Instituto Universitário de Lisboa

Departamento de Ciências e Tecnologias de Informação

Influência dos Sentimentos dos Turistas nos Social Media para o Desenvolvimento do Turismo

Marta Alpedrinha Ramos de Almeida Nave

Dissertação submetida como requisito parcial para obtenção do grau de

Mestre em Sistemas Integrados de Apoio à Decisão

Orientador(a):

Doutor Paulo Rita, Professor Catedrático, ISCTE-IUL

Co-orientador(a):

Doutor João Guerreiro, Professor Auxiliar, ISCTE-IUL

[Setembro, 2016]

AGRADECIMENTOS

Esta dissertação foi o resultado de todo o meu empenho e dedicação, mas também, de todos os importantes apoios e incentivos sem os quais não se teria tornado numa realidade e aos quais serei eternamente grata.

Agradeço ao Professor Doutor Paulo Rita e ao Professor Doutor João Guerreiro, pelas suas orientações, todo o apoio e disponibilidade para me esclareceremas dúvidas que me foram surgindo ao longo da realização deste trabalho; pela sabedoria que me transmitiram, pelas opiniões e críticas,pelo interesse demonstrado e pelo feedback em relação ao desenvolvimento e opções tomadas, não esquecendo também, todas as palavras de incentivo.

Agradeço também, aos meus amigos e colegas que estiveram ao meu lado durante esta fase, em especial à Carolina Santos, pelo apoio, incentivo, força e companheirismo.

Por último, tendo em conta que sozinha nada disto seria possível, dirijo os meus agradecimentos especiais aos meus familiares, principalmente aos meus pais e ao meu irmão, por serem o meu exemplo de força, garra, coragem e determinação, pelo apoio incondicional, pelo incentivo, pela amizade e principalmente pela paciência demonstrados e pela total ajuda na superação dos obstáculos que ao longo desta caminhada foram surgindo. Acima de tudo agradeço por nunca terem deixado de acreditar em mim.

Não posso terminar os agradecimentos sem te referenciar, “Aninhas dos quatro ovos”, estejas onde estiveres sei que estarás orgulhosa. Obrigada por teres acreditado que isto seria possível.

Dedico este trabalho aos meus pais, irmão e a ti “minha” Aninhas.

RESUMO

O turismo é hoje o motor para alcançar o crescimento e desenvolvimento económico, devendo ser apoiado para garantir a sustentabilidade ambiental. Segundo o *World Tourism Organization*, o turismo é uma das atividades económicas mais importantes do mundo, sendo este o reflexo da sua elevada participação no PIB e do alto nível de empregabilidade neste sector (Nations, 2012). Em 2013, este sector teve um crescimento de 5% em todo o mundo ((UNWTO), 2013). Esta realidade por si só justificou a necessidade de compreender as expectativas e as opiniões transmitidas pelos turistas em cada cidade que visitam.

Atualmente, com a evolução da *web* 2.0 e das redes sociais, a interação entre os clientes e as empresas passou a ser realizada de forma muito mais rápida e praticamente instantânea, trazendo vantagens e desvantagens. Os utilizadores passaram a poder difundir as suas opiniões de forma global e instantânea.As organizações passaram a poder tirar partido desta evolução através da análise das opiniões dos utilizadores relativamente aos bens/serviços prestados, reagindo de forma célere aos pontos apontados como negativos e confirmar a aceitação do público-alvo. Com a difusão desta prática de publicação das opiniões próprias, hoje quem usufrui de bens/serviços, sobretudo associados ao turismo (viagens, restaurantes, hotéis), não prescinde de publicar a sua opinião através dos *sites* das próprias organizações e nas redes sociais.No entanto, segundo a literatura, essa informação, não tem sido utilizada, apenas existem estudos centrados no sector hoteleiro e no planeamento de viagens. Por estes motivos achou-se que seria estritamente interessante analisar os comentários dos turistas, uma vez que é informação extremamente valiosa para este sector, através das técnicas de *Text Mining*(TM) e *Sentiment Analysis(SA)* e posteriormente criar um sistema de apoio à decisão(DSS). A criação deste último, será com o intuito de ser uma mais-valia para o turismo de cada cidade, percebendo o que os turistas apreciam em cada cidade, quais são as categorias de negócio mais e menos apreciadas, quais as categorias que têm que ser melhoradas, quais as que vão de encontro às expectativas dos turistas, e em que tipo de categoria de negócio se deve investir mais (tendo em conta os gostos dos turistas) para através dos comentários poder constatar quais as categorias que são menos atrativas e o seu porquê.

Através do *site*/aplicação *Yelp,*www.yelp.com, (um guia urbano de várias cidades, com o intuito de ajudar os utilizadores a encontrarem locais de lazer com base nas opiniões dos seus utilizadores), construímos um DSS, como foi referido anteriormente, utilizando previamente as técnicas tanto de TM e SA, que fosse capaz de avaliar os sentimentos/polaridade (positivo, negativo ou neutro) das opiniões/experiências e relacionar o modo como as experiências dos turistas em cada cidade, podem afetar os índices de satisfação dos turistas.

Palavras Chave:*Sentiment Analysis*, *Text Mining*, *Social Media*,*Web* 2.0, Turismo.

ABSTRACT

Tourism today is an engine for economic growth and development, and it must be supported to ensure envinonmental sustainability. According to the World Tourism Organization, tourism it's one of the most important economical activities world wide, which reflects in it's high GDP contribution and in the high employability level in the sector (Nations, 2012).

In 2013, this sector had a 5% growth world wide ((UNWTO), 2013). This reality just by itself, has justified the need to understand the expectations and opinions of tourists in each city that they visited.Today, with Web 2.0 and social network evolution, the interaction between clients and enterprises is much faster, almost instantaneous, which brigs advantages and disavantages. Now users can share their opinions in a global and instantaneous way and organizations can take advantage of thisevolution through exploring user's opinions.

However, according to the literature such valuable information is not being used and studies are currently focused on hotel management and in travel planning.Therefore it was found that it would be interesting to analize tourists comments by using Text Mining(TM) and Sentiment Analysis(SA) techniques to create a decision support system (DSS). Such system may allow city or hotel managers to know what tourists value in each city, which business categories are most and less valued, which categories need an improvement, which categories meet the tourists exceptations, and what's the business categories where investment should be bigger, according to the tourists needs and wants.

Through the website/app Yelp, www.yelp.com, (an urban guide to many cities,with the goal to help users find recreation places according to users opinions), we have built a decision support systemusing TM and SA techniques, that is capable of evaluatingthe sentiments and polarities (positive, negative,neutral) of tourists experiences. The DSS can be used to show how tourists experiences in each city may affect their satisfaction indexes.

Keywords:*Sentiment Analysis*, *Text Mining*, *Social Media*,*Web* 2.0, *Tourism*.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

1. INTRODUÇÃO

Neste capítulo pretende-se apresentar um primeiro enquadramento da presente investigação, identificando-se os principais objetivos e justificação a sua necessidade.

1.1. Enquadramento

As informações geradas nos *social media*, são fundamentais, porque refletem interpretações individuais das vivências dos utilizadores, conduzindo a "verdades" possivelmente não coincidentes com a realidade. Relativamente ao sector de turismo, as opiniões deixadas nos *social media*(comentários, avaliações, perceções dos destinos) são extremamente importantes uma vez que são transmitidas praticamente em tempo real e globalmente, ficando *online* para todos os utilizadores (Månsson, 2011). Segundo o *World Tourism Organization*, o turismo é uma das atividades económicas mais importante do mundo, sendo esta o reflexo da sua elevada participação no PIB e do alto nível de empregabilidade neste sector. Por outras palavras, o turismo é o motor do crescimento, do desenvolvimento económico e da sustentabilidade ambiental (Nations, 2012). Em 2013, o turismo teve um crescimento de cerca de 5% em todo o mundo (UNWTO, 2013).Os dados provisórios da *World Tourism Organization,* em 2014,apontavam para um aumento das chegadas de turistas internacionais de 4,4%, correspondendo a 1134,7 milhões de turistas(Estatística, 2015).

A *World Tourism Organization* obteve as seguintes projeções da evolução de fluxo de turistas estrangeiros no mundo e as previsões desde 2010 até 2020, como se pode observar de seguida na figura 1:

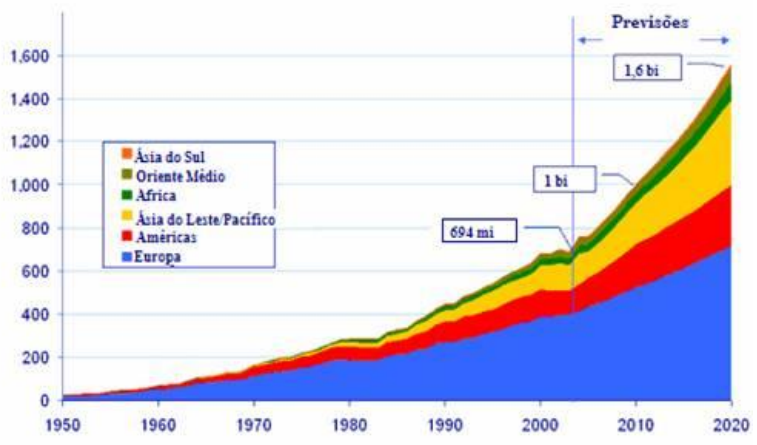


FIGURA 1: EVOLUÇÃO DO FLUXO DE TURISTAS ESTRANGEIROS NO MUNDO E PREVISÕES DE 2010 ATÉ 2020. (WORLD TOURISM ORGANIZATION).

Observando a figura 1, compreende-se que haverá um crescente crescimento no aumento de turistas, principalmente na Europa. Esperando de uma forma geral, em todo o mundo, que em 2020 as chegadas internacionais de turistas atinjam 1,6 bilhões.

Com a evolução da *web* 2.0 e o aparecimento das redes sociais, no final do século 20, a *internet*, teve uma enorme difusão, mudando de forma radical a maior parte da nossa vida económica e social (Milano, Baggio, & Piattelli, 2011). Com a aparição dos *social media*, os consumidores/clientes podem utilizar ativamente plataformas e *sites*, criando conteúdo *online* e partilhas de experiencias sobre produtos e serviços que acaba por ser informação extremamente importante, tanto para futuros clientes como para as diversas empresas (Ayeh, Leung, Au, & Law, 2012). Essa informação é extremamente valiosa para as empresas, uma vez que, reflete a "imagem" que os clientes têm delas, bem como os seus serviços.No entanto, as empresas não tem a capacidade de tratar desta informação, devido a ser informação não estruturada.

A utilização da*SA* de texto não estruturado (TM) tem aumentado nos últimos anos, uma vez que permite obter muita informação a partir das publicações colocadas na *internet,* nomeadamente os usos e preferências dos utilizadores/clientes dos mais variados serviços, tornando-se assim, uma vantagem estratégica para as organizações, uma vez que apreendem os riscos e ameaças da sua atividade (Bai, 2011).

A interpretação das emoções (SA), para além de poder promover uma vantagem estratégica, também pode afetar o comportamento individual e a tomada de decisão (Bollen, Mao, & Zeng, 2010).

Neste trabalho, pretende-se utilizar as técnicas de TM*(*dados não estruturados*)*,SA nos *social media* e *Correlated Topic Model* (CTM) para encontrar padrões de comportamento e segmentos de áreas de negócio para que possa ser usado para futura investigação epara posteriormente implementar um DSS, com o intuito de compreender como é que as experiencias (sendo atividades que os turistas tenham experimentado) nas cidades (usando casos de estudo dos Estados Unidos) podem afetar os índices de satisfação dos turistas. Este contributo pretende ajudar a indústria turística a sintonizar-se com as expectativas dos turistas e contribuir desse modo para o crescimento do sector através de um sistema de apoio à decisão. Este sistema de apoio à decisão avaliará em que medida os sentimentos positivos e negativos das experiências afetam a satisfação dos clientes (medido pelo rating).

1.2. Motivação

Analisando os dados não estruturados (através do TM) deixados nos *social media* é importante perceber as causas das preferências dos turistas em usufruir de experiências em certas cidades em detrimento de outras.

Tendo em conta a importância crescente do sector do turismo na economia das sociedades e dos países, através da entrada de receitas/divisas decorrentes dos alojamentos, deslocações, restauração, atividade socioculturais, eventos, museus, entre outros, é extremamente interessante e útil avaliar e interpretar os sentimentos e recomendações deixadas pelos turistas nos *social media*.

Com este tema, pretende-se identificar e evidenciar os aspetos mais valorizados e criticados pelos utilizadores nas suas experiências em cada cidade, percebendo como as experiências podem afetar o índice de satisfação dos turistas, e assim contribuir para uma melhor aplicação dos recursos e consequentemente garantir mais receitas, potenciando o crescimento económico (UNWTO, 2013).

1.3. Justificação de Investigação

O turismo, por razões diversas, nomeadamente a descida das idades de reforma em alguns países, e com desafogo económico, tem apresentado uma contínua expansão e diversificação ao longo das últimas seis décadas. Tornando-se um dos sectores de mais acelerado crescimento da economia mundial, transformando-se num motor para o crescimento económico e sustentabilidade ambiental. Em 2013 o turismo teve um crescimento de 5% em todo o mundo (UNWTO, 2013).

Tanto o TMcomo o SA são conceitos relativamente recentes.

Com o surgimento do *web* 2.0 e das redes sociais, passou a ser uma mais-valia para as empresas/organizações uma vez que os utilizadores deixam informações extremamente valiosas.

Ao observar estas três áreas (TM*,* SA e Turismo), observou-se que existia um *gap*, sobretudo relativamente à literatura sobre este tema, uma vez que os estudos existentes, focando estas três áreas, são centrados no sector hoteleiro e no planeamento de viagens. Por esse motivo, levanta-se a seguinte questão de investigação:

- Quais os maiores determinantes e limitações para os índices de satisfação das experiencias nas cidades em termos de opiniões nos *social media*?

Pretende-se responder a esta questão nesta dissertação.

Assim, pretende-se com esta investigação contribuir para complementar a pesquisa académica relacionada com esta lacuna, procedendo à análise dos *reviews* de um *dataset* fornecido pelo *Yelp,* sobre os Estados Unidos.

Este país foi escolhido por ser um dos países onde o sector do turismo é um dos maiores empregadores do país e por ser o que gera maiores receitas em todo o mundo.

A opção de abranger estas três áreas, TM, SA e Turismo, impôs-se como um desafio pela necessidade de dar resposta a uma lacuna detetada.

1.4. Objetivos da Investigação

O objetivo principal da investigação será compreender como é que as experiências positivas e negativas dos turistas (desde vida ativa, *shopping*, restauração, artes e entretenimento, automóveis, beleza e *spas*, educação, serviços de planeamento de eventos, serviços financeiros, comida, medicina, serviços domésticos, hotéis e viagens, vida noturna entre outros) relatados nos *social media*, *Yelp*, em cada cidade podem afetar o índice de satisfação dos turistas.

O conhecimento do objetivo principal permite ajudar a indústria turística a sintonizar-se com as expectativas dos utentes e contribuir desse modo para o crescimento do sector.

Através do *dataset*, que foi fornecido pelo *site*/aplicação *yelp* (guia urbano de várias cidades, que se propõe a ajudar os utilizadores a encontrarem locais de lazer com base nas opiniões dos seus utilizadores), www.yelp.com, obteve-se uma amostra, que contém 12371 *reviews*, 4615 negócios, 22 tipos de experiências em 51 cidades dos Estados Unidos.

Utilizando TM e SA pretende-se conseguir atingir os objetivos propostos, utilizando o *dataset* referido anteriormente, com base nas opiniões deixadas nos *social media* (neste caso no *Yelp*), identificando primeiro a polaridade dos sentimentos de cada experiência e criando um DSSque permita avaliar em que medida as experiencias em cada cidade podem afetar mais e menos o índice de satisfação de cada turista.

1.5. Metodologia

O *Design Science Research (DSR)*(Peffers, Tuunanen, Rothenberger, & Chatterjee, 2007), tem como principal objetivo resolver problemas práticos e teóricos com recursos a artefactos de Tecnologia de Informação(TI) destinados a resolver problemas organizacionais bem identificados (Peffers et al., 2007). Foi esta a metodologia escolhida para a resolução do problema detetado, uma vez que é de índole prático.

O *DSR,* tem uma abordagem em quatro tipos de análise (pontos de entrada) (Peffers et al., 2007):

- Centrado no Objetivo;
- Centrado no Problema;
- Centrado no Desenvolvimento;
- Projetado pelo Cliente.

No nosso caso a abordagem será efetuada através do segundo ponto de entrada referido, dado que a investigação resultou da observação do problema.

Esta metodologia é composta por seis fases:

1. Motivação e Identificação do Problema: Encontra-se descrito na secção de Motivação e na de Justificação de investigação.

2. Definição de Objetivos para a Solução: Encontram-se descritos na secção dos Objetivos de Investigação. No entanto, o principal objetivo é permitir relacionar como é que as experiências dos turistas em cada cidade podem afetar os seus índices de satisfação.

3. Desenho e Desenvolvimento: nesta fase, é proposto o desenho do DSSa ser desenvolvido e as suas explicações do desenvolvimento. No entanto, numa primeira instância antes da criação do DSS é necessário estruturar a informação não estruturada em termos, agregá-la em tópicos (***Bar***, ***Pizzeria***, ***Snack Bar***,***Brunch***,***Fast Food***, ***Cake Shop***, ***Tea House***, ***Restaurant***, ***Hairdresser***, ***Night Life***, ***Hotel***, ***Leisure***), posteriormente atribuir sentimentos à informação utilizada. Só após estes procedimentos é que iniciaremos o DSSpara explorar as reações dos turistas.

A definição e a conceção do DSS, assim como as técnicas utilizadas, tanto para TM como para SAserão aqui explicadas.

4. Demonstração: nesta fase, iremos aplicar o DSS criado utilizando o *dataset*pelo *site*/aplicação *Yelp*. Houve a necessidade de criar uma amostra aleatória do *dataset* referido*,* ficando assim, com uma amostra com 51 cidades dos Estados Unidos, 4615 negócios, 12371 *reviews* e 22 categorias de negócio (tipos de experiências). O *dataset* aplicado já contém a informação estruturada bem como a informação agregada em tópicos como foi referido anteriormente.

5. Avaliação: nesta etapa será descrita a avaliação do DSS. No fundo, consiste em comparar os resultados que o DSS nos oferece numa certa cidade em diferentes anos com o objetivo de se observar a evolução da interação dos utilizadores com os *social media*, bem como, a evolução dos gostos dos turistas, a evolução da própria cidade, a evolução do comportamento dos turistas e do desenvolvimento da cidade, avaliando em que medida os sentimentos positivos e negativos das experiências poderãoinfluenciar a satisfação dos turistas.

6. Comunicação: nesta última fase far-se-á a explicação do problema, da sua importância, utilidade e novidade (Peffers et al., 2007). Gera no fundo o objetivo da dissertação.

Esta dissertação está constituída pela Revisão Bibliográfica, Metodologia de Desenvolvimento e pela Conclusão. Na secção da Revisão Bibliográfica são abordados diversos temas, entre os quais, o turismo e *social media*, TM e DSS. Nasecção Metodologia do Desenvolvimento, contém todo o trabalho implementado para o propósito da dissertação enquadrado na metodologia *Cross Industry Standard Process for Data Mining* (CRISP-DM). Por último, na secção das Conclusões encontram-se os resultadosobtidos através do DSS criado.

2. REVISÃO BIBLIOGRÁFICA

Neste capítulo apresenta-se um enquadramento teórico da presente investigação, focado em seis pontos chaves: Turismo e *social media* na atualidade, *Text Mining,* aplicações em *Text Mining, Sentiment Analysis,* aplicações em *Sentiment Analysis e Decision Support System.*

2.1.Turismo e *Social Media*

Nesta secção pretende-se apresentar o estado atual do setor do turismo e a importância dos *Social media* para o desenvolvimento do turismo.

2.1.1. Turismo

O turismo é um fenómeno complexo e globalizado com uma grande importância socioeconómica, podendo ser definido como um fluxo monetário, devido aos gastos dos consumidores, aos rendimentos das empresas, às despesas e lucros e, devido, aos efeitos sobre a economia nacional e regional do país (Tribe, 1997; Darbellay & Stock, 2012). A importância do turismo tem crescido, chegando a um patamar de fenómeno/ global, afetando um número crescente de sectores, atraindo novos mercados com novas oportunidades para viagens, sendo um dos principais criadores de emprego a nível mundial (James et al., 2010; UNWTO, 2013).

O turismo ajuda diretamente as pequenas empresas, caso estejamos a falar de turismo rural, (atrações, hotéis...) e indiretamente, outros sectores como, postos de gasolina ou supermercados(Wilson, Fesenmaier, Fesenmaier, & Van Es, 2001).As companhias aéreas, por exemplo são fulcrais para o turismo, uma vez que, mais de metade dos turistas internacionais chegam ao seu destino por via aérea. O crescimento das companhias aéreas está intrinsecamente ligado à expansão do turismo (UNWTO, 2013).

O turismo tem tido uma contínua expansão e diversificação nas últimas seis décadas, tornando-se um dos maiores sectores e o que revela um crescimento mais acelerado no sector económico mundial. Em 2013, pode-se observar que cerca de 52% dos turistas internacionais fizeram viagens de lazer e férias, 14% para fins comerciais e profissionais e 27% por outros motivos, como visitar familiares, razões religiosas, tratamentos de saúde entre outros. As receitas obtidas do turismo internacional são os ganhos gerados nos países de destino derivados das despesas de alojamento, alimentação, bebidas, transportes locais, entretenimento, compras e outros bens e serviços (World Tourism Organization, 2013).

Existe uma grande diversificação de oferta de produtos turísticos nos destinos que os turistas visitam. Nesta indústria, a questão central é perceber o que faz com que uma experiência seja memorável para um turista (Xu, 2010).

O peso do turismo na economia global difere em função dos países, em função da dimensão e da abertura da economia, em termos de limitações de capacidade de produção, em função das tragédias/catástrofes (como foi no caso do ataque à América no 11 de Setembro 2001), bem

como em função do clima.Estes fatores influenciam fortemente o turismo, principalmente nos sectores que dependem fortemente dos recursos naturais (Dwyer, Forsyth, Madden, & Spurr, 2000; Goodrich, 2002C. Oh, 2005; H. Kim, Chen, & Jang, 2006; Dawson, Scott, & Mcboyle, 2009; Tang & Jang, 2009). Resumindo, o turismo traz grandes benefícios económicosàs indústrias que dele dependem e que poderão beneficiar desta situação económica favorável, oferecendo, melhores serviços, produtos e reforçando os fatores de atração do País como destino (Zhou, Yanagida, Chakravorty, & Leung, 1997;Tang & Jang, 2009; James et al., 2010).No caso do declínio deste sector, as repercussões sobre todos os sectores relacionados também serão graves (Zhou et al., 1997;James et al., 2010).

Segundo o *World Tourism Organization*, o turismo é o motor para o crescimento, desenvolvimento económico e sustentabilidade ambiental (Nations, 2012).

Em 2013, o turismo teve um crescimento de cerca de 5% em todo o mundo. Os resultados em 2013 foram superiores às expectativas e às previsões a longo prazo. Os resultados mais abonatórios, ou seja, com a procura mais forte para o turismo internacional, foram nas seguintes regiões (UNWTO, 2013):

- Ásia e Pacífico, mais de 6% de crescimento (principal sub-região é Sudoeste da Ásia com mais de 10%);
- África, mais de 6% de crescimento (principal sub-região é o Norte de África com mais de 6%);
- Europa, mais de 5% de crescimento (principal sub-região é Europa Oriental com mais de 7%; Europa Mediterrânea, Central e Sul com mais de 6%). A Europa continua a ser a região mais visitada do mundo.

Passando agora para as chegadas de turistas internacionais, isto é, visitantes que pernoitam, em 2013, este valor cresceu 5% em todo o mundo.O maior crescimento deveu-se a um aumento de 6% nas chegadas na Ásia e Pacífico, seguido pela Europa e África, ambos com um aumento de 5%, enquanto, nas Américas, as chegadas internacionais cresceram apenas 3% mantendo-se no Médio Oriente (UNWTO, 2013).

De seguida, apresenta-se a secção “Turismo nos Estados Unidos de América”. Esta enfatização é devido ao *dataset* (que será utilizado para o desenvolvimento desta dissertação) ser relativo a este país.

2.1.2. Turismo nos Estados Unidos de América

Em 2013, segundo a *World Tourism Organization*, nas Américas em concreto, houve um crescimento no turismo de cerca de 4%, para os 169 milhões de turistas.

Nos Estados Unidos da América, esta indústria, é uma das maiores empregadoras do país, e é a que gera maiores receitas no mundo (World Tourism Organization, 2006). Esta indústria contém uma variedade de tipos de empresas/organizações bastante diversificada, desde a

hotelaria, companhias aéreas, restauração e casinos (AMÉRICAN HOTEL & LODGING ASSOCIATION, 2006).

O turismo foi apontado como a indústria topo em muitos estudos e a terceira maior indústria de retalho nos EUA (NTTAC, 1991).

De uma forma geral estima-se que o número de turistas domésticos (turistas de um dado pais que visitam outro lugar nesse mesmo país) esteja em torno de 1,5 bilhões anualmente. Por outro lado, EUA são o terceiro país mais visitado por turistas estrangeiros, perdendo apenas para Espanha e França. É visitado anualmente por 65 milhões de turistas estrangeiros. Destes turistas estrangeiros, cerca de oito milhões vem do Canadá (fazendo gastos que rondam 6,2 bilhões de dólares) e sete milhões vem do México (fazendo gastos que rondam cinco bilhões de dólares). Os outros 50 milhões de turistas internacionais anuais, vem da Europa, Japão, Caribe, China, Brasil e Argentina.

A indústria turística nos Estados Unidos de América foi severamente afetada devido ao ataque terrorista do 11 de setembro, provocando quedas imediatas nas companhias aéreas de viagens de cerca de 50% e declínios idênticos nas ocupações nos hotéis. Esta tragédia teve dois tipos de implicações, o primeiro foi um aumento da preocupação da gestão com a segurança e proteção de locais de turismo e hospitalidade, e o segundo, em termos financeiros, provocou um aumento das despesas em novas medidas de segurança. A economia dos EUA, devido ao sucedido, foi “empurrada” para uma recessão, sendo as indústrias mais prejudicadas as de turismo e viagens (Goodrich, 2002).

2.1.3. Social Media

A importância de passar a palavra “boca-a-boca” sobre diversas áreas tem sido amplamente discutida e pesquisada, particularmente desde a adoção mundial das tecnologias de *internet*, que revolucionaram a forma de comunicar(Anderson, 1998; Goldenberg, Libai, & Muller, 2001; Stokes & Lomax, 2002; Zhu & Zhang, 2006). Através da *internet*, os indivíduos podem partilhar as suas ideias e opiniões, devido à facilidade de acessoà informação e aos meios digitais (Dellarocas, 2003).

Os *social media* são considerados como uma das ferramentas mais poderosas da rede *online,* que foi integrada na vida social e económica do mundo real. Estas estão constituídas por *sites* de redes sociais, *sites* de avaliação e recomendação de consumo, *wikis*, fóruns de *internet* entre outros. Os *social media* surgiram como uma nova forma de comunicação, através da integração de tecnologias de informação e comunicação (tecnologias móveis e *web-based*)(Zeng & Gerritsen, 2014).

O advento das tecnologias da *Web* 2.0 permitiu a criação eficiente e distribuição de conteúdo gerado pelo utilizador,*User Generated Content*–UGC– (consiste no conteúdo de *media* criado ou produzido pelo público em geral e não por profissionais pagos e distribuídos principalmente na *internet*(Daugherty, 2008)), resultando em grandes mudanças no cenário da *mediaonline*. Durante as últimas décadas, o panorama dos meios de comunicação evoluiu para um conglomerado complexo e dinâmico de ambos os meios de comunicação tradicionais e interativos que visam atender às necessidades do estilo de vida acelerado de hoje (Daugherty, 2008; Miguéns, Baggio, & Costa, 2008; Sigala, 2011).

Estas tecnologias da *Web* 2.0 e UGC mudaram, e provavelmente continuarão a mudar cada vez mais, a maneira como as pessoas pesquisam, encontram, leem, recolhem, contribuem com o seus próprio conteúdo, desenvolvem e consomem informação (Ye, Law, Gu, & Chen, 2011). Estas tecnologias, referidas anteriormente, permitem que qualquer indivíduo possa publicar o seu próprio conteúdo, opiniões, vídeos, áudio ou imagens para a *web* para que outros utilizadores possam ver e responder (Cox, Burgess, Sellitto, & Buultjens, 2009;Ayeh et al., 2012). O'Reilly (2005) sugere que a *Web* 2.0 é um conjunto de princípios que incluem a capacidade de integrar informações de novas maneiras, o desejo de aproveitar o conhecimento distribuído, e a necessidade de envolver os utilizadores como “coautores” (O'Reilly, 2005). *Sites*como o YouTube, MySpace e Wikipedia são plataformas que fornecem os chamados UGC, onde os cidadãos podem publicar seus próprios comentários, fotos, vídeos e muito mais *online* (Hermida & Thurman, 2008).

Estudos anteriores acerca de comunidades *online*(por exemplo, Hagel & Armstrong, 1997) classificaram os motivos que levam os utilizadores a gerarem os seus próprios conteúdos em duas grandes categorias: racional (por exemplo, troca de informações, transações) e emocional

(por exemplo, relações, fantasia). Expandindo a sua topologia, categorizamos UGC de acordo com a principal finalidade para a qual os utilizadores da *internet* participam em atividades de geração de conteúdos: motivações racionais, que podem incluir partilha de conhecimentos com o mundo, e, defendendo uma posição especial para um problema (defesa); motivações emocionais podem incluir a construção de conexões sociais com amigos, parentes ou outros utilizadores da *internet* (conexões sociais) ou entretenimento (autoexpressão).

Outros estudos recentemente realizados têm demonstrado que comentários gerados por utilizadores *online* têm uma influência significativa sobre as vendas de produtos de consumo (Chevalier & Mayzlin, 2006; Duan, Gu, & Whinston, 2008). Segundo os autores Park et al. (2007), as opiniões *online* de consumidores são muitas vezes consideradas mais confiáveis e credíveis do que as informações prestadas pelos fornecedores de produtos e serviços, uma vez que os consumidores são considerados mais honestos na sua opinião (Park, Lee, & Han, 2007). No entanto, uma das principais preocupações é que as empresas possam usar os funcionários para que estes "atuem" como consumidores para postar comentários positivos em nome da empresa ou para postar comentários negativos sobre a concorrência (Litvin, Goldsmith, & Pan, 2008) .

Os *sites* da *web* 2.0 e as redes sociais *online,* são caraterizados pela facilidade de interação, promovendo a formação de comunidades e a geração de conteúdo orientado aos consumidores (Miguéns et al., 2008). Estes *sites*, começam a afetar fortemente a maior parte das atividades, tendo um papel importante em muitos aspetos do turismo, sobretudo na procura de informações e na procura de comportamentos decisórios, na promoção do turismo e em se concentrar sobre as melhores práticas para interagir com os consumidores através dos canais dos *social media* (Ricci & Werthner, 2004; Milano et al., 2011;Zeng & Gerritsen, 2014). Os *sites* de avaliação *online*são cada vez mais são importantes fontes de informação sobre compra de produtos de turismo (Sparks, Perkins, & Buckley, 2013).

Os desenvolvimentos das tecnologia dos *social media* permitiu que os turistas partilhassem as experiências das suas viagens, difundindo informações que podem auxiliar no planeamento de viagens de outros turistas ou então, pode eventualmente influenciar potencialmente o destino das viagens, uma vez que, muitos turistas consultam os comentários *online* antes de fazer o planeamento das viagens (Schmallegger & Carson, 2008;Vermeulen & Seegers, 2009;Burgess, Sellitto, & Cox, 2009;Xiang & Gretzel, 2010; Månsson, 2011; Parra-López et al., 2011; Parra-López, Bulchand-Gidumal, Gutiérrez-Taňo, & Dĭaz-Armas, 2011; Amaral, Tiago, & Tiago, 2014; Zeng & Gerritsen, 2014). Cada turista interpreta um *site* na sua própria maneira individual que pode de facto não corresponder à realidade (Månsson, 2011).

Para esta indústria, é essencial compreender as mudanças no comportamento dos turistas e das novas tecnologias que afetam a distribuição e acessibilidade da informação relacionada com as viagens. A *internet* tem fundamentalmente reformulado a forma como a informação relacionada com o turismo é distribuída e a forma como os turistas planeiam e adquirem as viagens (Buhalis & Law, 2008; Xiang & Gretzel, 2010),

O sector do turismo tem assim sido influenciado pela *Web* 2.0 e UGC (Schmallegger & Carson, 2008;Akehurst, 2009; Xiang & Gretzel, 2010). Os efeitos positivos têm tido repercussões tanto em fenómenos quantificáveis (como e-commerce), como em questões intangíveis (como as relacionadas com a imagem ou lado informativo de produtos e/ou serviços específicos) (Milano et al., 2011). Os *sites* como o www.travelpod.com e www.Tripadvisor.com permitem que os consumidores troquem recomendações e opiniões sobre os vários destinos e produtos turísticos.

Desde o final dos anos noventa, a fonte preferida de informação para os viajantes foi sendo alterada, passando a ser a *internet* uma das principais fontes confiáveis de informação. Atualmente, *sites* de UGC são uma das principais fontes de conteúdo útil na *Web*,tornando-se uma fonte predominante de informações sobre dezenas de características dos destinos turísticos em várias geografias (ou seja, cidades, parques nacionais, monumentos, etc.) (Hecht & Gergle, 2010; Amaral et al., 2014).

Para os turistas,os *sites* de UGC tornaram-se uma fonte adicional de informação, fazendo parte do seu processo de procura de informação (Cox et al., 2009). Uma das preocupações levantadas sobre o uso de *sites* de UGC ao planear viagens é como o consumidor pode ter certeza que as opiniões visualizadas são, de facto independentes e, portanto, confiáveis(Gretzel, 2006).

Segundo o estudo realizado por Lu e Stechenkova (W. Lu & Stepchenkova, 2014), as plataformas de *social media*contém grandes volumes de UGC, sendo estes conteúdos amplamente utilizados por consumidores de serviços de hotelaria e turismo para a recuperação de informações no processo de decisão e partilha durante e após a experiência. Neste sector, no turismo, a *internet* é uma importante fonte de informações para os viajantes (Burgess et al., 2009).

Outro estudo sugere que o número de turistas que usam a *internet* para pesquisar informações sobre destinos e fazer reservas *online* aumentou nos últimos tempos (Amaral et al., 2014). A maioria dos consumidores preferem apreender as opiniões de outros consumidores sobre um hotel, em vez de confiar apenas em própria descrição do hotel sobre si *online*(Cox et al., 2009).

O turismo tem abraçado a tecnologia há mais de três décadas, começando pela evolução dos sistemas informatizados de reserva, comunicação com os clientes, interatividade, ferramentas para a pesquisa, armazenamento de dados em massa e apoio à gestão(Cooper & Hall, 2008).

Acompanhando a evolução da *Web* 2.0 veio Turismo 2.0, definido por William e Perez como uma revolução de negócios para a indústria do turismo estimulado pela adoção de uma nova plataforma - a *Web* social. Isto levou à construção de negócios e destinos usando o efeito de rede para melhorar a produtividade, à medida que mais empresas e indivíduos se tornaram criadores ativos (William & Perez, 2008).

Verificam-se alterações substanciais no processo de compra dos turistas. Os turistas tendem a comprar experiências, e, para minimizar os custos cognitivos, leem os comentários e as opiniões de outros turistas nas redes sociais(Neuhofer, Buhalis, & Ladkin, 2014).

Os *sites* de UGC sobre destinos de viagens, hotéis e serviços turísticos tornaram-se importantes fontes de informação para os viajantes, para além disso pode-se salientar que os processos de tomada de decisões de consumo são fortemente influenciados pelo "boca-a-boca" de outros consumidores (Pan, MacLaurin, & Crotts, 2007).

Com a crescente influência dos sites de UGC na assistência à tomada de decisão por parte dos clientes, os *social media* apresentam oportunidades sem precedentes e desafios para as empresas de turismo e hospitalidade (Ayeh et al., 2012).

Resumindo, os turistas/utilizadores ao usarem os *social media* podem obter benefícios funcionais, tais como (Purifoy, Yoo, & Gretzel, 2007;Hyan Yoo & Gretzel, 2008):

- ➔ As ferramentas de *social media* permitem ao viajante manter-se atualizado em relação a locais e atividades de interesse turístico.
- ➔ A atividade colaborativa na *internet* para organizar viagens pode ajudar os participantes a economizar as despesas e obter mais com os recursos investidos.
- ➔ As ferramentas fornecem benefícios funcionais mútuos aos participantes, uma vez que os utilizadores de *social medial* fornecem e recebem informações.

No entanto os turistas ao usarem os *social media* também podem obter benefícios sociais, como, o envolvimento dos turistas na troca de informações (comentários, ideias, opiniões) (Parra-López et al., 2011).

Está previsto um crescimento constante para o sector dos media, que afeta por sua vez o sector do turismo (Månsson, 2011).

Estudos anteriores demonstraram que as opiniões *online* sobre viagens podem influenciar as decisões dos viajantes (Gretzel & Yoo, 2008; Vermeulen & Seegers, 2009) .

A empresa *Compete Inc*. em 2006 realizou um estudo onde se pode constatar que, um em cada três compradores de viagens que acedem a informações dos *social media*, concordam que a informação os ajudou com a sua decisão de compra.

Noutro estudorealizado pela empresa *comScore* em 2007,pode-se verificar que, 84% dos utilizadores de avaliações de viagens diz que as opiniões de viagens influenciam significativamente a decisão de compra (Ayeh et al., 2012). Também em 2007, o *Country Brand Index* (CBI), mediu a atratividade dos países em diferentes áreas, e pode constatar que a *web* teve maior importância (67%) como canal para recolher informações sobre destinos turísticos (Milano et al., 2011).

A empresa *eMarketer*, em 2008 realizou um estudo que, mostrou que os motores de busca servem como fontes de informações, sendo a escolha número um, para as famílias americanas com o intuito do planeamento de férias (Xiang & Gretzel, 2010).

Em 2009, de acordo com o estudo realizado por*PhoCusWrigth*, nove em cada dez utilizadores dos *social media*, leram e confiaram nas opiniões *online* sobre produtos e serviços turísticos.

Constata-se que com o contínuo crescimento da influência dos utilizadores dos *social media*ecom a sua grande amplitude e profundidade (sendo que uma opinião se difunde em segundos nos *social media*), as opiniões nos *social media* são mais confiáveis do que as fontes oficiais (Vermeulen & Seegers, 2009; Milano et al., 2011).

Noutro estudo realizado por Ye, Law, Gu e Chen(2011) os autores revelam a influência dos comentários gerados pelo consumidor sobre as vendas *online* de quartos de hotel ao nível da empresa. Os resultados sugerem que comentários de utilizadores *online* têm um impacto significativo em reservas de hotéis *online*, e confirmam a importância do "boca-a-boca" *online* para a empresa de turismo (Ye et al., 2011). Outro estudo, conduzido por Dickinger e Mazanec (2008), mostrou que as recomendações de amigos e comentários *online* são os fatores mais importantes que influenciam reservas de hotéis *online*(Dickinger & Mazanec, 2008). Em 2008,foi referenciado por Chatterjee e Wang um estudo,onde sepode constatar que, 46,5% dos turistas pesquisam e selecionam os destinos de viagem e os hotéis via *internet*, 39,7% usam *web* para explorar e aprender sobre o destino de férias, 34,4% pesquisam na *internet* atracões no destino, 33,2%decidem a empresa aérea na *internet* e, por último, 31,8% usam a *internet* para saber sobre a cultura, eventos e respetivo património(Chatterjee & Wang, 2012).

Os *sites* de UGC, como o *TripAdvisor*, tornaram-se tão importantes que, 60% das pessoas que responderam a um estudo disseram que neste tipo de *sites*, leem os comentários *online* antes de comprar um novo produto ou um serviço, e 80% deles são influenciados por este comentário(O'Connor, 2010).

A *Compete Incorporated* (2007) estima que o consumidor é influenciado pelo conteúdo gerado por mais de US $ 10 bilhões por ano em viagens *online* devido ao aumento da credibilidade que este conteúdo tem em comparação com formas mais tradicionais de marketing de viagens. Quase 60% dos consumidores relataram que os *sites* de UGC tiveram um efeito positivo sobre a probabilidade de reserva de produtos e serviços relacionados com viagens (Compete Incorporated, 2007). Mais de 80% dos entrevistados também relataram que preferiram mais as opiniões dos consumidores sobre a descrição de um hotel do que a informação dos próprios hotéis. Uma das explorações mais abrangentes sobre o impacto das opiniões *online* de viagens sobre os consumidores é relatado num estudo realizado por Gretzel (2008), que foi apoiado pelo *TripAdvisor*. Uma pesquisa *online* com cerca de 1.500 utilizadores do *TripAdvisor* (www.tripadvisor.com) foi realizada para verificar como as opiniões *online* sobre viagens têm impacto sobre o comportamento no planeamento de viagens de lazer. A Forrester Research (2006) estimou que 34,7 por cento do total de gastos *online* está relacionado com viagens, e uma pesquisa recente indicou que mais de 74 por cento dos viajantes usam os comentários de outros consumidores como fontes de informação ao planejar viagens de recreio (Gretzel & Yoo, 2008).

Desta forma, torna-se fundamental a análise de todas estas opiniões formadas pelos clientes e dispersas nos *social media.* A análise desta informação não estruturada requer mecanismos de análise que vão além da leitura e interpretação das opiniões individuais. É necessária uma

abordagem semiautomática como por exemplo o TM que possa estruturar esta quantidade de informação cada vez maior para ajudar os decisores a encontrar os padrões positivos e negativos que afetam as suas organizações.

2.2. Text Mining

Nesta secção pretende-se apresentar as técnicas e metodologias presentes na literatura e que serão utilizadas e referenciadas no âmbito desta dissertação.

2.2.1. O que é o Text Mining

Text Mining(TM) é a descoberta de conhecimento a partir de fontes de Bases de dados que contêm texto livre, por outras palavras, é a descoberta e exploração de texto na procura de informação valiosa que esteja escondida (Kroeze, Matthee, & Bothma, 2003). Normalmente refere-se ao processo semiautomático de extração de padrões interessantes ou de conhecimentos interessantes que não são triviais, de documentos não estruturados (Tan, 1999; Sumathy & Chidambaram, 2013). É usado para representar qualquer sistema que analisa grande quantidade de textos em linguagem natural e deteta padrões linguísticos, tentando extrair informação útil (Pande & Khandelwal, 2014).

O TM é composto por duas fases genéricas. A primeira fase, *Text refining,* é a transformação de documentos de texto de forma livre para um formato estruturado ou semiestruturado escolhido, e a segunda fase, *Knowledge Distillation*, deduz padrões ou conhecimento desse formato intermédio. De seguida apresenta-se o esquema das duas fases genéricas do *Text Mining* (Tan, 1999; Sumathy & Chidambaram, 2013):

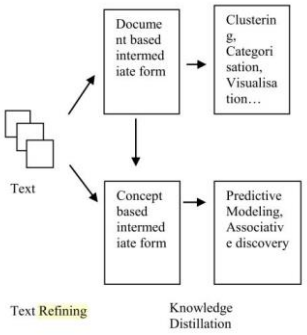


FIGURA 2: ESQUEMA DAS DUAS FASES GENÉRICAS DO TEXT MINING.
(TAN, 1999; SUMATHY & CHIDAMBARAM, 2013).

O TM pode ser visto como uma tarefa muito mais complexa que o *Data Mining*, uma vez que implica lidar com dados de texto não estruturados (Tan, 1999). Contudo, o *Text Mining* é visto como uma variação de *Data Mining* (Pande & Khandelwal, 2014). Por outras palavras, o *Text Mining* constitui um campo interdisciplinar que integra o *Data Mining*, *Web Mining*, recuperação de informação, extração de informação, linguística computacional e processamento de linguagem natural (Sumathy & Chidambaram, 2013).

O principal desafio surge a partir da complexidade da linguagem natural por esta não estar livre de ambiguidade (Sumathy & Chidambaram, 2013). Uma vez que a informação está em forma de texto não estruturado não é facilmente acessível para posteriormente ser usado por computador,

devido a ter sido escrito por humanos e por requerer interpretação de linguagem natural (Dijrre, Gerstl, & Seiffert, 1999).

Nesta área, TM, a maioria dos trabalhos que foram desenvolvidos em termos de descoberta de conhecimento centra-se em base de dados estruturadas, ignorando assim, os dados não estruturados, uma vez que, são difíceis de processar automaticamente e de representar (R. Feldman et al., 1998; Mostafa, 2013). Não obstante, os dados não estruturados, expressam uma quantidade de informação de grande utilidade para as organizações, tais como saber as opiniões dos clientes, saber como está a empresa no mercado, principalmente para as que lidam diariamente com grandes quantidades textuais, apesar da informação ser difícil de decifrar automaticamente, uma vez que, os textos possuem uma estrutura linguística destinada ao consumo de humanos e não de computadores (Provost & Fawcett, 2013; Sharda et al., 2013). Por estes motivos, a aparição de trabalhos sobre o TM surgiram muito mais tarde face ao *Data Mining*(Hearst, 1999).

O TMutiliza dados não estruturados para obter conhecimento capaz de resolver problemas do mundo real (Godbole, Shantanu; Bhattacharya, Indrajit;Gupta, 2010),
Para as empresas os *blogs* e as redes sociais são recursos extremamente importantes devido ao conhecimento que daí advém (para a gestão de relacionamento com os clientes, controlo de opinião publica, entre outros). Esse conhecimento é extremamente valioso, uma vez que, as opiniões desempenham um papel crucial, influenciando o comportamento da opinião pública, como por exemplo na compra de um determinado produto (Bai, 2011). As empresas podem olhar para os *blogs,* não só para os comentários sobre os seus produtos mas também para a comparação de produtos concorrentes e, até mesmo, sobre o ponto de vista da sua imagem corporativa (Gopal, Marsden, & Vanthienen, 2011).

A aplicabilidade de *Text Mining* tem maior frequência em sectores de (Sumathy & Chidambaram, 2013):

- Telecomunicações, energia e outros serviços;
- Tecnologia de informação e *internet*;
- Banca, seguradoras, mercados financeiros;
- Instituições políticas;
- Indústria farmacêutica e de investigação e cuidados de saúde;
- Bioinformática, *Business Intelligence* e segurança nacional.

Pode-se constatar que o TM pode ser aplicado em várias áreas. Por exemplo, pode ser aplicado a sistemas de gestão de crises naturais com o propósito de apoiar a tomada de decisão de um sistema de alerta de *tsunamis,* sendo este, capaz de monitorizar continuamente a situação e enviar um alerta oficial de aviso prévio ao público, analisando em tempo real os primeiros *tweets* de alarme que são geralmente publicados em poucos segundos após um evento de crise natural.

Neste caso o TMpermite classificar os *tweets* relevantes com o propósito de dar conhecimento da situação (Zielinski & Middleton, 2013)

Também pode ser aplicado à área da saúde. Uma vez que os utilizadores das redes sociais discutem as suas informações pessoais de saúde *online*, estas podem ser utilizadas para detetar potenciais padrões (Sokolova, Jafer, & Schramm, 2012;Oh & Park, 2013).

O TMem conjunto com o SAtambém pode ser utilizado no sector da hotelaria, devido ao grande volume de informação na *internet* sobre os hotéis. Ao aplicar estas tecnologias neste setor, o objetivo será garantir uma melhoria na gestão dos hotéis (uma vez que conseguem perceber o que os hóspedes consideram positivo ou negativo, podendo ajudar no desenvolvendo da competitividade e na formulação de novas estratégias para o negócio)(Lau, 2005; Kasper & Vela, 2011).

