

Trabalho Analisado	Número de palavras catalogadas	Palavras com maior frequência de ocorrência	Palavras de maior conteúdo semântico
FPGA Architectures: Survey and Challenges	35521	Logic (446) FPGA (363) That (307) Routing (288) Block (243) This (242) Architecture (217) Area (189) Programmable (169) Design (154)	**Architecture** **FPGA** **Block**
A decade of Reconfigurable Computing: A Visionary Retrospective	9428	Data (67) Reconfigurable (65) With (59) Memory (54) Array (51) Routing (40) Time (46) Architecture (38) Compiler (38) Design (37) From (36) Mapping (34) Communication (33) Based (32) Computing (30)	**Reconfigurable** **Architecture** **Computing**
Reconfigurable Computing: A survey of Systems and Software	23523	Reconfigurable (334) Hardware (226) That (206) This (178) Computing (155) Logic (152) Configuration (120) FPGA (127) Routing (122) Circuit (118) FPGA's (116) with (109) these (104) Systems (103) Time (103) Programmable (91)	**Reconfigurable** **Hardware** **Computing** **FPGA**

Figura 5. Análise bibliometrica realizada pela ferramenta ZipfTool

de apenas trabalhos com a máxima convergência para com os objetivos da pesquisa.

Resultados Quantitativos

Após a redefinição dos termos de busca, a Figura 6 ilustra o número de trabalhos relacionados, os quais foram selecionados utilizando os veículos de publicação de maior ex-

pressão no campo de estudo conforme análise qualitativa descrita anteriormente e as palavras definidas pela ferramenta ZipfTool como parâmetros de busca.

Para ser possível a redução do universo de análise, além dos parâmetros identificados pela ferramenta ZipfTool juntamente a utilização dos veículos de publicações de maior expressão, foi necessário aplicar alguns conceitos estatísticos, principalmente no que tange à amostragem estratificada, pois, realizando uma análise mais minimalista pode-se observar que trabalhos de um mesmo ano e de um mesmo veiculo de publicação possuem características homogêneas entre si, como por exemplo a aplicação de conceitos, os quais são, objetos de estudo da pesquisa.

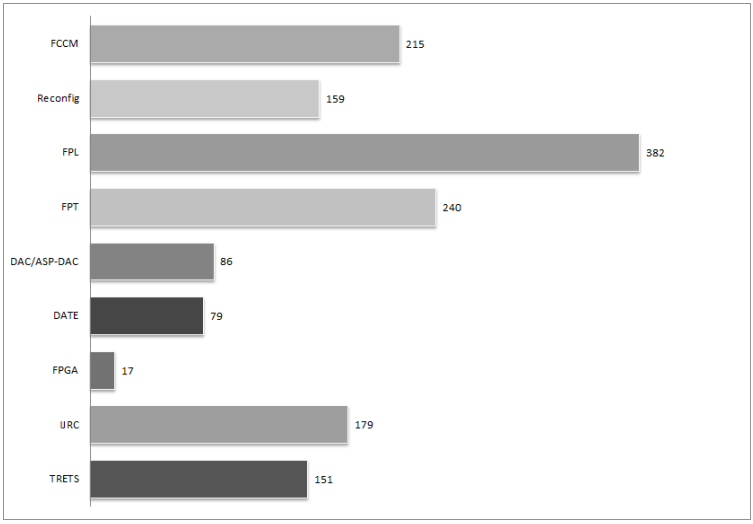

Figura 6. Número de Publicações nos principais veículos de divulgação

Através da Figura 6 pode-se observar uma grande redução do universo de análise. A ferramenta ZipfTool proporcionou a identificação dos principais termos que possuem relação direta os principais conceitos do campo de estudo, gerando assim, os parâmetros de busca necessário para seleção dos trabalhos impactantes para a pesquisa. Ainda, permitiu que os autores observassem características homogêneas entre determinados veículos de publicações, resultando na identificação de veículos mais expressivos para utilização da pesquisa, contribuindo diretamente para redução do universo amostral.

6 Conclusão

Neste trabalho foi apresentado o desenvolvimento da ferramenta ZipfTool, que trata-se de um mecanismo de análise de dados textuais com o principal objetivo de extrair do texto palavras de alto teor semântico, descritores e palavras chave. A ferramenta foi desenvolvida tendo como base os princípios da primeira e segunda Leis de Zipf's do campo da bibliometria, permitindo diversas opções de análise textual para o usuário.

