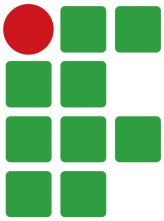

INSTITUTO FEDERAL
Goiás
Câmpus Anápolis

Pedro Henrique Silva Rodrigues

UMA ANÁLISE INICIAL DE BASES PARA MINERAÇÃO DE DADOS

Anápolis-GO

Maio - 2019

Pedro Henrique Silva Rodrigues

UMA ANÁLISE INICIAL DE BASES PARA MINERAÇÃO DE DADOS

Relatório Científico apresentado ao curso de Bacharelado em Ciências da Computação, como exigência da Disciplina de Mineração de Dados ao prof^{o} D^{r} $Daniel Xavier Sousa$

Instituto Federal de Goiás - Campus Anápolis

Anápolis-GO
Maio - 2019

Lista de ilustrações

Sumário

Resumo

Enormes quantidades de dados são gerados diariamente através de pesquisas na *web*, compartilhamento de informações pelas redes sociais, compras, imagens de câmeras e muitas outras. Retirar informações desses dados tem sido o carro-chefe de muitas empresas da atualidade. A esse processo de extrair informação a partir de um conjunto de dados dá-se o nome de Mineração de Dados. Para obter essas informações a Mineração de Dados utiliza uma infinidade de modelos matemáticos e estatísticos com o objetivo de entender o que esses dados estão informando. Para isso duas etapas nesse procedimento são extremamente importante: A preparação e limpeza dos dados pois existem, em meio a essa grande quantidade de informações, atributos que poluem os modelos e tornam as análises enviesadas. Devido a essa necessidade este estudo apresenta uma análise estatística inicial, procedimentos para limpeza, e classificação dos atributos de três coleções de dados: Um conjunto de *tweets* e uma coleção de textos de páginas *web* ambos relacionados ao assunto comida, obtidas por meio de uma ferramente de *Crawler*, e uma pesquisa realizada com docentes, técnicos administrativos e discentes do IFG (Instituto Federal de Goiás), sobre questões institucionais.

Introdução

O uso da internet cresceu de forma explosiva nas últimas décadas (CURRAN; FENTON; FREEDMAN, 2016). Acompanhando esse crescimento está a geração de dados em pesquisas web, compartilhamentos em redes sociais, divulgações de notícias, debates, gravações de câmeras, fotos de satélite, entre muitos outros. Muitas empresas enxergaram nesses dados uma possibilidade de negócio.

Essas empresas tem aplicado técnicas de busca e análise de dados. Devido à grande quantidade é inviável para um humano analisar todas essas informações (ROIGER, 2017). Dessa forma é necessário a utilização de ferramentas computacionais para extrair informações relevantes desses dados e gerar *insights*, que podem proporcionar as empresas vantagens de mercado, por exemplo.

Ao processo de automatizar buscas e fazer análises de dados dá-se o nome de Mineração de Dados (do inglês, *Data Mining*)(WITTEN et al., 2016). Os dados, principalmente na *web*, estão de carregados de textos irrelevantes, imagens, *tags HTML*, e gerar modelos baseados nesses dados sem nenhuma forma de tratamento ocasiona a obtenção de informações enviesadas ou inconsistentes dos dados. Para isso são aplicadas técnicas de limpeza aos dados.

Devido a essa necessidade este estudo apresenta uma análise estatística inicial, procedimentos para limpeza, e classificação dos atributos de três coleções de dados de situações reais: Um conjunto de *tweets* e uma coleção de textos de páginas *web* ambos relacionados ao assunto comida, obtidas por meio de uma ferramente de *Crawler* (método de obtenção de dados de páginas *web* (BULLOT; GUPTA; MOHANIA, 2003)), e uma pesquisa realizada com docentes, técnicos administrativos e discentes do IFG (Instituto Federal de Goiás), sobre questões institucionais.

1 Referencial Teórico

Nesse capítulo são apresentados definições sobre Mineração de Dados e seu objetivo, abordando suas principais etapas, e a importância da Análise Estatística durante esse processo.

1.1 Mineração de Dados

Mineração de Dados é a área que tem por objetivo compreender as entidades do mundo real através da formulação de modelos com base em um grupo de dados que se obtém durante o processo.