Também pode ser aplicado no planeamento de viagens, devido ao facto de na atualidade os turistas procurarem as opiniões e experiências publicadas por outros viajantes em diferentes plataformas *web* ao planear as suas próprias férias. No entanto turistas são sobrecarregadoscom um grande volume de informação na *web*e precisam de ferramentas para os ajudar na sua tomada de decisão. Por exemplo, Maresse-Tayloer et al. (2013) optaram por criar uma ferramenta, *OpinionZoom,* que fornece um conjunto de métodos de análise que pretende ajudar os utilizadores a digerir a informação disponível de forma facilitada.

Segundo a literatura, o TM envolve diversas etapas para a sua implementação. De uma forma geral este constituído pela extração de texto, transformação de texto, processamento de texto e os métodos de *clustering* e classificação.

2.2.2. Etapas de Text Mining

Nesta secção apresentam-se de seguida as etapas envolvidas no *Text Mining* (Sumathy & Chidambaram, 2013), bem como os métodos de cada etapa que foram utilizados nesta dissertação para o objetivo proposto:

Extração de Texto

Um documento, segundo a terminologia de TM*,* é uma peça de texto, que pode ser composta por uma simples frase ou um conjunto de várias páginas de um relatório. Um conjunto de documentos é designado por *corpus* (do latim “corpo”) (Sharda, Delen et. Al., 2013; Provost & Fawcett, 2013).

O*corpus* é normalmente um ficheiro de texto (ASCII), construído utilizando técnicas de reconhecimento de caracteres óticos (*optical character recognition techniques* **–**OCR**–**) para extrair palavras de documentos digitalizados mais antigos armazenados online, ou removendo todas as formatações ou estilos de texto de documentos recentes originalmente armazenados como conteúdo digital.

Transformação de Texto

O documento de texto é representado pelas palavras que contêm (Sumathy & Chidambaram, 2013). De seguida, apresenta-se uma abordagem para a representação do documento bem como algumas transformações:

Bag of words

É a representação de texto mais comum usada no *Text Mining* que representa cada documento como um vetor ponderado de termos, ignorando tanto a gramática como a ordem em que estas aparecem, tendo um peso associado a cada termo que será o número de ocorrências desse termo no documento(Blake, 2011; Sharda et. al., 2013).

Inicialmente as aplicações de TM apenas usavam este modelo para estruturar os dados, tentando classificá-los com base em duas ou mais classes pré-determinadas ou agrupando-as de forma natural (Sharda et. al., 2013). Com a evolução, começam a aparecer alguns estudos que alertam para o facto de este método não ser capaz de produzir conteúdo com qualidade suficiente para sustentar a aplicação das tarefas declassificação, associação, *clustering*, entre outras, uma vez que, os humanos não usam palavras sem uma certa ordem / estrutura, semântica e sintática. A tendência atual do TMestá virada para a inclusão de recursos e de técnicas mais avançadas como o processamento de língua natural (PLN) para colmatar este problema (Sharda et. al., 2013)

Document-by-term matrix (DTM)

A construção de uma DTM é uma das formas mais comuns para estruturar os dados contidos no*corpus*. É formada pelas linhas, que representam os documentos, pelas colunas que representam os termos, e, na sua intersecção, encontra-se o valor da frequência absoluta do termo no documento (Feinerer, Hornik, & Meyer, 2008).A DTM é uma representação *bag-of-words* dos documentos no corpus. Por esse motivo, a dimensionalidade da matriz é extremamente elevada quando se lida com vários documentos de várias fontes. Inúmeros documentos têm uma frequência zero para muitos termos na matriz e isto leva a uma matriz extremamente esparsa. A dispersão é muitas vezes reduzida pelo uso de um fator de ponderação diferente, tal como a frequência inversa do documento (TF-IDF) (Feinerer et al., 2008; Blei & Lafferty, 2009).

Term frequency-inverse Document frequency(TF-IDF)

O TF-IDF é utilizado como fator de ponderação sendo uma medida utilizada com o intuito de eliminar os termos que se repetem bastantes vezes num único documento e que se repetem muito poucas vezes nos outros documentos do *corpora*(Grün & Hornik, 2011).

A fórmula TF-IDF utiliza uma transformação logarítmicapara reduzir a influência de frequências *outlier*, e uma função de penalidade para diminuir a importância dos termos que ocorrem em todos os documentos (Delen & Crossland, 2008). De seguida apresenta-se a fórmula:

$$idf(i,j) = \begin{cases} 0 \; if \; wf_{ij} = 0 \\ \left(1 + \log\left(wf_{ij}\right)\right) \log \frac{N}{df_i} \; if \; wf_{ij} \geq 1 \end{cases}$$

FORMULA TF-IDF (DELEN & CROSSLAND 2008).

N: número total de documentos na coleção	I: termo
J: documento	Dfi: frequência do j em que aparece o i ao longo de toda a coleção
Wfij: frequência do i no j	

Processamento de Texto

Esta etapa de pré-processamento de texto é muito importante porque permite a fase inicial da estruturação dos dados.

Processamento de Linguagem Natural (PLN)

O PLN é um dos processos mais fulcrais do TMna fase de pré-processamento, proporcionando um primeiro nível de estruturação dos dados. Este processo é um subcampo da inteligência artificial e linguística computacional e é composto por um conjunto de técnicas teórico-computacionais que analisam e representam dados textuais(não estruturados), com o intuito de processar linguagem humana para vários tipos de tarefas e aplicações (Sharda et.al., 2013; Liddy et al., 2003).

Por norma, o PLN usa conceitos linguísticos, isto é, partes do discurso (substantivos, verbos, adjetivos, entre outros) e a estrutura gramaticalcom o objetivo de compreender a linguagem humana natural, tornando-os fáceis de serem manipulados pelos programas computacionais(Kao & Poteet, 2007; Sharda et.al., 2013).

A linguagem natural é bastante complexa, obrigandoa que o seu processamento tenha de lidar com muitas adversidades, como por exemplo, a ambiguidade, acabando por tornar o processamento excessivamente difícil, mas com uma grande relevância. Para combater essas dificuldades o processamento de linguagem natural utiliza diversas representações de conhecimento como a que se refere à análise do significado das palavras ou o *part-of-speech*, que no âmbito das análises de TM lida com o tratamento das palavras e as suas características gramaticais(S. Feldman, 1999; Kao & Poteet, 2007).

Tokenisation

Este processosepara todo o texto em palavras, removendo espaços em branco e virgulas (Sumathy & Chidambaram, 2013).Esta técnica é elaborada através da divisão de um conjunto de caracteres em tokens. A *Tokenisation* divide a sequência de caracteres descobrindo as

fronteiras de divisão das palavras. Nem todas as línguas produzem texto em forma de palavras puramente delimitadas por espaços. Na maioria das línguas Ocidentais o delimitador é o espaço, no entanto, em algumas línguas (Japonês, Chinês,Tailandês) não usam o espaço como delimitador. Estes casos requerem em primeiro lugar a aplicação de um processo de segmentaçãoque precisa de ser aplicado a uma *stream* de discurso contínuo de fala a fim de identificar as palavras que compõem uma expressão, dicção, elocução. Daí, poder-se dizer que a *tokenization* pode ser dividida em duas abordagens, linguagens em que o espaço é o delimitador e para linguagens em que o espaço não é o delimitador de *tokens* (Indurkhya & Damerau, 2010).

Part- of- speech

É considerado um processo muito complexo e difícil,uma vez que, tenta categorizar todos os termos com uma característica gramatical tais como, nomes, verbos, adjetivos, advérbios, pronomes, entre outros, sendo que, o *part-of-speech* não depende apenas da definição do termo, mas também depende do contexto em que está a ser usado (Indurkhya & Damerau, 2010;Sharda, Delen et. Al., 2013; Provost & Fawcett, 2013).

Este processo reduz o espaço de busca/procura para as palavras lexicalmente ambíguas e para a forma das palavras. Utiliza-se *Part-of-speech* devido à existência de muitas palavras com a mesma morfologia mas com significados diferentes ou com pronúncias distintas que podem dar azo a interpretações erradas (Gaowa, 2010).

Remoção Stopwords

Termos que ocorrem com frequência no texto, mas têm pouco poder de discriminação são chamados de *stopwords* (Blake, 2011). Aremoção de *Stopwords*é a eliminação das palavras, com a ajuda do dicionário de *stop words*, que não aportam conteúdo significativo para o documento ou não tem informação útil, ou seja, consiste na eliminação de tudo que sejam verbos auxiliares,determinantes, artigos, pronomes, preposições, interjeições e mais termos comuns e irrelevantes (Solka, 2008;Liu, 2008; Blake, 2011; Pande, 2011;Choy, 2012).

Existem váriosalgoritmos de remoção de *stopwords*. Ocálculo*standard*mais básico e direto ébaseado na frequência de palavras.As palavrascom maior frequência que são normalmente relativas a valores de alto ruido (*high noise*), são removidas visto servirem só para unir palavras numa frase e não influírem no seu sentido(Gaowa, 2010). A alta frequência de ocorrência dessas palavras no *Text Mining* apresenta obstáculos à compreensão do conteúdo dos documentos (Choy, 2012).

Stemming

É um processo de remoção de sufixos e prefixos, deixando a raiz da palavra, isto é, permite que palavras semelhantes sejam reduzidas aos seus radicais, com o intuito de não serem identificados como sendo palavras diferentes (Porter, 1980; Liu, 2008). Esteprocesso é muito utilizadono*TM*devido à redução da complexidade sem nenhuma perda grave de informação (Meyer, Hornik, & Feinerer, 2013). Para uma maior compreensão deste processo, pode-se

observar, por exemplo, as palavras *singer* e *singing*que ao serem alvo de transformação, são reduzidos ao seu radical *sing*(Porter, 1980; Liu, 2008). Tanto o *Stop Word Removal* como o *Stemming* reduzem o tamanho do léxico, poupando os recursos computacionais (Solka, 2008).

Métodos de *Clustering* de Texto

Esta etapa explica a utilização de métodos de *Data Mining* (*Clustering* e informação de classificação) aplicados ao texto não estruturado (Sumathy & Chidambaram, 2013).

Clustering

É um processo não supervisionado, que divide um determinado conjunto de documentos em grupos de documentos de conteúdos semelhantes (Dijrre et al., 1999) ou um conjunto de termos em grupos de termos utilizados de forma correlacionada. Estemétodo é usado para agrupar documentos semelhantes (Pande & Khandelwal, 2014).

O intuito deste algoritmo é agrupar documentos ou termos de tal forma que cada grupo de documentos tenha um elevado grau de similaridade intra-classe e reduzido grau de similaridade entre classes (Blake, 2011).

Este algoritmo, quando começou a ser aplicado nesta área, TM, era implementado com base nos algoritmos de *clustering* tradicionais (Guerreiro, Rita, & Trigueiros, 2015). Posteriormente, na literatura surgiram modelos que permitem uma análise de pertença *mixed*, ou seja, um termo pode estar em vários *clusters*, ao contrário da análise de pertença individual em que um termo pode estar apenas num único *cluster* (Blei & Lafferty, 2007). O *Topic Models* é um destes modelos, *mixed*, (Guerreiro et al., 2015), sendo um modelo probabilístico capaz de descobrir a estrutura semântica subjacente da coleção de documentos com base na análise Bayesiana hierárquica de textos (Blei & Lafferty, 2007; Blei & Lafferty, 2009). Tem sido aplicado em vários tipos de documentos, desde *e-mails*, trabalhos científicos até arquivos de jornais (Griffths & Steyvers, 2004; Wei & Croft, 2006). O modelo de*Topic Models*, permite que os termos pertençam a vários tópicos ao mesmo tempo (Grün & Hornik, 2011). Cada tópico representa uma distribuição de termos, em que, cada termo tem uma probabilidade diferente de pertencer a esse tópico(Blei & Lafferty, 2009).

Um dos algoritmos mais utilizados de *cluster* de *topic model* é o *Correlated Topic Model* (CTM) (Blei & Lafferty, 2007). Este último é baseado no *Latent Dirichlet Allocation* (LDA), que assume que os tópicos são independentes(Blei & Lafferty, 2009).No LDA os pressupostos baseiam-se num processo generativo que acredita-se ter sido utilizado para produzir o documento(Blei, Ng, & Jordan, 2003).Este algoritmo é baseado na indexação semântica latente e indexação semântica latente probabilística (Deerwester, Dumais, & Harshman, 1990; Hofmann, 1999) e assume que os documentos são escritos da seguinte forma:

Para cada tópico:

(a) desenha a distribuição de palavras $\vec{\beta}_k \sim \mathrm{Dir}_V(\eta)$

Para cada documento:

(a) desenha o vetor das proporções do tópico $\vec{\theta}_d \sim \mathrm{Dir}(\vec{\alpha})$

(b) para cada palavra:

(i) Define atribuição da palavra a um tópico $Z_{d,n} \sim \mathrm{Mult}(\dot{\theta}_d), Z_{d,n} \in \{1, \ldots, K\}$

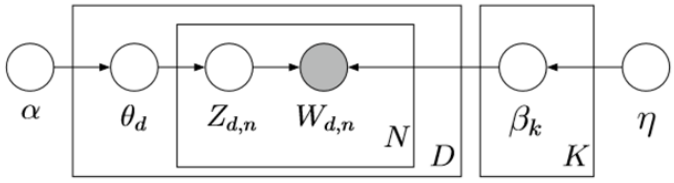


FIGURA 3: REPRESENTAÇÃO DO LDA. NÓS SOMBREADOS DENOTAM VARIÁVEIS ALEATÓRIAS E OS RESTANTES, VARIÁVEIS ALEATÓRIAS OCULTAS. LIMITES REPRESENTAM DEPENDÊNCIA ENTRE VARIÁVEIS ALEATÓRIAS; RETÂNGULOS DENOTAM REPLICAÇÃO (BLEI & LAFFERTY, 2009)**.**

De seguida apresenta-se a figura que mostra um exemplo que utiliza o artigo da revista Strahilevitz & Mylers (1998), que apresenta o processo generativo:

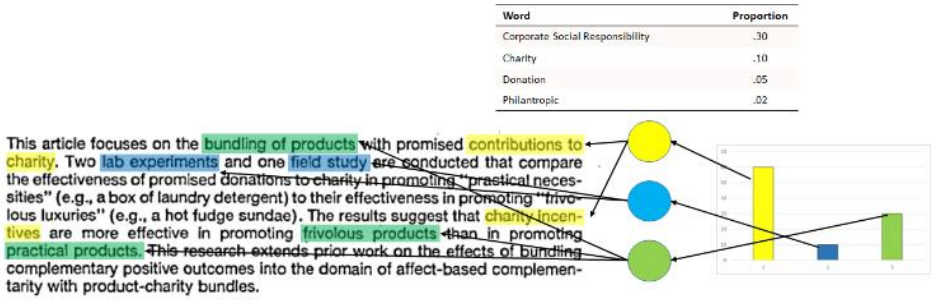


FIGURA 4: PROCESSO GENERATIVO.

Como se pode observar na figura 4 o histograma da direita que representa a distribuição dos tópicos sobre os documentos na tabela representa as proporções das palavras.

Embora este processo seja apenas um gerador de *bag-of-words* sem significado, é útil para o objetivo do algoritmo LDA, que serve para gerar um processo estocástico que representa o modelo escondido, e, posteriormente inverte-o usando probabilidades posteriores. Este algoritmo usa a distribuição de Dirichlet para a conceção do processo generativo devido a ter as propriedades de garantir que a distribuição de uma distribuição Dirichlet ainda é uma distribuição com as mesmas características, o que é útil para os cálculos (Blei & Lafferty, 2009).

Esta distribuição, Dirichlet, assume que os seus pontos dos vetores são quase independentes, o que significa que quando os tópicos são modelados usando essa distribuição, são assumidos como sendo independentes(Blei & Lafferty, 2009). No entanto, num documento *real-word*, os

temas geralmente são correlacionados.Já no CTM a distribuição usada para modelar as proporções de tópicos é uma distribuição normal logística, devido a esta incorporar a covariância entre tópicos (Atchison & Shen, 1980).

Sendo um algoritmo de procura de agrupamento/*clusters* hierárquico, o CTM não fornece o número exato de grupos que devem ser considerados. Em vez disso, o algoritmo fornece medições do modelo de variabilidade explicada como a média logarítmica da verossimilhança do modelo e sua perplexidade (a força do modelo para prever novas palavras após ajustamento de um modelo), para cada número possível de *clusters.* O número de *clusters*, ou seja, de tópicos, que melhor se adequam aos dados são, avaliados pela forma de como essas métricas mudam quando o número de clusters aumenta. O número ideal de *clusters* é atingido quando a variabilidade explicada não muda significativamente pela adição de mais clusters (Guerreiro et al., 2015).

O CTM é um dos algoritmos de *clustering*de *Topic Models*(Blei & Lafferty, 2007) e permite que sejam incorporadas correlações entre os vários tópicos o que faz dele um algoritmo mais adequado para análises de texto(Blei & Lafferty, 2007).Uma das mais relevantes vantagens é a eficiência em termos de *perplexity*, uma medida habitualmente aplicada com o intuito de determinar o quão bem o modelo prevê as restantes palavras que irão aparecer num determinado tópico depois de observar uma pequena parte dela (Guerreiro et al., 2015). Ao compararem os dois algoritmos (LDA e CTM), os autores do CTM descobriram que o CTM reduz a *perplexity* sobre o LDA pelo menos em 10% (Blei & Lafferty, 2007).

Este algoritmo utiliza-se com o intuito de conseguir uma sumarização/compactação eficiente e classificação dos *reviews* tendo em conta os seus temas(Guerreiro et al., 2015). Estes algoritmos de *clustering* tradicionais costumam ser usados para servir projetos de TM*,* como no caso deLu, Liu, & Yu, 2012*,* em que se pretendedescobrir temas relevantes de Ética sobre nanotecnologia que estão latentes em títulos, resumos e palavras-chave de pesquisa publicada.

Para conseguir chegar aos objetivos propostos para este projeto os modelos de CTM foram usados para encontrar *clusters/tópicos* entre a opinião dos consumidores.

Métodos de Classificação de Texto

Estes algoritmos têm como objetivo criar modelos (de classificação) que mapeiam com precisão cada documento a uma classe existente. Assim, o objetivo de um sistema de classificação em *Text Mining* é a criação de um classificador que quando fornecido com um conjunto de recursos de um novo documento é capaz de predizer a classe do novo documento. Estes algoritmos por norma são supervisionados, uma vez que os investigadores fornecem o algoritmo com exemplos de treino que inclui a classe correta (categoria) e os recursos usados para representar cada um dos documentos (tais como *vector-based representation*) (Blake, 2011). É estimado um modelo que é depois avaliado recorrendo a um conjunto de teste ou validação.

2.2.3. Sentiment Analysis

Sentiment Analysis (SA) pode servir para retirar as opiniões a partir da *internet* e avaliar as preferências dos clientes *online* bem como para o reconhecimento de emoções em texto(Bai, 2011; Cambria, Schuller, Xia, & Havasi, 2013).Segundo Mostafa (2013), SA é uma técnica de conhecimento automático que pretende encontrar padrões de sentimento escondidos nos inúmeros comentários, *blogs* ou em redes sociais. Segundo o autor, para se poder calcular o sentimento do texto, é necessário compará-lo com um léxico/dicionário que determine a força do sentimento. Também se pode dizer que SA é um processo com o objetivo de determinar se a polaridade de sentimento de um corpo textual tende para positivo, negativo ou neutro (Paltoglou & Thelwall, 2012; Kontopoulos, Berberidis, Dergiades, & Bassiliades, 2013).

Tem havido um súbito interesse crescente na área do PLN, mais concretamente no *Sentiment Analysis* nos últimos dez anos. As investigações variam entre classificações de níveis de documentos a estudos sobre a polaridade das palavras e frases (Pang & Lee, 2008; Liu, 2012; Sharda et al., 2013).

O *SA* em comentários, feedback ou críticas, no que diz respeito ao ambiente empresarial, particularmente na área de marketing e *Customer Relationship Management*(CRM)fornece indicadores úteis para muitas finalidades, podendo proporcionar a avaliação da aceitação de produtos ou a determinação de estratégia para melhorar a qualidade dos produtos. Permite também compreender as decisões políticas e analisar sentimentos sobre serviços públicos ou questões politicas (Prabowo & Thelwall, 2009; Liu, 2012). O sentimento detetado no texto pode ser assinalado como explícito e implícito. No caso do primeiro, o texto expressa diretamente uma opinião positiva ou negativa, enquanto que, no último, onde o texto implica um parecer positivo ou negativo (Liu, 2008).

Observa-se na literatura, muitos trabalhos desenvolvidos apontados para avaliar a polaridade do sentimento ao nível do documento (Pang & Lee, 2008).O *SA* de texto não estruturado (TM) tem aumentado nos últimos anos(Bai, 2011). As emoções (*SA*) podem afetar profundamente o comportamento individual e a tomada de decisão (Bollen et al., 2010). Pode servir para retirar opiniões a partir da *internet* e avaliar as preferências dos clientes *online*, sendo útil paraas áreas de marketing e economia, para promover uma vantagem estratégica para a empresa, ou para a deteção de risco e ameaças de segurança (Bai, 2011). Por exemplo, Bollen et. al 2010, aplicaram SA para prever o mercado de ações tendo em conta os estados de humor expressos num conjunto de *posts* do *twitter* diários (Bollen, Mao, & Zeng, 2010). Também pode ser aplicado na área de telecomunicações, com o intuito de as operadoras telefónicas perceberem o que os clientes mais valorizam nos serviços oferecidos pelas operadoras, para perceber o que leva os clientes a escolher uma operadora em detrimento de outra. Por outras palavras, a análise neste caso assenta na obtenção das características dos serviços mais valorizadas pelos clientes das operadoras (Setiawan, 2014).

Por último, no que diz respeito a área de turismo com a aplicabilidade destas técnicas, não foram encontrados muitos estudos suportados pela literatura. Em 2007 surgiram as primeiras publicações nesta área, onde três artigos abordam a influência inesperada dos *social media* nas empresas e na indústria do turismo, visto que, tanto as empresas como a indústria estavam a perder o controlo sobre o que estava a ser escrito sobre elas nas redes sociais (Dwivedi et al., 2007; Thevenot, 2007; Zeng & Gerritsen, 2014). Os autores, concluíram que, o crescimento e o impacto dos *social media* no turismo e na hospitalidade não deveria ser ignorada. Para além disso, os *social media* precisam de ser constantemente monitorizados pelo turismo com respostas e interações em tempo real (Grant-Braham, 2007).

Para a aplicabilidade de SA nas diversas áreas de negócio é necessário recorrer a ferramentas onde é predominante a utilização das técnicas baseadas em dicionários e de aprendizagem automática, ou seja, utilizam dados encontrados em recursos lexicográficos para atribuir sentimentos para um grande número de palavras determinando a polaridade de cada documento em positivo, neutro e negativo(Lawrence, 2014).Após a utilização de TMe SAobtém-se o conjunto de informação estruturada com base na informação não estruturada, no entanto, mantém-se a dificuldade emexplorar esta informação de forma a auxiliar melhor a tomada de decisão, uma vez que a quantidade de informação gerada é muito elevada. Para resolver essa lacuna optou-se por criar um *Decision Support System* (DSS).

2.3. Decision Support System (DSS)

A gestão eficiente da informação é cada vez mais relevante porque a dinâmica dos mercados em crescimento exigem que haja uma maior capacidade de armazenamento e de informação com qualidade(Wober & Gretzel, 2000).Tem sido geralmente aceite que as decisões das empresas deve ser baseada em informações, pois reduz a incerteza prevalecente nos processos de decisão e, portanto, torna-se um fator crítico de sucesso para empresas que estão a enfrentar crescente concorrência e mercados cada vez mais dinâmicos (Murdick & Munson, 1986; Glazer, 1991).

Um DSS é um sistema informatizado que contém o conhecimento de domínio específico e modelos de decisão analítica para auxiliar o tomador de decisão, apresentando informações que permitam a interpretação de várias alternativas (Wang, 1997).

Inevitavelmente quando se fala de DSS fala-se de dados, informação e conhecimento. Pode-se dizer que os dados são uma sequência de símbolos quantificados ou quantificáveis, ou seja, um texto são dados, uma vez que as letras são símbolos quantificados, visto que, o alfabeto por si só constitui uma base numérica. Além de textos, os dados também podem ser animações, imagens e som uma vez que podem ser quantificados. A informação é uma abstração informal, por outras palavras, a informação é algo significativo para alguém através de textos, imagens, sons ou animações (ou seja, não pode ser formalizada através de uma teoria lógica ou matemática). Por ultimo, o conhecimento é uma abstração interior, pessoal, de alguma coisa que foi experimentada por alguém. Por outras palavras, o conhecimento não pode ser descrito inteiramente porque se não seria apenas dados (se fosse descrito formalmente e não tivesse significado) ou informação (se descrito informalmente e tivesse significado) (Setzer, 2015).

Segundo Power, existem cinco tipos de DSS, DSS orientados a dados (*Data-driven* DSS), DSS orientado a modelos (*Model-driven* DSS), DSS orientado à comunicação (*Communication-driven* DSS), DSS orientado a documentos, (*Document-driven* DSS) e DSS orientado ao conhecimento (*Knowledge-driven* DSS) (Power, 2007). De seguida apresenta-se uma tabela com uma breve descrição dos tipos de DSS (Power, 2002; Power, 2007):

Data-driven DSS	Orientado aos dados.
	Este DSS normalmente envolve a integração, armazenamento e analise de grandes quantidades de dados usando ferramentas de *Data Warehousing* e OLAP (*online analytical processing*).
	Exemplos: sistemas de informação executivos, sistemas de suporte à decisão espacial.
Model-driven DSS	Orientado a modelos.
	Este DSS tem como base de funcionamento aplicações suportadas por modelos de dados.
	Exemplos: são utilizados por gestores, membros de uma organização para diferentes propósitos tendo em conta a área que se pretende dar respostas, finanças, comercial entre outras.
Communication-driven DSS	Orientado à comunicação.
	Este DSS é utilizado com o intuito de ajudar a conduzir uma reunião, mantendo a coordenação e a colaboração entre utilizadores. Por norma são implementados em ambientes *web* ou cliente-servidor. É direcionado para equipas internas e parceiros de organização.
	Exemplos: são utilizados para a comunicação entre membros de uma equipa, usando sistemas de mensagens, vídeo conferencia e whiteboard (ambiente de arquivos compartilhados) permitindo comunicação em tempo real.
Document-driven DSS	Orientado a documentos.
	O objetivo deste DSS centra-se na recolha, classificação e gestão de documentos não estruturados. A utilização de técnicas de *Text Mining* permite a classificação de documentos, filtragem de informação e é também utilizado na *web* ou para a deteção de *spam* de correio eletrónico.
Knowledge-Driven DSS	Orientado ao conhecimento.
	O propósito deste DSS é a resolução de problemas dentro do domínio específico e apoiam-se em tecnologias como sistemas de inteligência artificial ou sistemas periciais.

TABELA 1: TIPOS DE DSS (POWER, 2002; POWER, 2007).

De seguida apresenta-se a figura 5 com o intuito de demonstrar a evolução dos DSS.

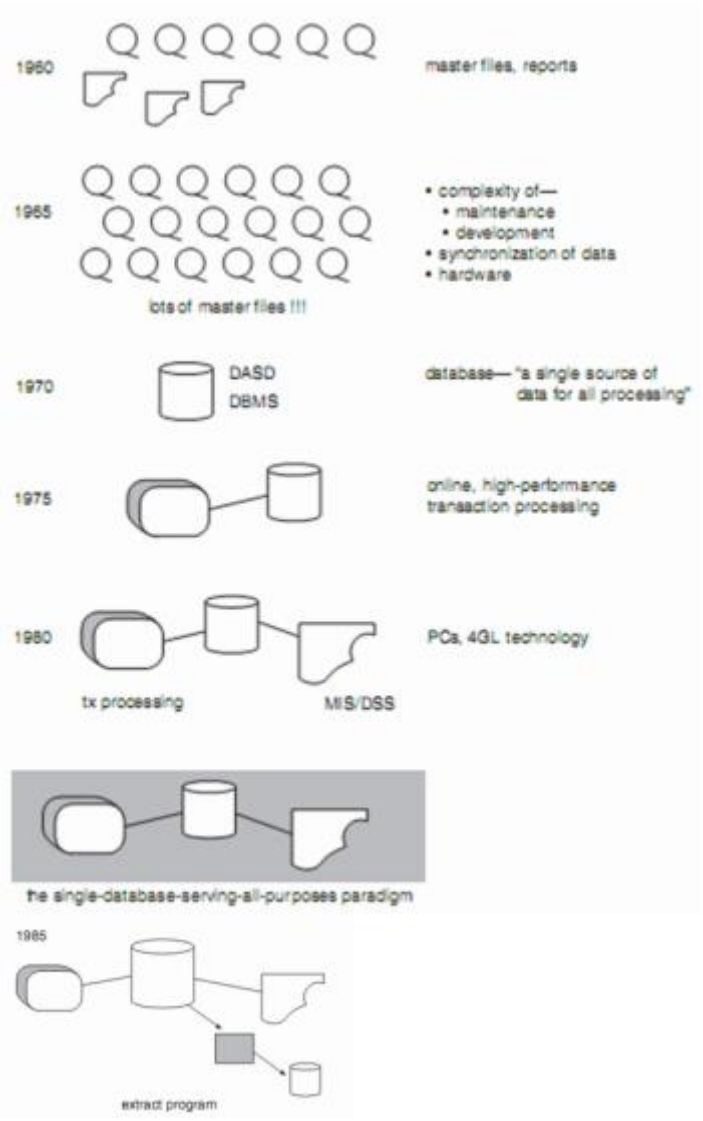


FIGURA 5: EVOLUÇÃO DOS DSS (INMON, 2002).

Observando a figura 5, pode-se dizer que esta evolução está sempre intrinsecamente relacionada e ao mesmo tempo é dependente da tecnologia.Até 1985 para além dos ficheiros tradicionais, o paradigma era uma base de dados que servia para todos os fins. As bases de dados só apareceram após o aparecimento dos discos magnéticos por volta dos anos 70. Também por volta dos anos 70 houve o surgimento dos PC que hoje em dia são vulgares. No ano1985 começaram a surgir os primeiros programas para extrair informação necessária através de alguns parâmetros. Acabaram por surgir novos conceitos como OLAP(*Online Analytical Processing*), ROLAP(*Relational* OLAP) e MOLAP (*Multidimensional* OLAP) e relacionados com a forma de estruturar os dados(Inmon, 2002).Só no meio de da década dos anos 90 é que se começou a perceber que as necessidades entre sistemas operacionais e os DSS em termos de base de dados eram diferentes, alterando assim o paradigma que até então vingava – uma base de dados para todos os fins–, ou seja, só nessa altura é que os fornecedores dos sistemas de base de dados reconheceram que havia diferenças entre sistemas de suporte à decisão dos sistemas tradicionais ou operacionais (Powell, 2001).

As investigações sobre DSS começaram na década de 1960,e destinaram-se a melhorar a tomada de decisão individual, fornecendo acesso mais fácil ao reconhecimento do problema, à estrutura de problemas, gestão de informações, ferramentas estatísticas, e a aplicação do conhecimento (Klein, 1990; Santana, 1995).

O conceito de DSS emergiu como um ambiente projetado para ajudar os decisores a responder a perguntas específicas, facilitando o uso de modelos e bancos de dados de uma forma interativa. Nos últimos anos, muitos exemplos podem ser encontrados na literatura referente à utilização de DSS em recursos de água, como pode ser visto no estudo de Labadie et al. (1989) ou Loucks & da Costa (1991), entre outros (Andreu, Capilla, & Sanchís, 1996).

Em quase todas as indústrias, os DSS estão a ser desenvolvidos a fim de apoiar o planeamento do investimento e comercialização (Wierenga & Van Bruggen, 2000).

As tecnologias da *web* estão a ser utilizadas para o desenvolvimento de ferramentas DSS pelos principais desenvolvedores de tecnologias de apoio à decisão, tais como a SAS Inc. (Cohen, Kelly, & Medaglia, 2001). A Oracle incentiva os seus clientes a portar os seus aplicativos DSS, como *Data Mining*, sistemas de processamento analítico (OLAP) gestão de relacionamento com clientes (CRM) e *online*, para um ambiente baseado na *web* usando o seu servidor de aplicativos da Oracle 9Ai (Oracle, 2001).

Os DSS baseados na *web* estão a ser empregues pelas organizações como auxiliares de decisão para os empregados/funcionários, bem como para os clientes. Uma aplicabilidade comum de DSS baseado na *web* tem sido com o intuito de ajudar os clientes a configurar produtos e serviços de acordo com as suas necessidades. Estes sistemas permitem que os clientes individuais para desenvolver seus próprios produtos, escolhendo entre um menu de atributos, componentes, preços e opções de entrega. Por exemplo, em *web-sites* da maioria dos fabricantes de computadores de mesa (www.dell.com, www.compaq.com e www.ibm.com), os indivíduos podem começar com uma configuração básica definida por um modelo de processador e velocidade, e depois ir para especificar a configuração completa com sua escolha do tamanho do disco rígido, memória e *add-ons*, como CD-ROMs, multimédia, monitores e impressoras. Pode-se encontrar casos de aplicabilidade de DSS baseados na *web* em diferentes indústrias, como, na indústria de vestuário (www.landsend.com, www.blair.com, www.weddingchannel.com), que permite que um utilizador de um modelo virtual possa projetar um vestido antes de pedir, na indústria financeira (www.calvertgroup.com), que permite ao utilizador experimentar vários planos de poupança-reforma, bem como na indústria de brinquedos (www.vermontteddybear.com), onde as crianças podem desenhar os ursos de peluche que eles desejam com diferentes cores, tamanhos e tipo de casaco (Bash, 2015) .

Um sistema de apoio à decisão de marketing (MDSS) pode ser de particular importância, uma vez que apoia as organizações na recolha, armazenamento, processamento e divulgação de informações, e no processo de tomada de decisão, fornecendo previsões e modelos de decisão acerca dos clientes de uma organização (Wöber, 2003).

Na indústria do turismo não há falta de dados de pesquisa de mercado, pelo contrário, existe um crescimento bastante descontrolado de várias fontes de dados, cada um com diferentes fins. Existem inquéritos sobre turismo de institutos nacionais e internacionais que são publicados em intervalos cada vez mais curtos eo nível de discriminação dos dados de mercado aumenta rapidamente (Wöber, 2003).

Nesta indústria do turismo tem-se desenvolvido DSS, e, as aplicações mais importantes são:

- Sistemas de apoio às decisões de marketing em organizações nacionais de turismo (Mazanec, 1986; Rita, 1993).
- Sistemas de aconselhamento de viagens para funcionários de transportes (Hruschka & Mazanec, 1990).
- Sistemas de apoio e planeamento regional sobre a seleção ideal de locais, para investir (Calantone & Benedetto, 1991; Walker, Greiner, McDonald, & Lyne, 1998).
- Sistemas que fornecem análises de portfolio turístico (Mazanec, 1994; Wöber, 1998).
- Ferramentas de simulação para o comportamento de viajantes e previsão de viagens em determinadas regiões (Middelkoop, 2001).

Para o desenvolvimento da investigação, optou-se por implementar um *Data-Driven* DSS devido a ser o que mais se adequa aoobjetivo proposto.

3. METODOLOGIA DE DESENVOLVIMENTO

Para todo o processo de desenvolvimento optou-se por utilizar a metodologia *Cross Industry Standard Process for Data Mining*(CRISP-DM), uma vez que, é considerada uma metodologia *standard* aplicada à extração de conhecimento dos dados (Sharda et al., 2013). A metodologia é composta por 6 fases, sendo um processo interativo e iterativo, não sequencial, dado que as fases podem ser executadas mais do que uma vez, dependendo dos resultados obtidos numa determinada fase, pode ser necessário retroceder para uma fase anterior (Chapman et al., 2000).

De seguida apresenta-se um esquema da metodologia seguida nesta dissertação:

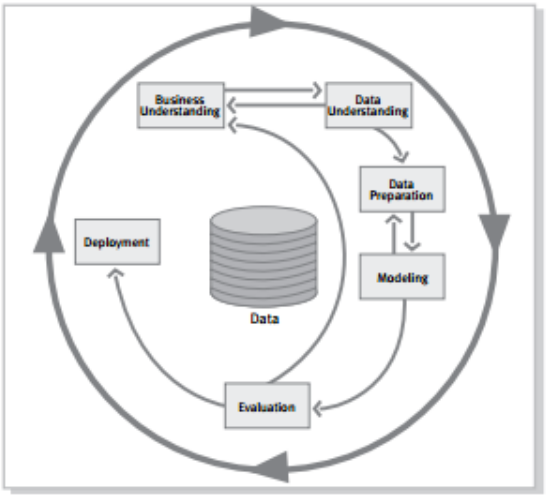


FIGURA 6: METODOLOGIA CRISP-DM. (CHAPMAN ET AL., 2000).

Optou-se por fazer algumas adaptações nas fases para o desenvolvimento deste projeto, uma vez que, neste caso faz mais sentido executar em conjunto as fases *Data Understanding* e a *Data Preparation* por um lado e o *Modeling*e o *Evaluation*por outro. De seguida apresentam-se as etapas:

3.1. Business Understanding

Esta etapa serviu para salientar a importância deste trabalho, enquadrando o ambiente da informação utilizada no contexto da amostra.

3.1.1. Compreenção do Negócio

O *Yelp* é um *site*/aplicação presente emwww.yelp.com, que funciona como um guia urbano eletrónico de várias cidades, para ajudar os utilizadores a encontrarem locais de lazer, tais como restaurantes, locais para fazer compras, relaxar, jogar, cinemas, teatros, museus, arte em geral, diversões noturnas, etc., baseado nas opiniões dos clientes que cooperam com esta plataforma/*site*.

Para os efeitos da presente dissertação foi recolhido um *dataset*disponibilizado pelo *Yelp*que contém informações sobre os *utilizadores*, os *reviews* e os *negócios* aí registados.

Achou-se que seria interessante, no âmbito desta dissertação, analisar as opiniões dos utilizadores (obtidas através da partilha de experiências dos utilizadores) utilizando TM e SA, para criar um DSS com o intuito de poder analisar os resultados obtidos com o TM e SM, podendo assim, ajudar no desenvolvimento do turismo nas cidades, identificando o que os turistas mais e menos valorizam, dando assim, um contributo importante para o crescimento económico e para o desenvolvimento da cidade.

3.2. Data Understanding + Data Preparation

Nesta etapa será apresentada a descrição dos dados, toda a parte exploratória dos dados a estruturar e todo o processo a percorrer até ficarem estruturados.

3.2.1. Descrição dos Dados

O*dataset* recolhido refere-se a uma amostra de dados de 16 estados dos Estados Unidos que contemplam 121 cidades, contendo 13490 negócios classificados em 22 categorias de negócio (***Active Life, Arts&Entertainment, Automotive, Beauty&Spas, Education, Event Planning&Service, Financional Services, Food, Health&Medical, Home Services, Hotel&Travel, Local Flavor, Local Services, Mass Media, Night Life, Pets, Professional Services, Public Services&Government, Real Estate, Religious Organizations, Restaurant, Shopping***) e 330071 *reviews* sobre as experiências dos clientes em cada uma destas cidades. Os dados obtidos através do *site* acima indicado foram recolhidos no formato JSON o que obrigou à sua conversão para CSV e separação em três ficheiros de forma a criar as tabelas da figura 7 (***Review, User e Business***), recorrendo ao atributo *type* e à linguagem *python*.

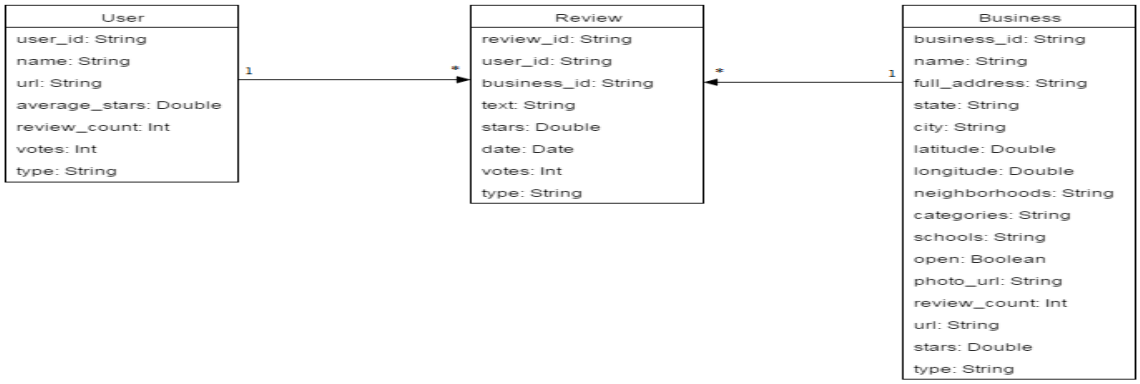


FIGURA 7: DIAGRAMA DE CLASSES DO *DATASET*.

Ao observar o diagrama de classes, que contém os atributos de cada classe bem como o tipo dos mesmos, constata-se que a opinião dos utilizadores está contida na tabela de ***Review***, tornando-a essencial para o objetivo desta dissertação. No entanto, as categorias de negócio, o estado, a cidade, as estrelas do negócio (stars: a classificação entre 1-5 atribuídaa cada experiência), longitude e latitude que serão essenciais futuramente para o sistema de apoio à

decisão, não se encontram desnormalizadas na mesma tabela que as opiniões dos utilizadores, como se pode observar no diagrama da Fig. 6. Estes atributos surgem somente na tabela/classe ***Business***, o que obrigou a recorrer ao atributo ***business_id*** para estabelecer a relação entre as tabelas/classes, e assim, relacionar as opiniões com os atributos necessários (***categories***, ***state***, ***city***, ***latitude*** e ***longitude***).

Uma vez separada a informação em três ficheiros, identificou-se os atributos necessários(dos dois ficheiros *Business* e *Review*) para a criação do *dataset* para a construção futura do sistema de apoio à decisão, como o ***review_id***, ***text***, ***stars***, ***date*** (do ficheiro***Review***), o ***business_id***, ***categories***, ***state***, ***city***, ***latitude***e***longitude*** (do ficheiro ***Business***). Cruzando a tabela ***Review*** com a tabela ***Business*** através da chave ***business_id***, criámos um único ficheiro no formato CSV com os atributos estritamente necessários para o objetivo desta dissertação.

De forma resumida, apresenta-se um diagrama do acima exposto:

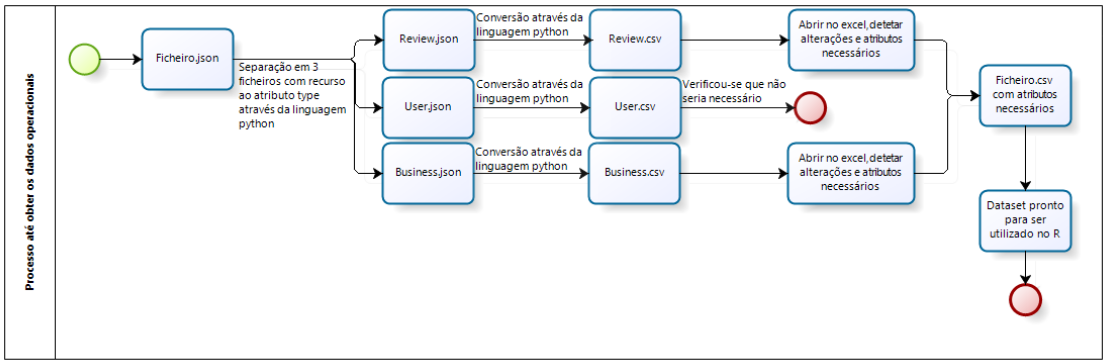


FIGURA 8: DIAGRAMA DO PROCESSO DE OBTENÇÃO DO DATASET PARA UTILIZAÇÃO.

Ao começar a explorar o *dataset* no programa *R* que foi utilizado para a análise e tratamento dos dados, deparámo-nos com algumas limitações devido à morosidade do processamento de dados, do equipamento utilizado, e à incapacidade de tratamento de grandes volumes de dados, nomeadamente na utilização de uma ferramentapara a análise de sentimentos nos textos, chamada *Semantria*, que fornece um *plug-in* para *excel* e que permite efetuar a análise até 15000 registos gratuitamente, obrigando a utilizar parcialmente o *dataset*.

Para resolver estas limitações, optou-se por fazer uma amostra aleatória de 14000 *reviews* obtendo assim, uma margem de segurança.

Após ter sido efetuada a amostra aleatória, ao observar os dados, verificou-se que existiam campos no *dataset* em branco, o que levou a retirar esses *reviews*, levando a uma redução da amostra para 12371 *reviews*.

A amostra final ficou composta por 12371 *reviews*, 4615 negócios, 51 cidades e 22 categorias.

3.2.2. Análise Exploratória

Para iniciar a estruturação dos dados não estruturados foi primeiro efetuada uma análise exploratória à amostra.Esta exploração, resumiu-se à contabilização do número total de

palavras, ao cálculo da média de avaliação de *reviews* por categoria (esta classificação é atribuída pelos utilizadores do *Yelp* a cada negócio e varia entre 1 e 5), a média de palavras por categoria de negócio, nº total de palavras e, por último média de palavras por *review*.

A tabela 2 apresenta os resultados obtidos:

Categoria de Negócio	Nº de Reviews	Media de Av. Reviews Stars	Nº Total de Palavras	Média de Palavras por Review
Active Life	164	4,0	23574	143,7
Arts & Entertainment	351	4,0	39614	112,9
Automotive	69	3,7	8986	130,2
Beauty & Spas	419	3,8	58240	139,0
Education	110	**4,2**	13669	124,3
Event Planning & Service	47	4,1	5631	119,8
Financial Services	16	**3,0**	**1826**	114,1
Food	**2430**	3,7	**268981**	110,7
Health & Medical	198	3,8	31810	160,7
Home Services	66	**3,4**	13767	**208,6**
Hotels & Travel	184	**3,4**	13727	**74,6**
Local Flavor	**8**	**4,3**	2170	**271,3**
Local Services	130	3,5	15606	120,0
Mass Media	**9**	**4,2**	**1026**	114,0
Night Life	**1215**	**3,5**	**157816**	129,9
Pets	31	3,8	4173	134,6
Professional Services	23	3,6	3211	139,6
Public Services & Government	25	3,7	3158	126,3
Real Estate	**14**	**2,5**	3508	**250,6**
Religious Organization	**4**	**4,4**	**162**	**40,5**
Restaurants	**6105**	3,6	**744617**	122,0
Shopping	753	3,8	84953	112,8
Total	12371	3,7	1500225	136,4

TABELA 2: REPRESENTAÇÃO DO *DATASET* E NÚMERO TOTAL DE PALAVRAS POR CATEGORIAS.

Na coluna 1 da tabela 2 pode observar-se como estão organizadas as categorias de negócio, em termos de quantidade de *reviews*. Pode-se observar como esta amostra se organiza por categoria de negócio, sendo as categorias com maior número de *reviews****Food*** (2430), ***Night Life*** (1215) e ***Restaurant*** (6105), e as com menos opiniões, ***Local Flavor*** (8), ***Mass Media*** (9), ***Real Estate*** (14) e ***Religious Organization***(4), o que denota maior preocupação dos utilizadores por questões relacionadas com a alimentação (***Food*** e ***Restaurant***) em detrimento das relacionadas com as características dos locais.