A ZipfTool é uma ferramenta parametrizável, que possui um arquivo de configuração a partir dos qual é possível que o usuário controle diversos aspectos, tais como: caracteres e palavras de descarte, análise unitária ou em bloco e a geração de gráficos para visualização dos resultados obtidos.

Diante do atual cenário técnico/científico, onde os campos de estudos possuem inúmeras publicações em inúmeros veículos, uma ferramenta de auxílio que possibilita o usuário a ter uma melhor análise sobre os termos estudados, verificando a consistência dos termos utilizados bem como o nível de repetição de determinadas palavras, torna-se bastante importante.

Para ilustrar uma das funcionalidades da ferramenta, um estudo de caso foi apresentado, onde a ferramenta ZipfTool auxiliou os autores a entender de forma mais consistente e integra os conceitos que estavam observando, bem como, foi utilizada com filtro, auxiliando no descarte de trabalhos os quais não possuíam relação direta com os objetivos da pesquisa.

A análise feita sobre a ferramenta ZipfTool apresentada neste artigo considerou sua versão produzida sobre a plataforma MATLAB. Esta escolha deu-se pela facilidade de programação fornecida pelo MATLAB, uma vez que muitas primitivas matemáticas necessárias já encontram-se definidas em sua linguagem. Essa abordagem permitiu o desenvolvimento mais rápido de uma prova de conceito do que seria possível em outras linguagens. Entretanto, para que a ferramenta possa ser amplamente utilizada pela comunidade científica, temos consciência que a implementação de uma versão $on - line$ é necessária. Nesse sentido, nos propomos, à guisa de trabalho futuro, a programar de uma nova versão da ZipfTool baseada na linguagem $Python$, que será disponibilizada para uso através de um navegador de internet, sem necessidade de instalação no computador do usuário e que será gratuita.

Contribuição dos autores:

- Diego Nunes Molinos: participou da elaboração do projeto, levantamento de requisitos, implementação e desenvolvimento da ferramenta, testes, análise de resultados, redação e revisão do artigo

- Daniel Gomes Mesquita: participou da elaboração do projeto e orientou todas as eta-

pas do trabalho incluindo o desenvolvimento da ferramenta, delineamento do estudo de caso, aprimoramento da revisão do estado-da-arte, redação e revisão final do artigo. Destaca-se a busca de referencial teórico que permitiu a adaptação da ZipfTool para o idioma português.

- Debora Nayar Hoff: contribuiu na definição dos argumentos relativos à construção do conhecimento e elementos constitutivos da revisão de literatura sobre bibliometria. Orientou todas as etapas de redação e revisão do artigo. Destaca-se neste processo sugestões de melhoria das descrições das experiências, análises e formação de quadros resumo dos resultados.

Referências

[1] N. Abbagnano. Tradução: Alfredo bosi - dicionário de filosofia. *Dicionário de filosofia*, 2, 1970.

[2] A. D. Booth. A "law"of occurrences for words of low frequency. *Information and control*, 10(4):386–393, 1967.

[3] P. Cabena, H. H. Choi, I. S. Kim, S. Otsuka, J. Reinschmidt, and G. Saarenvirta. Intelligent miner for data applications guide. *IBM RedBook SG24-5252-00*, 173, 1999.

[4] B. Camargo and A. Justo. Tutorial para uso do software de análise textual iramuteq. *Florianopolis-SC: Universidade Federal de Santa Catarina*, 2013.

[5] C. J. G. Campos. Método de análise de conteúdo: ferramenta para a análise de dados qualitativos no campo da saúde. *Rev Bras Enferm*, 57(5):611–4, 2004.

[6] Y.-S. Chen and F. F. Leimkuhler. Analysis of zipf's law: An index approach. *Information processing & management*, 23(3):171–182, 1987.

[7] K. Compton and S. Hauck. Reconfigurable computing: a survey of systems and software. *ACM Computing Surveys (csuR)*, 34(2):171–210, 2002.

[8] I. Dahlberg. Teoria do conceito. *Ciência da informação*, 7(2), 1978.

[9] S. de Amo. Técnicas de mineração de dados. *Jornada de Atualizaçao em Informatica*, 2004.