1.1.1 *Web Crawler*

A primeira etapa do processo de Mineração de Dados é a obtenção de um conjunto de dados, seja ele um conjunto de imagens, ou textos em páginas web, comentários em redes sociais, ou bancos de dados relacionais de aplicações comerciais. *Web Crawler* é o nome do procedimento de obtenção de dados a partir de páginas *web* (WITTEN et al., 2016).

A partir de uma página inicial o *Crawler* irá navegar pelo conteúdo da página buscando por links e associações com outras páginas. O *Crawler* então vai repetir esse procedimento pelas páginas dos links encontrados, gerando assim ao final uma grande rede contendo as associações entre as páginas e seus conteúdos (WITTEN et al., 2016).

Ao final do procedimento de *Crawling* um grande conjunto de páginas *web* são retornadas e os conteúdos e associações entre essas páginas podem ser utilizadas como parte do processo de mineração (WITTEN et al., 2016). É importante ressaltar que o processo de *Crawling* deve ser realizado respeitando as liberdades individuais impostas pelos donos das páginas.

1.1.2 Preparação dos Dados

A preparação e limpeza dos dados existem pois existem, em meio a uma grande quantidade de informações, atributos que poluem os modelos e tornam as análises enviesadas (WITTEN et al., 2016).

Ao obter páginas da *web* muito do conteúdo será composto por *tags HTML* que muitas vezes não representam informação importante para os processos de mineração. Nesse contexto entra a etapa de tratamento e preparação de dados. Essa etapa é responsável

pela limpeza dos dados que tem por objetivo facilitar o processo de análise automática dos dados pelos métodos computacionais.

Entre os métodos de preparação e limpeza dos dados estão: A radicalização que consiste em transformar as palavras de um texto em seus radicais. Remover de textos caracteres que dificultam o processo de aprendizado e palavras não significativas. Em casos de bases transacionais remover *outliers* e documentos com elementos faltantes pode melhorar a análise.

1.1.3 Formulação de Modelos

Diferentemente dos métodos de Aprendizado de Máquina, onde o modelo que faz predições ou classificação de novas ocorrências é o resultado final (MICHIE et al., 1994), os modelos finais de Mineração de Dados são a compreensão dos dados sob um aspecto que se deseja analisar.

1.1.4 Análise dos Modelos

A partir da exploração e análise dos modelos obtidos durante o processo de mineração, um entendimento das informações se cria. Esse entendimento é o resultado de todo o processo de Mineração. A partir desse entendimento novas assunções podem ser criadas suportando assim a geração de conhecimento. Com essas informações as empresas podem por exemplo mudar seu modo de funcionamento, se adaptar as necessidades de mercado e obter mais lucro. Essas informações são o objetivo desconhecido que se pretendia descobrir.

1.2 Análise Estatística

Durante o processo de Mineração de dados, uma das etapas mais importantes é a análise estatística, pois ela permite um entendimento inicial dos dados que será extremamente importante nas demais etapas.

As medidas de dispersão utilizadas para analisar as bases desse estudo são: mediana, moda, quantis, variância, desvio padrão e coeficiente de dispersão. Essas medidas representam a variabilidade dos dados, ou ainda sua tendência a certos valores (REIS et al., 1999). São valores que podem representar a probabilidade de um dado pertencer a um intervalo, ou de forma geral avaliam a distância de uma dada lista ao elemento de tendência central. Dados muito dispersos podem confundir modelos de agrupamento.

As distribuição de frequência (relativa ou acumulada) e histogramas permitem visualizar e comparar quantidade de ocorrências (REIS et al., 1999), o que é muito

importante pois identifica informações muito ou pouco recorrentes, auxiliando na remoção de *outliers* da coleção.

Uma outra análise importante em Mineração de Dados é a análise de correlação pode ser entendido como o grau de dependência entre duas variáveis (REIS et al., 1999) ou atributos em coleções.

2 Coleções de Dados Analisadas

As coleções de dados apresentadas nesse capítulo foram obtidas através de procedimentos de *Crawler*, ou em Bancos de Dados de situações reais, respeitando as regras de *Web Crawling* e Termos de Confidencialidade. Esse processo de obtenção de dados representa o estágio inicial da Mineração de Dados. Esse capítulo aborda os procedimentos para obtenção dessas coleções e suas descrições.