A coluna 2 da tabela 2, permite-nos fazer a exploração por categoria de negócio, identificando as classificações médias atribuídas pelos utilizadores em cada categoria. Constata-se que as categorias de negócio com pior avaliação média são ***Financial Services*** (3,0), ***Real Estate***(2,5), ***Home Services*** (3,4) e ***Hotels&Travel*** (3,4). As categorias de negócio com melhor avaliação média são ***Education*** (4,2), ***Local Flavor*** (4,3), ***Mass Media*** (4,2) e ***Religious Organization*** (4,4).

Numa segunda fase, pode-se observar as duas últimas colunas, onde se pode constatar que, na amostra, as categorias de negócio em que os clientes dão a sua opinião com maior número de palavras em média são ***Home Services*** (208,6), ***Local Flavor*** (271,3) e no ***Real Estate*** (250,6), sendo que a primeira e a terceira têm a pior avaliação média por categoria, e a segunda uma das melhores avaliações médias e o maior número médio de palavras como se pode verificar nas tabelas 2.

O gráfico da figura 9 apresenta as distribuições das categorias de negócio referidas:

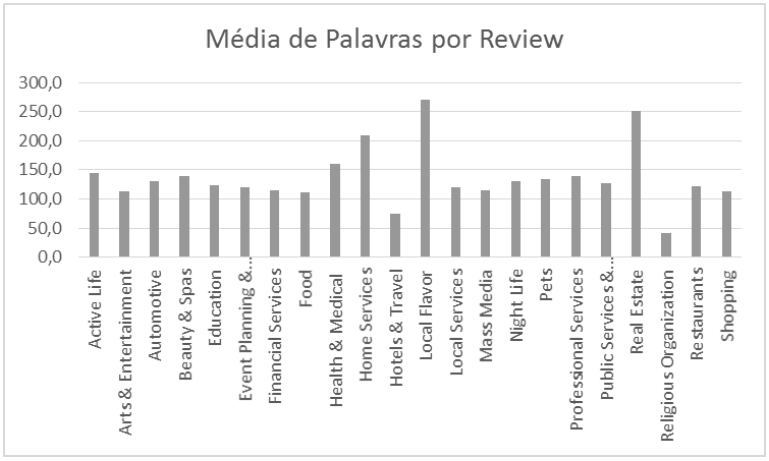


FIGURA 9: GRÁFICO DE MÉDIA DE PALAVRAS POR *REVIEW*.

3.2.3. Preparação dos Dados

Nesta secção, optou-se por preparar os dados e fazer duas análises, através do programa *R,*utilizando numa primeira fase a amostra como um todo, e numa segunda fase por categorias de negócio, tendo em conta dois cenários:unigramas (termos que ocorrem individualmente superiores ou igual a três caracteres) e bigramas (uma sequência de dois termos adjacentes), uma vez que, foram os que obtiveram as distribuições mais homogéneas após uma análise exaustiva e exploratória de vários cenários.

Nota: de salientar que quando são citadas as palavras termo e *entity* tem o mesmo significado para esta dissertação.

É nesta fase que, por norma, os processos de TM e os de DM se distinguem, visto que, num projeto de *data mining* os dados, normalmente encontram-se estruturados e aqui foi necessário proceder a esse tratamento.

De seguida apresentam-se os tratamentos dos dados que foram necessários utilizar para as diferentes análises/explorações e para os tópicos efetuados posteriormente seguindo as boas práticas da literatura (Graham, 2014):

- Remoções dos números, visto que, para as análises pretendidas os números não iriam alterar os resultados. Esta é uma prática corrente neste tipo de análises(Feinerer et al., 2008; Guerreiro et al., 2015);
- Remoção da pontuação (permite eliminar todos os sinais de pontuação, as chavetas, as aspas, os asteriscos, entre outros) uma vez que, para as análises pretendidas a pontuação não irá alterar os resultados(Feinerer et al., 2008; Guerreiro et al., 2015);
- *Stopwords* (remoção das palavras mais comuns consideradas irrelevantes (desde verbos auxiliares, artigos, pronomes, preposições, interjeições como termos mais comuns e irrelevantes) para a análise do texto (Liu, 2008; Feinerer et al., 2008; Guerreiro et al., 2015);

- Remoção de algumas palavras que com as*Stopwords* não estavam a ser removidas. Ao executar o código para esse efeito no R, verificou-se essa falha, pelo que, se teve de proceder a remoção das *stopwords* que faltavam utilizando código feito exclusivamente para esse efeito(Feinerer et al., 2008; Guerreiro et al., 2015);
- *Stemming*permite reduzir as palavras para a sua raiz (reduzindo variantes de palavras às suas respetivas raízes) (Liu, 2008; Porter, 1980);
- Reduzir espaços em branco (para melhorar a *performance*)(Feinerer et al., 2008; Guerreiro et al., 2015);
- Utilização de TF-IDF como fator de ponderação que dá maior peso aos termos que ocorrem com mais frequência num maior número de *reviews*(Grün & Hornik, 2011).

Na fig. 10, apresenta-se o diagrama da fase de tratamento dos dados, atrás descrito:

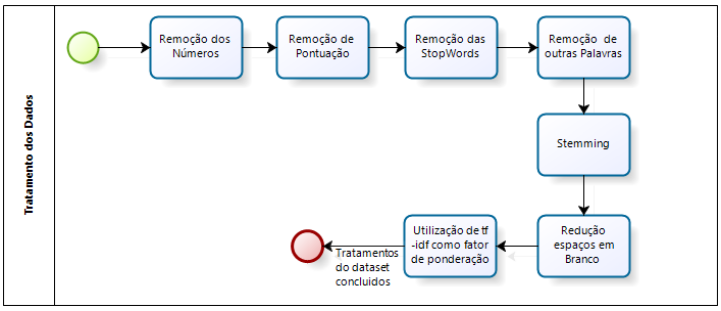


FIGURA 10: DIAGRAMA DO TRATAMENTO DOS DADOS.

No final, optou-se pela construção de uma matriz DTM, sendo uma das formas mais comuns para estruturar os dados contidos nos vários documentos ou *corpus*.

A DTM é formada pelas linhas, que representam os documentos, pelas colunas que representam os termos, e, na sua intersecção, encontra-se o valor da frequência absoluta do termo no documento. Para reduzir a esparsidade da matriz utiliza-se o factor de ponderação, tf-idf(Feinerer et al., 2008).A matriz final constituída por 12371 documentos ficou com 29894 termos identificados.

Para se ter uma maior perceção do comportamento dos dados, achou-se que seria interessante fazer uma análise exploratória observando os termos mais e menos frequentes (Corney, Vel, Anderson, & Mohay, 2002; Nahm & Mooney, 2002), uma vez que, indicam quantas vezes os utilizadores da plataforma/aplicação usam um determinado termo.

De seguida apresentam-se as análises exploratórias dos *reviews*, através do programa*R*, para as duas análises anteriormente referidas (com amostra total, e por categorias, nos cenários unigramas e bigramas) onde foram analisados os seguintes fatores:

- As palavras mais e menos frequentes (Forman, 2003; Matsuo & Ishizuka, 2004);
- Os termos com frequência mínima de ocorrência (tem que se ter em conta a quantidade de *reviews* e ter em conta os termos mais frequentes que foram obtidos na análise anterior);

- As *Wordcloud,* imagens formadas pelas palavras tendo em conta a frequência, (sendo uma forma mais representativa da frequência das palavras, para uma visualização mais simples e perspicaz de acordo com as frequências obtidas naanálise anterior)(Guerreiro et al., 2015).

A figura 11apresenta um diagrama com todo o processo de exploração por categoria de negócio, mencionado anteriormente:

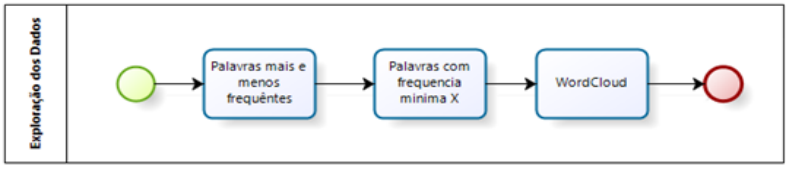


FIGURA 11: DIAGRAMA DE EXPLORAÇÃO DOS DADOS.

3.2.3.1. Análise da Amostra Completa

Nesta secção, apresentam-se os resultados das explorações do *dataset* global*,* que contemplam as opiniões sobre as 22 categorias de negócio mencionadas anteriormente, recorrendo ao *software* R.

Palavras mais e menos frequentes

Após os tratamentos dos dados, observou-se os termos mais e menos frequentes, como se pode verificar de seguida para os dois cenários:

Termos menos frequêntes	Quantidade	Termos mais frequêntes	Quantidade
aaaaaa	1	just	5314
aaaaaamaz	1	get	6161
aaaaand	1	like	6721
aaaaanywho	1	food	7085
aaaaawesom	1	good	7707
aacs	1	place	8512

TABELA 3: UNIGRAMAS MAIS E MENOS FREQUENTES.

Termos menos frequêntes	Quantidade	Termos mais frequêntes	Quantidade
high recommend	264	realli good	458
make sure	268	come back	565
place go	274	pretti good	596
you can	287	ice cream	639
next time	290	go back	679
if your	292	this place	814

TABELA 4: BIGRAMASMAIS E MENOS FREQUENTES.

Ao observar a tabela 3, num cenário de unigramas, constata-se que as palavras mais frequentes em todo o *dataset*,são: ***place***(8512), ***good***(7707),***food***(7085), ***like***(6721),***get*** (6161),***just***(5314).

Destaca-se que, uma das palavras mais frequente é a palavra ***food***, dado que as categorias ***Restaurant*** *e* ***Food***, contêm o maior número de *reviews*, como se pode observar na tabela 2.

Analisando agora a tabela 4, constata-se que os bigramas mais frequentes são, ***this place*** (814), ***go back*** (679), ***ice cream*** (639), ***pretti good*** (596), ***come back*** (565), ***realli good*** (458).

Nesta análise, contrariamente ao expectável as palavras relacionadas com ***Food*** e ***Restaurant*** não se destacam nas palavras mais frequentes, embora mais de metade da amostra se referir a estas duas categorias de negócio.

Termos com frequência mínima de ocorrência

Devido à exploração anterior apresentada nas tabelas3 e 4, e tendo em conta o elevado número de *reviews*, 12371, optou-se por observar as palavras que ocorrem no mínimo 3000 e 5000 vezes para o cenário de unigramas, e de 200 e 500 vezes para o cenário de bigramas. Estes *thresholds* foram escolhidos observando e analisando as tabelas anteriores das palavras mais frequentes.

As figuras12 e 13 apresentam o código utilizado em R para obter os termos com a frequência mínima para os dois cenário e os termos identificados:

```
> findFreqTerms(matriz, lowfreq=3000)#termos que aparecem no minimo 3000 vezes
 [1] "also"   "back"   "can"    "dont"   "food"   "friend" "get"    "good"   "great"  "ive"
[11] "just"   "like"   "love"   "one"    "order"  "place"  "realli" "servic" "time"   "tri"
> findFreqTerms(matriz, lowfreq=5000)#termos que aparecem no minimo 5000 vezes
[1] "food"  "get"   "good"  "just"  "like"  "place" "time"
```

FIGURA 12: UNIGRAMAS QUE APARECEM PELO MENOS 3000 E 5000 VEZES.

```
> findFreqTerms(rvwbie, lowfreq=200)#termos que aparecem no minimo 200 vezes
 [1] "can get"        "come back"      "dont know"      "even though"    "everi time"
 [6] "feel like"      "first time"     "food good"      "go back"        "great place"
[11] "high recommend" "ice cream"      "if your"        "im sure"        "ive ever"
[16] "look like"      "love place"     "make sure"      "next time"      "place go"
[21] "pretti good"    "realli good"    "tast like"      "they also"      "this place"
[26] "you can"
> findFreqTerms(rvwbie, lowfreq=500)#termos que aparecem no minimo 500
[1] "come back"   "go back"     "ice cream"   "pretti good" "this place"
```

FIGURA 13: BIGRAMAS QUE APARECEM PELO MENOS 200 E 500 VEZES.

Ao analisar as duas figuras 12 e 13 torna-se difícil identificar os assuntos mais abordados, uma vez que se encontram presentes todas as opiniões sobre todas as categorias de negócio, e também, porque os termos se encontram enviesados para as categorias com maior número de *reviews*.

Wordcloud

Optou-se por fazer as *Wordclouds* com as frequências mínimas referenciadas nas análises anteriores, obtendo assim uma visualização mais clara, simples e perspicaz.

De seguida apresentam-se as *Wordclouds* (em que o tamanho fonte das palavras é proporcional à frequência dos termos) das frequências mínimas tendo em conta os dois cenários:

A)

B)

C) D)

FIGURA 14: *WORDCLOUD*: A) UNIGRAMAS FREQUÊNCIA MÍNIMA 3000; B) UNIGRAMAS FREQUÊNCIA MÍNIMA 5000; C) BIGRAMAS FREQUÊNCIA MÍNIMA 200; D) BIGRAMAS FREQUÊNCIA MÍNIMA 500.

3.2.3.2. Análise por Categoria

Nesta secção, apresentam-se os resultados das explorações do *dataset* por categorias de negócio.

No entanto, por serem 22 categorias de negócio, optou-se por analisar apenas a categoria com maior número de *reviews,* ***Restaurant***, sendo as restantes categorias analisadas nos anexos desta dissertação.

Categoria *Restaurant*

A tabela 2mostra que esta categoria de negócio tem 6105 *reviews*. Nesta categoria encontram-se comentários sobre diferentes tipos de gastronomia (mexicano, chines, marroquino, indiano, japonês, latino-americano, francês, vietnamita, médio-oriente, espanhol, tailandês, sushi, americano (tradicional), coreano, italiano, grego, turco, britânico, irlandês, asiático de fusão entre outros) bem como sobre dieta vegetariana, pizzarias, bares de tapas, marisco, sanduíches, barraquinhas de comida, grelhados, pequenos-almoços, *snack-bar,* comida sem glúten, grelhados, saladas, cachorros quentes, *buffet* entre outros.

De seguida, apresentam-se as análises exploratórias nesta categoria de negócio tendo em conta os seguintes cenários:

Palavras mais e menos frequentes

Após os tratamentos dos dados, identificaram-se os termos mais e menos frequentes tendo em conta os cenários das tabelas 5 e 6:

Termos menos frequêntes	Quantidade	Termos mais frequêntes	Quantidade
aaaaawesom	1	just	2670
aahahah	1	get	2912
aahhhhh	1	like	3380
aarp	1	good	4458
aash	1	place	4537
abbrevi	1	food	5046

TABELA 5: UNIGRAMAS MAIS E MENOS FREQUENTES NA CATEGORIA *RESTAURANT*.

Termos menos frequêntes	Quantidade	Termos mais frequêntes	Quantidade
high recommend	264	realli good	458
make sure	268	come back	565
place go	274	pretti good	596
you can	287	ice cream	639
next time	290	go back	679
if your	292	this place	814

TABELA 6: BIGRAMASMAIS E MENOS FREQUENTES NA CATEGORIA *RESTAURANT*.

Ao observar a tabela 5 que apresenta o cenário de unigramas, pode-se constatar que as palavras mais frequentes nesta categoria de negócio são, ***food*** *(*5046), ***place*** (4537), ***good***(4458), ***like***(3380), ***get*** (2912) e ***just*** (2670).

Analisando o cenário de bigramas na tabela 6, nota-se que os bigramas mais frequentes nesta categoria de negócio são, ***this place***(814), ***go back***(679),***ice cream***(639), ***preatti good*** (596), ***come back*** (565) e ***realli good*** (458), mantendo a concordância dos temas abordados nesta categoria de negócio.

Constata-se neste caso, que a exploração feita com unigramas descreve melhor a categoria de negócio e consequentemente os assuntos abordados.

Termos com frequência mínima de ocorrência

Tal como para a análise da amostra completa, optou-se por fazer esta análise da frequência mínima dos termos, no caso de unigramas que ocorrem pelo menos 1000 e 2000 vezes, e que ocorrem pelo menos 100 e 300 vezes para o cenário de bigramas. Estes *thresholds* foram escolhidos observando e analisando as tabelas anteriores das palavras mais frequentes.

Apresentam-se a seguir as figuras 15 e 16 que contemplam os códigos e as análises acima referidas:

```
> findFreqTerms(matriz, lowfreq=1000)#termos que aparecem no minimo 1000 vezes
 [1] "also"    "alway"   "back"    "best"    "can"     "chicken"
 [7] "come"    "delici"  "dish"    "dont"    "eat"     "even"
[13] "food"    "fri"     "friend"  "get"     "good"    "got"
[19] "great"   "its"     "ive"     "just"    "like"    "littl"
[25] "look"    "love"    "lunch"   "make"    "menu"    "much"
[31] "nice"    "one"     "order"   "pizza"   "place"   "pretti"
[37] "price"   "realli"  "restaur" "salad"   "sandwich" "sauc"
[43] "servic"  "tast"    "they"    "thing"   "think"   "this"
[49] "time"    "tri"     "want"    "well"    "will"
> findFreqTerms(matriz, lowfreq=2000)#termos que aparecem no minimo 2000 vezes
 [1] "food"   "get"    "good"   "great"  "just"   "like"   "one"    "order"
 [9] "place"  "realli" "time"   "tri"
```

FIGURA 15: UNIGRAMAS QUE APARECEM PELO MENOS 1000 E 2000 VEZES NA CATEGORIA *RESTAURANT*.

```
> findFreqTerms(rvwbi, lowfreq=100)#termos que aparecem no mínimo 100 vezes
 [1] "can get"        "come back"      "dont know"      "dont think"
 [5] "even though"    "everi time"     "feel like"      "first time"
 [9] "food good"      "food great"     "go back"        "good food"
[13] "good place"     "great food"     "great place"    "happi hour"
[17] "high recommend" "ice cream"      "if your"        "im sure"
[21] "indian food"    "ive ever"       "ive never"      "ive tri"
[25] "last night"     "look like"      "love place"     "make sure"
[29] "much better"    "my favorit"     "my friend"      "next time"
[33] "pad thai"       "place go"       "pretti good"    "pretti much"
[37] "realli good"    "realli like"    "reason price"   "seem like"
[41] "sweet potato"   "tast like"      "thai food"      "they also"
[45] "this place"     "we order"       "will definit"   "you can"
> findFreqTerms(rvwbi, lowfreq=300)#termos que aparecem no mínimo 300 vezes
[1] "come back"   "go back"     "pretti good" "this place"
```

FIGURA 16: BIGRAMASQUE APARECEM PELO MENOS 100 E 300 VEZES NA CATEGORIA *RESTAURANT*.

A figura 15 mostra-nosalguns unigramasrelevantes, como, ***food, place, good,*** entre outros, que distinguem bem esta categoria. Por outro lado, na figura 16 ressaltam os bigramasrelevantes,

como, ***come back***, ***go back***, ***pretti good***, entre outros.Pode-se constatar, que nesta categoria (***Restaurant***), o cenário que melhor a representam é o de unigramas, uma vez que, as palavras a descrevem melhor.

Wordcloud

Optou-se por fazer as *Wordclouds* com as frequências mínimas referenciadas nas análises anteriorespara uma interpretação mais clara, simples e perspicaz.

FIGURA 17: *WORDCLOUD*DA CATEGORIA *RESTAURANT*: A) UNIGRAMAS FREQUÊNCIA MÍNIMA 1000; B) UNIGRAMAS FREQUÊNCIA MÍNIMA 2000; C) BIGRAMAS FREQUÊNCIA MÍNIMA 100; D) BIGRAMAS FREQUÊNCIA MÍNIMA 300.

Analisando as palavras mais frequentes verifica-se que tanto na amostra total como na amostra da categoria ***Restaurant*** são identificadas as mesmas palavras, embora com frequências diferentes. Esta constatação não surpreende dado que a esmagadora maioria de *reviews* na amostra total se refere a alimentação (categoria ***Restaurant***e ***Food***). É de salientar que nesta mesma análise na amostra global no cenário de unigramas a palavra mais frequente é ***place*** (8512) enquanto que na amostra da categoria ***Restaurant*** a palavra mais frequente é ***food*** (5046).

Por último, destaca-se que em ambos os casos no cenário de bigramas os resultados não descrevem bem a realidade da amostra.

3.3. Modeling + Evaluation

Nesta fase abordarei todos os procedimentos necessários para obter o DSSe qual a importância da sua criação.

3.3.1. Análise de Tópicos Latentes

Após a análise dos termos individuais que se encontram em cada *review* procedeu-se a uma segmentação desses termos em tópicos latentes. Embora o *Yelp* defina uma segmentação através das suas categorias de negócio, estas são categorias impostas pela gestão e não das preocupações e dos temas abordados pelos clientes aquando da sua opinião sobre a experiência. Achou-se por isso estritamente necessário criar tópicos para conseguir distinguir os temas dos *reviews*, através do algoritmo conhecido como *Correlated Topic Models* (CTM) (Blei & Lafferty, 2007). Este algoritmo utiliza-se com o intuito de conseguir uma sumarização/compactação eficiente dos termos presentes nos *reviews* criando *clusters* que podem ser úteis para a tomada de decisão(Guerreiro et al., 2015).

Sendo um algoritmo de procura de *clusters* hierárquico, o CTM não fornece o número exato de grupos que devem ser considerados. Em vez disso, o algoritmo permite analisar a variabilidade explicada como por exemplo a sua perplexidade (a força do modelo para prever novas palavras após ajustamento de um modelo) para cada número possível de *clusters.* O número de *clusters*/tópicos, que melhor se adequam aos dados são, avaliados pela forma de como essas métricas mudam quando o número de *clusters* aumenta. O número ideal de *clusters* é atingido quando a variabilidade explicada não muda significativamente pela adição de mais clusters (Guerreiro et al., 2015).

A figura 18 apresenta o gráfico da perplexidadedos modelos tendo em conta a amostra global de unigramas(com a pontuação entre os 5 e os 60 tópicos):

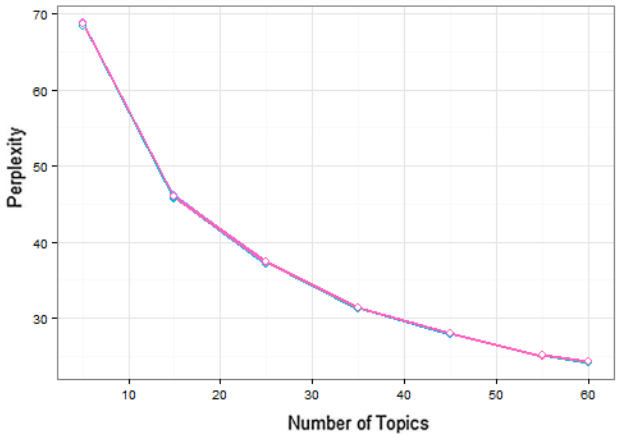


FIGURA 18: PERPLEXITY DA AMOSTRA GLOBAL UNIGRAMAS (COM A PONTUAÇÃO PARA OS 60 TÓPICOS MAIS SALIENTES).

Ao observar a métrica de perplexidade da amostra global através da figura 18, verifica-se que o número ideal de tópicos é de 15. No entanto,ao elaborar o *profilling* dos 15 tópicos constatou-se que eram de difícil categorização, uma vez que, existiam vários tópicos latentes que abordavam assuntos similares.Decidiu-se assim, proceder a uma análise de *part-of-speech* (POS), de forma a facilitar a categorização dos termos e a incluir apenas os termos relevantes para a análise. Neste caso, após uma análise exploratória optou-se por usar apenas os *Nouns* (*noun, singular or mass*), substantivos que resultaram da classificação POS.

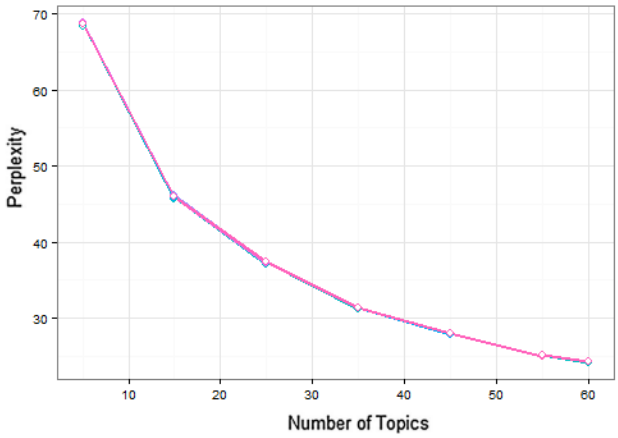


FIGURA 19: PERPLEXITY DOS 3 CENÁRIOS (COM A PONTUAÇÃO PARA OS 60 TÓPICOS MAIS SALIENTES).

Apesar do gráfico da perplexidadeda figura 19 manter o mesmo número de tópicos que na amostra global, a análise exploratória de *profilling* revelou que 12 tópicos latentes seriam mais adequados para explicar as categorias de forma mais adequada pelo que foi criado um modelo de tópicos assente em 12 *clusters.*

A tabela 7 apresenta a categorização dos tópicos, os termos e a descrição de cada tópico:

Tópicos	Termos	Categorização	Descrição
1	*food, place, time, servic, friend, way, noth, experi, burrito, day*	Bar	Os comentários que se encontram neste tópico são relacionados a bares que contem refeições.
2	*pizza, time, park, place, car, class, area, crust, street, slice*	Pizzeria	Os comentários que se encontram neste tópico são relativos a opiniões de pizarias.
3	*food, sandwich, place, coffe, salad, lunch, servic, meat, cafe, steak*	Snack Bar	Os comentários que se encontram neste tópico são relativos a opiniões sobre snack bares.
4	*place, order, lunch, food, breakfast, servic, atmospher, bacon, egg, chees*	Brunch	Os comentários que se encontram neste tópico são relativos a opiniões sobre estabelecimentos que servem pequenos-almoços.
5	*burger, cake, flavor, roll, cupcak, chocol, chees, bite, time, day*	Fast Food	Os comentários que se encontram neste tópico são referentes a opiniões sobre estabelecimentos de refeições rápidas.
6	*cream, ice, chocol, beer, bread, bar, dessert, even, cooki, someth*	Cake Shop	Os comentários que se encontram neste tópico são relativos a opiniões sobre pastelarias.
7	*tea, store, select, locat, night, school, stuff, shop, squar, campus*	Tea House	Os comentários que se encontram neste tópico são relativos a opiniões sobre chás e casas de chás.
8	*chicken, restaur, sauc, food, menu, rice, dish, spici, soup, beef*	Restaurant	Os comentários que se encontram neste tópico são relativos a opiniões sobre restaurantes.
9	*time, hair, staff, job, work, experi, custom, place, offic, care*	Hairdresser	Os comentários que se encontram neste tópico são relativos a cabeleireiros.
10	*bar, tabl, beer, time, bread, night, meal, bit, chees, area*	Night Life	Os comentários que se encontram neste tópico são relativos a opiniões sobre a vida noturna.
11	*room, busi, manag, staff, lot, food, store, buffet, park, day*	Hotel	Os comentários que se encontram neste tópico são relativos a opiniões sobre hotéis.
12	*food, place, time, servic, love, night, thing, price, everyth, area*	Leisure	Os comentários que se encontram neste tópico são relativos a atividades de lazer no geral.

TABELA 7: TÓPICOS, TERMOS E DESCRIÇÃO.

Os termos apresentados na tabela 7, são os 10 termos mais correlacionados com cada tópico, ordenados do mais correlacionado para o menos correlacionado com o tópico. Exemplificando,no tópico 1, o termo ***food*** é o mais correlacionado, enquanto que, o termo *day* é o menos correlacionado com o tópico.

3.3.2. Sentiment Analysis (SA)

Após a identificação/categorização dos tópicos, a ferramenta *Semantria* foi utilizada para a análise de sentimentos. O *Semantria* é um *software* da *Lexanalytics* e tem sido utilizado com sucesso na extração de conhecimento de dados não estruturados quer em documentos, sentenças ou frases (Lawrence, 2014).

A referida ferramenta analisa os sentimentos de uma forma automática alicerçando-se em algoritmos destinados a extrair os referidos sentimentos de um modo semelhante ao dos seres humanos. Isto é, o *software* "questionará/identificará" qual o sentimento presente em cada comentário, nos termos e nos tópicos elaborando uma classificação tendo em conta a sua polaridade (positivo,neutro e negativo).

A referida extração do sentimento de um documento une-se aos seguintes passos:

- Num primeiro passo o documento é separado em *part-of-speechtags*;
- Seguidamente o algoritmo irá realizar uma identificaçãodas frases que contêm sentimento;
- Posteriormente, incluirá uma escala logarítmica que vai de -10 a 10 pontos para cada frase que suporta o sentimento (Abeywardena, 2014);
- Por último, as pontuações são combinadas para determinar o sentimento geral.

Cada comentário, termo e tópico, através da inferência estatística, terá uma polaridade (positivo, neutro e negativo) e o seu valor de sentimento(compreendido de -2,0 e 2,0).

De seguida apresenta-se a tabela que ilustra as polaridades e os seus valores dos sentimentos correspondentes:

Polaridade	Valor de Sentimento
Positivo	Compreende o intervalo seguinte: [0,22;2]
Neutro	Compreende o intervalo seguinte: [-0,05;0,22 [
Negativo	Compreende o intervalo seguinte: [-2;-0,05 [

TABELA 8: POLARIDADE E OS VALORES CORRESPONDENTES.

Na figura 20 pode observar-se o funcionamento do *Semantria* segundo Lawrence (2014), funciona como uma *black box*, por outras palavras, funciona como um sistema que apenas permite visualizar o *input* e *output*, sendo que todo o processo interno (funcionalidades e características de transformação desde o *input*até chegar ao*output*) é desconhecido. (Lawrence, 2014).

FIGURA 20: MODELO DE ANÁLISE DE SENTIMENTOS. (LAWRENCE, 2014).

A utilização desta ferramenta permitiu explorar os 12371 *reviews* e classificar a polaridade, dos termos individuais identificados anteriormente,e procedertambém à classificação dos tópicos latentes descobertos com a utilização do modelo CTM bem como à identificação de novas categorias.Os 12371*reviews* analisados revelaram um conjunto de 83037 análises de sentimentos que permitiram explorar os seguintes pontos:

- A polaridade por documento, neste caso por *review*, e o valor do sentimento geral do documento (o sentimento de um documento resulta da classificação da polaridade sentimental dos vários termos e tópicos presentes no mesmo documento);
- A identificação dos termos (unigramas ou frases com n-gramas) detetados pelo*Semantria*, a polaridade e valor do sentimento;
  A identificação dos tópicos que o*Semantria* deteta, a polaridade e o valor da polaridade.

É de salientar que destas 83037 análises só foram identificadas as categorias definidas anteriormente em 33468, correspondentes a 10385 *reviews*, nos quais só foram identificados termos em 9924 *reviews*.

De seguida apresenta-se um exemplo para ilustrar os resultados obtidos no *Semantria*:

ID	Source Text	Document Sentiment	Document Sentiment +/-	Entity	Entity Type	Entity Sentiment	Entity Sentiment +/-	User Category	User Category Strength	User Category Sentiment	User Category Sentiment +/-
bb5a244e-8eaa-4572-822d-2e42c78eaca0	I'm surprised the overall rating of this place isn't higher! I think this is the best Indian restaurant in Central Square. The sauces are rich they serve decent portions (though the prices are higher than I'd like if you're not there for the lunch buffet). I feel like their chefs know their Indian food but what do I know - I'm Filipino. Their chicken tikka masala was always good but any of the lamb dishes are great. I also remember they made a mean thick mango lassi. The saag paneer was really good. All versions of their naan was hot fresh and oiled/buttered. The service was decent and they tried to make you happy.	0,60409	positive	place	choice	0,46865	neutral	Restaurant	0,77486	0,75	positive

TABELA 9: EXEMPLO DE RESULTADO DO SEMANTRIA.

3.3.3. Sistema de apoio à decisão para o Turismo

Tendo em conta que área do turismo tem um papel importante na economia devido as receitas diretas (alojamento/viagens) e indiretas (restauração passeios, entre outros) que acabam por potenciar o crescimento económico, um DSS, faz todo o sentido, devido à potencialidade do desenvolvimento do sector. Um DSS neste caso permite ajudar a identificar as cidades onde se pode desenvolver o turismo e o comércio, sabendo em que tipo de categorias de negócio é que se deve investir, quais as categorias de negócio mais fortes em cada cidade, bem como, perceber quais as categorias que se devem melhorar em certas cidades.

Um DSS é importante nesta área porque assim, poder-se-á perceber a afluência em certas cidades e não noutras, através dos sentimentos e das categorias em cada cidade, bem como, perceber o que é que os turistas valorizam mais em cada cidade.

Estas informações podem ser extremamente importantes para os institutos de turismo, visto que, pode ajudar nas decisões sobre que tipo de negócios que se devem expandir em cada cidade bem como os aspetos que os turistas mais valorizam em cada cidade, de acordo com o tipo de turistas que viajam para cada cidade.

3.3.3.1. Sistema de Apoio à Decisão – Modelo de Dados

De forma a apresentar um sistema de apoio à decisão para os gestores na área do turismo, e com os resultados da análise anterior que classificou os termos e tópicos em termos do seu valor sentimental, foram criados dois processos diferentes, com duas tabelas de factos assentes em dimensões conformes, de forma a poder ajudar os decisores a identificar os fatores com uma maior influênciapositiva ou negativa na satisfação dos clientes. O primeiro processo de negócio focou-se na Gestão de Categorias, analisando os 12 tópicos criados anteriormente com o CTM identificadas pelo *Semantria* e o outro focou-se na Gestão dos termos identificados (utilizando as palavras com frequência superior ou igual a 400, como foi referido anteriormente). Esta decisão de criar os dois processos foi motivada pela diferença de contexto em que ambas as análises ocorrem.

A escolha dos modelos dimensionais usados de gestão de categorias e gestão de termos justifica-se pela abrangência e complementaridade dos resultados, uma vez que, se por um lado permite “auscultar a saúde” empresarial dos diversos alvos, estado, cidade, de forma temporal por categorias de negócio, suas necessidades e expectativas, por outro permite obter o sentimento individual mais focado na definição das criticas e expectativas locais. Sendo que em conjunto refletem o espelho da realidade do sentimento das comunidades identificando os pontos fortes e fracos do turismo e consequente denúncia das áreas onde intervir.

De seguida abordaremos o modelo dimensional para o DSS:

Modelo dimensional

“O core de qualquer modelo dimensional é um conjunto de métricas de negócio que captura como um processo é avaliado e a descrição do contexto de cada medida” (Adamson,2010).

Para a criação do modelo dimensional é necessário saber quais os principais objetivos a serem analisados e assim modelar o que se deseja medir. Posto isto, para o desenvolvimento da modelação dimensional foi seguida a metodologia de Kimball. Esta metodologia contempla os seguintes 4 passos (Ross &Kimball,2002):

- Escolha do processo de negócio;
- Definição de grão do processo selecionado;
- Escolha das dimensões sobre as quais serão contextualizados os factos;

- Definição das medidas presentes nas tabelas de factos.

Passo 1, Escolha do Processo de Negócio:

Os processos escolhidos para a modelação foram a Gestão dos Sentimentos das Categorias e a Gestão dos Sentimentos dos Termos de Negócio, identificadas pelo *Semantria*, como mencionado anteriormente. Apresenta-se de seguida a tabela de processo de negócios:

	Country	Review	Business	Date	Polaridade	Entity	Category
Gestão de Categorias	x	x	x	x	x		x
Gestão de Termos	x	x	x	x	x	x	

TABELA 10: PROCESSO DE NEGÓCIO.

Passo 2, Definição do Grão do Processo:

O grão representa o nível de detalhe de uma tabela de factos, o que implica que, caso não esteja bem definido pode-se tornar prejudicial para o sucesso do modelo dimensional. A definição do grão é extremamente importante uma vez que dá a informação sobre a tabela de factos e define o detalhe em que todos os factos da tabela necessitam de estar para que haja correspondência. Desta forma, foram identificados os níveis de detalhe para a Gestão de Categorias e para a Gestão de Termos:

Tabela	Grão
TF_TR_*Category _Sentiment*	Uma categoria de negócio, com uma determinada polaridade, num determinado estado, numa determinada cidade, num determinado ano, mês, dia, por review e por negócio.
TF_TR_*Entity _Sentiment*	Um termo,com uma determinada polaridade, num determinado estado, numa determinada cidade, num determinado ano,mês, dia, por review e por negócio.

TABELA 11: DEFINIÇÃO DE GRÃO EM CADA TABELA DE FACTOS.

Passo 3, Escolha e definição das dimensões:

De seguida caracteriza-se as dimensões referidas anteriormente:

Dimensão *Date*:

A dimensão *Date* é utilizada para contextualizar os factos no tempo. Desta forma a dimensão *Date* contém neste caso todos os atributos desde o nível do dia até ao ano.

FIGURA 21: HIERARQUIA DA DIMENSÃO *DATE*.

De seguida apresenta-se a tabela com os atributos da dimensão *Date*:

Nr	Identificação	Descrição	Tipo	Chave(Tipo)
1	SK_Date(PK)	Identificador da data;	Integer	S
2	Date	Data;	Date	N
3	Day	Número do dia;	Integer	N
4	Month	Número do mês;	Integer	N
5	Year	Número do ano;	Integer	N

TABELA 12: ATRIBUTOS DA DIMENSÃO *DATE*.

Dimensão *Country*:

A dimensão Country é utilizada para contextualizar os factos geograficamente. De seguida apresenta-se a hierarquia da dimensão *Country*:

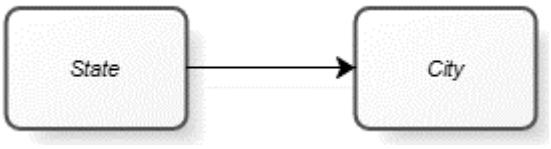


FIGURA 22: HIERARQUIA DA DIMENSÃO *COUNTRY*.

De seguida apresenta-se a tabela com os atributos da dimensão:

Nr	Identificação	Descrição	Tipo	Chave(Tipo)
1	SK_Country(PK)	Identificador de país	Integer	S
2	State	Nome do estado	String	N
3	City	Nome da cidade	String	N
4	Latitude	Latitude	Double	N
5	Longitude	Longitude	Double	N

TABELA 13: ATRIBUTOS DA DIMENSÃO *COUNTRY*.

Dimensão *Review*:

A dimensão *Review* diz respeito a toda a informação relacionada com os comentários dos utilizadores da plataforma/aplicação. Devido a possíveis alterações nos atributos, considerou-se que esta dimensão será uma *slowly changing dimension* de tipo 1, ou seja, é alterado por cima sem obter qualquer histórico.

De seguida apresenta-se a tabela com os atributos da dimensão:

Nr	Identificação	Descrição	Tipo	Chave(Tipo)
1	SK_Review(PK)	Identificador do *review*	Integer	S
2	NK_Review_id	Identificador do *review* do sistema fonte	String	NK
3	Source Text	Comentário sobre o negócio	String	N
4	Review_stars	Avaliação do negócio	Double	N
5	Number_of_words	Número de palavras por *review*	Integer	N
6	Value_of_Doc_Sentiment	Valor do sentimento do *review*	Double	N
7	Polarity_of_Sentiment	Polaridade do sentimento do *review*	String	N

TABELA 14: ATRIBUTOS DA DIMENSÃO *REVIEW*.

Dimensão *Business*:

Esta dimensão tem toda a informação relacionada com os negócios sobre os quais existem os comentários dos utilizadores na plataforma/aplicação. Uma vez que pode existir algum tipo de alteração nos atributos, considerou-se que será uma dimensão *slowly changing dimension* de tipo 1, uma vez que, se existir alteração em algum atributo, este será substituído sem necessita de histórico.

Seguidamente, apresenta-se a tabela com os atributos da dimensão referida:

Nr	Identificação	Descrição	Tipo	Chave(Tipo)
1	SK_Business(PK)	Identificador do *business*	Integer	S
2	NK_Business_id	Identificador do *business* do sistema fonte	String	NK
3	Business_categories	Categoria do negócio	String	N
4	Business_stars	Avaliação do *business*	Double	N

TABELA 15: ATRIBUTOS DA DIMENSÃO *BUSINESS*.

Dimensão Category:

Esta dimensão contém a descrição das doze categorias identificadas pelo CTM. Uma vez que pode existir algum tipo de alteração nos atributos, considerou-se que será uma dimensão *slowly changing dimension* de tipo 1, uma vez que, se existir alteração em algum atributo, este será substituído sem necessita de histórico.

Seguidamente, apresenta-se a tabela com os atributos da dimensão referida:

Nr	Identificação	Descrição	Tipo	Chave(Tipo)
1	SK_Category(PK)	Identificador da categoria	Integer	S
2	Descrição da Categoria	Nome da categoria	String	N

TABELA 16: ATRIBUTOS DA DIMENSÃO CATEGORIA

Dimensão Polaridade:

Esta dimensão contém a descrição da polaridade dos sentimentos (Positivo, Negativo e Neutro). Uma vez que pode existir algum tipo de alteração nos atributos, considerou-se que será uma dimensão *slowly changing dimension* de tipo 1, uma vez que, se existir alteração em algum atributo, este será substituído sem necessita de histórico.

Seguidamente, apresenta-se a tabela com os atributos da dimensão referida:

Nr	Identificação	Descrição	Tipo	Chave(Tipo)
1	SK_Polaridade(PK)	Identificador da polaridade	Integer	S
2	Descrição da Polaridade	Nome da polaridade	String	N

TABELA 17: ATRIBUTOS DA DIMENSÃO POLARIDADE

Passo 4, Definição das Medidas e Caracterização das Tabelas de Factos:

Finalmente o último passo da metodologia de Kimball, que diz respeito à definição dos factos das respetivas tabelas.

Para a apresentação dos factos, foi elaborada uma caracterização minuciosa das tabelas de factos:

- **TF_TR_*Category_Sentiment***

Descrição: Esta tabela contém todos os registos das categorias que foram identificadas bem como os seus sentimentos (polaridade e intensidade) através do *Semantria*;
DataMart: Gestão de Categorias;
Tipo: Tabela de Factos Transacional;
Utilidade estratégica: melhorar o desenvolvimento do turismo, indo de encontro às expectativas dos turistas, tendo em conta os sentimentos;

Atributos				
Dimensões				
Nr	**Identificação**	**Descrição**	**Tipo**	**Chave (Tipo)**
1	SK_Date(FK)	Data do comentário categorizado	Integer	S
2	SK_Review(FK)	Comentário categorizado	Integer	S
3	SK_Business(FK)	Negócio comentado	Integer	S
4	SK_Country(FK)	Sitio do negócio comentado e categorizado	Integer	S
5	SK_Category(FK)	Categoria identificada pela análise do CTM	Integer	S
6	SK_Polaridade(FK)	Indica a polaridade de sentimento	Integer	S
Métricas Elementares				
Nr	**Identificação**	**Descrição**	**Tipo**	**Incluir?**
1	Intensidade de Sentimentos	Indica intensidade do sentimento	#	S

TABELA 18: CARACTERIZAÇÃO DA TABELA DE FACTOS TF_TR_*CATEGORY_SENTIMENT*.

A **TF_TR_*Category_Sentiment*** regista os sentimentos das categorias de negócio por cada *review.*

Como se pode observar através da tabela anterior, a TF é constituída por 6 chaves estrangeiras. A métrica elementar tem natureza aditiva, uma vez que é possível agregá-los em todas as dimensões.

- **TF_TR_*Entity_Sentiment***

Descrição: Esta tabela contém todos os registos dos termos que foram detetadas através do *Semantria*;
DataMart: Gestão de Termos;
Tipo: Tabela de Factos Transacional;
Utilidade estratégica: melhorar o desenvolvimento do turismo, indo de encontro às expectativas dos turistas, tendo em conta os sentimentos;

Atributos				
Dimensões				
Nr	**Identificação**	**Descrição**	**Tipo**	**Chave (Tipo)**
1	SK_Date(FK)	Data do *review que contem as entities* identificadas	Integer	S
2	SK_Review(FK)	*Review* que contem as *entities* identificadas	Integer	S
3	SK_Business(FK)	Negócio que é avaliado que contem as *entities* identificadas	Integer	S
4	SK_Country(FK)	Sitio do negócio que é avaliado e que contem as *entities* identificadas	Integer	S
5	SK_Polaridade(FK)	Indica a polaridade de sentimento	Integer	S
6	*DD_Entity*	Descreve o nome da *entity*	String	N
7	DD_Root_Words	Descreve a raiz de cata *entity*	String	N
Métricas Elementares				
Nr	**Identificação**	**Descrição**	**Tipo**	**Incluir?**
1	Intensidade de Sentimentos	Indica intensidade do sentimento por *entity*	#	S

TABELA 19: CARACTERIZAÇÃO DA TABELA DE FACTOS TF_TR_*ENTITY_SENTIMENT*.

A **TF_TR_*Entity_Sentiment***regista todos os termos identificadas por cada *review.*

Como se pode observar através da tabela 19, a TF é constituída por 5 chaves estrangeiras. A métrica elementar têm uma natureza aditiva, uma vez que é possível agregá-la em todas as dimensões.

De seguida, após a descrição detalhada de todas as dimensões e factos é apresentado o modelo dimensional dos processos implementados, sendo que para efeitos de clareza são apresentados separados, muito embora as dimensões sejam iguais em cada um:

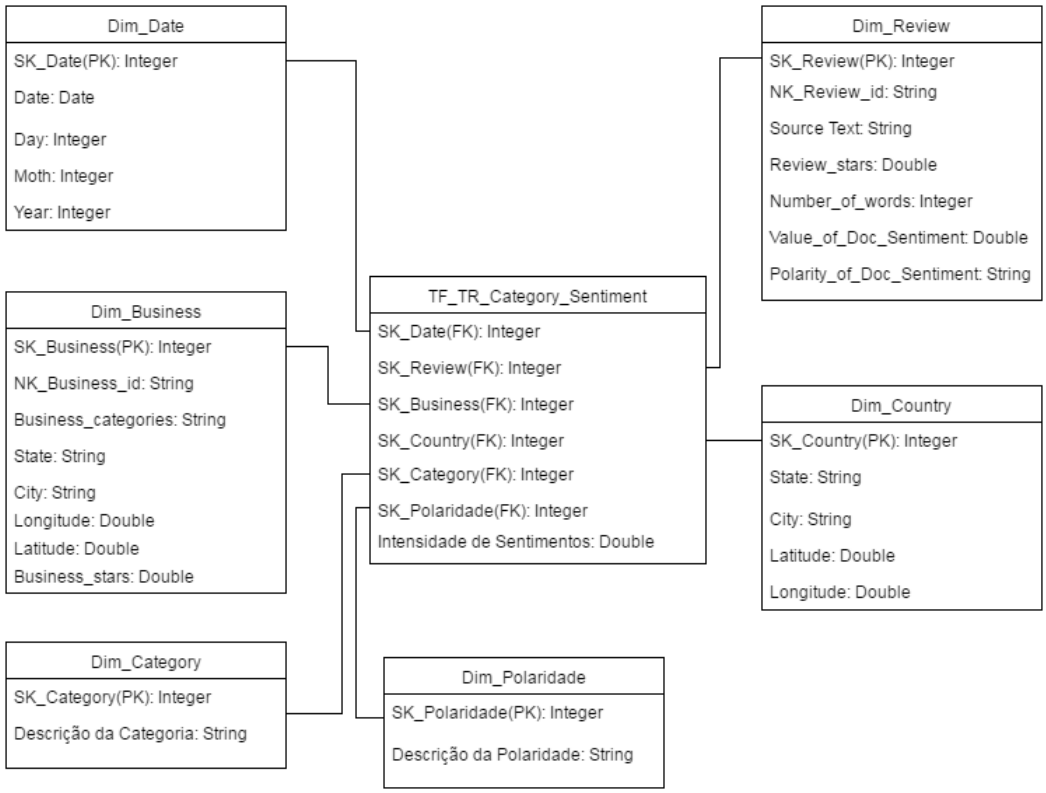


FIGURA 23: MODELO DIMENSIONAL FÍSICO TABELA DE FACTOS *TF_TR_CATEGORY_SENTIMENT*.