[10] R. Feldman and J. Sanger. *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge University Press, 2007.

[11] A. Gilat. *MATLAB com aplicações em Engenharia*. Bookman, 2006.

[12] W. Goffman and V. Newill. Generalization of epidemic theory. *Nature*, 204(4955):225–228, 1964.

[13] V. L. Guedes and S. Borschiver. Bibliometria: uma ferramenta estatá■stica para a gestão da informação e do conhecimento, em sistemas de informação, de comunicação e de avaliação cientí■fica e tecnologica. *Encontro Nacional de Ciência da Informação*, 6:1–18, 2005.

[14] R. Hartenstein. A decade of reconfigurable computing: a visionary retrospective. In *Proceedings of the conference on Design, automation and test in Europe*, pages 642–649. IEEE Press, 2001.

[15] IEEE. ieeexplore digital library. URL: http://www.ieeexplore.ieee.org/Xplore/home.jsp, 2013. Acesso em 17/05/2013.

[16] T. S. Jean. An introduction to informetrics. *Information processing management*, 28(1):1–3, 1992.

[17] M. Klemann, E. Reategui, and C. Rapkiewicz. Análise de ferramentas de mineração de textos para apoio a produção textual. In *Anais do Simpósio Brasileiro de Informática na Educação*, volume 1, 2011.

[18] D. Kuhn. Teaching and learning science as argument. *Science Education*, 94(5):810–824, 2010.

[19] E. Lima and S. Maia. Comportamento bibliométrico da língua portuguesa, como veículo de representação da informação. *Ciência da Informação*, 2(2), 1973.

[20] S. S. Lisboa. A importancia dos conceitos da geografia para a aprendizagem de conteúdos geográficos escolares. *CEP*, 36570:000, 2007.

[21] A. L. Macedo, E. Reategui, A. Lorenzatti, and P. Behar. Using text-mining to support the evaluation of texts produced collaboratively. In *Education and Technology for a better world*, pages 368–377. Springer, 2009.

[22] G. d. A. Martins. Sobre conceitos. definições e constructos nas ciências administrativas. *Gestão & Regionalidade*, 21(62), 2010.

[23] D. N. Molinos. Arcabouço conceitual para computação reconfigurável. URL: http://http://repositorio.ufu.br/handle/123456789/4550, 2014. Acesso em 03/01/2015.

[24] M. Z. Moretto and C. E. Rapkiewicz. Usando mineração de textos como suporte ao desenvolvimento de resumos no ensino médio. *RENOTE*, 11(3), 2013.

[25] S. Morgan Friedman. Wordcounter. URL: http://http://www.wordcounter.com, 2004. Acesso em 17/06/2014.

[26] P. Ratinaud. Iramuteq: Interface de r pour les analyses multidimensionnelles de textes et de questionnaires. *Téléchargeable à l?adresse: http://www.iramuteq.org*, 2009.

[27] J. Rose, I. Kuon, and R. Tessier. Fpga architecture: Survey and challenges. *Foundations and Trends® in Electronic Design Automation*, 2(2):135–253, 2008.

[28] M. d. S. Soares. Comparação entre metodologias áĄgeis e tradicionais para o desenvolvimento de software. *INFOCOMP Journal of Computer Science*, 3(2):8–13, 2004.

[29] D. Steinbock. Tagcrowd. URL: http://http://tagcrowd.com/, 2002. Acesso em 17/06/2014.

[30] J. H. Y. F. W. Wang, J. C. W. G. K. Koperski, D. Li, Y. L. A. R. N. Stefanovic, and B. X. O. R. Zaiane. Dbminer: A system for mining knowledge in large relational databases. In *Proc. Intl. Conf. on Data Mining and Knowledge Discovery (KDD?96)*, pages 250–255, 1996.

[31] V. R. Werneck. Sobre o processo de construção do conhecimento: o papel do ensino e da pesquisa. *Ensaio: Avaliação e Políticas Públicas em Educação*, 14(51):173–196, 2006.

[32] R. K. Yin. *Estudo de caso: Planejamento e métodos*, volume 4. Bookman Porto Alegre, 2005.

[33] G. K. Zipf. Relative frequency as a determinant of phonetic change. *Harvard studies in classical philology*, pages 1–95, 1929.