2.1 *Tweets* Relacionados ao Assunto: Comida

Essa coleção foi obtida através do uso da Interface de Programação de Aplicações (do inglês Application Programming Interface - API) disponibilizada pela rede social Twitter. Essa API faz buscas pelos comentários, chamados de *tweets* de seus usuários, que possuem palavras-chave ou termos associados a uma informação, tópico ou discussão explicitados com uma #. A API faz buscas por essas palavras-chave retornando os *tweets* que as incluem. os *Tweets* pesquisados foram #comida, #culinaria e #receita. Após passar pelo processo de limpeza, removendo palavras irrelevantes. Esses comentários passaram pelo processo *tokenizer* e radicalização onde se tornaram um conjunto de XXXXX atributos onde cada atributo representa um radical e seu valor para uma dado comentário é a quantidade de ocorrências desse radical no comentário, por ser contável esse atributo é discreto.

2.2 Textos de Páginas *Web* Relacionados ao Assunto: Comida

Essa coleção foi obtida através do uso da ferramenta de *Crawler*, *Scrapy* (Scrapy. . . ,). Essa ferramenta exige que sejam informadas as *URL's* de fronteira. Para isso foi utilizado um outro *Crawler* que faz buscas por tópicos e retorna um conjunto de páginas *Web* relacionadas ao(s) tópico(s) informados. Dessa forma fez-se uma pesquisa pelos tópicos Goiás, e Comida utilizando essa ferramenta e em seguida utilizou-se os *URL's* retornados como a fronteira inicial para o *Scrapy*. Ao tentar aplicar o processo de limpeza sobre esses dados, removendo *TAGs HTML*, textos irrelevantes e imagens os textos ficaram pouco representativos. Esses textos passaram pelo processo *tokenizer* e radicalização onde os textos se tornaram um conjunto de atributos onde cada atributo representa um radical e seu valor para uma dado documento é a quantidade de ocorrências desse radical no documento, por ser contável esse atributo é discreto. Mesmo com a limpeza a qualidade dos dados obtidos não gerou resultados satisfatória para essa análise.

2.3 Pesquisa Institucional do IFG

Essa coleção representa um questionário feito aos discentes, docentes, e técnicos administrativos dos campus do IFG (Instituto Federal de Goiás). No total 3701 pessoas responderam a esse questionário informando os seguintes dados:

Segmento ao qual pertence: É um atributo nominal pois é relativo a uma característica e consequentemente é discreto, pois assume um conjunto finito de valores. Ex.: Técnico Administrativo, Docente ou Aluno;

Campus que pertence: Também é nominal e discreto. Ex.: Jataí, Goiânia, Uruaçu entre outros.

Curso que pertence Também é nominal e discreto, porém é relativo somente ao discentes. Ex.: Bacharelado em Engenharia Civil

Questões: Também é nominal e discreto, inclui uma das três categorias: Não, Sim e Não Soube ou Não Respondeu. As questões são:

0. Tem conhecimento do último processo de autoavaliação Institucional?
1. Você percebe a utilização dos resultados da CPA?
2. Você participou do Planejamento do ano de 2018 na Pró-Reitoria a qual você está vinculado(a)-
3. Você considera satisfatória a divulgação do Planejamento anual do seu Câmpus?
4. Você participa da elaboração do Planejamento anual do seu Câmpus?
5. Os cursos ofertados no seu Câmpus atendem as demandas socioeconômicas da região?
6. De maneira geral, você considera que a formação que está recebendo é de boa qualidade?
7. Você acompanha os trabalhos do Conselho de Ensino Pesquisa e Extensão (CONEPEX)?
8. Você conhece ou participa de algum Projeto de Pesquisa do IFG?
9. Você conhece ou participa de algum Projeto de Extensão do IFG?
10. Você considera satisfatória a comunicação do IFG por meio do site e das redes sociais?
11. De maneira geral, você é bem atendido/a nos setores de atendimento ao/à discente/docente no IFG?
12. Você considera satisfatória a atuação do IFG para promoção da permanência e êxito dos/das estudantes?

13. Você conhece a função da ouvidoria do IFG?
14. Você conhece ou participa de algum Projeto de Ensino?

Deixe sua crítica ou sugestão para o IFG: Após passar pelo processo de *tokenizer* e radicalização as opiniões se tornam um conjunto de 2514 atributos onde cada atributo representa um radical e seu valor para uma dado documento é a quantidade de ocorrências desse radical na opinião/sugestão do entrevistado. Sendo portanto um atributo ordinal e por ser contável discreto.

As dimensões desse conjunto após o processo de limpeza e tratamento passam a ser 2544 atributos x 3701 documentos.