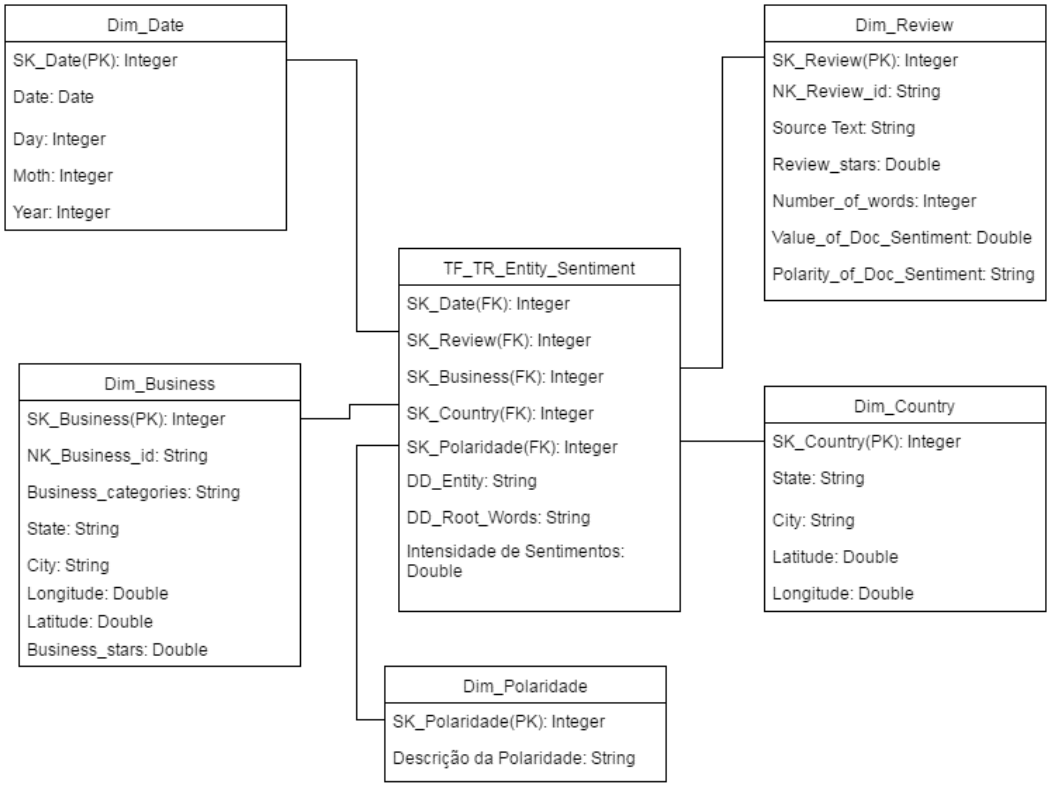


FIGURA 24: MODELO DIMENSIONAL FÍSICO TABELA DE FACTOS TF_TR_*ENTITY_SENTIMENT*.

4. RESULTADOS E DISCUÇÃO (DEPLOYMENT)

Nesta etapa irei apresentar o DSS (com os dois *dashboard* tendo por base os modelos dimensionais), bem como as análises de sensibilidade do DSS, demonstrando a utilidade que um DSS deste género pode ter, para identificar as áreas de turismo a desenvolver, e para identificar quais os fatores a corrigir nas cidades de acordo com as opiniões dos turistas. O *dashboard* do processo de gestão dos sentimentos das categorias permitirá perceber o estado das categorias nos diferentes estados e cidades com o intuito de tomar decisões tendo em conta as opiniões dos clientes e os seus sentimentos percebendo o que mais e menos valorizam em cada cidade e as suas expectativas motivando/potencializando a melhoria e o desenvolvimento das cidades. De seguida apresenta-se a tabela 20 que contém as métricas ea sua descrição. A tabela apresenta as análises (*queries*) incluídas no *dashboard* do processo de gestão de sentimentos. Estas análises derivam das métricas da tabela de factos TF_TR_Category_Sentiment e permitem apresentar uma visão das métricas sob várias perspetivas de análise, nomeadamente:

Métricas de Análise	Descrição/*Query*	Tipo	Fórmula
Percentagem de *Reviews por Ano*	Indica a percentagem de *reviews* por ano	%	nº total de *reviews* num ano/nº total de *reviews*
Quantidade de *Reviews*	Indica a quantidade de *reviews* categorizados por Estado e por Cidade	Count	nº total de *reviews*
Percentagem de *Reviews* por Categoria	Indica percentagem de *reviews* em cada categoria	%	nº total de *reviews* da Categora/nº total de *reviews*
Quantidade de *Reviews* por Polaridade	Indica a quantidade de *reviews* classificadas pelas diferentes polaridades	Count	nº total de *reviews* com a polaridade
Intensidade de Sentimentos por Categoria	Indica a intensidade de polaridade de sentimento(valor) em cada categoria	∑	somatório da intensidade de sentimento por categoria
Percentagem de Sentimento Positivo por Categoria	Indica a percentagem de opinioes positivas em cada categoria	%	nº total de polaridade positiva numa categoria/nº total de polaridades que foram identificadas numa categoria
Media das Avaliações das Categorias por Categoria	Indica a média da avaliação dada pelos clientes por categoria	AVG	somatório das Review_stars/quantidade de reviews
Distribuição de Avaliações por Categoria	Indica a quantidade de *reviews* classificados na escala de classificação (1-5) dado pelos clientes às categorias	Count	nº total de cada classificação de cada categoria
Polaridade de Sentimento por Categoria	Demonstra as distribuições das polaridades de cada categoria	∑	somatório dos value_of_sentiment tendo em conta a categoria

TABELA 20: MÉTRICAS DO *DASHBOARD* GESTÃO DE SENTIMENTOS CATEGORIAS.

Passando agora para o *dashboard* do processo de gestão dos sentimentos dos termos pode-se dizer que permitirá ter outra perspetiva para auxiliar o turismo, uma vez que permite obter o sentimento individual mais focado na definição das críticas e expectativas locais. Por outras palavras, permitirá um maior detalhe, focando-se nos termos individuais que possam gerar fatores de descontentamento ou de satisfação. De seguida apresenta-se a tabela 21 que contém as análises presentes no *dashboard* e a descrição das *queries* associadas às quais pretende responder:

Métricas de Análise	Descrição/*Query*	Tipo	Fórmula
Percentagem de *Reviews por Ano*	Indica a percentagem de *reviews* por ano	%	nº total de *reviews* num ano/nº total de *reviews*
Quantidade de *Reviews*	Indica a quantidade de *reviews* categorizados por Estado e por Cidade	Count	nº total de *reviews*
Percentagem de *Reviews* por Categoria	Indica percentagem de *reviews* em cada categoria	%	nº total de *reviews* da Categora/nº total de *reviews*
Quantidade de *Reviews* por Polaridade	Indica a quantidade de *reviews* classificadas pelas diferentes polaridades	Count	nº total de *reviews* com a polaridade
Intensidade de Sentimentos por Categoria	Indica a intensidade de polaridade de sentimento(valor) em cada categoria	Σ	somatório da intensidade de sentimento por categoria
Percentagem de Sentimento Positivo por Categoria	Indica a percentagem de opinioes positivas em cada categoria	%	nº total de polaridade positiva numa categoria/nº total de polaridades que foram identificadas numa categoria
Media das Avaliações das Categorias por Categoria	Indica a média da avaliação dada pelos clientes por categoria	AVG	somatório das Review_stars/quantidade de reviews
Distribuição de Avaliações por Categoria	Indica a quantidade de *reviews* classificados na escala de classificação (1-5) dado pelos clientes às categorias	Count	nº total de cada classificação de cada categoria
Polaridade de Sentimento por Categoria	Demonstra as distribuições das polaridades de cada categoria	Σ	somatório dos value_of_sentiment tendo em conta a categoria

TABELA 21: MÉTRICAS DO DSS GESTÃO DOS SENTIMENTOS DOS TERMOS.

Por último, é de salientar que em ambos os *dashboards*, em todas as análises é permitido fazer *drill down* tanto de ano para mês, como de estado para cidade.

4.1. *Dashboard* de Gestão de Sentimentos das Categorias

De seguida apresenta-se o *dashboard*da Gestão de Categorias que, numa perspetiva mais segmentada, permite uma análise de quais as categorias encontradas pelo CTM mais relevantes nas várias experiências pelas quais os clientes passaram em cada cidade:

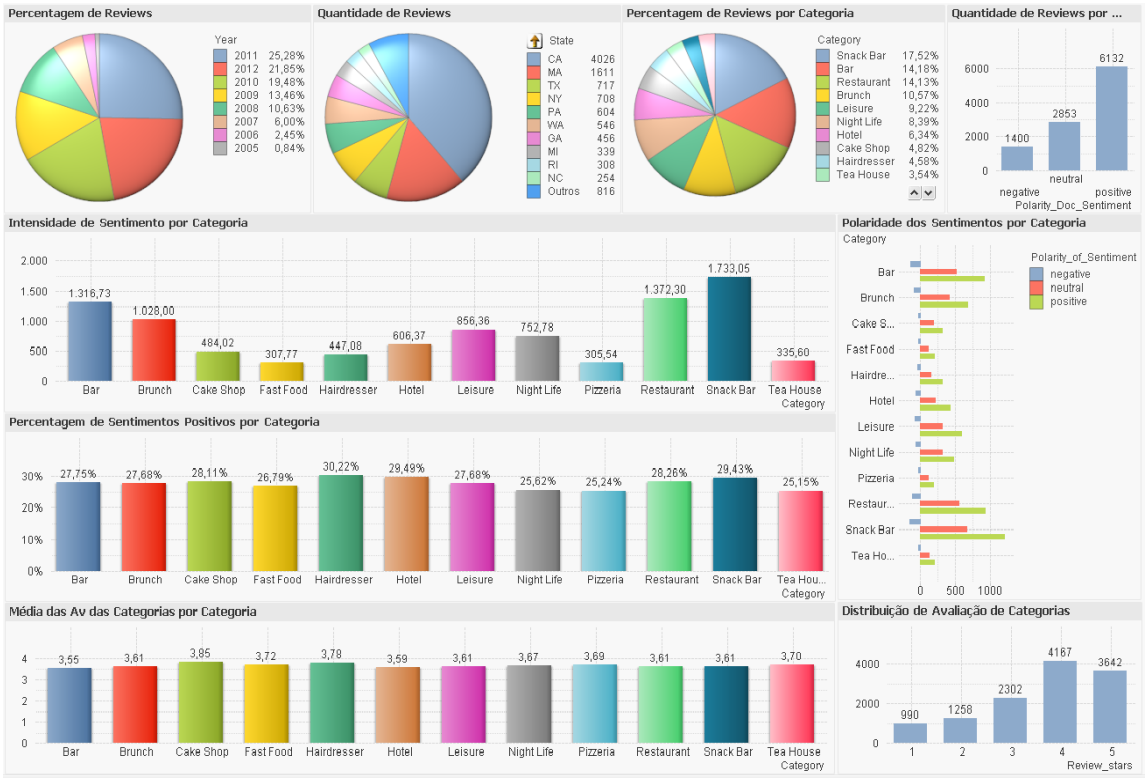


FIGURA 25: *DASHBOARD*DE GESTÃO DE SENTIMENTOS DAS CATEGORIASGERAL.

Atendendo a que o *dataset* disponível correspondia aos anos de 2005 a 2012, optou-se por escolher o primeiro e o último para esta análise com o objetivo de tentar compreender a evolução num período mais alargado possível e que no caso refletisse a realidade antes e após a crise financeira ocorrida nos Estados Unidos da América. Selecionei o estado e a cidade com maior número de *reviews* (CA e Berkeley). Embora esta análise seja parcial, pretende-se que possa demonstrar qual a utilidade de um *dashboard* deste género para o apoio à decisão de negócio.

Apresentamos nas figuras 26 e 27 as percentagens de *reviews* por ano e as quantidades de *reviews* por estado e cidade respetivamente:

FIGURA 26: ANÁLISES PERCENTAGEM DE *REVIEWS* POR ANO E QUANTIDADE DE *REVIEWS* POR ESTADO.

FIGURA 27: ANÁLISES PERCENTAGEM DE *REVIEWS* POR ANO E QUANTIDADE DE *REVIEWS* POR CIDADE.

A evolução das categorias de negócio nos anos considerados (2005 e 2012), serão apresentadas nas figuras 28 e 29:

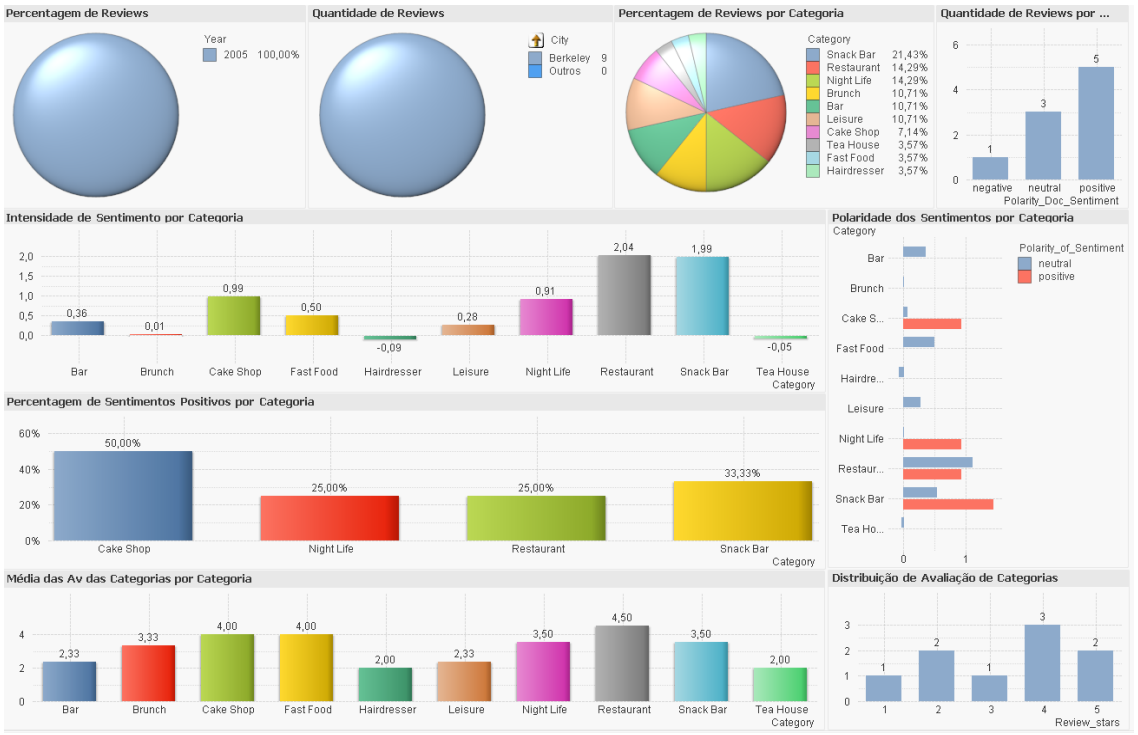


FIGURA 28: ANÁLISES *DASHBOARD* DE GESTÃO DE SENTIMENTOS DAS CATEGORIAS2005.

Em 2005, na figura 28, pode-se verificar no gráfico sobre a "Quantidade de *Reviews*"que existem apenas 9 *reviews* (o que já se tinha sido evidente na figura 26 que em 2005 no geral apenas detinha 0,84% de todas as *reviews* analisadas). Este reduzido número reflete a pouca utilização à época dos *social media* que iniciavam então a sua difusão marcada pelo lançamento de vários sites de *social media*. O *dashboard* da figura 28 permite-nos verificar em mais detalhe quais as especificidades destas *reviews,*

Em termos de polaridade, no gráfico "Quantidade de *Reviews* por Polaridade" da figura 28, dá-nos a indicação que destas 9 opiniões,5 são positivas, 3 são neutras e 1 negativa, o que indica que apesar de poucos opiniões registadas, a maioria detém um sentimento favorável.

No gráfico de "Percentagem de *Reviews* por Categoria", podemos observar que as categorias ***Snack Bar***, ***Restaurant*** e ***Night Life*** são as que têm maior percentagem de *reviews*, respetivamente, 21,43%, 14,29% e 14,29%, enquanto que, as categorias***Fast Food***, ***Hairdresser*** e ***Tea House***, são as que têm menor percentagem de *reviews* (cerca de 3,57%), o que indica que estas categorias não mereceram grande atenção por parte dos "clientes" que usava nesta altura os *social media*. Observa-se aqui também, que neste ano só existiamcomentários relativos a 10 categorias de negócio (***Snack Bar, Restaurant, Night Life, Brunch, Bar, Leisure, Cake Shop, Tea House, Fast Food e Hairdresser).***Depreende-se que um dos *reviews* está associado a mais do que uma categoria.

Nota-se também na figura 28, que as categorias com maior e menor intensidade de sentimento, são respetivamente, ***Restaurant*** (2,04), ***Snack Bar*** (1,99), e ***Tea House*** (-0,05), ***Hairdresser*** (-

0,09) e que as categorias ***Restaurant*** e ***Snack Bar*** têm maior número de *reviews* categorizados com uma polaridade positiva, e também, o maior número de *reviews* em absoluto.

Relativamente ao gráfico "Percentagem de Sentimento Positivo por Categoria", constata-se que as únicas categorias que obtiveram *reviews* positivos foram ***Cake Shop*** (50,00% dos *reviews* que estão categorizados nesta categoria são positivos), ***Snack Bar*** (33,33% dos *reviews* que estão categorizados nesta categoria são positivos), ***Night Life*** (25% dos *reviews* que estão categorizados nesta categoria são positivos) e por último, ***Restaurant*** (25% dos *reviews* que estão categorizados nesta categoria são positivos).

A análise da "Polaridade dos Sentimentos por Categoria" permite-nos concluir que das 10 categorias referenciadas anteriormente, as categorias ***Bar*** (0,357), ***Brunch*** (0,008), ***Fast Food*** (0,5), ***Hairdresser*** (-0,087), ***Leisure*** (0,276), ***Tea House*** (-0,050) contém uma polaridade neutra, tendo as categorias ***Cake Shop*** (0,067 e 0,927), ***Night Life*** (-0,016 e 0,927), ***Restaurant*** (1,110 e 0,927) e ***Snack Bar*** (0,542 e 1,445)polaridades neutra e positiva respetivamente.

O gráfico "Distribuição de Avaliação por Categoria" da figura 28 revela-nos que a classificação dos 9 *reviews* é a seguinte, 1 com 1 estrela, 2 com 2 estrelas, 1 com 3 estrelas, 3 com 4 estrelas e 2 com 5 estrelas.

Por último, observa-se no gráfico "Média das Avaliações das Categorias por Categoria", que as categorias com melhor avaliação são, ***Restaurant*** (4,50), ***Fast Food*** (4,00) e ***Cake Shop*** (4,00), e as categorias pior classificadas são, ***Hairdresser*** e ***Tea House*** (2,00). No entanto, estes valores não podem ser comparados diretamente, uma vez que a média destas avaliações esta relacionada com a quantidade de *reviews*, quantidade esta que apresenta uma grande discrepância entre as categorias.
Embora este quadro apresente para esta análise apenas 9 *reviews*, mostra a capacidade de análise que é possível ter para uma exploração anual dos sentimentos das categorias de negócio como poderemos analisar através de um ano mais recente na Figura 29.

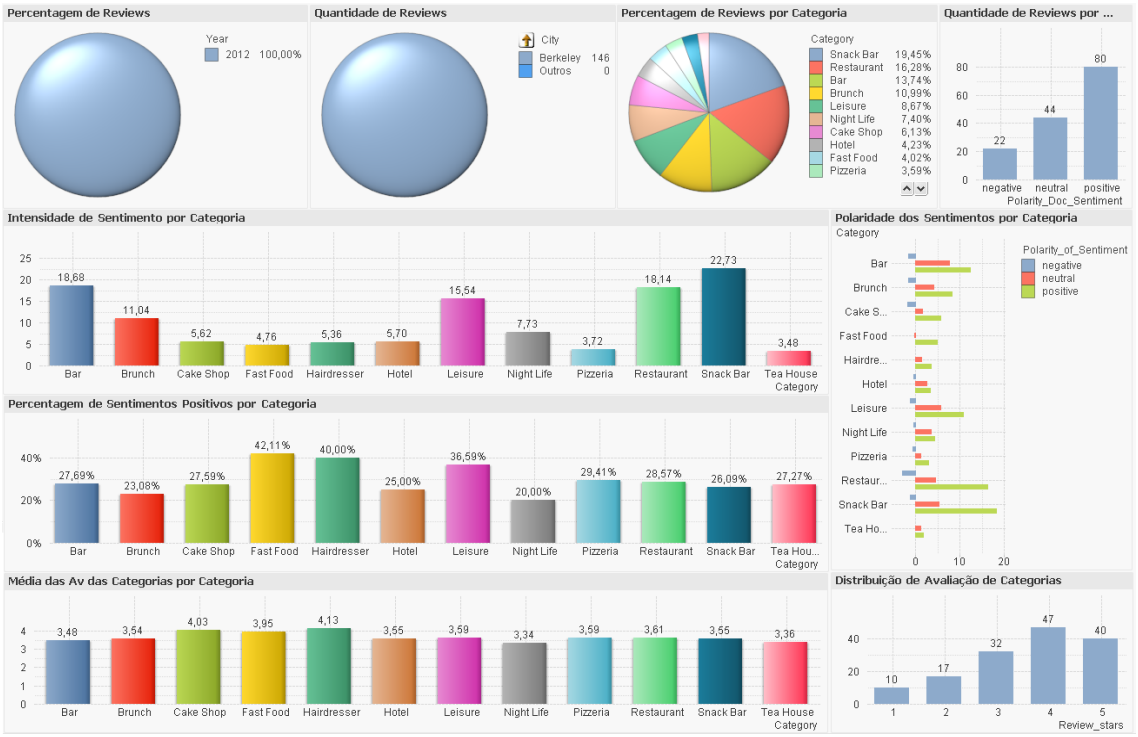


FIGURA 29: ANÁLISES DO *DASHBOARD* DE GESTÃO DE SENTIMENTOS DAS CATEGORIAS2012.

Analisando agora o ano de 2012 na figura 29, regista-se um enorme aumento na participação expressa das opiniões nos *social media*, de 9 comentários em 2005 passamos para 146 em 2012. Esta alteração assaz expressiva revela a par da introdução da utilização dos *social media*, o interesse que a comunidade reconheceu na divulgação e importância da sua própria opinião.

Em relação à polaridade verificou-se que desses 146 comentários, 22 são negativos, 44 neutros e 80 positivos. Nesta análise pode-se deduzir que mais de metade dos comentários registados detém um sentimento favorável.

Analisando as opiniões por categoria,gráfico “Percentagem de*Reviews* por Categoria”, constata-se que todas as categorias foram referenciadas em algum *review*, tendo a categoria***Snack Bar***obtido a maior percentagem, 19,45%, e a ***Tea House*** a menor percentagem, 2,33% de referências.

Observando a intensidade de sentimento, gráfico “Intensidade de Sentimento por Categoria”, figura 29, verifica-se que as categorias com maior intensidade de sentimento neste ano foram a categoria ***Snack Bar*** (22,73), ***Bar*** (18,68) e ***Restaurant*** (18,14), enquanto que com menor intensidade de sentimento podemos observar a categoria***Tea House*** (3,48) e ***Pizzeria*** (3,72). Pode-se concluir que, tanto as categorias ***Snack Bar***, ***Restaurant*** e ***Bar*** têm maior número de *reviews* categorizados com um valor de sentimento positivo, e também, devido a ter maior número de *reviews*.Por outro lado pode-se verificar, que as categorias que obtiveram uma maior intensidade de sentimento, isto é,que obtiveram uma maior satisfação tanto nos negócios, como nos serviços e no atendimento, optaram por dar o seu testemunhofavorável, por oposição às categorias com menor intensidade que optaram por deixar comentários menos

favoráveis.Analisando agora o gráfico "Percentagem de sentimento positivo por categoria", verifica-se que as categorias que obtiveram maior percentagem de *reviews* positivos, foram ***Fast Food*** (42,11%), ***Hairdresser*** (40,00%) e ***Leisure*** (36,59%), enquanto que com menos percentagem de *reviews* positivos podemos observar a categoria ***Hotel*** (25,00%), ***Brunch*** (23,08%) e por último, ***Night Life*** (20,00%). Pode-se constatar que, as categorias que obtiveram maior percentagem de *reviews* positivos, apesar de não serem das categorias com maior número de opiniões (***Fast Food*** 19, ***Hairdresser*** 15 e ***Leisure*** com 41), os seus comentários são maioritariamente positivos.

Analisando o gráfico "Polaridade dos Sentimentos por Categoria", obtêm-se os seguintes resultados:

	Polaridade dos Sentimentos		
Categorias	**Positivo**	**Neutro**	**Negativo**
Snack Bar	≈ 18,464	≈ 5,590	≈ -1,323
Restaurant	≈ 16,517	≈ 4,715	≈ -3,091
Bar	≈ 12,527	≈ 7,965	≈ -1,813
Leisure	≈ 10,968	≈ 5,983	≈ -1,323
Brunch	≈ 8,383	≈ 4,30	≈ -1,612
Cake Shop	≈ 5,882	≈ 1,711	≈ -1,976
Fast Food	≈ 5,143	≈ -0,378	
Night Life	≈ 4,466	≈ 3,862	≈ -0,600
Hairdresser	≈ 3,841	≈ 1,521	
Hotel	≈ 3,548	≈ 2,751	≈ -0,601
Pizzeria	≈ 3,111	≈ 1,342	≈ -0,736
Tea House	≈ 2,034	≈ 1,443	

TABELA 22: POLARIDADE DOS SENTIMENTOS POR CATEGORIA.

Olhando para o gráfico "Distribuição de Avaliação por Categoria", que também se encontra na figura 29, constata-se que da totalidade das opiniões/comentários (146), 40 obtiveram cinco estrelas, 47 obtiveram quatro estrelas, 32 obtiveram 3 estrelas, 17 obtiveram 2 estrelas e por ultimo, 10 obtiveram uma estrela o que revela que, mais de metade dos *reviews,* cerca de 119 *reviews*, tem uma classificação positiva (acima de 3 estrelas para cima). Por outro lado, pode-se dizer que a maioria das opiniões nesta cidade, **Berkeley**, sobre as categorias em relações aos serviços, aos sítios e ao atendimento sãosatisfatórios.

Por último, no gráfico "Média das Avaliações das Categorias por Categoria" da figura 29, observa-se que as categorias com melhor avaliação são, ***Hairdresser*** (4,13), ***Cake Shop*** (4,03) e ***Fast Food*** (3,95), enquanto que, com pior avaliação encontra-se a categoria ***Night Life*** (3,34). De relembrar, que estas medias não são muito fidedignas devido a diversidade de quantidade de *reviews* por categoria.

Ao observar o DSS pode-se apurar que no presente contexto as categorias (serviços, negócios,atendimento) tem bastante utilidade já que podem servir como apoio às decisões fulcrais de investimento para o desenvolvimento do turismo, conseguindo assim perceber os pontos fortes e os fracos, levando em consideração as expectativas dos clientes que juntamente

com os *social media* são peças fundamentais para o investimento devido à opinião e repercussão.

4.2. *Dashboard* de Gestão de Sentimentosdos Termos

De seguida apresenta-se o *dashboard* analisando os termos:

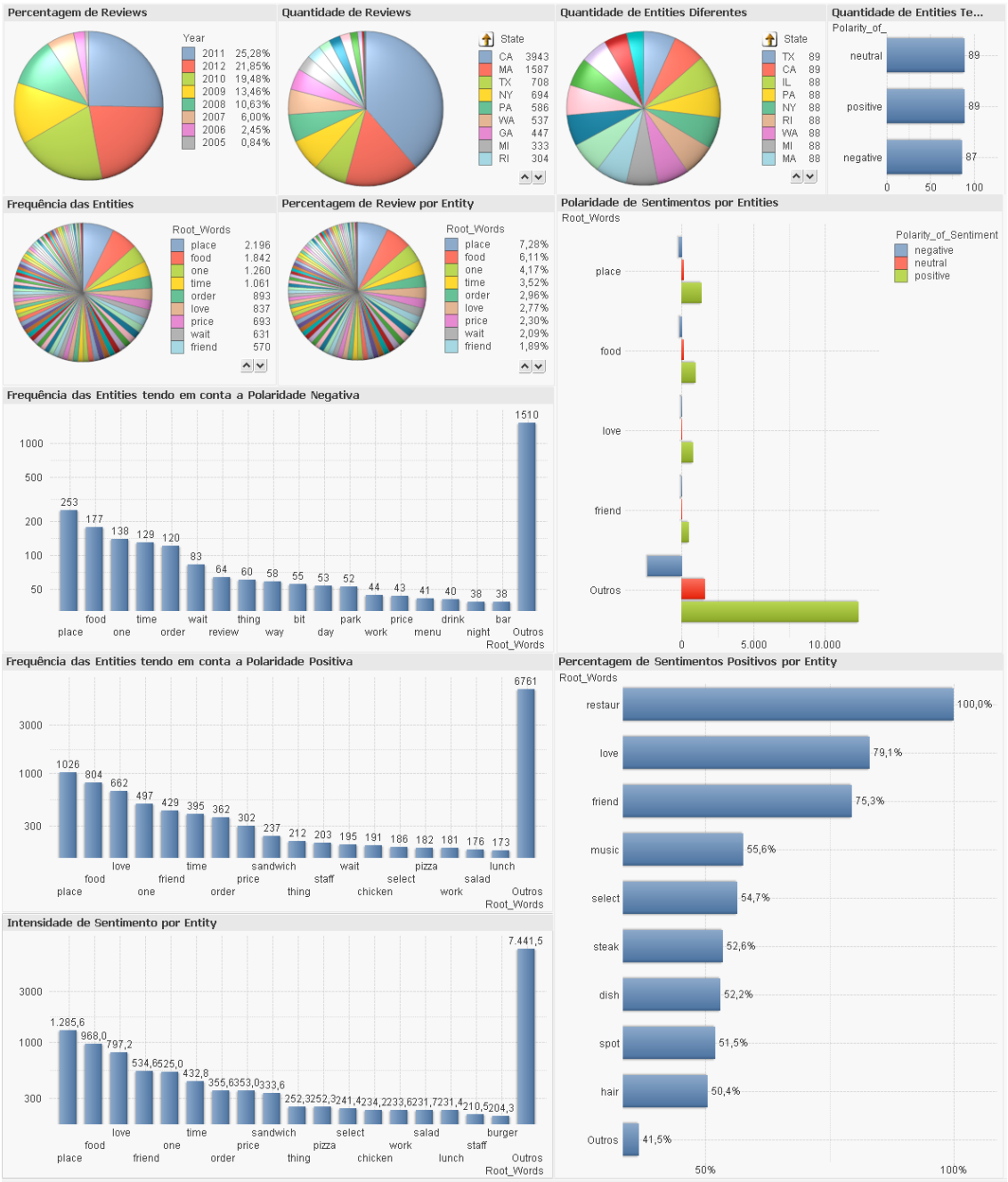


FIGURA 30: *DASHBOARD* DE GESTÃO DE SENTIMENTOS DOS TERMOS GERAL.

Ao observar os resultados de uma forma geral presentes na figura 30, optou-se por manter a coerência/concordância das análises feitas anteriormente, de 2005 a 2012, observando a evolução, bem como o estado com maior percentagem de *reviews*, neste caso **Califórnia (CA)**, analisando por sua vez a cidade com maior percentagem de *reviews*, **Berkeley**. As figuras31 e

32 apresentam a percentagem de *reviews* por ano e a quantidade de *reviews* por estado e por cidade:

FIGURA 31: ANÁLISES PERCENTAGEM DE *REVIEWS* POR ANO E QUANTIDADE DE *REVIEWS* POR ESTADO.

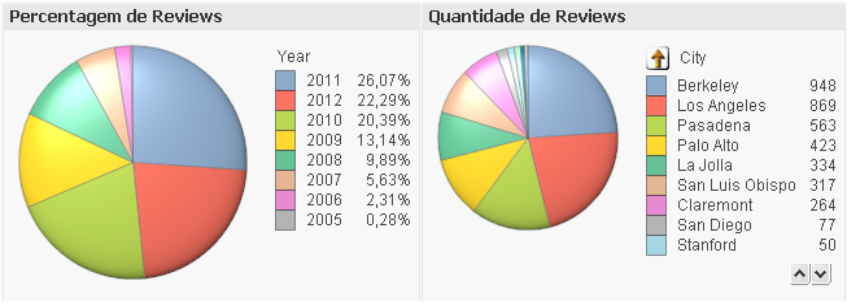


FIGURA 32: ANÁLISES PERCENTAGEM DE *REVIEWS* POR ANO E QUANTIDADE DE *REVIEWS* POR CIDADE.

De seguida apresentam-se as análises nos dois anos escolhidos (2005 e 2012), para se observar a evolução das categorias de negócio relativamente aostermos:

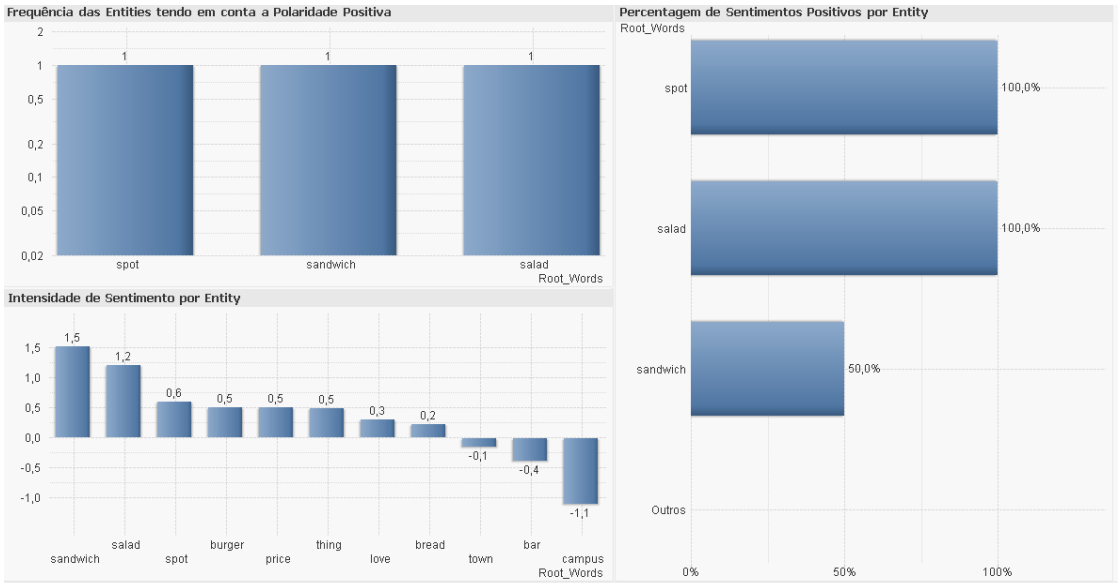


FIGURA 33: ANÁLISES DO *DASHBOARD* DE GESTÃO DE SENTIMENTOS DOS TERMOS2005.

Como indica a figura 32, em 2005, foram identificadas 8 *reviews* com termos identificados, como se pode observar pelo gráfico “Quantidade de *Reviews*”.

No gráfico “Quantidade de *Entities* Diferentes” observa-se a existência de 18 termos diferentes que foram detetados na cidade de**Berkeley**. Relativamente à quantidade de termos tendo em conta a sua polaridade, pode-se concluir através do respetivo gráfico, foram identificados 15 termosdiferentes com uma polaridade neutra, 3 com polaridade positiva e 1 com polaridade negativa. Observa-se ainda, que um dos termos obteve mais do que uma polaridade. Posto isto, ao visualizar os gráficos de “Frequência de *Entities*” e “Percentagem de *Review* por *Entities*”, verifica-se que, as três palavras mais frequentes e com maior percentagem são, ***bread*** (3vezes e 13,64%), ***bar*** (2vezes e 9,09%) e ***sandwich*** (2 vezes e 9,09%), as outras só foram identificadas uma única vez. Observando os 3 termos deduz-se que é a ***sandwich***que contém mais que uma polaridade.

De seguida observa-se as diferentes polaridades dotermo***sandwich,*** na figura 34:

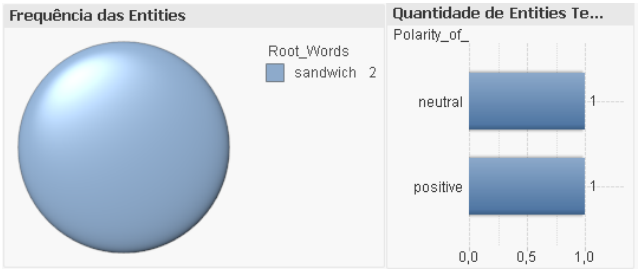


FIGURA 34: POLARIDADE DA ENTITY.

Analisando os gráficos “Polaridade de Sentimento por *Entities*”, “Frequência das *Entities* tendo em conta a Polaridade Negativa”, “Frequência das *Entities* tendo em conta a Polaridade Positiva” e a “Percentagem de Sentimentos Positivos por *Entity*”, constata-se que, dos 18 termosexiste uma onde foram identificadas duas polaridades, ***sandwich*** (obtendo uma percentagem de sentimento positivo de 50%), ***campus*** com uma polaridade negativa de 100%, e ***spot*** e ***salad*** com uma polaridade positiva com uma percentagem de sentimento positivo de 100% e restantes encontram-se com uma polaridade neutra.

Por último, analisando o gráfico de “Intensidade de Sentimento por *Entity*”, verifica-se que os termos com maior intensidade de sentimento, são, ***sandwich***(1,5) e ***salad***(1,2), e as que apresentam menor intensidade são,***bar***(-0,4) e ***campus***(-1,1), como se pode observar de seguida:

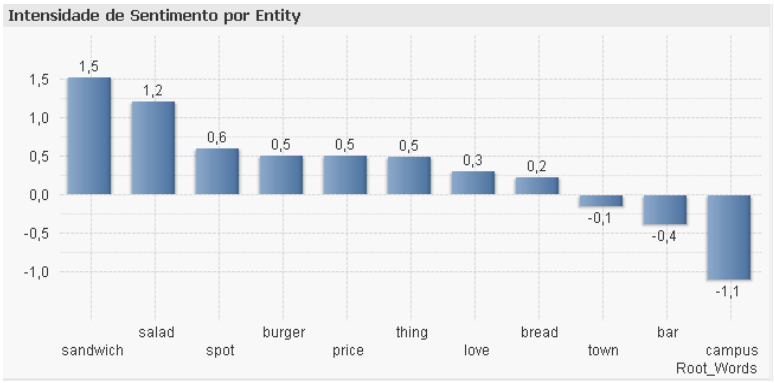


FIGURA 35: ANÁLISE DA INTENSIDADE DE SENTIMENTO.

Passando a analisar o ano de 2012, presente na figura 36, verifica-se que no gráfico de “Quantidade de *Reviews*”, existem 143 *reviews* nos quais foram identificados termos.Mais uma vez nota-se que o número de *reviews* existentes aumentou relativamente ao ano de 2005, tal como se tinha verificado na análise dos sentimentos das categorias.

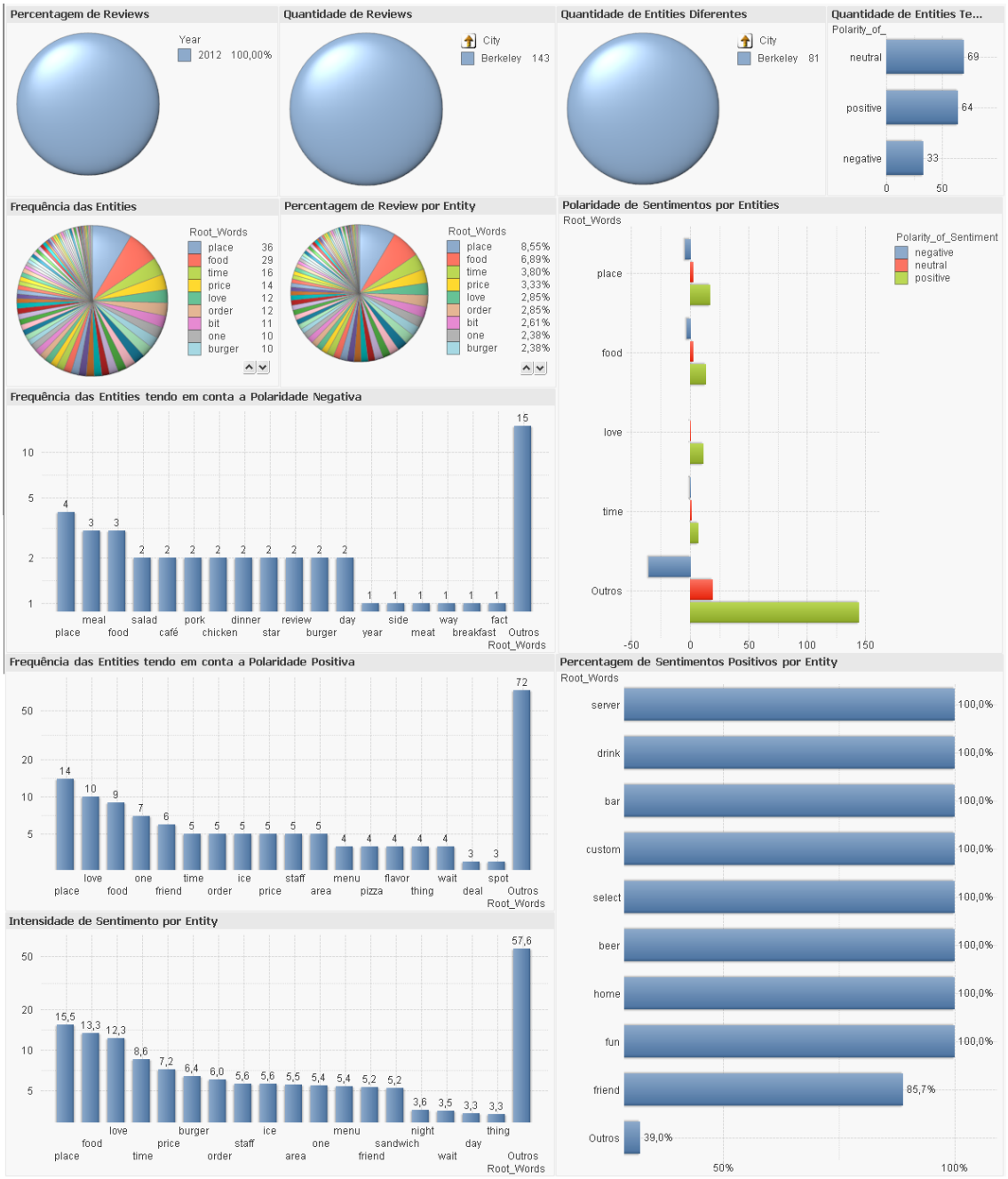


FIGURA 36: ANÁLISES DO *DASHBOARD* DE GESTÃO DE SENTIMENTOS DOS TERMOS2012.

No gráfico "Quantidade de *Entities* Diferentes" existem 81termos diferentes detetados em **Berkeley**. Observando agora o gráfico de "Quantidade de *Entities* tendo em conta a polaridade", verifica-se, que, foram identificados 69 termos com polaridade neutra, 64 com polaridade positiva e 33 com polaridade negativa. Existem também, 73 termos com mais do que uma polaridade. Por último, pode-se constatar que apesar de terem sido identificados poucos termos, 81, a maioria dos termos detém um sentimento satisfatório.

Posto isto, ao visualizar os gráficos de "Frequência de *Entities*" e "Percentagem de *Review* por *Entities*", verifica-se que, os três termos mais frequentes e com maior percentagem são, ***place***

(36 e 8,55%), ***food*** (29 e 6,89%) e ***time*** (16 e 3,80%), enquanto que, as que só foram identificadas uma única vez foram, ***server***, ***dish***, ***stuff***, ***steak, plate, water***, ***music***, ***home***, ***hour***, ***select***, ***point***, ***bar*** e ***custom*** com uma "percentagem de *review* por *Entity"* de 0,24%.

Os resultados obtidos devem-se, com naturalidade, ao facto da maioria dos comentários retirados da amostra deste ano estarem relacionados com as categorias/negócios ***Snack Bar***, ***Restaurant***, ***Bar*** e ***Brunch***.De seguida apresentam-se os gráficos onde se pode observar os termoscom menor frequência e percentagem (figura 37):

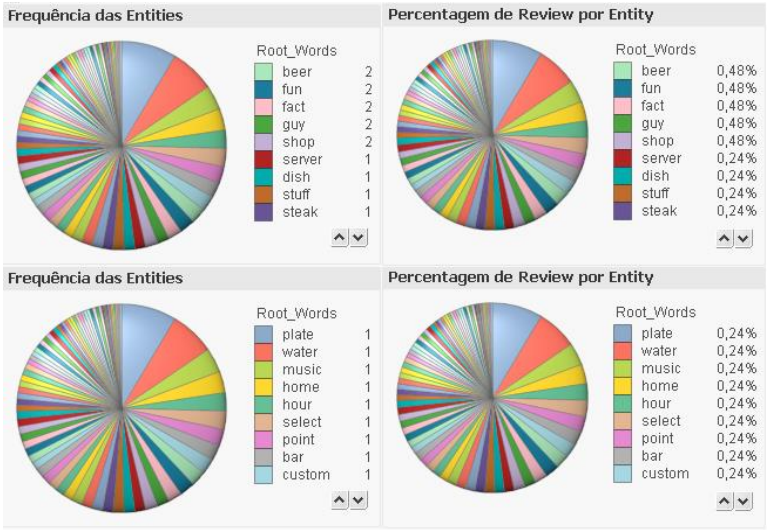


FIGURA 37: FREQUÊNCIA MÍNIMA DAS *ENTITIES* E AS PERCENTAGENS DE *REVIEWS*RESPETIVAS.

Observando os 3 termosmais frequentes, pode-se deduzir que contêm as três polaridades (devido naturalmente, a que se encontram em contextos diferentes).
Apresentam-se de seguida os gráficos de "Frequência das *Entities*" e "Quantidade de *Entities* tendo em conta a Polaridade" para os três termosmais frequentes:

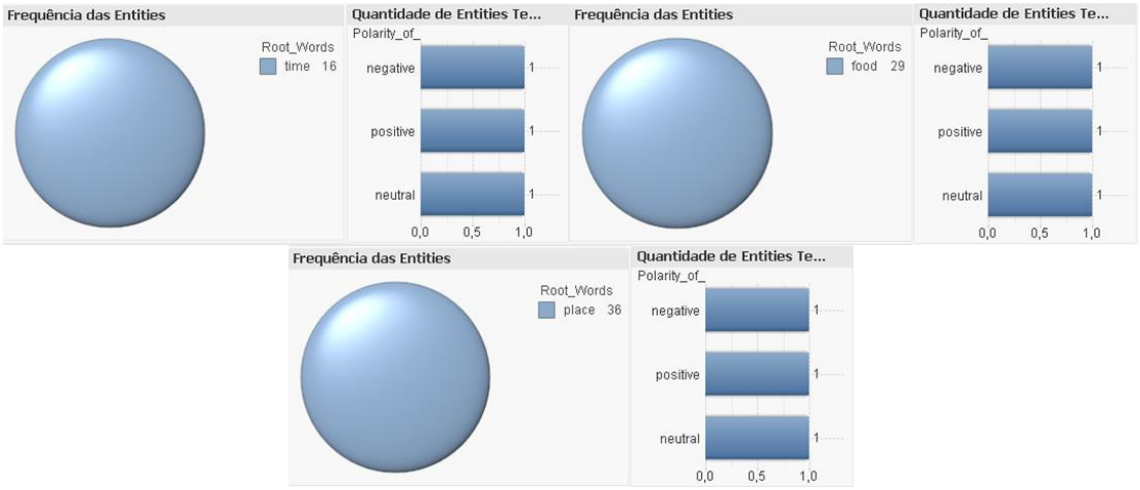


FIGURA 38: POLARIDADE DA *ENTITY*.