3 Metodologia Experimental

Os procedimentos de limpeza aplicados aos textos das coleções citadas em 2 foram *Parser* para os textos obtidos de páginas web, tokenizer e radicalização para todas as coleções.

O procedimento de *Parser* consiste em extrair o texto de tags, remover referências a imagens, e outros conteúdos *HTML*. Mesmo após aplicar esse procedimento no conteúdo das páginas web, ainda houve muitas informações poluindo esse conteúdo. Isso mostra a principal dificuldade do processo de Mineração de Dados, que são as tentativas de remover dados inconsistentes que acabam atrapalhando os resultados, pois o processo de limpeza pode não funcionar como o esperado. Realizar a análise estatística nesse caso não seria suficiente para obter informações de qualidade.

Tanto os tweets quanto os dados relacionais do IFG puderam ser melhor analisados. Essas coleções já disponibilizavam textos entendíveis e passíveis de ser aplicados diretamente o processo de tokenizer e radicalização. Esses métodos consistem em fazer uma quebra do texto separando as palavras em tokens, removendo caracteres especiais e repetições de letras não informativas. Em seguida cada token é transformado no seu radical que é o prefixo da palavra, onde por fim é avaliado a quantidade de ocorrências desse prefixo em cada texto.

A medida utilizada para avaliar a relevância desses radicais foi o TF-IDF. Uma medida calculada com base na equação 3.1.

$$TF - IDF_{doc_i,te_j} = \frac{TF_{doc_i,te_j} \cdot IDF_{te_j}}{\sum_{j=0}^{m} TF_{doc_i,te_j}} \tag{3.1}$$

Onde TF_{doc_i,te_j} é a frequência do termo dentro de um documento, e IDF_{te_j} é a frequência inversa dos termos nos documentos do corpus que o contém. Essa métrica normaliza a frequência do termo com base no tamanho do documento, evitando assim dar um peso elevado para uma palavra em textos muito grandes. É a métrica mais utilizada na literatura (RAMOS et al., 2003).

4 Resultados e Análise

Os resultados das análises estatísticas e e algumas asserções com base nessa análise estão contidas nos gráficos e textos desse capítulo, subdivido pelas coleções que foram analisadas. Com base nesses resultados é possível ter uma visão preliminar dos dados.

4.1 Pesquisa IFG

Utilizando o atributo cidade da coleção foi possível avaliar a frequência absoluta da quantidade de entrevistados da coleção. Do total de 3700 entrevistados cerca de 1200, aproximadamente $\frac{1}{3}$, pertencem a cidade de Goiânia (figura 4.1). Por ser a capital e possuir dois campus esse resultado já era esperado. Entretanto sabendo disso em futuras análises deve-se atentar para o impacto das opiniões desse campus não minimizar as informações dos demais. Desse forma sugere-se a utilização de amostragem com base em critérios de forma a normalizar os resultados. Os resultados relacionados a Reitoria e ao campus de Valparaíso teriam seus valores influenciados pelo campus de Goiânia sem o devido tratamento.

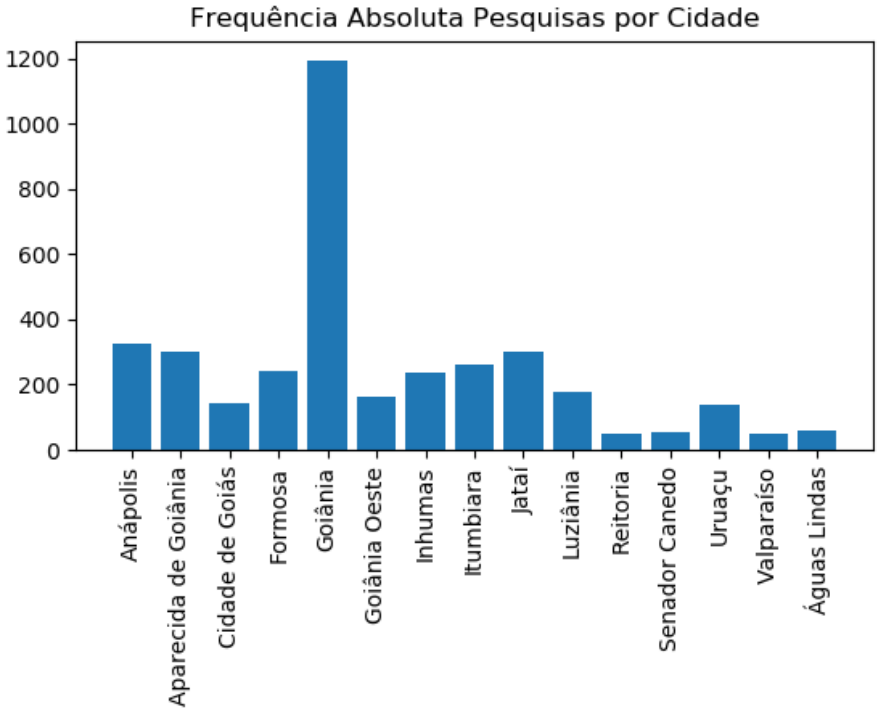

Figura 1 – Frequência Absoluta de Número de Pesquisas por Cidade

Avaliando a figura 2 e comparando as questões descritas em 2.3 é possível perceber

que os entrevistados em sua maioria acham satisfatória a divulgação do planejamento anual do seu campus, entretanto um ponto negativo é que muitos entrevistados não responderam a pergunta 12 que é se são bem atendidos/as nos setores de atendimento ao/à discente/docente no IFG, cabe ao IFG avaliar o que essa informação pode representar no contexto em que foi respondida. Com base nesse gráfico outras informações podem ser interpretadas, avaliando cada questão, nesse estudo não entraremos em detalhes de todas.

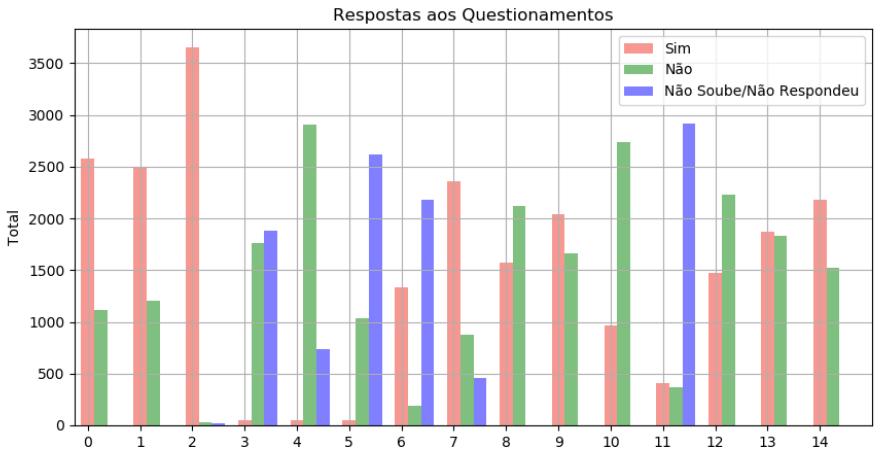

Figura 2 – Respostas aos questionamentos da pesquisa IFG

Devido a grande quantidade de radicais gerados no processo de radicalização iremos avaliar os cinco com maiores TF-IDF comparando-os. Dentre essas cinco, duas podem ser relacionadas a sugestão, indicando que existem aspectos que podem ser melhorados (radical melh no gráfico 3). Cabe ao processo de mineração responder: "O que pode ser melhorado?". Outros radicais sugerem a construção de algo, cabe ao processo de mineração responder: "O que pode está sendo sugerido para construção?". No gráfico 3 também mostra que o IFG recebeu muitos elogios, dois dos radicais mais relevantes são associados a um sentimento de aceitação. Houve ainda um quinto radical que demonstra o lado satírico de alguns estudantes pelo radical merd. Cabe ao IFG decidir em que podem ser usadas essas informações.

Os gráficos 4 e 5 mostram como esses radicais estão distribuídos pelos documentos, nota-se maior proximidade da médias os radicais melh, bom, e gost enquanto constru tem uma dispersão menor, e merd aparece como um possível *outlier*.

O gráfico de correlação 6 mostra uma sutil combinação dos termos melh, gost e bom, reafirmando a hipótese de que representam sentimentos favoráveis ao IFG. A ausência de correlação entre os demais sugere que essas palavras não ocorrem num mesmo documento.