Analisando os gráficos “Frequência das *Entities* tendo em conta a Polaridade Negativa”, “Frequência das *Entities* tendo em conta a Polaridade Positiva” e o “Percentagem de Sentimentos Positivos por *Entity*”, constata-se que, dos 81 termos, 3 termos tem maior frequência de polaridade negativa – ***place*** (4), ***meal*** (3) e ***food*** (3) –, e 3 termoscom maior frequência com polaridade positiva – ***place*** (14), ***love*** (10) e ***food*** (9) –.Isto deve-se naturalmente ao facto do tópico de assunto da maioria dos comentários serem sobre restauração, ***Snack Bar***, ***Bar*** e ***Brunch*** (cerca de 60,46% dos *reviews*), o que despoleta uma maior frequência destes termos tanto dum ponto de vista positivo como negativo.

Relativamente à “Percentagem de Sentimento Positivo por *Entity*”, existem 8 termos*com* a percentagem de100% – ***server***, ***drink***, ***bar***, ***custom***, ***select***, ***beer***, ***home*** e ***fun*** –, uma vez que todas as ocorrências de cada uma destes termossão somente positivas. A menor percentagem de sentimento positivo (9,1%) ocorre notermo***bit***, este valor reduzido deve-se a uma quantidade reduzida de polaridade positiva neste termo tendo em conta a quantidade de ocorrências. Neste caso o termo***bit*** é identificado cerca de 11 vezes, no entanto só uma vez é que é identificado com a polaridade positiva e outra negativa, o resto tem uma polaridade neutra o que acaba por provocar uma redução na percentagem. De seguida apresentam-se os gráficos “Frequência das *Entities* tendo em conta a Polaridade Negativa”, “Frequência das *Entities* tendo em conta a Polaridade Positiva” e o “Percentagem de Sentimentos Positivos por *Entity*” relativos ao termo***bit***:

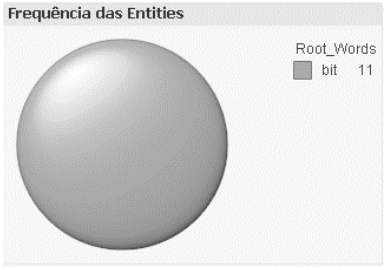


FIGURA 39: FREQUÊNCIA DA *ENTITY*.

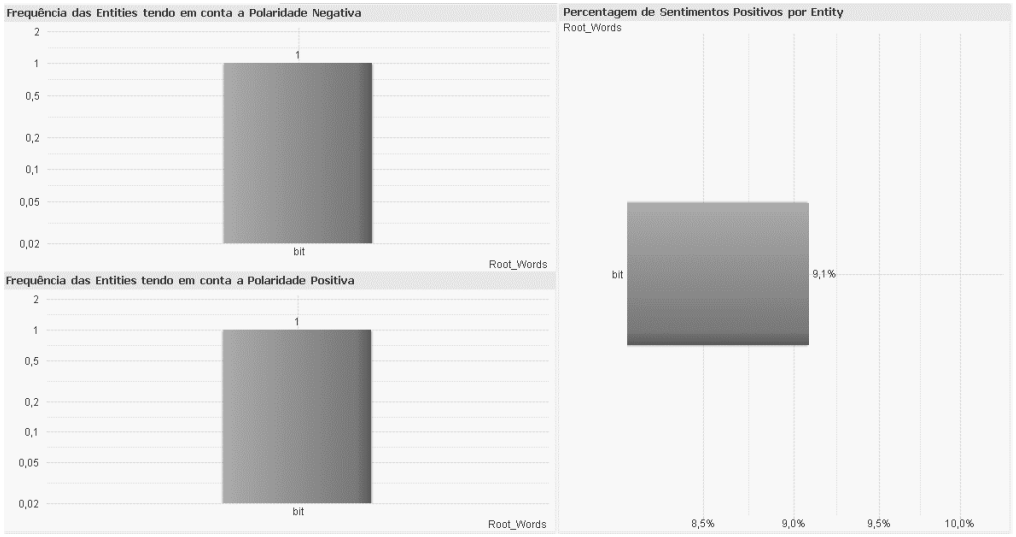


FIGURA 40: FREQUÊNCIA DA *ENTITY* SEGUNDO A POLARIDADE E A FREQUÊNCIA DAS *ENTITIES* TENDO EM CONTA A POLARIDADE POSITIVA.

De seguida, para analisar o gráfico de "Polaridade de Sentimento por *Entity*", foram utilizados os três termosmais frequentes, que foram identificadas anteriormente e que podem ser observadas nas figuras 41.

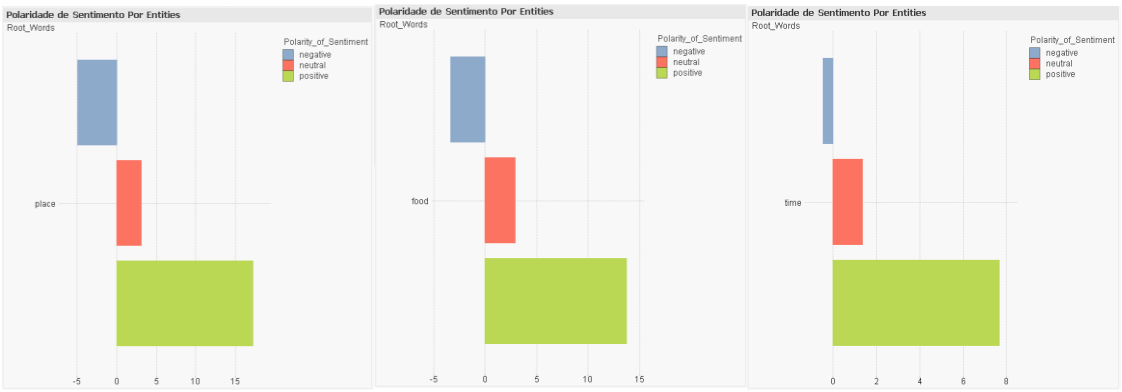


FIGURA 41: POLARIDADE DE SENTIMENTO POR *ENTITIES*.

Analisando a figura anterior, da análise "Polaridade de Sentimento por *Entities*",verificamos *que,* das 36 vezes que foi referida otermo***place*** nos *reviews*, foram-lhe identificadas as três polaridades de sentimento – positivo (17,27),neutro (3,20) e negativo (0,510) –. Conclui-se que apesar de ter sido identificado em diversas polaridades, na maioria das vezes que foi utilizada foi com um sentimento positivo.

Passando agora para a análise do termo***food***, constatou-se que, das 29 vezes que foi referida o termonos *reviews*, foram-lhe identificadas as três polaridades de sentimento – positivo (13,790) neutro (2,965) e negativo (-3,423) –. Desta forma, pode-se deduzir que apesar de ter sido identificado em diversas polaridades, na maioria foi identificado sentimento positivo.

Por último, analisando o termo***time***, das 16 vezes que foi referida nos *reviews*, também lhe foi identificado as três polaridades – positivo (7,690), neutro (1,390) e negativo (-0,490) –. Conclui-se que este termona maioria das ocorrências tem uma polaridade positiva.

Por último, analisando a figura 42de "Intensidade de Sentimento por *Entity*" pode-se visualizar os termoscom maior intensidade de sentimento, – ***place*** (15,3), ***food***(9,9) e ***love*** (8,6) – e os termoscom menor intensidade de sentimento são, –***meal***(-1,4), ***reviews*** (-1,4) e ***pork***(-1,7). Pode-se concluir que ***place***, ***food***e ***love***têm uma pequena frequência categorizada com um valor de sentimento negativo, ou seja, estes termos foram utilizados maioritariamente de forma positiva. Por outro lado pode-se verificar, que os termos que obtiveram uma menor intensidade de sentimento,***meal***, ***reviews*** e ***pork***é devido naturalmente às aparições desses termos de forma maioritariamente negativa.

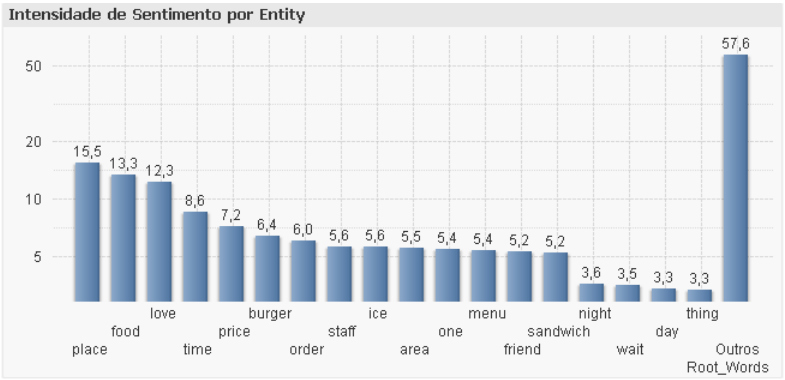


FIGURA 42: INTENSIDADE DE SENTIMENTO POR *ENTITY*.

Como conclusão dos resultados obtidos anteriormente no DSS, apresenta-se de seguida, as tabelas resumo dos resultados alcançados:

Resultados de 2005							
				Polaridade de Sentimento			
Categirias de Negócio	% de Reviews	Intensidade de Sentimento	% Sentimento Positivo	Positivo	Neutro	Negativo	Media das av. Categorias
Bar	10,71%	0,36			x		2,33
Pizzeria	-	-	-	-	-	-	-
Snack Bar	21,43%	1,99	33,33%	x	x		3,50
Brunch	10,71%	0,01			x		3,33
Fast Food	3,57%	0,50			x		4,00
Cake Shop	7,14%	0,99	50%	x	x		4,00
Tea House	3,57%	-0,05			x		2,00
Restuarant	14,29%	2,04	25%	x	x		4,50
Hairdresser	3,57%	-0,09			x		2,00
Night Life	14,29%	0,91	25%	x	x		3,50
Hotel	-	-	-	-	-	-	-
Leisure	10,71%	0,28			x		2,33

TABELA 23: TABELA RESUMO DOS RESULTADOS OBTIDOS NO *DASHBOARD* DE GESTÃO DE SENTIMENTOS DAS CATEGORIAS EM 2005.

Resultados de 2012							
				Polaridade de Sentimento			
Categirias de Negócio	% de Reviews	Intensidade de sentimento	% Sentimento Positivo	Positivo	Neutro	Negativo	Media das av. Categorias
Bar	13,74%	18,68	27,69%	x	x	x	3,48
Pizzeria	3,59%	3,72	29,41%	x	x	x	3,59
Snack Bar	19,45%	22,73	26,09%	x	x	x	3,55
Brunch	10,99%	11,04	23,08%	x	x	x	3,54
Fast Food	4,02%	4,76	42,11%	x	x		3,95
Cake Shop	6,13%	5,62	27,59%	x	x	x	4,03
Tea House	2,33%	3,48	27,27%	x	x		3,36
Restuarant	16,28%	18,14	28,57%	x	x	x	3,61
Hairdresser	3,71%	5,36	40,00%	x	x		4,13
Night Life	7,40%	7,73	20,00%	x	x	x	3,34
Hotel	4,23%	5,70	25,00%	x	x	x	3,55
Leisure	8,67%	15,54	36,59%	x	x	x	3,59

TABELA 24: TABELA RESUMO DOS RESULTADOS OBTIDOS NO *DASHBOARD* DE GESTÃO DE SENTIMENTOS DAS CATEGORIAS EM 2012.

Resultados 2005						
	Positivo	Neutro	Negativo	Total	Escala de Avaliação	Quantidade de Reviews avaliados
Quantidade de R*eviews*	5	3	1	9	1	1
					2	2
					3	1
					4	3
					5	2

Resultados 2012						
	Positivo	Neutro	Negativo	Total	Escala de Avaliação	Quantidade de Reviews avaliados
Quantidade de R*eviews*	80	44	22	146	1	10
					2	17
					3	32
					4	47
					5	40

TABELA 25: TABELA RESUMO DOS RESULTADOS OBTIDOS NO *DASHBOARD* DE GESTÃO DE SENTIMENTOS DAS CATEGORIAS EM 2005 E 2012.

Ao observar os resultados obtidos no *dashboard* de Gestão de Sentimentos das Categorias através da análise das tabelas 23, 24 e 25 observa-se um progresso da quantidade de *reviews* na cidade **Berkeley**, uma vez que, em 2005 apenas existiam 9 *reviews*, e em 2012 já foram identificados 146*reviews*na amostra utilizada. De uma forma geral em 2005os *reviews* identificados foram classificados na sua polaridade – positivo (5), neutro (3) e negativo (1) –. Já em 2012, dos 146 *reviews*, foram identificados – positivo (80), neutro (44) e negativo (22) –. Quanto à distribuição de avaliação das categorias, observa-se que em 2005, 1 *review* foi avaliado com 3 estrelas, 3 foram avaliados com 4 estrelas e 2 foram avaliados com 5 estrelas. Em 2012, 32 *reviews* foram avaliados com 3 estrelas, 47 *reviews* que foram avaliados com 4 estrelas e, por último, 40 *reviews* que foram avaliados com 5 estrelas. Pode-se deduzir que, houve uma evolução de 2005 para 2012 no sentido de haver um aumento significativo na utilização dos *social media*. Por outro lado, pode-se dizer que a maioria das opiniões nesta cidade, **Berkeley**, sobre as categorias em relações aos serviços, aos sítios e ao atendimento são satisfatórios. Nesta análise, ao escolher a categoria que se pretende analisar minuciosamente, pode-se observar a distribuição de todos os comentários obtidos. É útil esta análise, uma vez que, pode-se observar a distribuição das opiniões numa categoria de negócio, podendo melhorar caso a distribuição esteja com piores resultados.

Verifica-se que, em 2005 pouco mais de metade dos *reviews*, cerca de 66,67%, foram avaliados com uma pontuação positiva, enquanto que, em 2012, cerca de 81,51% dos *reviews* foram avaliados positivamente. Para além do aumento do número de *reviews*, nota-se uma maior incidência nos valores de avaliação das categorias de 3,4 e 5, o que significa que mais de metade dos *reviews* sobre as categorias tem uma avaliação positiva.

Pode-se constatar com esta análise, que essa evolução, positiva, demonstra que as categorias de negócio cada vez mais se aproximam das expectativas dos turistas.

Em 2005, só existiam *reviews* de 10 categorias de negócio, enquanto que, em 2012 já existiam *reviews* para as 12 categorias. Pode-se também observar, que os dois cenários (2005 e 2012) têm em comum as categorias com maior percentagem de *reviews* ***Snack Bar*** e

Restaurant.Conclui-se que, de uma forma geral, os turistas que viajam para Berkeley apreciammaioritariamente os serviços de ***restauração*** e ***snack bar***.

Para se identificar as categorias que vão mais de encontro às expectativas dos turistas, têm de se ter em conta a percentagem de sentimento positivo e a percentagem de *reviews*, uma vez que, a percentagem do sentimento positivo advém da percentagem de *reviews* em cada categoria.

Observando de uma forma geral as análises de 2005, poder-se-ia dizer que uma das categorias que os turistas mais valorizam seria ***Cake Shop***, uma vez que, tem a percentagem de sentimento positivo mais elevada, de cerca de 50,00%, no entanto, esses 50,00% advém de uma percentagem pequena de *reviews*, isto é, 7,34%, o que revela que esta não é a categoria de negócio que os turistas mais valorizam.

Posto isto, em 2005, de uma forma sucinta, pode-se constatar que as categorias que os turistas mais valorizam e que vão mais de encontro às suas expectativas são, ***Snack Bar*** (com 21,43% de *reviews* em que 33,33% tem sentimento positivo), ***Restaurant*** (com 14,29% de *reviews* em que 25,00% tem sentimento positivo) e ***Night Life*** (14,29% de *reviews* em que 25,00% tem sentimento positivo) –.

Passando agora para a análise de 2012, de uma forma resumida, deduz-se, que as categorias que os turistas mais valorizam, ou que vão mais de encontro com as expectativas dos turistas, são –***Snack Bar*** (19,45% de *reviews* em que 26,09% contém sentimento positivo), ***Restaurant*** (16,28% de *reviews* em que 28,57% tem sentimento positivo) e ***Bar*** (13,75% *reviews* em que contém 27,69% de sentimento positivo) –.

Relativamente ao referido, constata-se que, as categorias que vão de encontro às expectativas em 2005, duas delas, mantem-se em 2012, –***Restaurant*** e ***Snack Bar***–, surgindo uma terceira em 2012, – ***Bar***–, tendo também bastante força em relação às expectativas dos turistas.

Em relação às médias relativas dos comentários, não se consegue ter uma perceção clara, uma vez que, a quantidade de *reviews* por categorias é muito díspar, o que provoca resultados não muito fiáveis.

Por último, relativamente à intensidade de sentimentos, está de acordo com o que foi referido anteriormente, sendo em 2005, –***Snack Bar*** (1,99), ***Restaurant*** (2,04) e ***Night Life*** (0,91) — e em 2012, –***Snack Bar*** (22,73), ***Restaurant*** (18,68) e ***Bar*** (18,14) –, havendo uma clara evolução de 2005 para 2012.

Resultados de 2005									
			Polaridade de Sentimento						
Entities Detetadas	Qtdd	% de Review	Positivo	Neutro	Negativo	Freq. Sentimento Negativa	Freq. Sentimento Positiva	% Sentimento Positivo	Intensidade de Sentimento
bread	3	13,64%		x					0,20
bar	2	9,09%		x					-0,40
sandwich	2	9,09%	x	x			1	50,00%	1,50
place	1	4,55%		x					
café	1	4,55%		x					
thing	1	4,55%		x					0,50
spot	1	4,55%	x				1	100,00%	0,60
campus	1	4,55%			x	1			-1,10
price	1	4,55%		x					0,50
plate	1	4,55%		x					
salad	1	4,55%	x				1	100,00%	1,20
stuff	1	4,55%		x					
town	1	4,55%		x					-0,10
breakfast	1	4,55%		x					
love	1	4,55%		x					0,30
lot	1	4,55%		x					
meal	1	4,55%		x					
burger	1	4,55%		x					0,50

TABELA 26: TABELA RESUMO DOS RESULTADOS OBTIDOS NO DASHBOARD DE GESTÃO DE SENTIMENTOS DOS TERMOS EM 2005.

Resultados de 2012									
			Polaridade de Sentimento						
Entities Detetadas	Qtdd	% de Review	Positivo	Neutro	Negativo	Freq. Sentimento Negativa	Freq. Sentimento Positiva	% Sentimento Positivo	Intensidade de Sentimento
place	36	8,55%	x	x	x	4	14	38,90%	15,50
food	29	6,89%	x	x	x	3	9	31,00%	9,00
time	16	3,80%	x	x	x	1	5	31,30%	8,60
price	14	3,33%	x	x			5	35,70%	7,20
love	12	2,85%	x	x			10	83,30%	12,30
order	12	2,85%	x	x	x	1	5	41,70%	6,00
bit	11	2,61%	x	x	x	1	1	9,10%	1,40
one	10	2,38%	x	x	x	1	7	70,00%	5,40
burger	10	2,38%	x	x	x	2	3	30,00%	6,40
area	10	2,38%	x	x			5	50,00%	5,50
sandwich	10	2,38%	x	x			2	20,00%	5,20
menu	9	2,14%	x	x			4	44,40%	5,40
meal	8	1,90%	x	x	x	3	2	25,00%	-1,40
flavor	8	1,90%	x	x	x	1	4	50,00%	3,30
day	7	1,66%	x	x	x	2	3	42,90%	3,30
chicken	7	1,66%	x	x	x	2	1	14,30%	-0,70
friend	7	1,66%	x	x			6	85,70%	5,20
wait	7	1,66%	x	x			4	57,10%	3,50
dinner	7	1,66%	x	x	x	2	1	14,30%	-0,50
pizza	7	1,66%	x	x	x	1	4	57,10%	2,50
work	7	1,66%	x	x			3	42,90%	2,60
side	6	1,43%	x	x	x	1	1	16,70%	-0,20
ice	6	1,43%	x	x			5	83,30%	5,60
staff	6	1,43%	x	x		5		83,30%	5,60
salad	6	1,43%	x	x	x	2	2	33,30%	-0,10
star	6	1,43%	x		x	2	2	33,30%	1,10
year	5	1,19%	x	x	x	1	2	40,00%	3,20
thing	5	1,19%	x		x	1	4	80,00%	3,30
way	5	1,19%	x	x	x	1	2	40,00%	1,90
pork	5	1,19%			x	2			-1,70
deal	5	1,19%	x	x			3	60,00%	2,40
meat	5	1,19%	x	x	x	1	2	40,00%	1,70
night	5	1,19%	x	x			3	60,00%	3,60
spot	5	1,19%	x	x	x	1	3	60,00%	2,20
park	4	0,95%	x	x	x	1	1	25,00%	1,30
door	4	0,95%	x		x	1	3	75,00%	3,20
kind	4	0,95%	x	x			2	50,00%	2,10
café	4	0,95%		x	x	2			-1,70
cream	4	0,95%	x	x			2	50,00%	2,10
reason	4	0,95%	x				3	75,00%	2,80
dessert	4	0,95%	x	x			2	50,00%	1,50
line	4	0,95%	x	x	x	1	1	25,00%	1,60
rice	3	0,71%	x	x			1	33,30%	0,50
street	3	0,71%	x	x			1	33,30%	0,90
campus	3	0,71%	x	x			2	66,70%	1,90
week	3	0,71%	x	x			1	33,30%	1,60
money	3	0,71%	x	x			1	33,30%	0,70
fish	3	0,71%	x	x	x	1	1	33,30%	-0,30
breakfast	3	0,71%		x	x	1			-0,80
lot	3	0,71%		x					0,80
review	3	0,71%			x	2			-1,40
sushi	3	0,71%	x	x			1		2,00
part	2	0,48%	x	x			1	50,00%	1,50
decor	2	0,48%	x	x			1	50,00%	0,60
person	2	0,48%			x	1			-0,60
seat	2	0,48%	x	x			1	50,00%	1,00
tea	2	0,48%	x	x			1	50,00%	0,50
fan	2	0,48%	x	x			1	50,00%	1,00
bread	2	0,48%		x					0,70
drink	2	0,48%	x				2	100,00%	2,00
room	2	0,48%		x					0,90
store	2	0,48%	x				1	50,00%	0,60
soup	2	0,48%	x				1	50,00%	1,90
beer	2	0,48%	x				2	100,00%	1,90
fun	2	0,48%	x				2	100,00%	1,90
fact	2	0,48%	x		x	1	1	50,00%	0,00
guy	2	0,48%		x	x	1			-0,30
shop	2	0,48%	x				1	50,00%	0,70
server	1	0,24%	x				1	100,00%	0,90
dish	1	0,24%			x	1			-1,10
stuff	1	0,24%		x					
steak	1	0,24%		x					0,40
plate	1	0,24%		x					-0,20
water	1	0,24%		x					0,50
music	1	0,24%		x					0,50
home	1	0,24%	x				1	100,00%	1,90
hour	1	0,24%		x					0,20
select	1	0,24%	x				1	100,00%	0,70
point	1	0,24%		x					
bar	1	0,24%	x				1	100,00%	1,30
custom	1	0,24%	x				1	100,00%	0,70

TABELA 27: TABELA RESUMO DOS RESULTADOS OBTIDOS NO DASHBOARD DE GESTÃO DE SENTIMENTOS DOS TERMOS EM 2012.

Resultados 2005			
	Positivo	Neutro	Negativo
Quantidade de *Entities*	3	15	1
Quantidade de Reviews	8		
Quantidade de *Entities*	18		

Resultados 2012			
	Positivo	Neutro	Negativo
Quantidade de *Entities*	64	69	33
Qunatidade de *Reviews*	143		
Quantidade de *Entities*	81		

TABELA 28: TABELA RESUMO DOS RESULTADOS OBTIDOS NO *DASHBOARD* DE GESTÃO DE SENTIMENTOS DOS TERMOS EM 2005 E 2012.

Ao observar as análises, derivadas do sistema de apoio à decisão dos termosatravés dastabelas 26,27 e 28 na cidade de **Berkeley**, para os dois anos que já foram referidos (2005 e 2012), pode-se constatar que, em 2005 (tabela 26), nos 8 *reviews*existentes foram identificados 18 termosdiferentes, que contém uma polaridade – positiva (3), neutra (15) e negativa (1)–. O facto do número de "Quantidade de *Entities* por Polaridade" revela que existem termoscom mais do que uma polaridade. Já em 2012, foram detetados 143, dentro destes foram identificados 81 termosdiferentes. Esses termos foram categorizados segundo a polaridade – positivo (64), neutra (69) e negativo (33) –, ocorrendo exatamente o que foi referido anteriormente. Pode-se constatar que houve uma evolução na utilização dos *social media*, como referido anteriormente, e como se pode observar no aumento dos *reviews*, e no aumento dos termos. Além disso, pode-se constatar que, em ambos os casos, apesar de terem sido identificados poucos termos, a maioria dos termos detém um sentimento satisfatório/neutro. Constatando-se assim, que na primeira análise, em 2005, os "clientes" tinham maior pudor em dar a sua opinião negativamente do que na análise mais recente, 2012,ou seja, houve uma alteração de comportamento dos "clientes" passando a serem mais colaborativos com os *social media*, ajudando assim, para uma melhor decisão, uma vez que a opinião é mais sincera e se consegue perceber melhor o que mais e menos valorizam em cada cidade.É de salientar que tanto em 2005 como em 2012 a polaridade que predomina é a neutra.

Em 2005, observando a frequência dos termos e a "Percentagem de *Reviews* por *Entity"*, constata-se que as com maior frequência são, –***bread*** (3 e 13,64%), ***bar*** (2 e 9,09%) e ***sandwich*** (2 e 9,09) –, enquanto que, em 2012, –***place*** (36 e 8,55%), ***food*** (29 e 6,89%) e***time*** (16 e 3,80) –. Ambas as palavras mais frequentes vão de encontro às categorias de negócio mais frequentes de cada ano.

Em 2005, as três palavras mais frequentes têm uma polaridade neutra, só a ***sandwich*** é que também tem uma polaridade positiva. Já em 2012, os três termos mais frequentes têm os três tipos de polaridade.Os termos com maior frequência e polaridade positiva mantem assim, a

concordância com as categorias com maior intensidade de sentimento em ambos os casos (em 2005 –***Restaurant***(2,04) ***e Snack Bar***(1,99) – e em 2012 –***Snack Bar***(22,73) e ***Restaurant***(18,68) –), uma vez que, observando os comentários onde se encontram estes termos abordam opiniões sobre estas categorias.

Em relação às frequências de sentimento tanto positivo como negativo, constata-se que em 2005, só existe uma palavra positiva e nenhuma com a polaridade negativa, enquanto que, em 2012, ***place*** aparece 4 vezes de forma negativa e 14 vezes positiva, ***food*** é identificada 3 vezes negativa e 9 vezes positiva e, por último, ***time*** é identificada 1 negativamente e 5 positivas. Realça-se que os termos ***place*** e ***food*** são as que tem maior frequência negativa e positiva.

Em relação à percentagem de sentimento positivo em 2005 (tabela 26), a palavra ***sandwich*** aparece com uma percentagem de cerca de 50,00%, no entanto, tem que se entender que esta elevada percentagem está relacionada com o número de vezes que esta ocorre, neste caso ocorre apenas 2 vezes. Já em 2012 (tabela 27), as três palavras mais frequentes, –***place*** (38,90% das 36 vezes que ocorre são positivas), ***food*** (31,00% das 29 ocorrências são positivas) ***time (***31,30% das 16 ocorrências são positivas) –. Reforçando assim, que as categorias e a sua intensidade de sentimento estão intrinsecamente relacionadas com os termos mais frequentes e a sua polaridade. Por outras palavras, pode-se dizer que vão de encontro às expectativas que se criaram nas análises das categorias, maior número de *reviews* e maior intensidade de sentimento –em 2005 ***Snack Bar***(com intensidade de sentimento 1,99), ***Restaurant***(2,04) e em 2012 ***Snack Bar***(com intensidade de sentimento 22,73), ***Restaurant*** (com intensidade de sentimento 18,14) e ***Bar*** (com intensidade de sentimento 18,68)–.

Por último, observando a intensidade de sentimento das palavras mais frequentes, na análise de 2005, nota-se que: –***bread*** (0,20, tendo uma polaridade neutra), ***bar*** (-0,40, tendo polaridade neutra, sendo uma das intensidades mais baixas) e ***sandwich*** (1,5 dos valores mais elevados tendo uma polaridade positiva e neutra) –. Passando para análise de 2012, observa-se que as palavras mais frequentes contêm a seguinte intensidade de sentimento, – ***place*** (15,5, das intensidades mais elevadas), ***food*** (9,00 também sendo uma das mais elevadas) e por último ***time*** (contendo uma intensidade de sentimento 8,6) –.Mantém-se assim a concordância com as análises efetuadas às categorias.

Concluindo a análise, pode-se deduzir que em 2012 as categorias, em Berkeley, que vão encontro às expectativas dos turistas são –***Snack Bar***, ***Restaurant*** e ***Bar***– correspondentes aos termos mais frequentes e com maior intensidade de sentimento.

5. CONCLUSÕES

Esta dissertação centrou-se fundamentalmente no estudo do desenvolvimento do turismo em certas cidades, com o intuito de melhorar a sua oferta, tendo em conta as opiniões dos turistas nos *social media*. O estudo realizado comprova a importância das técnicas de ***Sentiment Analysis*** e do ***Text Minin*****g**para a extração de conhecimento de dados não estruturados e do suporte ao apoio à decisão.

As duas técnicas acima referidas são utilizadas nesta dissertação com o intuito de identificar as preferências dos turistas em cada cidade.

Foi realizada uma revisão da literatura que se focou em nove grandes temas, ***Text Mining***, ***Sentiment Analysis***, ***Tourism***, ***Sentiment Analysis and social media***, ***Text mining****and****social media***, ***text mining and tourism***,***sentiment analysis and tourism, DSS e DSS and tourism***. Após essa revisão, detetou-se a dificuldade de identificação dos maiores determinantes e limitações para os índices de satisfação das experiências nas cidades em termos das opiniões nos *social media*.

Para responder a este problema, foi implementado um sistema de apoio à decisão para facilitar as tomadas de decisões para o desenvolvimento das cidades tendo em conta os gostos dos turistas em termos de categorias de negócio. Achou-se que seria interessante terminar esta dissertação com um DSS, de forma a poder ser uma mais-valia, para os institutos de turismo.

Para esse DSS, foram utilizados os resultados obtidos com as técnicas de ***text mining*** e ***sentiment analysis,*** que permitiram identificar as preferências dos turistas.Os resultados foram obtidos recorrendo, numa primeira fase, ao *software***R** para a etapa do pre-processamento, onde foram realizadas as limpezas, transformações, estruturação dos dados e algumas análises descritivas,como por exemplo a frequência dos termos, que resultaram na construção dos *wordclouds* e a correlação dos termos, o que constitui o primeiro passo para a construção dos tópicos.

Após a etapareferida, optou-se por fazer a análise de sentimentos utilizando o *softwareSemantria*. Este *software* é de grande utilidade nesta área, tendo funcionalidades interessantes e uma fácil utilização e entendimento.O software identificou, ao utiliza-lo, o sentimento de cada comentário, de cada entidade e os tópicos relacionados com cada comentário. Estes resultados que o *Semantria* fornece têm um detalhe fulcral, uma vez que, o sentimento de um comentário de uma forma geral pode ser positivo, no entanto, as entidades ou tópicos identificados no mesmo comentário não o são.

No entanto, foram detetadas algumas lacunas neste *software* a nível de classificações da negação e a nível do *part-of-speech.*

Posteriormente, para a elaboração do DSS (recorrendo posteriormente ao software *QlikView*), foi necessário implementardois modelos dimensionais, um relacionado com os tópicos (Gestão

de Sentimentos das Categorias) e outro com os termos (Gestão de Sentimentos dos Termos), recorrendo aos quatro passos de Kimball.

EsteDSS vai ajudar no desenvolvimento do turismo e do comércio em cada cidade (que tipo de negócio é que se deve investir, quais as categorias de negócio mais fortes em cada cidade, quais as categorias que devem ser melhoradas em cada cidade), melhorando e diversificandoa sua oferta, tendo em conta as opiniões, as suas experiências e expectativas dos turistas nos *social media*. Com o DSS pode-se também, perceber a afluência e quais as razões manifestadas através dos sentimentos e das categorias.

Existiram algumas limitações neste estudo, uma delas foi relativamente à amostra uma vez que a maioria dos *reviews* eram relacionados com a alimentação, devido a esse motivo também existiram bastantes dificuldades em fazer o *profilling* dos tópicos.

Por último, para trabalhos futuros, achou-se que seria extremamente interessante e importante a sua implementação em institutos de turismo, como no Instituto do Turismo de Portugal,para ajudar nas decisões sobre que tipo de negócio que se devem expandir em cada cidade bem como identificar os aspetos que os turistas mais valorizam em cada cidade, de acordo com o tipo de turistas que viajam para cada cidade. Por outras palavras, a implementação de um sistema destes potencializará o turismo e aumentará o seu desenvolvimento, uma vez que, este sector está em contínua ascensão.

6. BIBLIOGRAFIA

(UNWTO), W. T. O. (2013). *UNWTO Annual Report 2013* (pp. 1–45).

Abeywardena, I. S. (2014). Public Opinion on OER and MOOC : A Sentiment Analysis of Twitter Data, 1–6.

Akehurst, G. (2009). User generated content: The use of blogs for tourism organisations and tourism consumers. *Service Business*, *3*(1), 51–61.

Amaral, F., Tiago, T., & Tiago, F. (2014). User-generated content : tourists ' profiles on TripAdvisor, *01*, 137–147.

Anderson, E. W. (1998). Customer Satisfaction and Word of Mouth. *Journal of Service Research*, *1*(1), 5–17.

Andreu, J., Capilla, J., & Sanchís, E. (1996). AQUATOOL, a generalized decision-support system for water-resources planning and operational management. *Journal of Hydrology*, *177*(3-4), 269–291.

Aston, N., Liddle, J., & Hu, W. (2014). Twitter Sentiment in Data Streams with Perceptron. *Journal of Computer and Communications*, *2014*(February), 11–16.

Atchison, J. & Shen, S. M. (1980). Logistic-normal distributions: Some properties and uses. *Biometrika*, *67*(2), 261–272.

Ayeh, J. K., Leung, D., Au, N., & Law, R. (2012). Perceptions and strategies of hospitality and tourism practitioners on *social media*: An exploratory study. In *Information and Communication Technologies in Tourism 2012: Proceedings of the International Conference in Helsingborg, Sweden, January 25–27, 2012* (pp. 1–12).

Ba, S., Kalakota, R., & Whinston, A. B. (1995). Executable documents as the basis for DSS. In *Proceedings of the Third ISDSS Conference* (Vol. 2, pp. 373-381).

Bai, X. (2011). Predicting consumer sentiments from online text. *Decision Support Systems*, *50*(4), 732–742.

Bash, E. (2015). An Empirical Investigation of Decision-Making Satisfaction in *Web*-Based Decision Support Systems. *PhD Proposal*, *1*, 1–29.

Bermingham, M. L., Pong-Wong, R., Spiliopoulou, a, Hayward, C., Rudan, I., Campbell, H., … Haley, C. S. (2015). Application of high-dimensional feature selection: evaluation for genomic prediction in man. *Scientific Reports*, *5*, 10312.

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of Science. *The Annals of Applied Statistics*, *1*(1), 17–35.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*(4-5), 993–1022.

Blei, D. M., & Lafferty, J. D. (2009). Topic models. In A. N. Srivastava & M. Sahami (Eds.), Text mining: classification, clustering, and applications (pp. 71–94). Boca Raton: CRC Press.

Blake, C. (2011). Information. *Annual Review Of Information Science And Technology*, *45*, 123–155.

Bollen, J., Mao, H., & Zeng, X.-J. (2010). Twitter mood predicts the stock market, 1–8. Computational Engineering, Finance, and Science; Computation and Language; Physics and Society.

Buhalis, D., & Law, R. (2008). Progress in information technology and tourism management: 20 years on and 10 years after the Internet-The state of eTourism research. *Tourism Management*, *29*(4), 609–623.

Burgess, S., Sellitto, C., & Cox, C. (2009). User-generated content (UGC) in tourism: benefits and concerns of online consumers. *17th European Conference on Information Systems*, (c7815a9a-b206-4076-72c7-4da7cc027e52).

Butler, R. W. (1980). The Concept of a Tourist Area Cycle of Evolution: Implications for Management of Resources. *The Canadian Geographer/Le Géographe Canadien*, *24*(1), 5–12.

Calantone, R. J., & Benedetto, C. A. (1991). KNOWLEDGE ACQUISITION. *Annals of Tourism Research*, *18*, 202–212.

Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, (April), 15–21.

Chatterjee, P., & Wang, Y. (2012). Online Comparison Shopping Behavior of Travel Consumers. Journal of Quality Assurance in Hospitality & Tourism, 13(1), 1–23.

Chen, J., Huang, H., Tian, S., & Qu, Y. (2009). Feature selection for text classification with Naïve Bayes. *Expert Systems with Applications*.

Chevalier, J. a, & Mayzlin, D. (2006). The Effect of Word of Mouth on Sales: Online Book Reviews. *Journal of Marketing Research*, *43*(3), 345–354.

Choy, M. (2012). Effective Listings of Function Stop words for Twitter. *arXiv Preprint arXiv:1205.6396*. Retrieved from

Cohen, M. D., Kelly, C. B., & Medaglia, A. L. (2001). Decision support with Web-enabled software. *Interfaces*, *31*(2), 109–129.

Compete Incorporated. (2007). Compete consumer generated content study reveals opportunities for travel marketers. PR Newswire.

Cooper, C., & Hall, C. M. (2008). Contémporary tourism an international approach, *45*.

Corney, M., Vel, O. De, Anderson, a, & Mohay, G. (2002). Gender-preferential text mining of e-mail discourse. *18th Annual Computer Security Applications Conference 2002 Proceedings*, *13*, 282–289.

Cox, C., Burgess, S., Sellitto, C., & Buultjens, J. (2009). The Role of User-Generated Content in Tourists' Travel Planning Behavior. *Journal of Hospitality Marketing & Management*, *18*(8), 743–764.

Darbellay, F., & Stock, M. (2012). Tourism as complex interdisciplinary research object. *Annals of Tourism Research*, *39*(1), 441–458.

Daugherty, T. (2008). Exploring Consumer Motivations for Creating User-Generated Content. *Journal of Interactive Advertising.*, *8*(2), 1–24. 25p. 1 Diagram.

Dawson, J., Scott, D., & Mcboyle, G. (2009). Climate change analogue analysis of ski tourism in the northeastern USA. *Climate Research*, *39*(June), 1–9.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the Américan society for information science*, *41*(6), 391.

Delen, D. & Crossland, M. D. (2008). Seeding the survey and analysis of research literature with text mining. *Expert Systems with Applications*, *34*(3), 1707–1720.

Dellarocas, C. (2003). The Digitization of Word of Mouth: Promise and Challenges of Online Feedback Mechanisms. *Management Science*, *49*(10), 1407–1424.

Dickinger, A., & Mazanec, J. (2008). Consumers' Preferred Criteria for Hotel Online Booking. *Information and Communication Technologies in Tourism 2008*, 244–254.

Dijrre, J., Gerstl, P., & Seiffert, R. (1999). Finding Text Mining : Nuggets in Mountains of Textual Data, 398–401.

Domingos, P., & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, *130*, 103–130. Retrieved from

Duan, W., Gu, B., & Whinston, A. B. (2008). Do online reviews matter? - An empirical investigation of panel data. *Decision Support Systems*, *45*(4), 1007–1016.

Dwivedi, M., Shibu, T. P., & Venkatesh, U. (2007). Social software practices on the Internet: Implications for the hotel industry. *International journal of contémporary hospitality management*, *19*(5), 415-426.

Dwyer, L., Forsyth, P., Madden, J., & Spurr, R. (2000). Economic impacts of inbound tourism under different assumptions regarding the macroeconomy. *Current Issues in Tourism*, *3*, 325–363.

Feldman, S. (1999). NLP Meets the Jabberwocky: Natural Language Processing in Information Retrieval. *Online - Weston Then Wilton*, *23*(May), 62–73.

Feldman, R., Fresko, M., Hirsh, H., Aumann, Y., Liphstat, O., Schler, Y., & Rajman, M. (1998). Knowledge Management: A Text Mining Approach. *Proc of the 2nd Int Conf on Practical Aspects of Knowledge Management (PAKM98*, *Basel, Swi*(April 2016), 1–10.

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text Mining Infrastructure in R. *Journal Of Statistical Software*, *25*(5), 1–54.

Forman, G. (2003). An Extensive Empirical Study of Feature Selection Metrics for Text Classification. *Journal of Machine Learning Research*, *3*, 1289–1305.

Gaowa, G. (2010). The Selection of Mongolian Stop Words. *IEEE*, 71–74.

Gartner, W. C. (2004). Rural tourism development in the USA. *International Journal of Tourism Research*.

Glazer, R. (1991). Marketing in information intensive environments: strategic implications of knowledge as an asset. *Journal of Marketing*, *55*(4), 1–19.

Godbole, Shantanu; Bhattacharya, Indrajit;Gupta, A. (2010). Building Re-usable Dictionary Repositories for Real-world Text Mining. *Em Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, 1189–1198

Goldenberg, J., Libai, B., & Muller, E. (2001). A Complex Systems Look at the Talk of the Network : Process of Word-of-Mouth Underlying. *Marketing Letters*, *12*(3), 211–223.

Goodrich, J. (2002). September 11, 2001 attack on América: a record of the immediate impacts and reactions in the USA travel and tourism industry. *Tourism Management*.

Gopal, R., Marsden, J. R., & Vanthienen, J. (2011). Information mining — Reflections on recent advancements and the road ahead in data, text, and media mining. *Decision Support Systems*, *51*(4), 727–731.

Graham, W. (2014). Hands-On Data Science with R Text Mining.

Grant-Braham, B. (2007). The Social Media and Travel Chatter, (September), 2007–2009.

Gretzel, U. (2006). Consumer Generated Content - Trends and Implications for Branding. *E-Review of Tourism Research*, *4*(3), 9–11.

Gretzel, U., & Yoo, K.-H. (2008). Use and impact of online travel reviews. *ResearchGate*, (November 2015).

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*(suppl 1), 5228-5235.

Grün, B., & Hornik, K. (2011). topicmodels : An R Package for Fitting Topic Models. *Journal of Statistical Software*, *40*(13), 1–30.

Guerreiro, J., Rita, P., & Trigueiros, D. (2015). A Text Mining-Based Review of Cause-Related Marketing Literature. *Journal of Business Ethics*, (Adkins 1999), 1–18.

Hearst, M. (1999). Untangling text data mining. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics.*

Hagel, J., & Armstrong, A. (1997). Net gain: Expanding markets through virtual communities. Harvard Business Press.

Hecht, B., & Gergle, D. (2010). On the " Localness " of User-Generated Content. *In Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, 229–232.

Hermida, A., & Thurman, N. (2008). a Clash of Cultures. *Journalism Practice*, *2*(3), 343–356.

Hofmann, T. (1999, August). Probabilistic latent semantic indexing. In:*Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 50-57). ACM.

Hruschka, H., & Mazanec, J. (1990). Computer-assisted travel counseling. *Annals of Tourism Research*, *17*(2), 208-227.

Hyan Yoo, K., & Gretzel, U. (2008). The Influence of Perceived Credibility on Preferences for Recommender Systems as Sources of Advice. *Information Technology & Tourism*.

Inmon, W. H. Building the data warehouse. 2002. *United States of América: Robert Ibsen*

Indurkhya, N., & Damerau, F. J. (2010). Handbook of Natural Language Processing, Second Edition. *CRC Press*, *2*, 1–676.

James, R. K., Acquisitions, S., Marquita, E., & Boes, A. V. (2010). *Licensed to : iChapters User Licensed to : iChapters User*. *Business* (p. 957).

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2006). *An Introduction to Statistical Learning*. *Design* (Vol. 102).

Kao, A., & Poteet, R. S. (Eds). (2007). *Natural language processing and text mining*. *Springer Science & Business Media*.

Kasper, W., & Vela, M. (2011). Sentiment analysis for hotel reviews. *Computational Linguistics-Applications Conference*, *231527*, 45–52.

Kim, H., Chen, M., & Jang, S. (2006). Tourism expansion and economic development: the case of Taiwan. *Tourism Management*, *27*, 925–933. Retrieved from

Kim, S., & Han, K. (2006). Some effective techniques for naive bayes text classification. *Knowledge and Data Engineering,* 18(11).

Kimball, R., & Ross, M. (2013). The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling (3rd Edition). John Wiley & Sons, Inc.

Klein, M. (1990). *Expert systems: A decision support approach: With applications in management and finance*. Addison Wesley Publishing Company.

Kontopoulos, E., Berberidis, C., Dergiades, T., & Bassiliades, N. (2013). Ontology-based sentiment analysis of twitter posts. *Expert Systems with Applications*, *40*(10), 4065–4074.

Kroeze, J. H., Matthee, M. C., & Bothma, T. J. D. (2003). *Differentiating data- and text-mining terminology*. Paper presented at the Proceedings of the 2003 annual research conference of the South African institute of computer scientists and information technologists on Enablement through technology, Johannesburg, South Africa. Kumar, J. A., & Abirami, S. (2015). an Experimental Study of Feature Extraction Techniques in Opinion, *4*(1), 15–21.

Labadie, J. W., Brazil, L. E., Corbu, I., & Johnson, L. E. (1989). *Computerized decision support systems for water managers*. ASCE.

Lamkanfi, A., & Demeyer, S. (2011). Comparing mining algorithms for predicting the severity of a reported bug. *Software Maintenance and Reengineering (CSMR).*

Lau, K.-N. (2005). *Text Mining for the Hotel Industry*. Cornell Hotel and Restaurant Administration Quarterly, *46*(3), 344–362.

Lawrence, L. (2014). Reliability of Sentiment Mining Tools: A comparison of Semantria and Social Mention, University of Twente, 1–13.

Lee, S., & Baker, J. (2010). *An empirical comparison of four text mining methods.* System Sciences (HICSS)*,* 1–10.

Li, N., & Wu, D. D. (2010). Using text mining and sentiment analysis for online forums hotspot detection and forecast. *Decision Support Systems*, *48*(2), 354–368.

Liddy, L., Hovy, E., Lin, J., Prager, J., Radev, D., Vanderwende, L., & Weischedel, R. (2003). Natural Language Processing. *Encyclopedia of Library and Information Science*, *2*, 1–16.

Linus, A., & Lawrence, P. (2014). Reliability of Sentiment Mining Tools : A comparison of Semantria and Social Mention, 1–13.

Litvin, S. W., Goldsmith, R. E., & Pan, B. (2008). Electronic word-of-mouth in hospitality and tourism management. *Tourism Management*, *29*(3), 458–468.

Liu, B. (2008). Web Data Mining: Exploring Hyperlinks,Contents, and Usage Data. (2nd ed.). Springer Berlin Heidelberg New York.

Liu, B. (2012). Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies*, *5*(1), 1–167.

Loucks, D. P., & da Costa, J. R. (Eds.). (2013). *Decision support systems: Water resources planning* (Vol. 26). Springer Science & Business Media.

Lu, L. Y. Y., Lin, B. J. Y., Liu, J. S., & Yu, C.-Y. (2012). Ethics in Nanotechnology: What's Being Done? What's Missing? *Journal of Business Ethics*, *109*(4), 583–598.