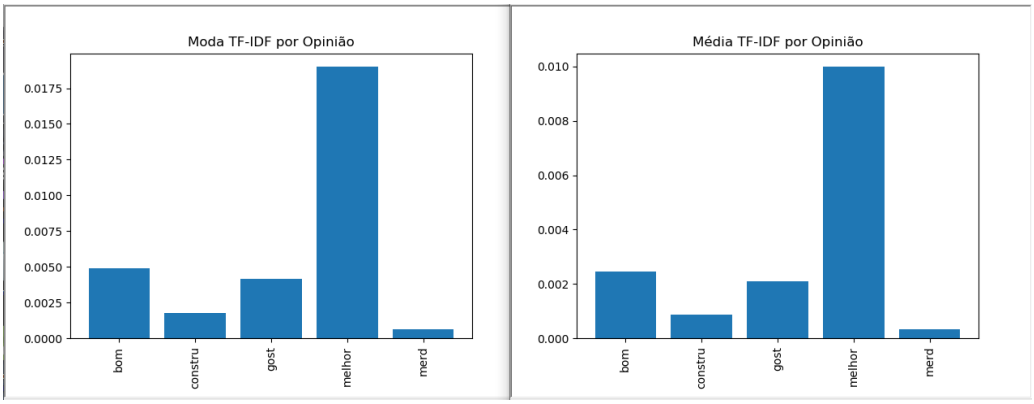

Figura 3 – Medidas de Tendência Central do TF-IDF para os 5 radicais com maior relevância das opiniões e sugestões dos entrevistados.

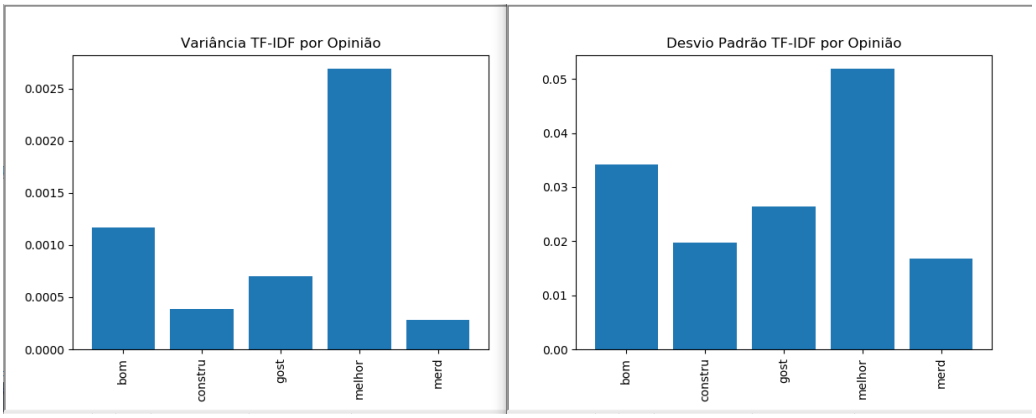

Figura 4 – Medidas de Dispersão do TF-IDF para os 5 radicais com maior relevância das opiniões e sugestões dos entrevistados.

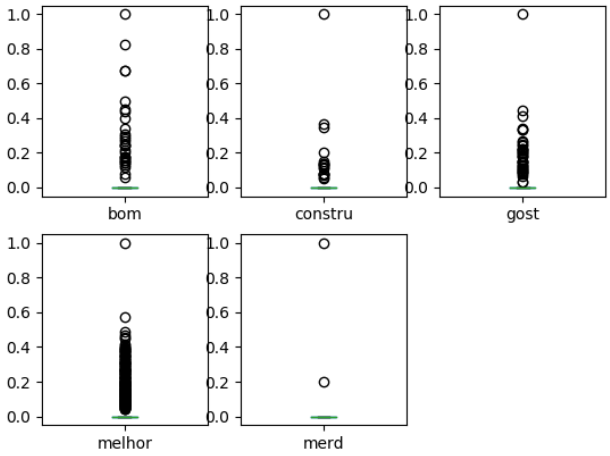

Figura 5 – BoxPlot das ocorrências dos 5 radicais mais relevantes da pesquisa

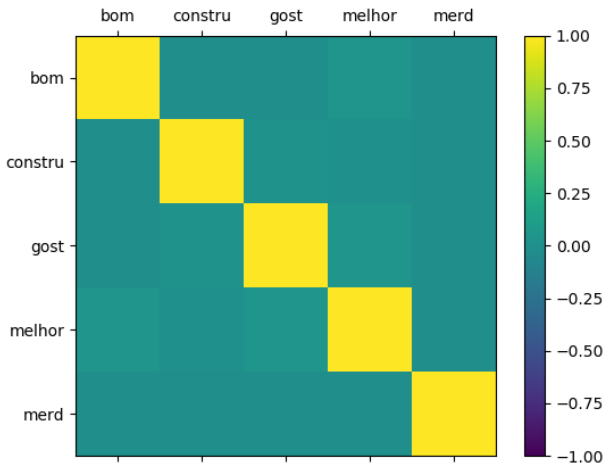

Figura 6 – Correlação entre os radicais mais presentes na pesquisa.

4.2 Coleção de *Tweets*

Como os *tweets* contêm a #comida, espera-se que os resultados mostrem uma grande ocorrência dessa palavra nos documentos. Um ponto importante a se avaliar observando esses resultados é a presença de termos em inglês comoo food, e recet, que é o prefixo para recetas (receitas em português). O ponto de maior destaque observando o gráfico 7 é a presença dos radicais cas, e cocin, que são relativos a casa e cozinha. Não a uma explicação no momento para essa ocorrência, o que nos leva a questão: Porque comida foi associado a esses dois termos?

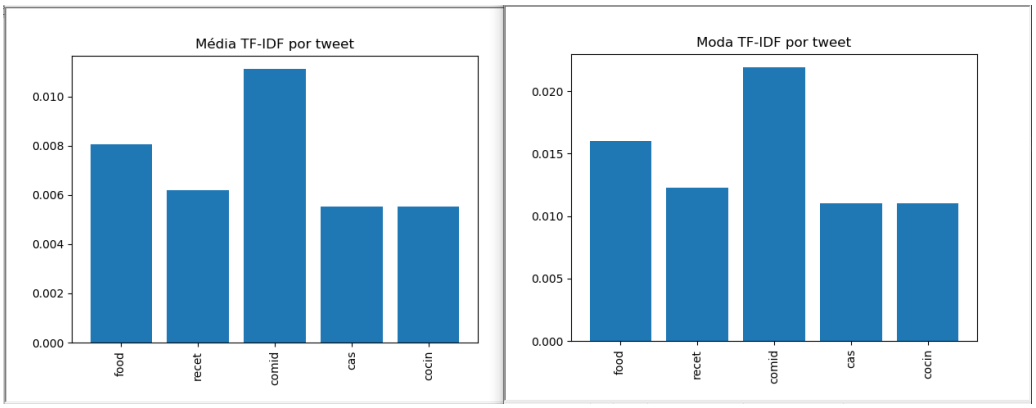

Figura 7 – Medidas de Tendência Central do TF-IDF para os 5 radicais com maior relevância dos tweets.

Nota-se uma ligeira diferença entre o radical comid e os demais no boxplot, porém esse gráfico está pouco informativo e portanto não se devem fazer alegações acerca de seu conteúdo.

No gráfico de correlação, nota-se uma forte correlação entre os radicais cas e cocin,

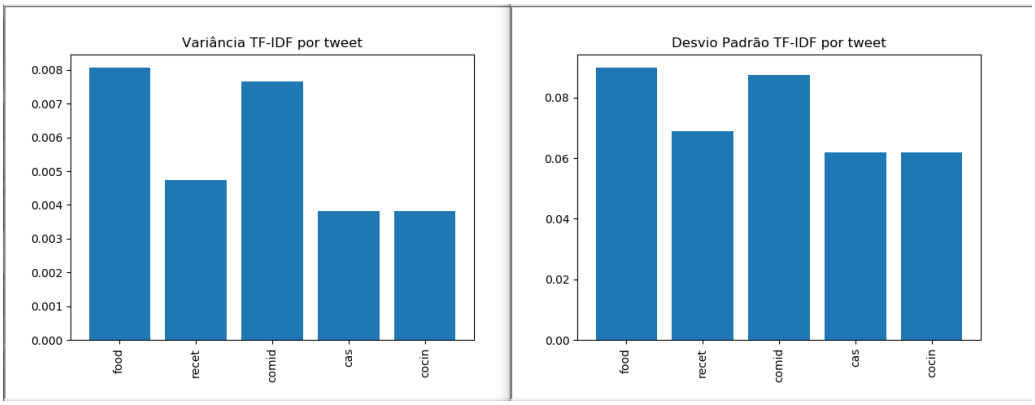

Figura 8 – Medidas de Dispersão do TF-IDF para os 5 radicais com maior relevância dos tweets.