Lu, W., & Stepchenkova, S. (2014). User-Generated Content as a Research Mode in Tourism and Hospitality Applications: Topics, Methods, and Software. *Journal of Hospitality Marketing & Management*, *24*(2), 119–154.

Månsson, M. (2011). Mediatized tourism. *Annals of Tourism Research*, *38*(4), 1634–1652.

Marrese-Taylor, E., Velasquez, J. D., & Bravo-Marquez, F. (2013). Opinion Zoom: A Modular Tool to Explore Tourism Opinions on the Web. *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, 261–264.

Matsuo, Y., & Ishizuka, M. (2004). Keyword Extraction From a Single Document Using Word Co-Occurrence Statistical Information. *International Journal on Artificial Intelligence Tools*, *13*(01), 157–169.

Mazanec, J. A. (1994). International tourism marketing-adapting the growth share matrix. *Marketing in Europe, Case Studies. Sage Publications: London*, 184-203.

Meyer, D., Hornik, K., & Feinerer, I. (2013). Institutional Repository, (October).

Middelkoop, M. (2001). *MERLIN: A decision support system for outdoor leisure planning: Development and test of a rule-based microsimulation model for the evaluation of alternative scenarios and planning options*. Technische Universiteit Eindhoven, Faculteit Bouwkunde, Capaciteitsgroep Stedebouw.

Miguéns, J., Baggio, R., & Costa, C. (2008). Social media and Tourism Destinations : TripAdvisor Case Study. *Advances in Tourism Research*.

Milano, R., Baggio, R., & Piattelli, R. (2011). The effects of online social media on tourism websites. In *Information and Communication Technologies in Tourism 2011* (pp. 471–483).

Mostafa, M. M. (2013). More than words: Social networks' text mining for consumer brand sentiments. *Expert Systems with Applications*, *40*(10), 4241–4251.

Murdick, R. G., and J. C. Munson (1986). *MIS Concepts & Design*, 2d ed. New York: Prentice Hall.

Nahm, U. Y., & Mooney, R. J. (2002). Text mining with information extraction. *In AAAI 2002 Spring Symposium on Mining Answers from Texts and Knowledge Bases*, *1*.

Nations, U. (2012). Did you know ?, 2012–2013.

Neuhofer, B., Buhalis, D., & Ladkin, A. (2014). A typology of technology-enhanced tourism experiences. *International Journal of Tourism Research*, *16*, 340–350.

O'Connor, P. (2010). Managing a Hotel's Image on TripAdvisor. *Journal of Hospitality Marketing & Management*, *19*(7), 754–772.

Oh, C. (2005). The contribution of tourism development to economic growth in the Korean economy. *Tourism Management*, *26*, 39–44.

Oh, S., & Park, M. S. (2013). Text mining as a method of analyzing health questions in social Q&A. *Proceedings of the Américan Society for Information Science and Technology*, *50*(1), 1–4.

Oracle White Paper: Oracle 9i Application Server (Dec. 2001).

O'Reilly, T. (2005). What is Web 2.0.

Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, *2*(1-2), 1–135.

Paltoglou, G., & Thelwall, M. (2012). Twitter, MySpace, Digg: Unsupervised sentiment analysis in social media. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *3*(4), 66.

Pan, B., MacLaurin, T., & Crotts, J. C. (2007). Travel Blogs and the Implications for Destination Marketing. *Journal of Travel Research*, *46*(August), 35–45.

Pande, V., & Khandelwal, A. (2014). A Survey of Different Text Mining Techniques. *IBMRD's Journal of Management & …*, (1), 125–133.

Park, D.-H., Lee, J., & Han, I. (2007). The Effect of On-Line Consumer Reviews on Consumer Purchasing Intention: The Moderating Role of Involvement. *International Journal of Electronic Commerce*.

Parra-López, E., Bulchand-Gidumal, J., Gutiérrez-Taňo, D., & Dĭaz-Armas, R. (2011). Intentions to Use Social Media in Organising and Taking Vacation Trips. *Computers in Human Behaviour*, *27*(2), pp640–654.

Peffers, K. E. N., Tuunanen, T., Rothenberger, M. a, & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, *24*(3), 45–77.

Pete, C., Julian, C., Randy, K., Thomas, K., Thomas, R., Colin, S., & Wirth, R. (2000). Crisp-Dm 1.0. *CRISP-DM Consortium*, 76.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program: Electronic Library and Information Systems*.

Power, D. J. (1998). Web-based decision support systems. *DSstar, The On-Line Executive Journal for Data-Intensive Decision Support*, *2*(33).

Powell, Ron. (2001). Anniversary Special: A 10 Year Journey.

Power, D. J. (2002). Decision Support Systems: Concepts and Resources for Managers. *Studies in Informatics and Control*, *11*(4), 349-350.

Power, D. J. (2007). A brief history of decision support systems (2007).

Prabowo, R., & Thelwall, M. (2009). Sentiment analysis: A combined approach. *Journal of Informetrics*, *3*(2), 143–157.

Purifoy, M. ;, Yoo, K. H., & Gretzel, U. (2007). OnlineTravelReviewReport. *Laboratory for Intelligent Systems in Tourism Texas A&M University*.

Provost, F., & Fawcett, T. (2013). *Data Science for Business: What you need to know about data mining and data-analytic thinking*. " O'Reilly Media, Inc.".

Ricci, F., & Werthner, H. (2004). E-Commerce and Tourism. *Communications of the ACM*, *47*(12), 101–105.

Rita, P. (1993). *A knowledge-based system for promotion budget allocation by national tourism organizations* (Doctoral dissertation, Doctoral thesis, University of Wales, College of Cardiff).

Salton, G., Wong, A., & Yang, C. (1975). A vector space model for automatic indexing. *Communications of the ACM*.

Santana, M. (1995). Managerial learning: a neglected dimension in decision support systems. *Proceedings of the Twenty-Eighth Hawaii International Conference on System Sciences*, *4*, 82–91.

Schmallegger, D., & Carson, D. (2008). Blogs in tourism: Changing approaches to information exchange. *Journal of Vacation Marketing*, *14*(2), 99–110.

Setiawan, J. (2014). Using Text Mining to Analyze Mobile Phone Provider Service Quality (Case Study: Social Media Twitter). *International Journal of Machine Learning and Computing*, *4*(1), 106–109.

Setzer, V. W. (2015). Dado, Informação, Conhecimento e Competência. *Instituto de Matemática E Estatística Da Universidade de São Paulo*, 1–14.

Sharda, R., Delen, D., & Turban, E. (2013). *Business Intelligence: A Managerial Perspective on Analytics*. Prentice Hall Press.

Shen, C., & Kuo, C.-J. (2015). Learning in massive open online courses: Evidence from social media mining. *Computers in Human Behavior*.

Shih, C. (2009). *The Facebook era: Tapping online social networks to build better products, reach new audiences, and sell more stuff*. *Electronic Commerce, T. Nash, ed*.

Sigala, M. (2011). WEB 2 . 0 , Social Marketing Strategies and Distribution Channels for City Destinations : Enhancing the Participatory Role of Travelers and Exploiting their Collective Intelligence. *Information Communication Technologies and City Marketing*, 1249–1251.

Sokolova, M., Jafer, Y., & Schramm, D. (2012). Text Mining for Personal Health Information on Twitter. *Proceedings of the 2012 IEEE …*, *61*(3), 127–30.

Solka, J. (2008). Text data mining: theory and methods. *Statistics Surveys*, *2*, 94–112.

Stokes, D., & Lomax, W. (2002). Taking Control of Word-of-Mouth Marketing: The Case of an Entrepreneurial Hotelier. *Journal of Small Business and Enterprise Development*, *9*(4), 349–357.

Sumathy, K., & Chidambaram, M. (2013). Text Mining: Concepts, Applications, Tools and Issues–An Overview. *International Journal of Computer …*, *80*(4), 29–32.

travel reviews as persuasive communication: The effects of content type, source, and certification logos on consumer behavior. *Tourism Management*.

Tan, A. (1999). Text Mining : The state of the art and the challenges Concept-based. *In Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases*, *8*(April), 65–70.

Tang, C.-H. (Hugo), & Jang, S. (Shawn). (2009). The tourism–economy causality in the United States: A sub-industry level examination. *Tourism Management*, *30*(4), 553–558.

Thevenot, G. (2007). Blogging as a social media. *Tourism and Hospitality Research*, *7*(3-4), 287–289.

Tribe, J. (1997). The indiscipline of tourism. *Annals of Tourism Research*.

Vermeulen, I. E., & Seegers, D. (2009). Tried and tested: The impact of online hotel reviews on consumer consideration. *Tourism Management*.

Walker, P. A., Greiner, R., McDonald, D., & Lyne, V. (1998). The Tourism Futures Simulator: A systems thinking approach. *Environmental Modelling and Software*, *14*(1), 59–67.

Wang, H. (1997). Intelligent agent-assisted decision support systems: Integration of knowledge discovery, knowledge analysis, and group decision support. *Expert Systems with Applications*, *12*(3), 323–335.

Wei, X., & Croft, W. B. (2006). LDA-based document models for ad-hoc retrieval. *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval SIGIR 06*, *pages*, 178–185.

Wierenga, B., & Van Bruggen, G. H. (2000). *Marketing management support systems: Principles, tools, and implementation* (Vol. 10). Springer Science & Business Media.

William, E., & Perez, E. (2008). Tourism 2.0. the social web as a platform to develop a knowledge-based ecosystem. Networks and Tourism.

Wilson, S., Fesenmaier, D. R., Fesenmaier, J., & Van Es, J. C. (2001). Factors for Success in Rural Tourism Development. *Journal of Travel Research*, *40*(2), 132–138.

Wöber, K. W. (1998). TourMIS: An adaptive distributed marketing information system for strategic decision support in national, regional, or city tourist offices. *Pacific Tourism Review*, *2*, 273-286.

Wober, K., & Gretzel, U. (2000). Tourism Managers' Adoption of Marketing Decision Support Systems. *Journal of Travel Research*, *39*(November), 172–181.

Wöber, K. W. (2003). Information supply in tourism management by marketing decision support systems. *Tourism Management*, *24*(3), 241–255.

World Tourism Organization, U. (2013). *Tourism in the world : key figures* (p. 16).

Xiang, Z., & Gretzel, U. (2010). Role of social media in online travel information search. *Tourism Management*, *31*(2), 179–188.

Xu, J. B. (2010). Perceptions of tourism products. *Tourism Management*, *31*(5), 607–610.

Yang, Y., & Pedersen, J. (1997). A comparative study on feature selection in text categorization. *ICML*, *97*(July), 412–420.

Ye, Q., Zhang, Z., & Law, R. (2009). Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert Systems with Applications*, *36*(3), 6527–6535.

Ye, Q., Law, R., Gu, B., & Chen, W. (2011). The influence of user-generated content on traveler behavior: An empirical investigation on the effects of e-word-of-mouth to hotel online bookings. *Computers in Human Behavior*, *27*(2), 634–639.

Zeng, B., & Gerritsen, R. (2014). What do we know about social media in tourism? A review. *Tourism Management Perspectives*.

Zhou, D., Yanagida, J. F., Chakravorty, U., & Leung, P. (1997). Estimating economic impacts from tourism. *Annals of Tourism Research*, *24*(1), 76–89.

Zhu, F., & Zhang, X. (2006). The Influence of Online Consumer Reviews on the Demand for Experience Goods: The Case of Video Games. *ICIS 2006 Proceedings*, 17.

Zielinski, A., & Middleton, S. (2013). Social Media Text Mining and Network Analysis for Decision Support in Natural Crisis Management. *10th International Conference on Information Systems for Crisis Response and Management*, (May 2013), 840–845.

ANEXOS

Anexo A – análise das explorações das categorias de negócio

Nesta secção, apresentam-se as análises exploratórias dos *reviews* das restantes categorias de negócio que não foram abordadas anteriormente devido a serem demasiadas categorias (22 no total). Estas explorações contemplam as seguintes análises tendo em conta os dois cenários (unigramas e bigramas):

- As palavras mais e menos frequentes (Forman, 2003; Matsuo & Ishizuka, 2004);
- Os termos com frequência mínima de ocorrência (tem que se ter em conta a quantidade de *reviews* e ter em conta os termos mais frequentes que foram obtidos na análise anterior);
- As *Wordcloud,* imagens formadas pelas palavras tendo em conta a frequência, (sendo uma forma mais representativa da frequência das palavras, para uma visualização mais simples e perspicaz de acordo com as frequências obtidas na análise anterior)(Guerreiro et al., 2015).

De seguida apresentam-se as análises das categorias:

Active Life

Como se pode observar na tabela 2, esta categoria de negócio está constituída por 164 *reviews*. Esta categoria contempla tudo o que é ao ar livre, desde golfe, escalada, equitação, pesca, *karts*, fitness, bólingue, ténis, até parques de diversão e infantis, jardim zoológico, aluguer de bicicletas, praias, campos de férias, aluguer de material desportivo entre outros.
Foi feita uma exploração de dados para se compreender e para ter uma perceção mais clara sobre os dados tendo em conta os dois cenários referidos anteriormente.

De seguida, apresentam-se as explorações nesta categoria de negócio:

Palavras mais e menos frequentes em cada categoria

Após os tratamentos dos dados, observou-se os termos mais e menos frequentes, como se pode observar de seguida:

Termos menos frequêntes	Quantidade	Termos mais frequêntes	Quantidade
abc	1	place	86
aboot	1	great	87
abroad	1	can	88
abund	1	like	94
abus	1	class	114
academ	1	get	114

TABELA 29: UNIGRAMAS MAIS E MENOS FREQUENTES NA CATEGORIA *ACTIVE LIFE*.

Termos menos frequêntes	Quantidade	Termos mais frequêntes	Quantidade
high recommend	264	realli good	458
make sure	268	come back	565
place go	274	pretti good	596
you can	287	ice cream	639
next time	290	go back	679
if your	292	this place	814

TABELA 30: BIGRAMAS MAIS E MENOS FREQUENTESNA CATEGORIA DE *ACTIVE LIFE*.

Os unigramas mais frequentes identificam-se com a categoria de negócio *Active Life,* como se pode observar na figura a cima. Nesta categoria de negócio as palavras mais frequentes são ***get***(114), ***class***(114), ***like***(94), ***can***(88), ***great***(87), ***place***(86), enquanto que, as menos frequentes, ***abc*** (1), ***aboot*** (1), ***abroad*** (1), ***abund*** (1), ***abus*** (1) e ***academ***(1).

Observando a tabela 30, verifica-se que os bigramas mais frequentes são**, *this place*** (814), ***go back*** (679), ***ice cream*** (639), ***pretti good*** (596), ***come back*** (565) e ***realli good*** (458).

Termos com frequência mínima de ocorrência

Devido à exploração feita anteriormente nos dois cenários, como se pode observar na tabela 29 e 30, e tendo em conta o número de *reviews*, 164, optou-se por observar os termos que ocorrem no mínimo de 20 e 50 vezes, no caso de unigramas, e no caso de bigramas, quando ocorrem no mínimo 300 e 500.

De seguida, apresenta-se a figura 43 e 44, onde apresenta-se o código utilizado bem como os termos nos dois cenários.

```
findFreqTerms(matriz, lowfreq=20)#termos que aparecem no minimo 20 vezes
  [1] "also"     "alway"    "and"      "area"     "around"   "awesom"
  [7] "back"     "best"     "better"   "bike"     "bit"      "but"
 [13] "can"      "cant"     "class"    "clean"    "come"     "day"
 [19] "definit"  "dont"     "equip"    "even"     "experi"   "feel"
 [25] "find"     "first"    "fit"      "free"     "friend"   "fun"
 [31] "get"      "give"     "good"     "got"      "great"    "gym"
 [37] "help"     "high"     "hour"     "instructor" "its"    "ive"
 [43] "just"     "keep"     "kid"      "know"     "let"      "like"
 [49] "littl"    "locat"    "look"     "lot"      "love"     "machin"
 [55] "make"     "mani"     "month"    "much"     "need"     "never"
 [61] "new"      "nice"     "now"      "one"      "park"     "peopl"
 [67] "person"   "place"    "pretti"   "realli"   "review"   "right"
 [73] "room"     "run"      "say"      "see"      "space"    "staff"
 [79] "star"     "still"    "studio"   "take"     "there"    "they"
 [85] "thing"    "think"    "this"     "time"     "tri"      "two"
 [91] "use"      "walk"     "want"     "way"      "week"     "well"
 [97] "went"     "will"     "work"     "workout"  "year"     "yoga"
[103] "you"      "youll"    "your"     "zoo"

> findFreqTerms(matriz, lowfreq=50)#termos que aparecem no minimo 50 vezes
 [1] "also"   "can"    "class"  "get"    "good"   "great"  "gym"    "just"   "like"
[10] "one"    "park"   "peopl"  "place"  "realli" "run"    "there"  "time"   "will"
[19] "work"   "your"   "zoo"
```

FIGURA 43: UNIGRAMAS QUE APARECEM NO MÍNIMO 20 E 50 VEZES NA CATEGORIA *ACTIVE LIFE*.

```
> findFreqTerms(rvwbie, lowfreq=300)#termos que aparecem no mínimo 300 vezes
 [1] "can get"     "come back"   "dont know"   "feel like"   "first time"
 [6] "go back"     "great place" "ice cream"   "im sure"     "ive ever"
[11] "look like"   "pretti good" "realli good" "tast like"   "they also"
[16] "this place"
> findFreqTerms(rvwbie, lowfreq=500)#termos que aparecem no mínimo 500 vezes
[1] "come back"   "go back"     "ice cream"   "pretti good" "this place"
```

FIGURA 44: BIGRAMAS QUE APARECEM NO MÍNIMO 300 E 500 VEZES NA CATEGORIA *ACTIVE LIFE*.

Esta categoria contempla desportos, na exploração de unigramas, na figura 43, pode-se observar algumas palavras relevantes, como, ***gym, bike, fit, yoga, zoo, time, park***, entre outras, que distinguem bem esta categoria.

No caso do cenário dos bigramas, na figura 44, nesta exploração pode-se observar algumas palavras relevantes, como, ***this place, come back, go back,*** entre outras, que distinguem bem esta categoria.

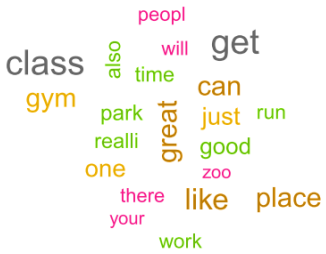
Ao observar os dois cenários, pode-se constatar, que esta categoria encontra-se melhor representada no cenário dos bigramas, uma vez que, é nesse cenário que as palavras mais frequentes descrevem melhor a categoria de negócio.

Wordcloud

Optou-se por fazer os *Wordclouds* com as frequências mínimas referenciadas nas análises anteriores, figura 43 e 44, para obter uma visualização mais clara, simples e perspicaz.

De seguida apresentam-se os *Wordclouds* (em que o tamanho fonte das palavras é proporcional à frequência dos termos) das frequências mínimas tendo em conta os dois cenários:

A)



B)

C)

D)

FIGURA 45: *WORDCLOUD* DA CATEGORIA *ACTIVE LIFE*: A) UNIGRAMAS FREQUÊNCIA MÍNIMA 20; B) UNIGRAMAS FREQUÊNCIA MÍNIMA 50; C) BIGRAMAS FREQUÊNCIA MÍNIMA 300; E) BIGRAMAS FREQUÊNCIA MÍNIMA 500.

Arts & Entertainment

Como se pode observar na tabela 2, esta categoria de negócio está constituída por 351 *reviews*.

Esta categoria constitui opiniões sobre cinemas, casas de espetáculos, casinos, *Jazz e blues*, estádios, galerias de arte, museus, festivais, arte urbana, castelos, ópera e *ballet*, bilheteiras, centros culturais entre outros.

Foi feita uma exploração de dados para se compreender e para ter uma perceção mais clara sobre os dados.

De seguida, apresentam-se as explorações nesta categoria de negócio:

Palavras mais e menos frequentes

Após os tratamentos dos dados, observou-se os termos mais e menos frequentes, como se pode observar de seguida:

Termos menos frequêntes	Quantidade	Termos mais frequêntes	Quantidade
aacs	1	show	151
abduction	1	see	152
abil	1	get	161
aborigin	1	like	162
about	1	great	167
abus	1	place	185

TABELA 31: UNIGRAMAS MAIS E MENOS FREQUENTES NA CATEGORIA *ARTS&ENTERTAINMENT*.

Termos menos frequêntes	Quantidade	Termos mais frequêntes	Quantidade
a cool	1	ive seen	14
a day	1	feel like	15
a decent	1	great place	17
a dream	1	love place	17
a fair	1	go back	18
a fell	1	you can	19

TABELA 32: BIGRAMAS MAIS E MENOS FREQUENTES NA CATEGORIA *ARTS&ENTERTAINMENT*.

Ao observar a exploração de unigramas, tabela 31, pode-se verificar que as palavras mais frequentes nesta categoria, são ***place*** (185), ***great*** (167), ***like***(162), ***get*** (161), ***see***(152) e ***show***(151). No entanto, as menos frequentes que foram detetadas nos *reviews* categorizados por *Arts&Entertainment* foram ***aacs*** (1), ***abduction*** (1), ***abil*** (1),***aborigin*** (1), ***about***(1), ***abus*** (1).

Na tabela 32, observa-se os bigramas mais frequentes nesta categoria que são os seguintes, ***you can*** (18), ***go back*** (18), ***love place*** (17), ***great place*** (17), ***feel like***(15), ***ive seen***(14). Os bigramas com menos frequência foram ***a cool*** (1), ***a day*** (1), ***a decent***(1), ***a dream*** (1), ***a fair*** (1) e ***a fell*** (1).

Ao analisar os dois cenários, pode-se deduzir que os bigramas descrevem melhor esta categoria de negócio.

Termos com frequência mínima de ocorrência

Por consequência das análises anteriores, tabela 31 e 32, e pelo número de *reviews* 351, optou-se por analisar os termos que tem uma frequência mínima nos dois cenários, nos unigramas frequência mínima de 50 e 100 vezes e em relação aos bigramas frequência mínima 10 e 15 vezes.

A continuação encontra-se a figura 46 e 47 com os respetivos resultados dos dois cenários:

```
> findFreqTerms(matriz, lowfreq=50)#termos que aparecem no minimo 50
 [1] "also"    "around"  "back"    "can"     "come"    "dont"    "enjoy"   "even"
 [9] "everi"   "exhibit" "experi"  "feel"    "film"    "find"    "first"   "food"
[17] "friend"  "fun"     "get"     "good"    "great"   "its"     "ive"     "just"
[25] "kid"     "like"    "littl"   "look"    "lot"     "love"    "make"    "movi"
[33] "much"    "museum"  "night"   "old"     "one"     "park"    "peopl"   "place"
[41] "play"    "pretti"  "price"   "realli"  "seat"    "see"     "show"    "theater"
[49] "theatr"  "there"   "they"    "thing"   "this"    "ticket"  "time"    "venu"
[57] "walk"    "want"    "way"     "went"    "will"    "year"    "your"
> findFreqTerms(matriz, lowfreq=100)#termos que aparecem no minimo 100
 [1] "can"   "get"   "good"  "great" "just"  "like"  "movi"  "one"
 [9] "place" "realli" "seat"  "see"   "show"  "theater" "time"
```

FIGURA 46: UNIGRAMAS QUE APARECEM NO MÍNIMO 50 E 100 VEZES NA CATEGORIA *ARTS&ENTERTAINMENT*.

```
> findFreqTerms(rvwbi, lowfreq=10)#termos que aparecem no minimo 10 vezes
 [1] "dont know"   "feel like"   "first time"  "go back"     "good time"
 [6] "great place" "if your"     "ive seen"    "love place"  "make sure"
[11] "middl east"  "movi theater" "this place" "ticket price" "you can"
> findFreqTerms(rvwbi, lowfreq=15)#termos que aparecem no minimo 15 vezes
[1] "feel like"  "go back"    "great place" "love place" "you can"
```

FIGURA 47: BIGRAMAS QUE APARECEM NO MÍNIMO 10 E 15 VEZES NA CATEGORIA *ARTS&ENTERTAINMENT*.

Como se pode observar na figura a cima, quanto maior a frequência das palavras maior é a ligação com o tema, neste caso, com a categoria de negócio.

Esta categoria contempla opiniões sobre arte e entretenimento, observando a exploração dos unigramas, pode-se observar algumas palavras relevantes, como, ***show, movi, good, like, place, time***, entre outras, que distinguem bem esta categoria.

Analisando agora o cenário de bigramas, os termos mais frequentes são, ***feel like, go back, great place***, entre outras.

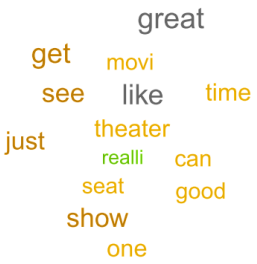
Após a análise dos dois cenários, constata-se que os melhores resultados se obtêm quando se analisa o cenário de unigramas, uma vez que, com estes vai mais de encontro à categoria de negócio.

Wordcloud

Optou-se por fazer os *Wordclouds* com as frequências mínimas referenciadas nas análises anteriores, figura 46 e 47, para obter uma visualização mais clara, simples e perspicaz.

De seguida apresentam-se os *Wordclouds* (em que o tamanho fonte das palavras é proporcional à frequência dos termos) das frequências mínimas tendo em conta os dois cenários:

A)



B)

C) D)

FIGURA 48: *WORDCLOUD* DA CATEGORIA *ARTS&ENTERTAINMENT*: A) UNIGRAMAS FREQUÊNCIA MINÍMA 50; B) UNIGRAMAS FREQUÊNCIA MÍNIMA 100; C) BIGRAMAS FREQUÊNCIA MÍNIMA 10; D) BIGRAMAS FREQUÊNCIA MÍNIMA 15.

Automotive

Como se pode observar na tabela 2, esta categoria de negócio está constituída por 69 *reviews*.

Esta categoria contempla opiniões sobre mecânicos, estacionamentos, concecionários de motas e de automóveis, lavagens de automóveis, pneus, postos de abastecimentos, reboques, peças de automóveis entre outros.

Foi feita uma exploração de dados para se compreender e para ter uma perceção mais clara sobre os dados.

De seguida, apresentam-se as explorações nesta categoria de negócio:

Palavras mais e menos frequentes

Após os tratamentos dos dados, observou-se os termos mais e menos frequentes nos dois cenários, como se pode observar de seguida:

Termos menos frequêntes	Quantidade	Termos mais frequêntes	Quantidade
aamco	1	they	32
about	1	get	42
absurd	1	work	43
acceler	1	time	44
accident	1	servic	50
accomod	1	car	122

TABELA 33: UNIGRAMAS MAIS E MENOS FREQUENTES NA CATEGORIA *AUTOMOTIVE*.

Termos menos frequêntes	Quantidade	Termos mais frequêntes	Quantidade
a coupl	1	servic depart	6
a honest	1	get car	7
a lot	1	go back	7
a mechan	1	take car	7
a quarter	1	car wash	8
aaa busi	1	oil chang	14

TABELA 34: BIGRAMAS MAIS E MENOS FREQUENTES NA CATEGORIA *AUTOMOTIVE*.

Como se pode observar no cenário de unigramas, tabela 33, as palavras mais frequentes nesta categoria de negócio, são, ***car***(122), ***servic*** (50), ***time***(44), ***work*** (43), ***get***(42), ***they*** (32). As menos frequentes ***aamco*** (1), ***about*** (1), ***absourd***(1), ***acceler*** (1), ***accident***(1), ***accomod*** (1).

Visualizando a exploração da frequência de bigramas, tabela 34, que os bigramas mais frequentes são, ***oil chang***(14), ***car wash*** (8), ***take car***(7), ***go back*** (7), ***get car*** (7) e ***servic depart*** (6), mantendo a concordância dos temas abordados nesta categoria de negócio. No entanto, os bigramas menos frequentes ***a coupl*** (1), ***a honest***(1), ***a lot*** (1), ***a mechan*** (1), ***a quarter*** (1) e ***aaa busi***(1).

Termos com frequência mínima de ocorrência

Após a análise feita anteriormente, para os dois cenários, tabela 33 e 34, e tendo em conta o número de *reviews* nesta categoria de negócio, 69, decidiu-se explorar as palavras que ocorrem, para unigramas no mínimo 20 vezes e para bigramas no mínimo 5 vezes.

Na figura 49 e 50, que se apresentam de seguida, expõe-se o código utilizado e os termos nos dois cenários.

```
> findFreqTerms(matriz, lowfreq=20)#termos que aparecem no minimmo 20 vezes
 [1] "back"    "car"     "get"     "good"    "got"     "just"    "like"    "need"
 [9] "one"     "place"   "price"   "problem" "repair"  "said"    "servic"  "they"
[17] "time"    "tire"    "took"    "will"    "work"
```

FIGURA 49: UNIGRAMAS QUE APARECEM NO MÍNIMO 20 VEZES NA CATEGORIA *AUTOMOTIVE*.

```
> findFreqTerms(rvwbi, lowfreq=5)#termos que aparecem no minimo 5 vezes
 [1] "car back"       "car wash"       "car zone"       "get car"
 [5] "go back"        "high recommend" "oil chang"      "palo alto"
 [9] "servic depart"  "take car"
```

FIGURA 50: BIGRAMAS QUE APARECEM NO MÍNIMO 5 VEZES NA CATEGORIA *AUTOMOTIVE*.

Esta categoria contempla tudo relativo a automóveis e motas, como foi referido anteriormente, Na exploração de unigramas, figura 49, pode-se observar algumas palavras relevantes nesta categoria, como, ***car,problem, repair, servic*****,** entre outras, que distinguem bem esta categoria.

No cenário de bigramas, figura 50, pode-se observar alguns bigramas relevantes, como, ***car back, car wash, car zone, get car,*** entre outras, que distinguem bem esta categoria.

Por último, pode-se concluir, que nesta categoria, a análise de frequência mínima de termos tem melhor expressão nos bigramas, uma vez que, descreve melhor a categoria de negócio.

Wordcloud

Optou-se por fazer os *Wordclouds* com as frequências mínimas referenciadas nas análises anteriores, figura 49 e 50, para obter uma visualização mais clara, simples e perspicaz.

De seguida apresentam-se os *Wordclouds* (em que o tamanho fonte das palavras é proporcional à frequência dos termos) das frequências mínimas tendo em conta os dois cenários:

A) B)

FIGURA 51: *WORDCLOUD* DA CATEGORIA *AUTOMOTIVE*: A) UNIGRAMAS FREQUÊNCIA MÍNIMA 20; B) BIGRAMAS FREQUÊNCIA MÍNIMA 5.

Beauty & Spas

Como se pode observar na tabela 2, esta categoria de negócio está constituída por 419 *reviews*.

As opiniões nesta categoria referem-se desde cuidados com a pele, cosmética e produtos de beleza, salões de beleza, massagistas, barbeiros, spas, solário, depilação, maquilhadores entre outros.

Foi feita uma exploração de dados para se compreender e para ter uma perceção mais clara sobre os dados.

De seguida, apresentam-se as explorações nesta categoria de negócio:

Palavras mais e menos frequentes

Após os tratamentos dos dados, observou-se os termos mais e menos frequentes nos dois cenários, como se pode observar de seguida com o seu respetivo código:

Termos menos frequêntes	Quantidade	Termos mais frequêntes	Quantidade
aaaaaa	1	great	216
aaaaaamaz	1	like	222
aaaaand	1	place	266
aback	1	time	268
abot	1	get	336
above	1	hair	373

TABELA 35: UNIGRAMAS MAIS E MENOS FREQUENTES NA CATEGORIA *BEAUTY&SPA*.

Termos menos frequêntes	Quantidade	Termos mais frequêntes	Quantidade
a believer	1	come back	25
a chang	1	this place	28
a client	1	get hair	30
a coupl	1	cut hair	35
a facial	1	go back	44
a friend	1	hair cut	47

TABELA 36: BIGRAMAS MAIS E MENOS FREQUENTES NA CATEGORIA *BEAUTY&SPA*.

Ao analisar de forma minuciosa a tabela 35, unigramas, pode-se observar que as palavras mais frequentes desta categoria de negócio são, ***hair***(373), ***get***(336), ***time*** (268), ***place*** (266), ***like***(222),***great*** (216).

Fazendo uma exploração da frequência de bigramas, tabela 36, pode-se observar, que os bigramas mais frequentes são, ***hair cut***(47), ***go back*** (44), ***cut hair***(35), ***get hair*** (30), ***this place*** (28) e ***come back*** (25), mantendo a concordância dos temas abordados nesta categoria de negócio.

Termos com frequência mínima de ocorrência

Devido à exploração feita anteriormente, como se pode observar nas tabelas 35 e 36, e tendo em conta o número de *review*s, 419, achou-se pertinente observar as palavras que ocorrem, no cenário de unigramas, no mínimo 150 e 200 vezes e no cenário de bigramas, no mínimo 10 e 20.

De seguida, apresentam-se as figuras 52 e 53, onde se encontra o código utilizado para obter os termos que aparecem com uma frequência mínima indicada, tendo em conta os doiscenários.

```
> findFreqTerms(matriz, lowfreq=150)#termos que aparecem no minimo 150 vezes
 [1] "back"   "cut"    "friend" "get"    "good"   "got"    "great"  "hair"   "ive"
[10] "just"   "like"   "look"   "one"    "place"  "realli" "salon"  "she"    "time"
[19] "want"
> findFreqTerms(matriz, lowfreq=200)#termos que aparecem no minimo 200 vezes
[1] "cut"   "get"   "great" "hair"  "just"  "like"  "place" "time"
```

FIGURA 52: UNIGRAMAS QUE APARECEM NO MÍNIMO 150 E 200 VEZES NA CATEGORIA *BEAUTY&SPA*.

```
> findFreqTerms(rvwbi, lowfreq=10)#termos que aparecem no mínimo 10 vezes
 [1] "call back"      "came back"      "color hair"     "come back"
 [5] "custom servic"  "cut color"      "cut hair"       "deep tissu"
 [9] "didnt even"     "dont know"      "even though"    "everi time"
[13] "feel like"      "first time"     "get hair"       "get haircut"
[17] "gift card"      "go back"        "good job"       "great job"
[21] "great place"    "hair cut"       "hair stylist"   "high recommend"
[25] "ive ever"       "ive go"         "last minut"     "long time"
[29] "look forward"   "look good"      "look great"     "look like"
[33] "made appoint"   "made feel"      "make appoint"   "make feel"
[37] "make sure"      "nail salon"     "never go"       "place go"
[41] "pretti much"    "price reason"   "realli like"    "she also"
[45] "staff friend"   "take time"      "they also"      "this place"
[49] "tissu massag"   "took time"      "wait time"      "will go"
[53] "year ago"       "year now"
> findFreqTerms(rvwbi, lowfreq=20)#termos que aparecem no mínimo 20 vezes
 [1] "come back"    "cut hair"     "first time"   "get hair"     "go back"
 [6] "great job"    "hair cut"     "hair stylist" "look like"    "make sure"
[11] "this place"
```

FIGURA 53: BIGRAMAS QUE APARECEM NO MÍNIMO 10 E 20 VEZES NA CATEGORIA *BEAUTY&SPA*.

Na exploração de unigramas, figura 52, pode-se observar algumas palavras relevantes, como*,* ***cut, hair, time, place****,* entre outras, que distinguem bem esta categoria.

No cenário de bigramas, pode-se observar alguns bigramas relevantes, como, ***cut hai, get hair, hair cut*****,** entre outras, que distinguem bem esta categoria.

Por último, relativamente a esta análise de frequências, pode-se constatar que, o cenário que melhor representa esta categoria de negócio é o de bigramas, uma vez que, os termos são mais representativos.

Wordcloud

Optou-se por fazer os *Wordclouds* com as frequências mínimas referenciadas nas análises anteriores, figura 52 e 53, para obter uma visualização mais clara, simples e perspicaz.

De seguida apresentam-se os *Wordclouds* (em que o tamanho fonte das palavras é proporcional à frequência dos termos) das frequências mínimas tendo em conta os dois cenários:

FIGURA 54: *WORDCLOUD*DA CATEGORIA *BEAUTY&SPA*:A) UNIGRAMAS FREQUÊNCIA MÍNIMA 150; B) UNIGRAMAS FREQUÊNCIA MÍNIMA 200; C) BIGRAMAS FREQUÊNCIA MÍNIMA 10; D) BIGRAMAS FREQUÊNCIA MÍNIMA 20.

Education

Como se pode observar na tabela 2, esta categoria de negócio está constituída por 110 *reviews*.

Esta categoria de negócio contempla opiniões sobre universidades e ensino superior, infantários, escolas especializadas, escolas primárias, básicas e secundárias, centros de explicações, serviços educativos, professores particulares, educação especial, preparação para exames entre outros.

Foi feita uma exploração de dados para se compreender e para ter uma perceção mais clara sobre os dados.

De seguida, apresentam-se as explorações nesta categoria de negócio:

Palavras mais e menos frequentes

Após os tratamentos dos dados, observou-se os termos mais e menos frequentes, como se pode observar de seguida:

Termos menos frequêntes	Quantidade	Termos mais frequêntes	Quantidade
abort	1	love	54
about	1	student	55
abroad	1	great	56
absorb	1	campus	58
abstract	1	class	76
abund	1	school	80

TABELA 37: UNIGRAMAS MAIS E MENOS FREQUENTES NA CATEGORIA *EDUCATION*.

Termos menos frequêntes	Quantidade	Termos mais frequêntes	Quantidade
a answer	1	final week	6
a got	1	great place	6
a great	1	high recommend	6
a name	1	beauti campus	7
a take	1	feel like	7
aa school	1	this place	7

TABELA 38: BIGRAMAS MAIS E MENOS FREQUENTES NA CATEGORIA *EDUCATION*.

Nas figuras acima, podem-se observar os termos mais e menos frequentes nesta categoria de negócio em cada cenário.

Pode-se constatar, na exploração de unigramas, tabela 37, que as palavras mais frequentes desta categoria de negócio são, ***school***(80), ***class***(70), ***campus***(58), ***great***(56), ***student***(55),***love*** (54).

Fazendo uma exploração da frequência de bigramas, tabela 38, pode-se observar que os bigramas mais frequentes são, ***this place*** (7), ***feel like***(7), ***beauti campus*** (7), ***high recommend***(6), ***great place*** (6) e ***final week*** (6), mantendo a concordância dos temas abordados nesta categoria de negócio.

Termos com frequência mínima de ocorrência

Tendo em conta as análises efetuadas anteriormente, para ver as palavras mais e menos frequentes nos dois cenários, tabela 37 e 38, e o número de *reviews* desta categoria, 110, optou-se por observar as palavras que ocorrem pelo menos, no caso dos unigramas, 20 e 50 vezes e no caso dos bigramas 2 e 5 vezes.

De seguida, apresentam-se as figuras 55 e 56, onde apresenta-se o código utilizado para obter os termos que aparecem com o mínimo referido anteriormente bem como os termos associados.

```
> findFreqTerms(matriz, lowfreq=20)#termos que aparecem no minimo 20 vezes
 [1] "beauti"    "best"      "build"     "campus"    "can"       "class"
 [7] "colleg"    "come"      "dont"      "even"      "experi"    "friend"
[13] "get"       "good"      "graduat"   "great"     "hair"      "help"
[19] "just"      "like"      "lot"       "love"      "make"      "one"
[25] "park"      "peopl"     "place"     "professor" "program"   "realli"
[31] "say"       "school"    "student"   "take"      "there"     "they"
[37] "think"     "this"      "time"      "want"      "will"      "year"
[43] "your"
> findFreqTerms(matriz, lowfreq=50)#termos que aparecem no minimo 50 vezes
[1] "campus"  "class"   "get"     "great"   "love"    "school"  "student"
```

FIGURA 55: UNIGRAMAS QUE APARECEM NO MÍNIMO 20 E 50 VEZES NA CATEGORIA *EDUCATION*.

```
> findFreqTerms(rvwbi, lowfreq=5)#termos que aparecem no minimo 5 vezes
 [1] "around campus"  "aveda institut" "beauti campus"  "cal poli"
 [5] "even though"    "feel like"      "final week"     "great place"
 [9] "high recommend" "law school"     "mba program"    "of cours"
[13] "student bodi"   "there lot"      "this place"
```

FIGURA 56: BIGRAMAS QUE APARECEM NO MÍNIMO 5 VEZES NA CATEGORIA *EDUCATION*.

Analisando a exploração no cenário de unigramas, figura 55, pode-se observar algumas palavras relevantes, como, ***class, campus, school, student***, entre outras, que distinguem bem esta categoria.

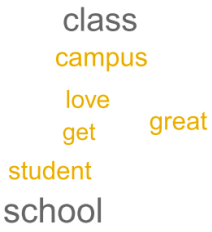
Ao observar o cenário de bigramas, figura 56, nesta exploração pode-se observar algumas palavras relevantes, como, ***around campus, aveda institut, beauti campus***, entre outras, que distinguem bem esta categoria.

Wordcloud

Optou-se por fazer os *Wordclouds* com as frequências mínimas referenciadas nas análises anteriores, figuras 55 e 56, para obter uma visualização mais clara, simples e perspicaz.

De seguida apresentam-se os *Wordclouds* (em que o tamanho fonte das palavras é proporcional à frequência dos termos) das frequências mínimas tendo em conta os dois cenários:

A)



B)

C)

FIGURA 57: *WORDCLOUD*DA CATEGORIA *EDUCATION*: A) UNIGRAMAS FREQUÊNCIA MÍNIMA 20; B) UNIGRAMAS FREQUÊNCIA MÍNIMA 50; C) BIGRAMAS FREQUÊNCIA MÍNIMA 5.

Event Planning & Service

Como se pode observar na tabela 2, esta categoria de negócio está constituída por 47 *reviews*.

Estes 47 *reviews* desta categoria de negócio contemplam opiniões desde papelarias, espaços para eventos, artigos de festas, aluguer de barcos, fotógrafos entre outros.

Foi feita uma exploração de dados para se compreender e para ter uma perceção mais clara sobre os dados.

De seguida, apresentam-se as explorações nesta categoria de negócio:

Palavras mais e menos frequentes

Após os tratamentos dos dados, observou-se os termos mais e menos frequentes, como se pode observar de seguida:

Termos menos frequêntes	Quantidade	Termos mais frequêntes	Quantidade
about	1	look	20
accommod	1	place	20
accomod	1	get	21
accord	1	good	27
across	1	work	27
actor	1	wed	30

TABELA 39: UNIGRAMAS MAIS E MENOS FREQUENTES NA CATEGORIA *EVENT PLANNING&SERVICE*.

Termos menos frequêntes	Quantidade	Termos mais frequêntes	Quantidade
a a	1	let parti	3
a even	1	make sure	3
a videograph	1	month old	3
a woman	1	wed photograph	3
abl forg	1	will pay	3
abl quick	1	high recommend	4

TABELA 40: BIGRAMAS MAIS E MENOS FREQUENTES NA CATEGORIA *EVENT PLANNING&SERVICE*.

Como se pode observar, nas duas figuras a cima, encontram-se as palavras mais e menos frequentes desta categoria de negócio.

Observando o cenário de unigramas, tabela 39, pode-se dizer que as palavras mais frequentes são, ***wed***(30), ***work*** (27), ***good***(27), ***get***(21), ***place***(20) e ***look***(20).

Ao fazer uma exploração da frequência de bigramas, tabela 40, os bigramas mais frequentes são, ***high recommend*** (4), ***will pay*** (3), ***wed photograph*** (3), ***moth old*** (3), ***make sure*** (3), ***let parti*** (3), mantendo a concordância dos temas abordados nesta categoria de negócio.

Pode-se constatar, que neste caso, a exploração feita com bigramas obteve melhores resultados, uma vez que, as palavras descrevem melhor a categoria de negócio e consequentemente os assuntos abordados.

Termos com frequência mínima de ocorrência

Depois das análises anteriores, tabela 39 e 40, e tendo em conta a o número de *reviews* desta categoria, 47, decidiu-se observar as palavras que ocorreram no mínimo para os unigramas 10 e 20 vezes e para os bigramas 2 e 3 vezes.

De seguida, apresenta-se a figura 58 e 59, onde se encontra o código utilizado para obter os termos que aparecem no mínimo referido anteriormente para os dois cenários:

```
> findFreqTerms(matriz, lowfreq=10)#termos que aparecem no mínimo 10 vezes
 [1] "also"       "and"        "can"        "day"        "dont"       "even"
 [7] "event"      "food"       "friend"     "get"        "good"       "great"
[13] "just"       "like"       "look"       "love"       "make"       "much"
[19] "nice"       "one"        "parti"      "photo"      "photograph" "pictur"
[25] "place"      "profession" "realli"     "recommend"  "she"        "shoot"
[31] "sure"       "take"       "they"       "time"       "use"        "want"
[37] "wed"        "well"       "will"       "work"
> findFreqTerms(matriz, lowfreq=20)#termos que aparecem no mínimo 20 vezes
[1] "get"   "good"  "great" "like"  "look"  "place" "wed"   "work"
```

FIGURA 58: UNIGRAMAS QUE APARECEM NO MÍNIMO 10 E 20 VEZES NA CATEGORIA *EVENT PLANNING&SERVICE*.

```
> findFreqTerms(rvwbi, lowfreq=2)#termos que aparecem no mínimo 2 vezes
 [1] "alli arts"       "anyon look"      "arts guild"      "balloon structur"
 [5] "bread crumb"     "bridal engag"    "can make"        "captur special"
 [9] "chicken salad"   "curri chicken"   "definit plan"    "dream hummus"
[13] "engag session"   "even better"     "everyth went"    "face face"
[17] "fantast work"    "feel like"       "fleurimond cater" "get marri"
[21] "get perfect"     "great place"     "great thing"     "great time"
[25] "help us"         "high qualiti"    "high recommend"  "im sure"
[29] "laura work"      "let parti"       "like your"       "love work"
[33] "make money"      "make sure"       "meet face"       "memori well"
[37] "month old"       "month wed"       "my husband"      "old timey"
[41] "palam rural"     "perfect shot"    "photo shoot"     "photograph came"
[45] "place good"      "purchas proof"   "put togeth"      "recommend anyon"
[49] "rural centr"     "six month"       "special day"     "take photo"
[53] "take pictur"     "ten thousand"    "they time"       "thousand villag"
[57] "top notch"       "tweed ride"      "us look"         "via email"
[61] "we got"          "wed photo"       "wed photograph"  "went without"
[65] "will pay"        "without hitch"   "work good"       "work palam"
[69] "work super"      "yam casserol"    "you can"         "you will"
> findFreqTerms(rvwbi, lowfreq=3)#termos que aparecem no mínimo 3 vezes
 [1] "anyon look"      "feel like"       "fleurimond cater" "help us"
 [5] "high recommend"  "let parti"       "make sure"       "month old"
 [9] "wed photograph"  "will pay"
```

FIGURA 59: BIGRAMAS QUE APARECEM NO MÍNIMO 2 VEZES NA CATEGORIA *EVENT PLANNING&SERVICE*.

Observando agora a exploração de unigramas, pode-se visualizar algumas palavras relevantes,como, *great, good, like, look,* entre outras, que distinguem bem esta categoria.

Observando a figura 58, cenário de unigramas, verifica-se que os unigramas mais frequentes são, ***good, like, look, place, work,*** entre outros.

Analisando agora a figura 59, análise de bigramas, pode-se observar que os bigramas mais frequentes, são, ***anyon look. feel like, fleurimon cater,*** entre outros.

Conclui-se que, esta categoria não esta bem representada nos termos, tanto em unigramas como em bigramas, como se pode observar nas análises efetuadas.

Wordcloud

Optou-se por fazer os *Wordclouds* com as frequências mínimas referenciadas nas análises anteriores, figura 58 e 59, para obter uma visualização mais clara, simples e perspicaz.