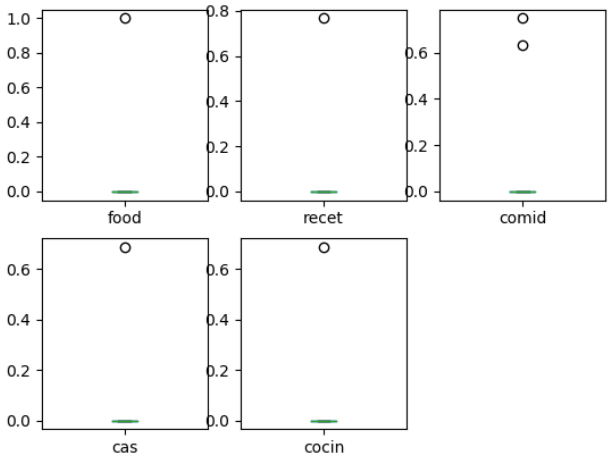

Figura 9 – BoxPlot das ocorrências dos 5 radicais mais relevantes nos tweets.

o que sugere que eles estão presentes nos mesmo documentos. Não há relação entre os demais, o que se espera visto que os dados estão sendo provenientes de três idiomas distintos. Avaliar a linguagem natural se torna ainda mais complexo quando são vários idiomas, portanto sugere-se fazer uma filtragem dos dados previamente selecionando um idioma específico ou utilizar algum método de tradução.

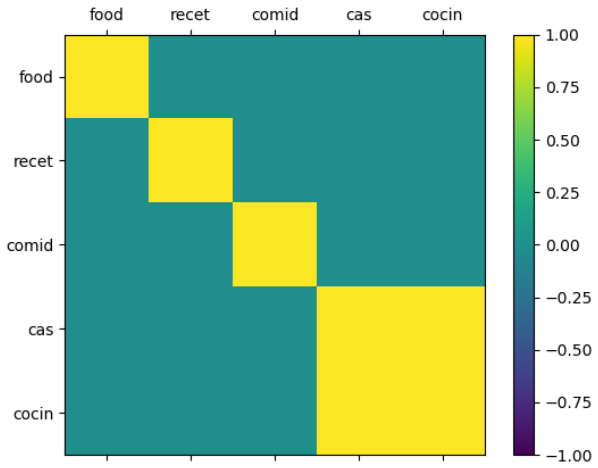

Figura 10 – Correlação entre os radicais mais presentes nos tweets.

5 Conclusão

Conforme visto a análise estatística pode explicar de forma sucinta padrões e ocorrências nos dados. Devido a isso ela futuramente nos fornecerá uma base no processo de Mineração de Dados para as coleções de tweets e Pesquisa do IFG.

Cabe ainda trabalhar no processo de limpeza e obtenção de dados de páginas web, pois esses dados também representam parte importante do conteúdo sobre comida disponível publicamente. Com esses resultados foi possível fazer afirmações inicias sobre as coleções e gerar questionamentos sobre o que mais elas podem informar.

Referências

BULLOT, H.; GUPTA, S. K.; MOHANIA, M. K. A data-mining approach for optimizing performance of an incremental crawler. In: IEEE. *Proceedings IEEE/WIC International Conference on Web Intelligence (WI 2003)*. [S.l.], 2003. p. 610–615. Citado na página 5.

CURRAN, J.; FENTON, N.; FREEDMAN, D. *Misunderstanding the internet*. [S.l.]: Routledge, 2016. Citado na página 5.

MICHIE, D. et al. Machine learning. *Neural and Statistical Classification*, Technometrics, v. 13, 1994. Citado na página 7.

RAMOS, J. et al. Using tf-idf to determine word relevance in document queries. In: PISCATAWAY, NJ. *Proceedings of the first instructional conference on machine learning*. [S.l.], 2003. v. 242, p. 133–142. Citado na página 12.

REIS, E. et al. Estatística aplicada. *Lisboa: Edições Sílabo*, 1999. Citado 2 vezes nas páginas 7 e 8.

ROIGER, R. J. *Data mining: a tutorial-based primer*. [S.l.]: Chapman and Hall/CRC, 2017. Citado na página 5.

Scrapy A Fast and Powerful Scraping and Web Crawling Framework. <https://scrapy.org/>. Accessed: 2019-04-25. Citado na página 9.

WITTEN, I. H. et al. *Data Mining: Practical machine learning tools and techniques*. [S.l.]: Morgan Kaufmann, 2016. Citado 2 vezes nas páginas 5 e 6.