De seguida apresentam-se os *Wordclouds* (em que o tamanho fonte das palavras é proporcional à frequência dos termos) das frequências mínimas tendo em conta os dois cenários:

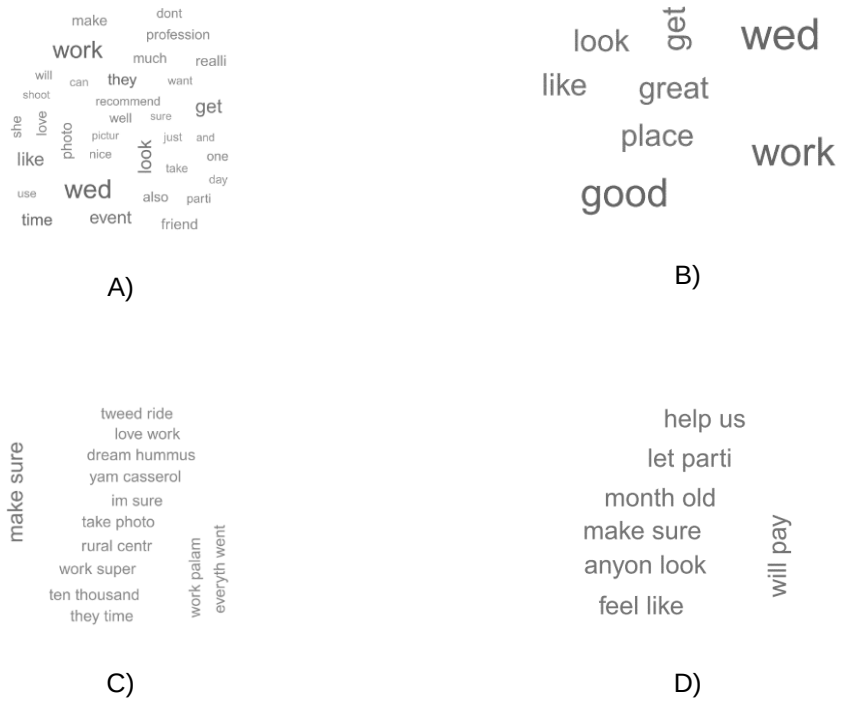


FIGURA 60: *WORDCLOUDEVENT PLANNING&SERVICE*: A) UNIGRAMAS FREQUÊNCIA MÍNIMA 10; B) UNIGRAMAS FREQUÊNCIA MÍNIMA 20; C) BIGRAMAS FREQUÊNCIA MÍNIMA 2; D) BIGRAMAS FREQUÊNCIA MÍNIMA 3.

Financial Services

Como se pode observar na tabela 2, esta categoria de negócio está constituída por *16reviews*.

Esta categoria contempla opiniões sobre seguros, investimentos, bancos, serviços de créditos, casas de câmbio, entre outros temas similares.

Foi feita uma exploração de dados para se compreender e para ter uma perceção mais clara sobre os dados.

De seguida, apresentam-se as explorações nesta categoria de negócio nos dois cenários:

Palavras mais e menos frequentes

Após os tratamentos dos dados, observou-se os termos mais e menos frequentes, como se pode observar de seguida:

Termos menos frequêntes	Quantidade	Termos mais frequêntes	Quantidade
acct	1	locat	9
acknowledg	1	time	9
adida	1	account	10
agre	1	servic	10
aircondit	1	branch	12
all	1	bank	18

TABELA 41: UNIGRAMAS MAIS E MENOS FREQUENTES NA CATEGORIA *FINANCIAL SERVICES*.

Termos menos frequêntes	Quantidade	Termos mais frequêntes	Quantidade
a not	1	atm fee	3
abl call	1	busi account	3
abl save	1	credit card	3
account credit	1	credit union	3
account line	1	wf locat	3
account mild	1	well fargo	4

TABELA 42: BIGRAMAS MAIS E MENOS FREQUENTES NA CATEGORIA *FINANCIAL SERVICES*.

Ao observar as duas figuras anteriores, pode-se deduzir que, as palavras mais frequentes são as que melhor descrevem esta categoria de negócio abordada nos 16 *reviews*.

Ao analisar a figura de unigramas, tabela 41, pode-se visualizar as palavras mais e menos frequentes nesta categoria de negócio, podendo concluir-se que as palavras mais frequentes são, ***bank***(18), ***branch*** (12), ***servic***(10), ***account*** (10), ***time*** (9) e ***locat*** (9).

Ao fazer exploração da frequência com bigramas, tabela 42, os bigramas mais frequentes são, ***well fargo*** (4), ***wf locat*** (3), ***credit union*** (3), ***credit card*** (3), ***busi accoount*** (3), ***atm fee*** (3), mantendo a concordância dos temas abordados nesta categoria de negócio.

Termos com frequência mínima de ocorrência

Devido à exploração feita anteriormente, como se pode observar nas tabelas 41 e 42, e tendo em conta o número de *reviews*, 16, optou-se por observar as palavras que ocorrem no mínimo 5 e 10 vezes, no cenário de unigramas, e no mínimo 2 vezes, no cenário de bigramas.

De seguida, apresenta-se as figuras 61 e 62, onde se encontra o código utilizado para obter os termos que aparecem pelo menos 5 e 10 vezes para o cenário de unigramas e, 2 vezes para os bigramas, juntamente com os termos associados.

```
> findFreqTerms(matriz, lowfreq=5)#termos que aparecem no mínimo 5 vezes
 [1] "account" "alway"   "appli"   "ask"     "atm"     "bank"    "bill"    "branch"
 [9] "busi"    "call"    "can"     "card"    "check"   "credit"  "custom"  "get"
[17] "good"    "help"    "know"    "locat"   "love"    "need"    "never"   "nice"
[25] "one"     "open"    "pay"     "person"  "servic"  "teller"  "time"    "wait"
[33] "well"    "work"
> findFreqTerms(matriz, lowfreq=10)#termos que aparecem no mínimo 10 vezes
[1] "account" "bank"    "branch"  "servic"
```

FIGURA 61: UNIGRAMAS QUE APARECEM NO MÍNIMO 5 E 10 VEZES NA CATEGORIA *FINANCIAL SERVICES*.

```
> findFreqTerms(rvwbi, lowfreq=2)#termos que aparecem no mínimo 2 vezes
 [1] "atm fee"       "bank compani"  "busi account"  "car loan"      "chang sinc"
 [6] "credit card"   "credit union"  "custom servic" "debit card"    "dont know"
[11] "fargo locat"   "form letter"   "get cash"      "la resid"      "need done"
[16] "open account"  "pay atm"       "place get"     "pretti good"   "save account"
[21] "util bill"     "well fargo"    "wf locat"      "you will"
```

FIGURA 62: BIGRAMAS QUE APARECEM NO MÍNIMO 2 VEZES NA CATEGORIA *FINANCIAL SERVICES*.

Como se pode observar na figura da exploração de unigramas, figura 61, as palavras relevantes que distinguem a categoria, são,***account, bank, servic, branch, card***, entre outras.

Observando agora a exploração do cenário de bigramas, figura 62, pode-se constatar que os bigramas vão de encontro à categoria de negócio, podendo enaltecer alguns bigramas que se identificam com esta categoria, como*, **atm fee, back compani, busi account**,* entre outros.

Wordcloud

Optou-se por fazer os *Wordclouds* com as frequências mínimas referenciadas nas análises anteriores, figura 61 e 62, para obter uma visualização mais clara, simples e perspicaz.

De seguida apresentam-se os *Wordclouds* (em que o tamanho fonte das palavras é proporcional à frequência dos termos) das frequências mínimas tendo em conta os dois cenários:

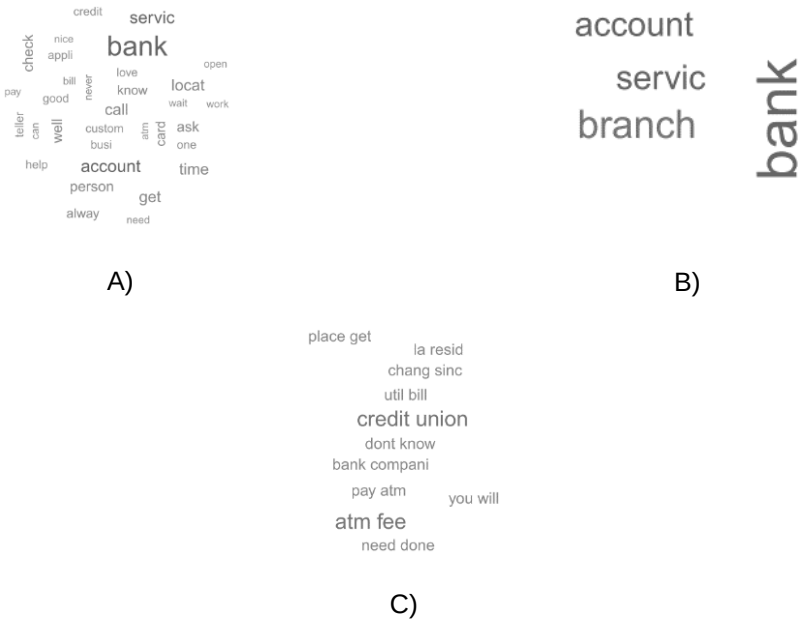


FIGURA 63: *WORDCLOUD* DA CATEGORIA *FINANCIAL SERVICES*: A) UNIGRAMAS FREQUÊNCIA MÍNIMA 5; B) UNIGRAMAS FREQUÊNCIA MÍNIMA 10; C) BIGRAMAS FREQUÊNCIA MÍNIMA 2.

Food

Como se pode observar na tabela 2, esta categoria de negócio está constituída por 2430 *reviews*.

Esta categoria reflete opiniões sobre chá e café, lojas de conveniência, gelatarias, padarias, comida especializada, salões de chá, mercearias, cervejeiras, doçaria, mercados, sumos e batidos, quiosques, garrafeiras, talho, peixaria entre outros.

Foi feita uma exploração de dados para se compreender e para ter uma perceção mais clara sobre os dados.

De seguida, apresentam-se as explorações nesta categoria de negócio:

Palavras mais e menos frequentes

Após os tratamentos dos dados, observou-se os termos mais e menos frequentes, como se pode observar de seguida:

Termos menos frequêntes	Quantidade	Termos mais frequêntes	Quantidade
aaaand	1	just	1005
aahh	1	food	1066
aalbeit	1	get	1164
abandon	1	like	1381
abbrevi	1	good	1423
abc	1	place	1603

TABELA 43: UNIGRAMAS MAIS E MENOS FREQUENTES NA CATEGORIA *FOOD*.

Termos menos frequêntes	Quantidade	Termos mais frequêntes	Quantidade
a beguil	1	can get	109
a big	1	go back	110
a bohemian	1	pretti good	115
a book	1	red velvet	150
a buddi	1	this place	153
a bunch	1	ice cream	477

TABELA 44: BIGRAMAS MAIS E MENOS FREQUENTESNA CATEGORIA *FOOD*.

Pode-se observar na análise de unigramas, tabela 43, as palavras mais e menos frequentes nesta categoria de negócio, são, ***place***(1603), ***good***(1423), ***like***(1381), ***get*** (1164), ***food***(1066) e ***just***(1005).

Ao fazer exploração da frequência com bigramas, tabela 44, os bigramas mais frequentes, que são, ***ice cream*** (477), ***this place*** (153), ***red velvet*** (150), ***pretti good*** (115), ***go back*** (110), ***can get*** (109), mantendo a concordância dos temas abordados nesta categoria de negócio.

Termos com frequência mínima de ocorrência

Devido à exploração feita anteriormente, tendo em conta os dois cenários, tabela 43 e 44, e tendo em conta o número de *reviews*, 2430, optou-se por observar as palavras que ocorrem no mínimo 700 e 1000 vezes, no cenário dos unigramas, e que ocorrem no mínimo 50 e 100, bigramas.

De seguida, apresenta-se a figura 64 e 65, onde se encontra o código utilizado para obter os termos que ocorrem no mínimo, como foi referido anteriormente para os 2 cenários.

```
> findFreqTerms(matrizesp, lowfreq=700)#termos que aparecem no mínimo 700 vezes
  [1] "actual"    "also"     "alway"    "amaz"     "and"      "anoth"
  [7] "anyth"     "area"     "around"   "ask"      "atmospher" "away"
 [13] "awesom"    "back"     "bad"      "bar"      "beer"     "best"
 [19] "better"    "big"      "bit"      "bread"    "burger"   "busi"
 [25] "but"       "call"     "came"     "can"      "cant"     "cheap"
 [31] "check"     "chees"    "chicken"  "chocol"   "clean"    "close"
 [37] "coffe"     "come"     "cook"     "cream"    "custom"   "day"
 [43] "decent"    "definit"  "delici"   "dessert"  "didnt"    "differ"
 [49] "dinner"    "dish"     "dont"     "drink"    "eat"      "els"
 [55] "end"       "enjoy"    "enough"   "even"     "ever"     "everi"
 [61] "everyth"   "expect"   "experi"   "favorit"  "feel"     "find"
 [67] "first"     "flavor"   "food"     "for"      "found"    "free"
 [73] "fresh"     "fri"      "friend"   "get"      "give"     "good"
 [79] "got"       "great"    "guy"      "happi"    "help"     "high"
 [85] "home"      "hot"      "hour"     "howev"    "huge"     "ice"
 [91] "ill"       "item"     "its"      "ive"      "just"     "kind"
 [97] "know"      "last"     "like"     "line"     "littl"    "live"
[103] "locat"     "long"     "look"     "lot"      "love"     "lunch"
[109] "made"      "make"     "mani"     "mayb"     "meal"     "meat"
[115] "menu"      "minut"    "much"     "need"     "never"    "new"
[121] "next"      "nice"     "night"    "not"      "noth"     "now"
[127] "offer"     "one"      "open"     "option"   "order"    "park"
[133] "pay"       "peopl"    "perfect"  "person"   "pizza"    "place"
[139] "plate"     "portion"  "pretti"   "price"    "probabl"  "put"
[145] "qualiti"   "quick"    "quit"     "realli"   "reason"   "recommend"
[151] "restaur"   "review"   "rice"     "right"    "roll"     "room"
[157] "said"      "salad"    "sandwich" "sauc"     "say"      "seat"
[163] "see"       "seem"     "select"   "serv"     "servic"   "shop"
[169] "side"      "sinc"     "sit"      "small"    "someth"   "special"
[175] "spot"      "staff"    "star"     "start"    "still"    "stop"
[181] "store"     "student"  "super"    "sure"     "sweet"    "tabl"
[187] "take"      "tast"     "tasti"    "tea"      "that"     "there"
[193] "they"      "thing"    "think"    "this"     "though"   "thought"
[199] "time"      "took"     "top"      "tri"      "two"      "use"
[205] "usual"     "visit"    "wait"     "walk"     "want"     "wasnt"
[211] "way"       "week"     "well"     "went"     "when"     "will"
[217] "wine"      "work"     "worth"    "year"     "you"      "your"
> findFreqTerms(matriz, lowfreq=1000)#termos que aparecem no mínimo 1000 vezes
[1] "food"  "get"   "good"  "just"  "like"  "place"
```

FIGURA 64: UNIGRAMAS QUE APARECEM NO MÍNIMO 700 E 1000 VEZES NA CATEGORIA *FOOD*.

```
> findFreqTerms(rvwbi, lowfreq=50)#termos que aparecem no mínimo 50 vezes
 [1] "behind counter" "bubbl tea"     "can get"       "chocol chip"
 [5] "coffe shop"     "come back"     "cream chees"   "cream sandwich"
 [9] "dont know"      "everi time"    "feel like"     "first time"
[13] "frozen yogurt"  "go back"       "great place"   "ice cream"
[17] "if your"        "im sure"       "ive ever"      "late night"
[21] "look like"      "mexican food"  "much better"   "my favorit"
[25] "next time"      "peanut butter" "pretti good"   "realli good"
[29] "red velvet"     "tast like"     "they also"     "this place"
[33] "you can"
> findFreqTerms(rvwbi, lowfreq=100)#termos que aparecem no mínimo 100 vezes
[1] "can get"      "frozen yogurt" "go back"      "ice cream"    "pretti good"
[6] "red velvet"   "tast like"     "this place"
```

FIGURA 65:BIGRAMAS QUE APARECEM NO MÍNIMO 50 E 100 VEZES DA CATEGORIA *FOOD*.

Ao analisar a exploração de unigramas, figura 64, pode-se visualizar algumas palavras relevantes, como, ***food, like, place*****,** entre outras, que distinguem bem esta categoria.

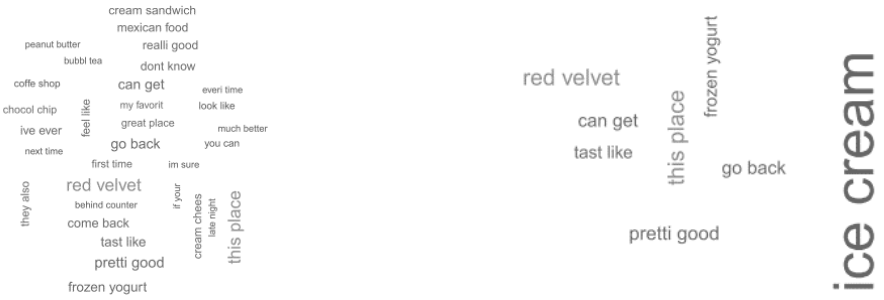
Tanto em unigramas como em bigramas as palavras mais frequentes não distinguem bem a categoria de negócio.

Wordcloud

Optou-se por fazer os *Wordclouds* com as frequências mínimas referenciadas nas análises anteriores, figura 64 e 65, para obter uma visualização mais clara, simples e perspicaz.

De seguida apresentam-se os *Wordclouds* (em que o tamanho fonte das palavras é proporcional à frequência dos termos) das frequências mínimas tendo em conta os dois cenários:

A) B)



C) D)

FIGURA 66: *WORDCLOUD*DA CATEGORIA *FOOD*: A) UNIGRAMAS FREQUÊNCIA MÍNIMA 700; B) UNIGRAMAS FREQUÊNCIA MÍNIMA 1000; C)BIGRAMAS FREQUÊNCIA MÍNIMA 50; D) BIGRAMAS FREQUÊNCIA MÍNIMA 100.

Health & Medical

Como se pode observar na tabela 2, esta categoria de negócio está constituída por *198reviews*.

Esta categoria constitui opiniões dos utilizadores sobre médicos, fisioterapias, optometristas, dentistas, clinicas médicas, hospitais, urgências, nutricionistas, lares de idosos, farmácias, enfermeiros entre outros.

Foi feita uma exploração de dados para se compreender e para ter uma perceção mais clara sobre os dados.

De seguida, apresentam-se as explorações nesta categoria de negócio:

Palavras mais e menos frequentes.

Após os tratamentos dos dados, observou-se os termos mais e menos frequentes, como se pode observar de seguida:

Termos menos frequêntes	Quantidade	Termos mais frequêntes	Quantidade
aaa	1	offic	101
abil	1	one	103
abscess	1	care	115
accent	1	doctor	123
achill	1	get	146
acid	1	time	155

TABELA 45: UNIGRAMAS MAIS E MENOS FREQUENTES NA CATEGORIA *HEALTH&MEDICAL*.

Termos menos frequêntes	Quantidade	Termos mais frequêntes	Quantidade
a doctor	1	from desk	13
a friend	1	see dr	13
a great	1	feel like	14
a half	1	wait room	16
a help	1	dr tabsh	19
a hour	1	high recommend	21

TABELA 46: BIGRAMAS MAIS E MENOS FREQUENTES NA CATEGORIA *HEALTH&MEDICAL*.

Ao observar a análise de unigramas, pode-se constatar, com a tabela 45, as palavras mais frequentes nesta categoria de negócio são, ***time***(155), ***get*** (146), ***doctor***(123), ***care*** (115), ***one*** (103) e ***offic*** (101).

Ao fazer exploração da frequência com bigramas, como se pode observar na figura a cima, os bigramas mais frequentes, que são*,* ***high recommend*** *(21*), ***dr tabsh*** (19), ***wait room*** (16), ***feel like*** (14), ***see dr*** (13) e ***front desk*** (13), mantendo a concordância dos temas abordados nesta categoria de negócio.

Termos com frequência mínima de ocorrência

Tendo em conta as análises feitas anteriormente, bem como o número de *reviews* nesta categoria, 198, optou-se por observar as palavras que ocorrem no mínimo 50 e 100 vezes, no cenário de unigramas, e no mínimo 5 e 10 para os bigramas.

Na continuação apresentam-se as figuras67 e 68, onde se pode observar o código utilizado para obter os termos que aparecem no mínimo tendo em conta os dois cenários.

```
> findFreqTerms(matriz, lowfreq=50)#termos que aparecem no mínimo 50 vezes
 [1] "also"     "appoint"  "back"     "best"     "call"     "can"
 [7] "care"     "day"      "dentist"  "doctor"   "dont"     "even"
[13] "experi"   "eye"      "feel"     "friend"   "get"      "good"
[19] "great"    "hour"     "insur"    "ive"      "just"     "know"
[25] "like"     "look"     "make"     "need"     "never"    "offic"
[31] "one"      "patient"  "place"    "realli"   "recommend" "see"
[37] "she"      "staff"    "they"     "time"     "wait"     "well"
[43] "will"     "work"     "year"
> findFreqTerms(matriz, lowfreq=100)#termos que aparecem no mínimo 100 vezes
[1] "care"   "doctor" "get"    "offic"  "one"    "time"
```

FIGURA 67: UNIGRAMAS QUE APARECEM NO MÍNIMO 50 E 100 VEZESNA CATEGORIA *HEALTH&MEDICAL*.

```
> findFreqTerms(rvwbi, lowfreq=5)#termos que aparecem no mínimo 5 vezes
 [1] "answer question" "appoint dr"      "appoint time"    "ask question"
 [5] "bedsid manner"   "bill insur"      "call back"       "call offic"
 [9] "can get"         "care patient"    "come back"       "coupl day"
[13] "custom eye"      "custom servic"   "dental work"     "desk staff"
[17] "do not"          "doctor dr"       "doctor offic"    "dont think"
[21] "dr choy"         "dr eshom"        "dr johnson"      "dr mathew"
[25] "dr most"         "dr perlman"      "dr szpiro"       "dr tabsh"
[29] "entir staff"     "even though"     "everi day"       "everi singl"
[33] "everi time"      "explain everyth" "eye exam"        "feel like"
[37] "friend famili"   "front desk"      "get appoint"     "go back"
[41] "go dr"           "great experi"    "he also"         "high recommend"
[45] "if want"         "im sure"         "in addit"        "insur compani"
[49] "ive come"        "ive ever"        "ive never"       "last year"
[53] "long time"       "look like"       "love love"       "made appoint"
[57] "made feel"       "make appoint"    "make feel"       "make sure"
[61] "month later"     "offic staff"     "oral surgeon"    "primari care"
[65] "recommend dr"    "root canal"      "see doctor"      "see dr"
[69] "somewher els"    "staff friend"    "they also"       "this place"
[73] "time ive"        "took time"       "tri get"         "urgent care"
[77] "wait hour"       "wait room"       "wait time"       "want go"
[81] "went dr"         "wisdom tooth"    "work done"       "year ago"
[85] "year later"      "year now"        "year old"
> findFreqTerms(rvwbi, lowfreq=10)#termos que aparecem no mínimo 10 vezes
 [1] "answer question" "come back"       "dr tabsh"        "even though"
 [5] "everi time"      "feel like"       "front desk"      "go back"
 [9] "high recommend"  "ive ever"        "look like"       "make feel"
[13] "recommend dr"    "see dr"          "wait room"       "year ago"
```

FIGURA 68: BIGRAMAS QUE APARECEM NO MÍNIMO 5 E 10 VEZES NA CATEGORIA *HEALTH&MEDICAL*.

Esta categoria contempla opiniões sobre a área da saúde, na exploração no cenário de unigramas, pode-se observar algumas palavras relevantes, como, ***doctor, offic, dentist****,* entre outras, que diferenciam bem a categoria.

No entanto, pode-se constatar que, o cenário que vai mais de encontro com a categoria de negócio é o de unigramas, uma vez que, está melhor representado nos termos.

Wordcloud

Optou-se por fazer os *Wordclouds* com as frequências mínimas referenciadas nas análises anteriores, figura 67 e 68, para obter uma visualização mais clara, simples e perspicaz.

De seguida apresentam-se os *Wordclouds* (em que o tamanho fonte das palavras é proporcional à frequência dos termos) das frequências mínimas tendo em conta os dois cenários:

FIGURA 69: *WORDCLOUD*DA CATEGORIA *HEALTH&MEDICAL*: A) UNIGRAMAS FREQUÊNCIA MÍNIMA 50; B) UNIGRAMAS FREQUÊNCIA MÍNIMA 100; C) BIGRAMAS FREQUÊNCIA MÍNIMA 5; D) BIGRAMAS FREQUÊNCIA MÍNIMA 10.

Home Services

Como se pode observar na tabela 2, esta categoria de negócio está constituída por 66 *reviews*.

Home Services está constituído por opiniões sobre decoração de interiores, canalizadores, chaves e serralharia, mudanças, imobiliário, limpeza domestica, jardineiros, cortinas e estores, pintores, eletricistas, sistemas de segurança entre outros.

Foi feita uma exploração de dados para se compreender e para ter uma perceção mais clara sobre os dados.

De seguida, apresentam-se as explorações nesta categoria de negócio nos dois cenários:

Palavras mais e menos frequentes

Após os tratamentos dos dados, observou-se os termos mais e menos frequentes, como se pode observar de seguida:

Termos menos frequêntes	Quantidade	Termos mais frequêntes	Quantidade
abandon	1	live	45
about	1	time	46
abund	1	one	48
accent	1	they	49
access	1	place	54
accessori	1	work	63

TABELA 47: UNIGRAMAS MAIS E MENOS FREQUENTES NA CATEGORIA *HOME SERVICES*.

Termos menos frequêntes	Quantidade	Termos mais frequêntes	Quantidade
a function	1	properti manag	5
a lifestyl	1	real estat	5
a lot	1	apart complex	6
a resid	1	do not	6
a scam	1	this place	6
a sent	1	garag door	15

TABELA 48: BIGRAMAS MAIS E MENOS FREQUENTES NA CATEGORIA *HOME SERVICES*.

Pode-se constatar segundo a análise de unigramas, tabela 47, que as palavras mais frequentes nesta categoria de negócio são, ***work***(63), ***place***(54), ***they***(49), ***one***(48), ***time*** (46) e***live*** (45).

Ao fazer exploração da frequência com bigramas, na tabela 48, os bigramas mais frequentes, que são,***garag door***(15),***this place***(6), ***do not*** (6), ***apart complex*** (6), ***real estat*** (5) e ***properti manag*** (5), mantendo a concordância dos temas abordados nesta categoria de negócio.

Termos com frequência mínima de ocorrência

Tendo em conta as explorações anteriores e o número de *reviews*, 66, optou-se por observar as palavras que ocorrem no mínimo 20 e 30 vezes, no caso de unigramas, e no caso de bigramas 3 e 5.

De seguida, apresenta-se a figura 70 e 71, onde se encontra o código utilizado para obter os termos tendo em conta os termos mínimos, referidos anteriormente nos dois cenários.

```
> findFreqTerms(matriz, lowfreq=20)#termos que aparecem no minimo 20 vezes
 [1] "also"   "apart"  "around" "call"   "can"    "car"    "clean"  "day"
 [9] "dont"   "door"   "even"   "floor"  "garag"  "get"    "good"   "great"
[17] "hous"   "just"   "key"    "know"   "like"   "live"   "look"   "make"
[25] "manag"  "month"  "move"   "much"   "need"   "never"  "new"    "nice"
[33] "now"    "old"    "one"    "park"   "peopl"  "place"  "price"  "room"
[41] "servic" "they"   "thing"  "this"   "time"   "use"    "walk"   "want"
[49] "well"   "will"   "work"   "year"
> findFreqTerms(matriz, lowfreq=30)#termos que aparecem no minimo 30 vezes
 [1] "also"  "apart" "clean" "day"   "dont"  "door"  "get"   "great" "just"
[10] "key"   "like"  "live"  "look"  "move"  "need"  "one"   "park"  "place"
[19] "room"  "they"  "time"  "use"   "will"  "work"
```

FIGURA 70: UNIGRAMAS QUE APARECEM PELO MENOS 20 E 30 VEZES NA CATEGORIA *HOME SERVICES*.

```
> findFreqTerms(rvwbi, lowfreq=3)#termos que aparecem no mínimo 3 vezes
 [1] "abl get"       "almost year"    "apart complex"  "bus stop"
 [5] "car broken"    "clean bathroom" "clean servic"   "colleg kid"
 [9] "colleg student" "come back"     "costa verd"     "custom servic"
[13] "do not"        "dont think"     "dont want"      "dream dinner"
[17] "dri wall"      "estat agent"    "even though"    "everi month"
[21] "faa took"      "first time"     "front door"     "garag door"
[25] "gave us"       "give us"        "he friend"      "high recommend"
[29] "im sure"       "in addit"       "just feel"      "leas offic"
[33] "live room"     "look like"      "mainten staff"  "melior maid"
[37] "never call"    "next day"       "open window"    "park spot"
[41] "peopl work"    "pest control"   "phone call"     "place live"
[45] "pretti good"   "pretti much"    "properti manag" "real estat"
[49] "reason price"  "rental hous"    "right next"     "secur guard"
[53] "show us"       "thing will"     "this place"     "time one"
[57] "use melior"    "vacuum cleaner" "want live"
> findFreqTerms(rvwbi, lowfreq=5)#termos que aparecem no mínimo 5 vezes
[1] "apart complex"  "bus stop"       "custom servic"  "do not"
[5] "garag door"     "leas offic"     "properti manag" "real estat"
[9] "this place"
```

FIGURA 71: BIGRAMAS QUE APARECEM PELO MENOS 3 E 5 VEZES NA CATEGORIA *HOME SERVICES*.

Ao analisar o cenário unigrama, na figura 70, deduz-se que esta categoria não se encontram bem descrita com a frequência das palavras.

Ao fazer exploração da frequência com bigramas, na figura 71, os bigramas mais frequentes, que são,***garag door***(15),***this place***(6), ***do not*** (6), ***apart complex*** (6), ***real estat*** (5) e ***properti manag*** (5), mantendo a concordância dos temas abordados nesta categoria de negócio.

Wordcloud

Optou-se por fazer os *Wordclouds* com as frequências mínimas referenciadas nas análises anteriores, figura 70 e 71, para obter uma visualização mais clara, simples e perspicaz.

De seguida apresentam-se os *Wordclouds* (em que o tamanho fonte das palavras é proporcional à frequência dos termos) das frequências mínimas tendo em conta os dois cenários:

A)

B)

C) D)

FIGURA 72: *WORDCLOUD*DA CATEGORIA *HOME SERVICES*: A) UNIGRAMAS FREQUÊNCIA MÍNIMA 20; B) UNIGRAMAS FREQUÊNCIA MÍNIMA 30; C) BIGRAMAS FREQUÊNCIA MÍNIMA 3; D) BIGRAMAS FREQUÊNCIA MÍNIMA 5.

Hotels & Travel

Como se pode observar na tabela 2, esta categoria de negócio está constituída por 184 *reviews*.

Esta categoria contém opiniões sobre visitas guiadas, pensões, aluguer de carros, aeroportos, hotéis, transportes, agência de viagens, parques de campismo, hostels, aluguer de motas, arrendamentos para férias, termas medicinais entre outros.

Foi feita uma exploração de dados para se compreender e para ter uma perceção mais clara sobre os dados.

De seguida, apresentam-se as explorações nesta categoria de negócio:

Palavras mais e menos frequentes

Após os tratamentos dos dados, observou-se os termos mais e menos frequentes tendo em conta os dois cenários, como se pode observar de seguida:

Termos menos frequêntes	Quantidade	Termos mais frequêntes	Quantidade
aaa	1	place	103
aarp	1	just	104
abid	1	get	118
abil	1	stay	136
aboard	1	hotel	160
absent	1	room	220

TABELA 49: UNIGRAMAS MAIS E MENOS FREQUENTES NA CATEGORIA *HOTELS&TRAVEL*.

Termos menos frequêntes	Quantidade	Termos mais frequêntes	Quantidade
a bar	1	feel like	9
a cab	1	across street	10
a littl	1	dont know	10
a must	1	custom servic	11
a nice	1	come back	14
a question	1	front desk	20

TABELA 50: BIGRAMAS MAIS E MENOS FREQUENTES NA CATEGORIA *HOTELS&TRAVEL*.

Pode-se deduzir; após observar a tabela 49, as palavras mais frequentes nesta categoria de negócio são, ***room*** *(*220), ***hotel***(160), ***stay***(136), ***get*** (118), ***just***(104) e ***place***(103).

Após a exploração da frequência de bigramas, como se pode observar na tabela 50, os bigramas mais frequentes são, ***front desk***(20),***come back***(14), ***custom servic*** (11), ***dont know*** (10), ***across street*** (10) e **feel like** (9), mantendo a concordância dos temas abordados nesta categoria de negócio.

Pode-se constatar, que neste caso, que a exploração feita com unigramas obteve melhores resultados, uma vez que, as palavras descrevem melhor a categoria de negócio e consequentemente os assuntos abordados.

Termos com frequência mínima de ocorrência

Devido às explorações feitas anteriormente, tabela 49 e 50, e tendo em conta o número de *reviews,* 184, optou-se por analisar as palavras que ocorrem no mínimo 30, 50 e 100 vezes, para os unigramas e 5 e 10 para os bigramas.

De seguida, apresentam-se as figuras 73 e 74, com o código utilizado para obter os termos que aparecem pelo menos 30 e 50 vezes, unigramas, e 5 e 10, bigramas, bem como os termos associados.

```
> findFreqTerms(matriz, lowfreq=30)#termos que aparecem no minimo 30
 [1] "also"     "area"     "around"   "back"     "bad"      "bar"
 [7] "bathroom" "bed"      "breakfast" "bus"     "call"     "can"
[13] "car"      "clean"    "come"     "day"      "desk"     "didnt"
[19] "dont"     "even"     "feel"     "food"     "free"     "friend"
[25] "front"    "get"      "good"     "got"      "great"    "help"
[31] "hotel"    "hour"     "its"      "just"     "know"     "like"
[37] "littl"    "locat"    "look"     "love"     "make"     "minut"
[43] "much"     "need"     "nice"     "night"    "one"      "park"
[49] "peopl"    "place"    "pretti"   "price"    "realli"   "right"
[55] "room"     "servic"   "staff"    "stay"     "suit"     "take"
[61] "there"    "this"     "time"     "trip"     "two"      "use"
[67] "walk"     "want"     "way"      "well"     "will"     "work"
[73] "your"
> findFreqTerms(matriz, lowfreq=50)#termos que aparecem no minimo 50 vezes
 [1] "back"   "bed"    "can"    "dont"   "get"    "good"   "great"  "hotel"
 [9] "just"   "like"   "nice"   "night"  "one"    "park"   "place"  "realli"
[17] "room"   "servic" "staff"  "stay"   "time"
```

FIGURA 73: UNIGRAMAS QUE APARECEM PELO MENOS 30 E 50 VEZES NA CATEGORIA *HOTELS&TRAVEL*.

```
> findFreqTerms(rvwbi, lowfreq=5)#termos que aparecem pelo menos 5 vezes
 [1] "across street" "around corner" "bed comfort"   "book room"
 [5] "can get"       "come back"     "comfort bed"   "conveni locat"
 [9] "custom servic" "desk staff"    "dont know"     "even though"
[13] "faculti club"  "feel like"     "felt like"     "first time"
[17] "flat screen"   "free wifi"     "front desk"    "front door"
[21] "get room"      "give star"     "go back"       "good servic"
[25] "great deal"    "half hour"     "hot water"     "hotel room"
[29] "im sure"       "ive stay"      "last week"     "live room"
[33] "look like"     "make sure"     "night stay"    "park car"
[37] "place stay"    "pool area"     "pretti good"   "restaur bar"
[41] "return car"    "room clean"    "room nice"     "rush hour"
[45] "second time"   "staff friend"  "stay hotel"    "th floor"
[49] "this hotel"    "this place"    "valet park"    "walk distanc"
[53] "water pressur" "within walk"   "you can"
> findFreqTerms(rvwbi, lowfreq=10)#termos que aparecem pelo menos 10 vezes
[1] "across street" "come back"     "custom servic" "dont know"
[5] "front desk"
```

FIGURA 74: BIGRAMAS QUE APARECEM PELO MENOS 5 E 10 VEZES NA CATEGORIA *HOTELS&TRAVEL*.

Como se pode observar nas figuras 73e 74, quanto maior a frequência das palavras melhor está representada a categoria de negócio.

Na exploração no cenário de unigramas, pode-se observar algumas palavras relevantes, como, *hotel, place, room*, entre outras, que distinguem bem esta categoria.

Constata-se que nesta categoria de negócio, o cenário de unigramas representa melhor a categoria como se pode observar.

Wordcloud

Optou-se por fazer os *Wordclouds* com as frequências mínimas referenciadas nas análises anteriores, figura 73 e 74, para obter uma visualização mais clara, simples e perspicaz.

De seguida apresentam-se os *Wordclouds* (em que o tamanho fonte das palavras é proporcional à frequência dos termos) das frequências mínimas tendo em conta os dois cenários:

A)

B)

C)

D)

FIGURA 75: *WORDCLOUD*DA CATEGORIA *HOTELS&TRAVEL*:A) UNIGRAMAS FREQUÊNCIA MÍNIMA 30; B) UNIGRAMAS FREQUÊNCIA MÍNIMA 50; C)BIGRAMAS FREQUÊNCIA MÍNIMA 5; D) BIGRAMAS FREQUÊNCIA MÍNIMA 10.

Local Flavor

Como se pode observar na tabela 2, esta categoria de negócio está constituída por *8 reviews*.

Nesta categoria, as opiniões que se podem encontrar são sobre sítios únicos em cada cidade.

Foi feita uma exploração de dados para se compreender e para ter uma perceção mais clara sobre os dados.

De seguida, apresentam-se as explorações nesta categoria de negócio tendo em conta os dois cenários:

Palavras mais e menos frequentes

Após os tratamentos dos dados, observou-se os termos mais e menos frequentes, como se pode observar de seguida:

Termos menos frequêntes	Quantidade	Termos mais frequêntes	Quantidade
abl	1	food	8
across	1	realli	9
activ	1	tabl	9
actual	1	time	10
addit	1	one	12
allerg	1	event	17

TABELA 51: UNIGRAMAS MAIS E MENSO FREQUENTES NA CATEGORIA *LOCAL FLAVOR*.

Termos menos frequêntes	Quantidade	Termos mais frequêntes	Quantidade
a great	1	stick around	2
a strang	1	will definit	2
abbi great	1	bay area	3
abbi organ	1	elit event	3
abbi whi	1	jeremiah innocent	3
abl eat	1	red wine	3

TABELA 52: BIGRAMAS MAIS E MENOS FREQUENTES NA CATEGORIA *LOCAL FLAVOR*.

Após observar a tabela 51, exploração de unigramas, pode-se depreender que as palavras mais frequentes nesta categoria de negócio são, ***event*** *(*17), ***one*** (12), ***time***(10), ***tabl***(9), ***realli*** (9) e ***food*** (8).

Também se pode constatar, após observar a tabela 52, que os bigramas mais frequentes nesta categoria de negócio, que são os seguintes, ***red wine***(3),***jeremiah innocent***(3), ***elit event*** (3), ***bay area*** (3), **will definit** (2) e ***stick around*** (2), mantendo a concordância dos temas abordados nesta categoria de negócio.

Termos com frequência mínima de ocorrência

Optou-se por fazer uma exploração para se observar as palavras com uma frequência mínima para os dois cenários, tendo em conta o número de *reviews* bem como as análises anterior. Nestas explorações a frequência mínima para unigramas é de 5 e 10 vezes e para bigramas optou-se pela frequência mínima de 2 e 3 vezes.

De seguida, apresenta-se as figuras 76 e 77, tendo em conta os dois cenários, com o código utilizado para obter a frequência dos termos:

```
> findFreqTerms(matriz, lowfreq=5)#termos que aparecem no mínimo 5 vezes
 [1] "can"      "chees"    "definit"  "eat"      "event"    "food"
 [7] "good"     "got"      "home"     "hot"      "know"     "like"
[13] "music"    "one"      "realli"   "red"      "rosemari" "sangria"
[19] "shrimp"   "sinc"     "tabl"     "think"    "time"     "went"
[25] "wine"     "year"
> findFreqTerms(matriz, lowfreq=10)#termos que aparecem no mínimo 10 vezes
[1] "event" "one"   "time"
```

FIGURA 76: UNIGRAMAS QUE APARECEM PELO MENOS 5 E 10 VEZES NA CATEGORIA *LOCAL FLAVOR*.

```
> findFreqTerms(rvwbi, lowfreq=2)#termos que aparecem no mínimo 2 vezes
 [1] "arm arm"          "bay area"          "chees nut"
 [4] "didnt know"       "dont think"        "elit event"
 [7] "gumbo fest"       "heard event"       "henrietta tabl"
[10] "hi how"           "hot day"           "isidor ida"
[13] "jeremiah innocent" "kaitlynn s"       "mumbo gumbo"
[16] "my favorit"       "oh man"            "realli enjoy"
[19] "realli stood"     "red wine"          "rosemari simpl"
[22] "shrimp crawfish"  "simpl syrup"       "stay bay"
[25] "stick around"     "will definit"
> findFreqTerms(rvwbi, lowfreq=3)#termos que aparecem no mínimo 3 vezes
[1] "bay area"          "elit event"        "jeremiah innocent"
[4] "red wine"
```

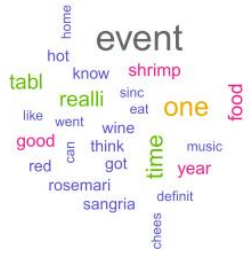
FIGURA 77: BIGRAMAS QUE APARECEM PELO MENOS 2 E 3 VEZES NA CATEGORIA *LOCAL FLAVOR*.

Esta categoria contempla sítios únicos das cidades, por esse motivo é difícil distinguir esta categoria através das palavras mais frequentes, visto que, esta categoria tem um leque muito variado, desde restaurantes, monumentos, concertos, eventos, parques, museus, entre outros.

Wordcloud

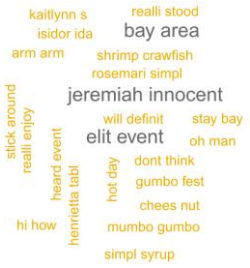
Optou-se por fazer os *Wordclouds* com as frequências mínimas referenciadas nas análises anteriores, figura 76 e 77, para obter uma visualização mais clara, simples e perspicaz.

De seguida apresentam-se os *Wordclouds* (em que o tamanho fonte das palavras é proporcional à frequência dos termos) das frequências mínimas tendo em conta os dois cenários:

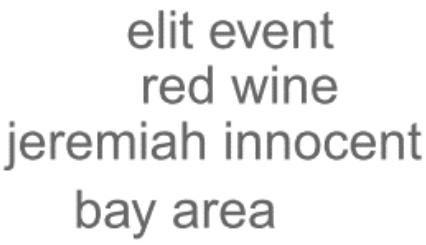


A)

B)



C)



D)

FIGURA 78: *WORDCLOUD*DA CATEGORIA *LOCAL FLAVOR*: A) UNIGRAMAS FREQUÊNCIA MÍNIMA 5; B) UNIGRAMAS FREQUÊNCIA MÍNIMA 10; C) BIGRAMAS FREQUÊNCIA MÍNIMA 2; D) BIGRAMAS FREQUÊNCIA MÍNIMA 3.

Local Services

Como se pode observar na tabela 2, esta categoria de negócio está constituída por 130 *reviews*.

Esta categoria contém opiniões dos utilizadores sobre informática e reparação de computadores, lavandarias e limpeza a seco, arranjos de costura, eletrodomésticos e reparações, amas e cuidados infantis, sapateiros, fotocópias e impressões, correios e serviços de entregas, notários, desinfestação, remoção de lixo entre outros.

Foi feita uma exploração de dados para se compreender e para ter uma perceção mais clara sobre os dados.

De seguida, apresentam-se as explorações nesta categoria de negócio:

Palavras mais e menos frequentes

Após os tratamentos dos dados, observou-se os termos mais e menos frequentes, como se pode observar de seguida:

Termos menos frequêntes	Quantidade	Termos mais frequêntes	Quantidade
abandon	1	ive	54
abil	1	price	54
above	1	servic	57
abroad	1	time	61
absurd	1	get	67
academ	1	place	68

TABELA 53: UNIGRAMAS MAIS E MENOS FREQUENTES NA CATEGORIA *LOCAL SERVICES*.

Termos menos frequêntes	Quantidade	Termos mais frequêntes	Quantidade
a b	1	go back	7
a day	1	this place	7
a great	1	come back	9
a kickass	1	great job	9
a night	1	custom servic	14
a nope	1	dri clean	18

TABELA 54: BIGRAMAS MAIS E MENOS FREQUENTES NA CATEGORIA *LOCAL FLAVOR*.

Após observar a tabela 53, pode-se depreender que os unigramas mais frequentes nesta categoria de negócio são, ***place*** *(*68), ***get***(67), ***time*** (61), ***servic*** (57), ***price*** (54) e ***ive*** (54).

Com a observação da tabela 54, pode-se depreender que os bigramas mais frequentes nesta categoria de negócio são os seguintes, ***dri clean***(18),***custom servic***(14), ***geat job*** (9), ***come back*** (9), **this place** (7) e ***go back*** (7), mantendo a concordância dos temas abordados nesta categoria de negócio.

Termos com frequência mínima de ocorrência

Devido às explorações anteriores, como se pode observar na tabela 53 e 54, e tendo em conta o número de *reviews*, 130, optou-se por observar as palavras que ocorrem no mínimo 20 e 30 vezes, no cenário de unigramas, e 3 e 5 no cenário bigramas.

De seguida, apresenta-se a figura 79 e 80, onde apresenta-se o código utilizado bem como os termos associados tendo em conta os dois cenários.

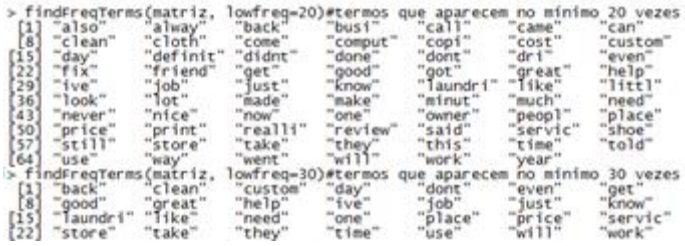


FIGURA 79: UNIGRAMAS QUE APARECEM PELO MENOS 20 E 30 VEZES NA CATEGORIA *LOCAL FLAVOR*.

FIGURA 80: BIGRAMAS QUE APARECEM PELO MENOS 3 E 5 VEZES NA CATEGORIA *LOCAL FLAVOR*.

Ao observar a figura da análise de unigramas, figura 79, os unigramas que se podem destacar são, ***place, servic, time e work,*** entre outras que distinguem bem esta categoria de negócio.

Ao fazer uma análise sobre a figura de bigramas, figura 80, pode-se salientar os bigramas ***custom servic, dri clean, servic*****,** ***time ive***, entre outras que distinguem esta categoria.

Pode-se deduzir que em ambos os cenários esta categoria de negócio está bem representada.

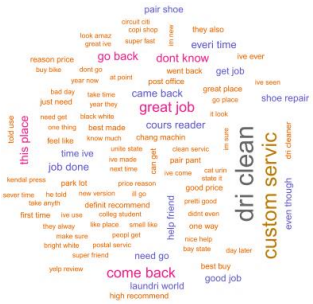
Wordcloud

Optou-se por fazer os *Wordclouds* com as frequências mínimas referenciadas nas análises anteriores, figuras79 e 80, para obter uma visualização mais clara, simples e perspicaz.

De seguida apresentam-se os *Wordclouds* (em que o tamanho fonte das palavras é proporcional à frequência dos termos) das frequências mínimas tendo em conta os dois cenários:

A)

B)



C)

D)

FIGURA81: *WORDCLOUD*DA CATEGORIA *LOCAL FLAVOR*: A) UNIGRAMAS FREQUÊNCIA MÍNIMA 20; B) UNIGRAMAS FREQUÊNCIA MÍNIMA 30; C) BIGRAMAS FREQUÊNCIA MÍNIMA 3;D) BIGRAMAS FREQUÊNCIA MÍNIMA 5.

Mass Media

Como se pode observar na tabela 2, esta categoria de negócio está constituída por 9 *reviews*.

Nesta categoria encontram-se opiniões sobre a imprensa escrita, canais de televisão e estações de radio.

Foi feita uma exploração de dados para se compreender e para ter uma perceção mais clara sobre os dados.

De seguida, apresentam-se as explorações nesta categoria de negócio:

Palavras mais e menos frequentes

Após os tratamentos dos dados, referenciados anteriormente, observou-se os termos mais e menos frequentes tendo em conta os dois cenários, como se pode observar de seguida:

Termos menos frequêntes	Quantidade	Termos mais frequêntes	Quantidade
across	1	kind	6
add	1	morn	6
addit	1	new	7
advertis	1	play	7
ago	1	love	8
aielli	1	music	9

TABELA 55: UNIGRAMAS MAIS E MENOS FREQUENTES NA CATEGORIA *MASS MEDIA*.

Termos menos frequêntes	Quantidade	Termos mais frequêntes	Quantidade
a feeling	1	half hour	2
a friend	1	hour block	2
a true	1	kut kut	2
across board	1	play list	2
add fact	1	sunday morn	2
addit great	1	ulrich schnauss	2

TABELA 56: BIGRAMAS MAIS E MENOS FREQUENTES NA CATEGORIA *MASS MEDIA*.

Ao observar a figura do cenário de unigramas, tabela 55, pode-se constatar que as palavras mais frequentes nesta categoria de negócio são, ***music*** *(*9), ***love*** (8), ***play***(7), ***new*** (7), ***morn***(6) e ***kind***(6).

Pode-se também, depreender que no caso dos bigramas, tabela 56, os mais frequentes nesta categoria de negócio são os seguintes, ***ulrich schnauss*** (2), ***sunday morn***(2),***play list***(2), ***kut kut*** (2), ***hour back*** (2) e ***half hour*** (2), mantendo a concordância dos temas abordados nesta categoria de negócio.

Termos com frequência mínima de ocorrência

Tendo em conta as explorações anteriores, tabela 55 e 56, e o número de *reviews*, 9, optou-se por analisar as palavras que ocorrem no mínimo 2, 3 e 5 vezes, no cenário de unigramas, e, no de bigramas uma frequência mínima de 2.

De seguida, apresenta-se a figura 82 e 83, onde se encontra o código utilizado para obter os termos com a frequência mínima referenciada anteriormente, tendo em conta os cenários, bem como os termos associados.

```
> findFreqTerms(matriz, lowfreq=3)#termos que aparecem no mínimo 3 vezes
 [1] "anyth"    "articl"   "can"      "cant"     "dan"      "favorit"
 [7] "get"      "great"    "hour"     "ill"      "interest" "just"
[13] "kind"     "know"     "kut"      "like"     "listen"   "local"
[19] "love"     "member"   "morn"     "much"     "music"    "new"
[25] "noon"     "often"    "old"      "onair"    "one"      "place"
[31] "play"     "program"  "radio"    "say"      "show"     "song"
[37] "station"  "thank"    "thing"    "wxpn"     "xpn"
> findFreqTerms(matriz, lowfreq=5)#termos que aparecem no mínimo 5 vezes
[1] "great" "kind"  "love"  "morn"  "music" "new"   "one"   "play"  "xpn"
```

FIGURA82: UNIGRAMAS QUE APARECEM PELO MENOS 3 E 5 VEZES NA CATEGORIA *MASS MEDIA*.

```
> findFreqTerms(rvwbi, lowfreq=2)#termos que aparecem no mínimo 2 vezes
 [1] "almost like"     "at noon"         "block generic"   "free at"
 [5] "go back"         "great music"     "half hour"       "hour block"
 [9] "kut kut"         "play list"       "sunday morn"     "ulrich schnauss"
```

FIGURA 83: BIGRAMAS QUE APARECEM PELO MENOS 2 VEZES NA CATEGORIA *MASS MEDIA*.

Nesta exploração, cenário de unigramas, pode-se observar algumas palavras relevantes, como, ***music, play***, entre outras, que se distinguem bem esta categoria.

Analisando a exploração do cenário de bigramas, figura 83, pode-se visualizar alguns bigramas relevantes, como**, *play list,great music****,*** entre outras, que distinguem bem esta categoria.

No entanto, pode-se concluir que esta categoria não se encontra bem representada nos dois cenários, poderá ser por existir um número muito pequeno de *reviews* nesta categoria.

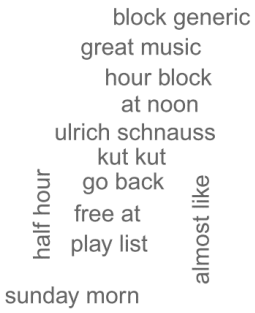
Wordcloud

Optou-se por fazer os *Wordclouds* com as frequências mínimas referenciadas nas análises anteriores, figura 82 e 83, para obter uma visualização mais clara, simples e perspicaz.

De seguida apresentam-se os *Wordclouds* (em que o tamanho fonte das palavras é proporcional à frequência dos termos) das frequências mínimas tendo em conta os dois cenários:

A)

B)



C)

FIGURA 84: *WORDCLOUD*DA CATEGORIA *MASS MEDIA*: A)UNIGRAMAS FREQUÊNCIA MÍNIMA 3; B) UNIGRAMAS FREQUÊNCIA MÍNIMA 5; C) BIGRAMAS FREQUÊNCIA MÍNIMA 2.

Night Life

Como se pode observar na tabela 2, esta categoria de negócio está constituída por 1215 *reviews*. Nesta categoria, encontram-se opiniões sobre clubes de comedia, casas de espetáculos, bares, snooker e bilhar, *jazz blues*, discotecas, entretenimento para adultos, *karaoke*, piano bar e *coffeeshops*.

Foi feita uma exploração de dados para se compreender e para ter uma perceção mais clara sobre os dados.

De seguida, apresentam-se as explorações nesta categoria de negócio nos dois cenários:

Palavras mais e menos frequentes

Após os tratamentos dos dados, observou-se os termos mais e menos frequentes, como se pode observar de seguida:

Termos menos frequêntes	Quantidade	Termos mais frequêntes	Quantidade
aaaaanywho	1	drink	629
aaron	1	like	702
aback	1	bar	720
abandon	1	food	807
abbb	1	good	928
abound	1	place	1063

TABELA 57: UNIGRAMAS MAIS E MENOSFREQUENTES NA CATEGORIA *NIGHT LIFE*.

Termos menos frequêntes	Quantidade	Termos mais frequêntes	Quantidade
a babi	1	great okace	62
a bar	1	pretti good	72
a bdub	1	go back	73
a beer	1	beer select	77
a breath	1	this place	85
a celeb	1	happi hour	146

TABELA 58: BIGRAMAS MAIS E MENOS FREQUENTES NA CATEGORIA *NIGHT LIFE*.

Ao observar a tabela 57, sobre a exploração de unigramas, pode-se constatar que as palavras mais frequentes nesta categoria de negócio são, ***place*** *(*1063), ***good***(928), ***food***(807), ***bar***(720), ***like*** (702) e ***drink***(629).

Com a observação da tabela58, pode-se depreender que os bigramas mais frequentes nesta categoria de negócio são os seguintes, ***happi hour*** (146), ***this place***(85),***beer select***(77), ***go back*** (73), ***pretto good*** (72) e ***great place*** (62), mantendo a concordância dos temas abordados nesta categoria de negócio.

Termos com frequência mínima de ocorrência

Devido à exploração feita anteriormente, tabela 57 e 58, e tendo em conta o número de *reviews*, 1215, optou-se por observaros termos com uma frequência mínima de 200 e 500, no cenário de unigramas, e de 20 e 50 de bigramas.

De seguida, apresenta-se a figura 85 e 86, onde se encontra o código utilizado para obter termos mínimos referenciados anteriormente, tendo em conta os cenários, bem como os termos associados:

```
> findFreqTerms(matriz, lowfreq=200)#termos que aparecem no minimo 200 vezes
 [1] "also"   "back"   "bar"    "beer"   "best"   "burger" "can"    "come"
 [9] "crowd"  "dont"   "drink"  "even"   "food"   "fri"    "friend" "get"
[17] "good"   "got"    "great"  "hour"   "its"    "ive"    "just"   "know"
[25] "like"   "littl"  "look"   "love"   "make"   "menu"   "much"   "nice"
[33] "night"  "one"    "order"  "peopl"  "place"  "pretti" "price"  "realli"
[41] "select" "servic" "tabl"   "they"   "thing"  "think"  "this"   "time"
[49] "tri"    "wait"   "want"   "well"   "will"
> findFreqTerms(matriz, lowfreq=500)#termos que aparecem no minimo 500 vezes
 [1] "bar"   "beer"  "drink" "food"  "get"   "good"  "great" "just"  "like"
[10] "one"   "place" "time"
```

FIGURA 85: UNIGRAMAS QUE APARECEM PELO MENOS 200 E 500 VEZES NA CATEGORIA *NIGHT LIFE*.

```
> findFreqTerms(rvwbi, lowfreq=20)#termos que aparecem no mínimo 20 vezes
 [1] "bar food"      "beer select"   "beer tap"      "can get"
 [5] "come back"     "danc floor"    "dive bar"      "dont know"
 [9] "dont like"     "dont think"    "drink good"    "even though"
[13] "everi time"    "feel like"     "first time"    "food drink"
[17] "food good"     "friday night"  "get drink"     "give place"
[21] "go back"       "good beer"     "good drink"    "good food"
[25] "good place"    "good select"   "good thing"    "good time"
[29] "great beer"    "great place"   "happi hour"    "high recommend"
[33] "if your"       "im sure"       "ive ever"      "ive never"
[37] "last night"    "late night"    "like place"    "littl bit"
[41] "look like"     "love place"    "mac chees"     "much better"
[45] "my friend"     "next time"     "noth special"  "one best"
[49] "one favorit"   "place get"     "place go"      "place great"
[53] "pool tabl"     "pretti good"   "pretti much"   "realli good"
[57] "realli like"   "realli nice"   "reason price"  "saturday night"
[61] "seem like"     "tast like"     "they also"     "this place"
[65] "wait staff"    "we also"       "we order"      "you can"
[69] "your look"
> findFreqTerms(rvwbi, lowfreq=50)#termos que aparecem no mínimo 50 vezes
[1] "beer select" "come back"   "go back"     "great place" "happi hour"
[6] "pretti good" "this place"
```

FIGURA 86: BIGRAMAS QUE APARECEM PELO MENOS 20 E 50 VEZES NA CATEGORIA *NIGHT LIFE*.

No caso da exploração de unigramas, figura 85, pode-se observar algumas palavras relevantes, como, ***drink, bar, beer,*** entre outras, que distinguem bem esta categoria.

Ao observar a exploração de bigramas, figura 86, pode-se constatar que esta categoria de negócio encontra-se melhor representada no cenário de unigramas, devido a que, os termos são mais ajustados à categoria de negócio.

Wordcloud

Optou-se por fazer os *Wordclouds* com as frequências mínimas referenciadas nas análises anteriores, figura 85e 86, para obter uma visualização mais clara, simples e perspicaz.

De seguida apresentam-se os *Wordclouds* (em que o tamanho fonte das palavras é proporcional à frequência dos termos) das frequências mínimas tendo em conta os dois cenários:

A)

B)

C) D)

FIGURA 87: *WORDCLOUD*DA CATEGORIA *NIGHT LIFE*: A) UNIGRAMAS FREQUÊNCIA MÍNIMA 200; B) UNIGRAMAS FREQUÊNCIA MÍNIMA 500; CBIGRAMAS FREQUÊNCIA MÍNIMA 20; D) BIGRAMAS FREQUÊNCIA MÍNIMA 50.

Pets

Como se pode observar na tabela 2, esta categoria de negócio está constituída por 31 *reviews*.

Esta categoria, esta composta por opiniões sobre serviços para animais, canis, lojas de animais e adoção de animais.

Foi feita uma exploração de dados para se compreender e para ter uma perceção mais clara sobre os dados.De seguida, apresentam-se as explorações nesta categoria de negócio tendo em conta os dois cenários:

Palavras mais e menos frequentes

Após os tratamentos dos dados, observou-se os termos mais e menos frequentes, como se pode observar de seguida:

Termos menos frequêntes	Quantidade	Termos mais frequêntes	Quantidade
accommod	1	care	19
accut	1	love	19
activ	1	just	21
adequ	1	get	23
admit	1	pet	29
adopte	1	dog	46

TABELA 59: UNIGRAMAS MAIS E MENOS FREQUENTES NA CATEGORIA *PETS*.

Termos menos frequêntes	Quantidade	Termos mais frequêntes	Quantidade
a bit	1	she said	3
a great	1	teeth clean	3
a major	1	two vet	3
abl easili	1	year now	3
abl get	1	human societi	4
absolut explan	1	dr mcdowel	5

TABELA 60: BIGRAMAS MAIS E MENSO FREQUENTES NA CATEGORIA *PETS*.

Ao observar o cenário de unigramas, tabela 59, pode-se constatar que as palavras mais frequentes nesta categoria de negócio são, ***dog*** (46), ***pet*** (29), ***get***(23), ***just*** (21), ***love*** (19) e ***care*** (19).

Após o estudo do cenário de bigramas, tabela 60, pode-se compreender que os bigramas mais frequentes nesta categoria de negócio são os seguintes, ***dr mcdowel*** (5), ***human societi***(4),***year now***(3), ***two vet*** (3), ***teeth clean*** (3) e ***she said*** (3), mantendo a concordância dos temas abordados nesta categoria de negócio.

Pode-se constatar, que neste caso, que a exploração feita com unigramas obteve melhores resultados, uma vez que, as palavras descrevem melhor a categoria de negócio e consequentemente os assuntos abordados.

Termos com frequência mínima de ocorrência

Tendo em conta as análises realizadas anteriormente, tabela 59 e 60, bem como o número de *reviews,* 31, optou-se por visualizar os termos com frequência mínima de, no caso de unigramas, de 10 e 20 vezes e no caso de bigramas, de 2 e 3 vezes.

De seguida, apresenta-se as figuras 88 e 89, onde se pode encontrar o código utilizado para obter os termos com a frequência mínima referida anteriormente, tendo em conta os cenários, bem como os termos associados:

```
> findFreqTerms(matriz, lowfreq=10)#termos que aparecem no minimo 10 vezes
 [1] "adopt"  "alway"  "anim"   "can"    "care"   "cat"    "dog"    "get"
 [9] "just"   "know"   "love"   "make"   "need"   "one"    "peopl"  "pet"
[17] "place"  "pup"    "puppi"  "realli" "staff"  "store"  "take"   "they"
[25] "time"   "took"   "vet"    "will"
> findFreqTerms(matriz, lowfreq=20)#termos que aparecem no minimo 20 vezes
[1] "dog"  "get"  "just" "pet"
```

FIGURA 88: UNIGRAMAS QUE APARECEM PELO MENOS 10, 20 E 30 VEZES NA CATEGORIA *PETS*.

```
> findFreqTerms(rvwbi, lowfreq=2)#termos que aparecem no minimo 2 vezes
 [1] "ador dog"       "animal hospit"  "atlanta human"  "babi sparrow"
 [5] "banfield locat" "can get"        "care anim"      "care receiv"
 [9] "cat dog"        "cat will"       "custom servic"  "didnt know"
[13] "dog atlanta"    "dog food"       "dog kibbl"      "dog rescu"
[17] "dog train"      "dog you"        "doggi hotel"    "dr mcdowel"
[21] "dr o"           "even though"    "feel comfort"   "foster parent"
[25] "friend staff"   "full grown"     "go educ"        "good hand"
[29] "heart worm"     "hour emerg"     "human societi"  "know dog"
[33] "love anim"      "love pet"       "made us"        "make sure"
[37] "my dog"         "nail she"       "offic visit"    "one night"
[41] "peopl petco"    "pet hospit"     "pet hotel"      "pet owner"
[45] "pet store"      "pet will"       "pug puppi"      "pup peopl"
[49] "recommend place" "shaggi dog"    "she said"       "teeth clean"
[53] "thank shaggi"   "time explain"   "took care"      "trust manu"
[57] "two vet"        "well care"      "will go"        "will well"
[61] "year now"       "you can"        "youll get"
> findFreqTerms(rvwbi, lowfreq=3)#termos que aparecem no minimo 3 vezes
 [1] "atlanta human" "care anim"     "cat dog"       "doggi hotel"
 [5] "dr mcdowel"    "human societi" "love pet"      "make sure"
 [9] "pet owner"     "pet store"     "shaggi dog"    "she said"
[13] "teeth clean"   "two vet"       "year now"
```

FIGURA 89: BIGRAMAS QUE APARECEM PELO MENOS 2 E 3 VEZES NA CATEGORIA *PETS*.

Observando a figura 88, cenário de unigramas, pode-se dizer que as palavras mais relevantes são, ***pet, dog, adopt, anim, cat***, entre outras, que distinguem bem esta categoria.

Analisando agora o cenário de bigramas, figura 89, nesta exploração pode-se observar alguns bigramas relevantes, como, ***care anim, cat dog, doggi hotel, pet owner***, entre outras, que distinguem bem esta categoria.

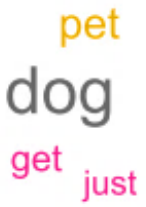
Constata-se que este cenário representa melhor a categoria de negócio.

Wordcloud

Optou-se por fazer os *Wordclouds* com as frequências mínimas referenciadas nas análises anteriores, figura 88 e 89, para obter uma visualização mais clara, simples e perspicaz.

De seguida apresentam-se os *Wordclouds* (em que o tamanho fonte das palavras é proporcional à frequência dos termos) das frequências mínimas tendo em conta os dois cenários:

A)



B)

C)

D)

FIGURA 90: *WORDCLOUD*DA CATEGORIA *PETS*: A) UNIGRAMAS FREQUÊNCIA MÍNIMA 10; B) UNIGRAMAS FREQUÊNCIA MÍNIMA 20; C) BIGRAMAS FREQUÊNCIA MÍNIMA 2; D) BIGRAMAS FREQUÊNCIA MÍNIMA 3.

Professional Services

Como se pode observar na tabela 2, esta categoria de negócio está constituída por 23 *reviews*.

Nesta categoria encontram-se opiniões dos utilizadores da plataforma *yelp* sobre advogados, contabilistas, agências de emprego, *web desing*, *desing* gráfico, arquitetos, operadores de *internet*, marketing, publicidade, relações públicas, produção de vídeo, limpeza de escritórios, investigadores e detetives privados, *life coach*, orientação profissional, serviços de segurança, reparação de barcos, produção musical entre outros.

Foi feita uma exploração de dados para se compreender e para ter uma perceção mais clara sobre os dados.

De seguida, apresentam-se as explorações nesta categoria de negócio, tendo em conta os dois cenários:

Palavras mais e menos frequentes

Após os tratamentos dos dados, observou-se os termos mais e menos frequentes, como se pode observar de seguida:

Termos menos frequêntes	Quantidade	Termos mais frequêntes	Quantidade
abl	1	call	20
account	1	they	20
accus	1	time	21
across	1	help	22
actual	1	phone	22
admit	1	work	29

TABELA 61: UNIGRAMAS MAIS E MENOS FREQUENTES NA CATEGORIA *PROFESSIONAL SERVICES*.

Termos menos frequêntes	Quantidade	Termos mais frequêntes	Quantidade
a oh	1	take care	3
a scam	1	tech support	3
abl manag	1	time warner	3
access mani	1	took time	3
access web	1	they said	4
account go	1	custom servic	6

TABELA 62: BIGRAMAS MAIS E MENOS FREQUENTES NA CATEGORIA *PROFESSIONAL SERVICES*.

Ao observar a figura do cenário de unigramas, tabela 61, pode-se constatar que as palavras mais frequentes nesta categoria de negócio são, ***work*** *(*29), ***phone***(22), ***help***(22), ***time***(21), ***they***(20) e ***call*** (20).

Após o estudo da figura do cenário de bigramas, tabela 62, pode-se compreender que os bigramas mais frequentes nesta categoria de negócio são os seguintes, ***custom servic***(6), ***they said***(4),***took time***(3), ***time warner*** (3), ***tech support*** (3) e ***take care*** (3), mantendo a concordância dos temas abordados nesta categoria de negócio.

Termos com frequência mínima de ocorrência

Tendo em conta exploração feita anteriormente, nas tabelas 61 e 62, e o número de *reviews*, 23, optou-se por observar as palavras que ocorrem no mínimo 5 e 10 vezes no cenário de unigramas, e para o outro cenário, optou.se por analisar 2 e 3 vezes de frequência mínima.

De seguida, apresentam-se as figuras 91 e 92 onde apresenta-se o código utilizado para obter os termos com a frequência mínima referenciada anteriormente para os dois cenários.

```
> findFreqTerms(matriz, lowfreq=5)#termos que aparecem no minimo 5 vezes
 [1] "also"      "ask"       "att"       "back"      "basic"     "call"
 [7] "can"       "care"      "chang"     "come"      "connect"   "custom"
[13] "day"       "didnt"     "doesnt"    "dont"      "even"      "ever"
[19] "experi"    "explain"   "find"      "first"     "found"     "get"
[25] "got"       "help"      "home"      "hour"      "internet"  "issu"
[31] "ive"       "just"      "look"      "mani"      "month"     "need"
[37] "never"     "new"       "nice"      "now"       "number"    "one"
[43] "peopl"     "phone"     "problem"   "question"  "realli"    "rebat"
[49] "recommend" "said"      "see"       "servic"    "set"       "someon"
[55] "still"     "store"     "support"   "take"      "talk"      "they"
[61] "time"      "told"      "tri"       "use"       "walk"      "way"
[67] "will"      "wireless"  "work"      "year"      "you"
> findFreqTerms(matriz, lowfreq=10)#termos que aparecem no minimo 10 vezes
 [1] "ask"     "att"     "back"    "call"    "can"     "custom"  "even"    "get"
 [9] "got"     "help"    "issu"    "month"   "need"    "phone"   "rebat"   "said"
[17] "servic"  "store"   "take"    "they"    "time"    "way"     "work"
```

FIGURA 91: UNIGRAMAS QUE APARECEM PELO MENOS 5 E 10 VEZES D

NA CATEGORIA *PROFESSIONAL SERVICES*.

```
> findFreqTerms(rvwbi, lowfreq=2)#termos que aparecem no minimo 2 vezes
 [1] "another month"   "appear work"     "ask can"         "ask question"
 [5] "att dsl"         "bever connect"   "call back"       "call number"
 [9] "call tech"       "chang password"  "come mail"       "conn store"
[13] "custom servic"   "custom support"  "didnt work"      "doesnt work"
[17] "dsl internet"    "each time"       "even though"     "explain problem"
[21] "first time"      "get rebat"       "havent reciev"   "help find"
[25] "help they"       "high recommend"  "im sorri"        "im sure"
[29] "just complet"    "mail short"      "make call"       "meet room"
[33] "mod studio"      "mr evan"         "never call"      "new iphon"
[37] "new phone"       "park harvard"    "phone call"      "phone servic"
[41] "qualifi rebat"   "rebat mail"      "rebat she"       "rebat told"
[45] "reciev rebat"    "right away"      "set phone"       "she said"
[49] "step chang"      "still rebat"     "take care"       "take peopl"
[53] "talk custom"     "tech support"    "they said"       "time warner"
[57] "took time"       "walk step"       "way back"        "westwood store"
[61] "wireless router" "work it"         "work paul"       "work space"
[65] "you cant"
> findFreqTerms(rvwbi, lowfreq=3)#termos que aparecem no minimo 3 vezes
 [1] "call number"   "custom servic"  "didnt work"    "doesnt work"
 [5] "even though"   "help find"      "help they"     "high recommend"
 [9] "take care"     "tech support"   "they said"     "time warner"
[13] "took time"
```

FIGURA 92: BIGRAMAS QUE APARECEM PELO MENOS 2 E 3 VEZES NA CATEGORIA *PROFESSIONAL SERVICES*.

Ao analisar o cenário de unigramas, figura 91, pode-se observar algumas palavras relevantes, como, ***call, help, phone, time, work*, *time***, entre outras, que distinguem bem esta categoria.

Observando os resultados do cenário de bigramas, figura 92, pode-se examinar algumas palavras relevantes, como, ***custom servic, help find, cal number***, entre outras, que distinguem bem esta categoria.

Wordcloud

Optou-se por fazer os *Wordclouds* com as frequências mínimas referenciadas nas análises anteriores, figura 91 e 92, para obter uma visualização mais clara, simples e perspicaz.

De seguida apresentam-se os *Wordclouds* (em que o tamanho fonte das palavras é proporcional à frequência dos termos) das frequências mínimas tendo em conta os dois cenários:

A) B)

C) D)

FIGURA 93: *WORDCLOUD*DA CATEGORIA *PROFESSIONAL SERVICES*: A) UNIGRAMAS FREQUÊNCIA MÍNIMA 5; B) UNIGRAMAS FREQUÊNCIA MÍNIMA 10; C) BIGRAMAS FREQUÊNCIA MÍNIMA 2; D) BIGRAMAS FREQUÊNCIA MÍNIMA 3.

Public Services & Government

Como se pode observar na tabela 2, esta categoria de negócio está constituída por 25 *reviews*.

Nesta categoria encontram-se opiniões sobre bibliotecas, correios, monumentos históricos e pontos de interesse, departamento de transportes, esquadras, embaixadas, serviços de finanças, tribunais, centros comunitários e bombeiros.

Foi feita uma exploração de dados para se compreender e para ter uma perceção mais clara sobre os dados.

De seguida, apresentam-se as explorações nesta categoria de negócio:

Palavras mais e menos frequentes

Após os tratamentos dos dados, observou-se os termos mais e menos frequentes, como se pode observar de seguida:

Termos menos frequêntes	Quantidade	Termos mais frequêntes	Quantidade
abl	1	floor	14
accident	1	librari	15
accommod	1	place	15
address	1	realli	17
advantag	1	get	19
adventur	1	can	25

TABELA 63: UNIGRAMAS MAIS E MENOS FREQUENTES NA CATEGORIA *PUBLIC SERVICES&GOVERNMENT*.

Termos menos frequêntes	Quantidade	Termos mais frequêntes	Quantidade
abl purchas	1	want go	2
academ journal	1	wouldnt give	2
academ year	1	year can	2
access build	1	buffalo ny	3
access databas	1	nd floor	3
access studi	1	post offic	4

TABELA 64: BIGRAMAS MAIS E MENOS FREQUENTES NA CATEGORIA *PUBLIC SERVICES&GOVERNMENT*.

Após analisar a tabela 63, cenário de unigramas, pode-se verificar que as palavras mais frequentes nesta categoria de negócio são, ***can*** *(*25), ***get*** (19), ***realli***(17), ***place*** (15), ***librari***(15) e ***floor*** (14).

Após o estudo da tabela 64, cenário de bigramas, pode-se verificar que os bigramas mais frequentes nesta categoria de negócio são os seguintes, ***post offic***(4), ***nd floor***(3),***buffalo ny***(3), ***year can*** (2), ***wouldnt give*** (2) e ***want go***(2), mantendo a concordância dos temas abordados nesta categoria de negócio.

Termos com frequência mínima de ocorrência

Após as análises feitas anteriormente, e tendo-a em conta bem como o número de *reviews*, 25, decidiu-se observar a frequência mínima dos termos, no cenário de unigramas, 5 e 10 vezes, e no cenário de bigramas 2 e 3 vezes.

De seguida, apresentam-se as figuras 94 e 95, que contém o código utilizado para obter os termos com as frequências mínimas referenciadas anteriormente bem como os termos associados.

```
> findFreqTerms(matriz, lowfreq=5)#termos que aparecem no mínimo 5 vezes
 [1] "across"  "also"    "alway"   "ask"     "big"     "book"    "can"
 [8] "complet" "day"     "doheni"  "dont"    "even"    "find"    "floor"
[15] "free"    "get"     "give"    "good"    "help"    "ive"     "just"
[22] "know"    "librari" "like"    "look"    "lot"     "mail"    "make"
[29] "minut"   "much"    "need"    "now"     "one"     "park"    "place"
[36] "post"    "realli"  "stamp"   "student" "studi"   "system"  "take"
[43] "tape"    "that"    "there"   "they"    "thing"   "this"    "time"
[50] "use"     "want"    "week"    "well"    "will"    "work"    "year"
[57] "yes"     "your"
> findFreqTerms(matriz, lowfreq=10)#termos que aparecem no mínimo 10 vezes
 [1] "book"    "can"     "dont"    "even"    "floor"   "get"     "just"
 [8] "librari" "place"   "realli"  "there"   "time"    "want"
```

FIGURA 94: UNIGRAMAS QUE APARECEM PELO MENOS 5 E 10 VEZES NA CATEGORIA *PUBLIC SERVICES&GOVERNMENT*.

```
> findFreqTerms(rvwbi, lowfreq=2)#termos que aparecem no mínimo 2 vezes
 [1] "alway friend"      "buffalo ny"          "can get"
 [4] "can much"          "can use"             "didnt know"
 [7] "dont even"         "dont know"           "dude buffalo"
[10] "film foreign"      "find place"          "first time"
[13] "foreign languag"   "interlibrari loan"   "intern stamp"
[16] "just year"         "librari card"        "midterm final"
[19] "nd floor"          "owe big"             "pack tape"
[22] "place they"        "post offic"          "realli need"
[25] "sinc dont"         "stamp want"          "stumbl across"
[28] "there never"       "they also"           "they great"
[31] "this place"        "touch histori"       "want go"
[34] "wouldnt give"      "year can"
> findFreqTerms(rvwbi, lowfreq=3)#termos que aparecem no mínimo 3 vezes
[1] "buffalo ny" "nd floor"   "post offic"
```

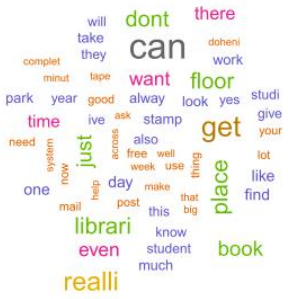
FIGURA 95: BIGRAMAS QUE APARECEM PELO MENOS 2 E 3 VEZES NA CATEGORIA *PUBLIC SERVICES&GOVERNMENT*.

Como se pode observar no cenário de unigramas, figura 94, os termos que se destacam, são, ***book, floor,realli, librari, place***, entre outras, que distinguem bem esta categoria.

Na exploração de bigramas, pode-se observar alguns bigramas relevantes, como, ***post offic, year can, realli need***, entre outras, que distinguem bem esta categoria.

Wordcloud

Optou-se por fazer os *Wordclouds* com as frequências mínimas referenciadas nas análises anteriores, figura 94 e 95, para obter uma visualização mais clara, simples e perspicaz.
De seguida apresentam-se os *Wordclouds* (em que o tamanho fonte das palavras é proporcional à frequência dos termos) das frequências mínimas tendo em conta os dois cenários:



A)

B)

C)

D)

FIGURA 96: *WORDCLOUD*DA CATEGORIA *PUBLIC SERVICES&GOVERNMENT*: A) UNIGRAMAS FREQUÊNCIA MÍNIMA 5; B) UNIGRAMAS FREQUÊNCIA MÍNIMA 10; C) BIGRAMAS FREQUÊNCIA MÍNIMA 2; D) BIGRAMAS FREQUÊNCIA MÍNIMA 3.

Real Estate

Como se pode observar na tabela 2, esta categoria de negócio está constituída por14 *reviews.*

Nesta categoria encontram-se opiniões dos utilizadores da plataforma sobre agentes imobiliários, serviços imobiliários, apartamentos, gestão de propriedades, corretores hipotecários, alojamento

universitário, imobiliário comercial, escritórios partilhados, casas pré-fabricadas e liquidação do património.

Foi feita uma exploração de dados para se compreender e para ter uma perceção mais clara sobre os dados.

De seguida, apresentam-se as explorações nesta categoria de negócio tendo em conta os cenários:

Palavras mais e menos frequentes

Após os tratamentos dos dados, observou-se os termos mais e menos frequentes, como se pode observar de seguida:

Termos menos frequêntes	Quantidade	Termos mais frequêntes	Quantidade
accept	1	time	13
accord	1	leas	16
account	1	move	18
acknowledg	1	hous	21
across	1	manag	22
act	1	apart	29

TABELA 65: UNIGRAMAS MAIS E MENOS FREQUENTES NA CATEGORIA *PUBLIC REAL ESTATE*.

Termos menos frequêntes	Quantidade	Termos mais frequêntes	Quantidade
a coupl	1	help us	3
abl get	1	long time	3
abl put	1	month leas	3
absolut doubt	1	secur deposit	3
absolut privaci	1	silton properti	3
absolut rude	1	sign leas	5

TABELA 66: BIGRAMAS MAIS E MENOS FREQUENTES NA CATEGORIA *PUBLIC REAL ESTATE*.

Depois de observar a tabela 65, cenário de unigramas, pode-se constatar que as palavras mais frequentes nesta categoria de negócio são, ***apart*** *(*29), ***manag*** (22), ***hous***(21), ***move*** (18), ***leas***(16) e ***time***(13).

Após analisar a tabela 66, bigramas, constata-se que os bigramas mais frequentes nesta categoria de negócio são os seguintes, ***sign leas***(5), ***silton properti***(3),***secur deposit***(3), ***month leas*** (3), ***long time*** (3) e ***help us*** (3), mantendo a concordância dos temas abordados nesta categoria de negócio.

Termos com frequência mínima de ocorrência

Devido às explorações feitas anteriormente, nas tabelas65 e 66, e também pelo número de *reviews*, 14, optou-se por observar a frequência mínima de termos, no caso de unigramas, 5 e 10 vezes, e no caso de bigramas, 3 vezes.

De seguida, apresentam-se as figuras 97 e 98, onde se encontra o código utilizado para obter os termos com as frequências anteriormente referidas:

```
> findFreqTerms(matriz, lowfreq=5)#termos que aparecem no mínimo 5 vezes
 [1] "also"     "and"      "apart"    "austin"   "back"     "call"
 [7] "can"      "charg"    "clean"    "come"     "complet"  "day"
[13] "deposit"  "dont"     "even"     "experi"   "find"     "first"
[19] "found"    "front"    "get"      "give"     "great"    "help"
[25] "home"     "hous"     "just"     "know"     "leas"     "leav"
[31] "letter"   "like"     "live"     "long"     "look"     "made"
[37] "mail"     "make"     "manag"    "mani"     "messag"   "month"
[43] "move"     "never"    "new"      "offic"    "one"      "peopl"
[49] "permiss"  "place"    "pool"     "process"  "properti" "put"
[55] "realli"   "receiv"    "rent"     "review"   "say"      "see"
[61] "show"     "sign"     "silton"   "someon"   "star"     "tenant"
[67] "they"     "think"    "this"     "time"     "took"     "tri"
[73] "well"     "will"     "work"     "you"
> findFreqTerms(matriz, lowfreq=10)#termos que aparecem no mínimo 10 vezes
 [1] "apart"  "clean"  "day"    "dont"   "even"   "experi" "get"    "hous"
 [9] "leas"   "manag"  "month"  "move"   "time"   "will"
```

FIGURA 97: UNIGRAMAS QUE APARECEM PELO MENOS 5 E 10 VEZES NA CATEGORIA *PUBLIC REAL ESTATE*.

```
> findFreqTerms(rvwbi, lowfreq=3)#termos que aparecem no mínimo 3 vezes
[1] "dont know"      "everi day"      "help us"        "long time"
[5] "month leas"     "secur deposit"  "sign leas"      "silton properti"
```

FIGURA 98: BIGRAMAS QUE APARECEM PELO MENOS 3 VEZES NA CATEGORIA *PUBLIC REAL ESTATE*.

Ao analisar a figura 97, unigramas, pode-se observar algumas palavras relevantes, como, ***apart, hous, manag***, entre outras, que distinguem bem esta categoria.

Na exploração de bigramas, figura 98, pode-se observar alguns bigramas relevantes, como, ***secur deposit, help us, dont know***, entre outras, que distinguem bem esta categoria.

Wordcloud

Optou-se por fazer os *Wordclouds* com as frequências mínimas referenciadas nas análises anteriores, figura 97 e 98, para obter uma visualização mais clara, simples e perspicaz.
De seguida apresentam-se os *Wordclouds* (em que o tamanho fonte das palavras é proporcional à frequência dos termos) das frequências mínimas tendo em conta os dois cenários:

A)

B)

C)

FIGURA 99: *WORDCLOUD*DA CATEGORIA *PUBLIC REAL ESTATE*: A) UNIGRAMAS FREQUÊNCIA MÍNIMA 5; B) UNIGRAMAS FREQUÊNCIA MÍNIMA 10; C) BIGRAMAS FREQUÊNCIA MÍNIMA 3.

Religious Organization

Como se pode observar na tabela 2, esta categoria denegócio é constituída por4 *reviews*.

Nela se encontram comentários sobre igrejas, mesquitas, sinagogas, templos budistas e templos hinduístas.

Foi feita uma exploração de dados para se compreender e para ter destes uma perceção mais clara.

De seguida, apresentam-se as explorações nesta categoria de negócio tendo em conta os dois cenários:

Palavras mais e menos frequentes

Após os tratamentos dos dados, observou-se os termos mais e menos frequentes, como se pode observar de seguida:

Termos menos frequêntes	Quantidade	Termos mais frequêntes	Quantidade
activ	1	this	2
affirm	1	though	2
alleluia	1	week	2
also	1	love	3
alway	1	music	3
atmospher	1	peopl	3

TABELA 67: UNIGRAMAS MAIS E MENOS FREQUENTES NA CATEGORIA *PUBLIC RELIGIOUS ORGANIZATION*.

Termos menos frequêntes	Quantidade	Termos mais frequêntes	Quantidade
activ involv	1	week it	1
affirm peopl	1	week week	1
alleluia sing	1	welcom group	1
also lot	1	workship quit	1
alway pleasant	1	zomg hella	1
attend differ	1	love love	2

TABELA 68: BIGRAMAS MAIS E MENOS FREQUENTES NA CATEGORIA *PUBLIC RELIGIOUS ORGANIZATION*.

Analisando o cenário de unigramas, na tabela67, pode-se constatar que as palavras mais frequentes nesta categoria de negócio são, ***peopl*** *(*3), ***music***(3), ***love***(3), ***week*** (2), ***though*** (2) e ***this*** (2).

Na tabela 68, verifica-se que os bigramas mais frequentes nesta categoria de negócio são os seguintes, ***love love***(2), ***zomg hella***(1),***worship quit***(1), ***welcomgroup*** (1), ***week week*** (1) e ***week it*** (1), mantendo a concordância dos temas abordados nesta categoria de negócio.

Termos com frequência mínima de ocorrência

Devido à exploração anterior, que se encontra nas tabelas67 e 68, e tendo em conta o número de *reviews*, 4, optou-se por observar as palavras que ocorrem no mínimo 2 e 3 vezes para os unigramas, e, 2 vezes para os bigramas.

De seguida, apresentam-se as figuras 100 e 101, que contém o código utilizado para obter os termos com a frequência mínima, no caso de unigramas, 2 e 3 vezes, no caso de bigramas, 2 vezes, bem como os termos associados.

```
> findFreqTerms(matriz, lowfreq=2)#termos que aparecem no mínimo 2 vezes
[1] "communiti" "lot"       "love"      "meet"      "music"     "peopl"
[7] "this"      "though"    "week"
> findFreqTerms(matriz, lowfreq=3)#termos que aparecem no mínimo 3 vezes
[1] "love"  "music" "peopl"
```

FIGURA100: UNIGRAMAS QUE APARECEM PELO MENOS 2 E 3 VEZES NA CATEGORIA *PUBLIC RELIGIOUS ORGANIZATION*.

```
> findFreqTerms(rvwbi, lowfreq=2)#termos que aparecem no mínimo 2 vezes
[1] "love love"
```

FIGURA101: BIGRAMAS QUE APARECEM PELO MENOS 2 VEZES NA CATEGORIA *PUBLIC RELIGIOUS ORGANIZATION*.

No cenário de unigramas figura 100, podemos observar algumas palavras relevantes, como, ***love***, ***peopl***, ***music***, entre outras, que distinguem bem esta categoria.

Como se pode observar na figura 101, com a frequência mínima de 2 só se encontra o bigrama ***love love***, que distinguem bem esta categoria.

Wordcloud

Optou-se por fazer os *Wordclouds* com as frequências mínimas referenciadas nas análises anteriores, figuras 100 e 101, para obter uma visualização mais clara, simples e perspicaz.

De seguida apresentam-se os *Wordclouds* (em que o tamanho fonte das palavras é proporcional à frequência dos termos) das frequências mínimas tendo em conta os dois cenários:

FIGURA 102: *WORDCLOUD*DA CATEGORIA *PUBLIC RELIGIOUS ORGANIZATION*: A) UNIGRAMAS *Sh* FREQUÊNCIA MÍNIMA 2; B) UNIGRAMAS FREQUÊNCIA MÍNIMA 3; C) BIGRAMAS FREQUÊNCIA MÍNIMA 2.

Como se pode observar na tabela 2, esta categoria de negócio está constituída por 753 *reviews*.

Nesta categoria de negócio encontram-se opiniões sobre variadíssimas áreas desde tabacarias, cosmética e produtos de beleza, moda, material de escritório, antiquários, lojas de brinquedos, centros comerciais. Eletrónica, material desportivo, galerias de arte, óticas e oculistas, flores, computadores, lojas de fotografia, relógios entre outros.

Foi feita uma exploração de dados para se compreender e para ter uma perceção mais clara sobre os dados.

De seguida, apresentam-se as explorações nesta categoria de negócio:

Palavras mais e menos frequentes

Após os tratamentos dos dados, observou-se os termos mais e menos frequentes, tendo em conta os dois cenários, como se pode observar de seguida:

Termos menos frequêntes	Quantidade	Termos mais frequêntes	Quantidade
aardvarkor	1	just	294
aaskin	1	get	301
abbas	1	one	314
abcs	1	place	360
abod	1	like	411
abroad	1	store	562

TABELA 69: UNIGRAMAS MAIS E MENOS FREQUENTES*NA CATEGORIA SHOPPING*.

Termos menos frequêntes	Quantidade	Termos mais frequêntes	Quantidade
a berkeley	1	they also	22
a big	1	go back	23
a bit	1	can find	28
a blockbust	1	great select	31
a broken	1	this place	32
a bunch	1	custom servic	71

TABELA 70: BIGRAMAS MAIS E MENOS FREQUENTES*NA CATEGORIA SHOPPING*.

Pode-se constatar, ao observar a tabela 69, que as palavras mais frequentes nesta categoria de negócio são, ***store*** (562), ***like***(411), ***place***(360), ***one***(314), ***get*** (301) e***just*** (294).

Com a tabela 70, pode-se deduzir que os bigramas mais frequentes nesta categoria de negócio são os seguintes, ***custom servic***(71), ***this place***(32),***great select***(31), ***can find*** (28), ***go back*** (23) e ***they also*** (22), mantendo a concordância dos temas abordados nesta categoria de negócio.

Pode-se constatar, que neste caso, que a exploração feita com unigramas obteve melhores resultados, uma vez que, as palavras descrevem melhor a categoria de negócio e consequentemente os assuntos abordados.

Termos com frequência mínima de ocorrência

Tendo em conta as explorações anteriores, tabela 69 e 70, bem como o número de *reviews* nesta categoria, 753, optou-se por observar a frequência mínima de palavras, no caso de unigramas, a frequência mínima de 100 e 300 vezes, no caso de bigramas, frequência mínima de 15 e 20 vezes.

De seguida, apresentam-se as figuras 103 e 104, onde se encontra o código utilizado para obter os termos com a frequência mínima referenciada anteriormente, tendo em conta o cenário, bem como os termos associados.

```
> findFreqTerms(matriz, lowfreq=100)#termos que aparecem no minimo 100 vezes
 [1] "also"   "alway"  "and"    "around" "back"   "book"   "buy"    "can"
 [9] "cloth"  "come"   "custom" "cute"   "day"    "dont"   "even"   "find"
[17] "found"  "friend" "get"    "good"   "got"    "great"  "help"   "item"
[25] "its"    "ive"    "just"   "know"   "like"   "littl"  "look"   "lot"
[33] "love"   "make"   "much"   "need"   "never"  "new"    "nice"   "now"
[41] "one"    "peopl"  "place"  "pretti" "price"  "realli" "sale"   "see"
[49] "select" "servic" "shoe"   "shop"   "someth" "staff"  "store"  "stuff"
[57] "there"  "they"   "thing"  "think"  "this"   "time"   "tri"    "use"
[65] "walk"   "want"   "well"   "will"   "work"   "your"
> findFreqTerms(matriz, lowfreq=300)#termos que aparecem no minimo 300 vezes
[1] "get"   "like"  "one"   "place" "store"
```

FIGURA 103: UNIGRAMAS QUE APARECEM PELO MENOS 100 E 300 VEZES*NA CATEGORIA SHOPPING*.

```
> findFreqTerms(rvwbi, lowfreq=15)#termos que aparecem no minimo 15 vezes
 [1] "can find"      "can get"      "come back"     "custom servic"
 [5] "dont know"     "even though"  "everi time"    "feel like"
 [9] "go back"       "great deal"   "great place"   "great select"
[13] "harvard squar" "help find"    "if your"       "look like"
[17] "love place"    "make sure"    "place go"      "pretti much"
[21] "realli nice"   "reason price" "store like"    "they also"
[25] "this place"    "this store"   "thrift store"  "you can"
[29] "your look"
> findFreqTerms(rvwbi, lowfreq=20)#termos que aparecem no minimo 20 vezes
 [1] "can find"     "custom servic" "feel like"    "go back"
 [5] "great place"  "great select"  "look like"    "love place"
 [9] "place go"     "they also"     "this place"   "you can"
[13] "your look"
```

FIGURA 104: BIGRAMAS QUE APARECEM PELO MENOS 15 E 20 VEZES*NA CATEGORIA SHOPPING*.

Observando a exploração com os unigramas, figura 103, constata-se que as palavras mais relevantes, como, ***like, place, store***, entre outras, que distinguem esta categoria.

Analisando os resultados que se obtiveram no cenário de bigramas, figura 104, visualiza-se alguns bigramas relevantes, como, ***custom servic, great select, this place***, entre outras, que distinguem esta categoria.

No entanto, esta análise confirma a conclusão obtida na exploração anterior, onde se deduz que os melhores resultados nesta categoria se obtêm no cenário de unigramas.

Wordcloud

Optou-se por fazer os *Wordclouds* com as frequências mínimas referenciadas nas análises anteriores, figura 103 e 104, para obter uma visualização mais clara, simples e perspicaz.

De seguida apresentam-se os *Wordclouds* (em que o tamanho fonte das palavras é proporcional à frequência dos termos) das frequências mínimas tendo em conta os dois cenários:

A)

B)

C)

D)

FIGURA 181: *WORDCLOUDDA CATEGORIA SHOPPING*: A) UNIGRAMAS FREQUÊNCIA MÍNIMA 100; B) UNIGRAMAS FREQUÊNCIA MÍNIMA 300; C) BIGRAMAS FREQUÊNCIA MÍNIMA 15; D) BIGRAMAS FREQUÊNCIA MÍNIMA 20.