II Workshop de Matemática, Estatística e Computação Aplicadas à Indústria

Um Método para Comparar Palavras entre Categorias utilizando Word2Vec e Redução de Dimensionalidade no Problema de Categorização de Produtos pela Descrição

Gilsiley H. Darú,[1] Philipe D. Almeida,[2] Francisco J. Pigato,[3] Antônio Castelo Filho[4]
ICMC-USP

1 Introdução

Com o advento da pandemia e com as restrições impostas em muitos países, a área de e-commerce cresceu significativamente. Com isto, as empresas tiveram que aprimorar suas plataformas de comércio eletrônico. Plataformas estas que precisam entregar uma experiência adequada para o usuário de forma simples e objetiva. No entanto, encontram-se dificuldades na busca de informações adequadas, o que gera um longo trabalho de classificação e preparação de dados para entrega de um serviço de atendimento ao usuário condizente com sua necessidade. O objetivo desta pesquisa é desenvolver técnicas e algoritmos que permitam extrair e classificar dados e informações de produto de forma automática a fim de entregar uma melhor qualidade nas consultas dos indivíduos. Temos então um problema de qualificação de dados, e mais especificamente no nosso caso, a rotulação ou classificação dos itens.

Um exemplo de dificuldades neste processo de categorização, suponha que se deseja classificar as descrições para a categoria biscoito. Assim, uma primeira forma que surge na mente é buscar por palavras chaves. Isto permitiria encontrar descrições do tipo ”BISCOITO DUX SALGADO 648G”e ”BISCOITO CRACKERS DUX 216G”. Porém existem casos em que a palavra biscoito está abreviada como ”BISC LEITE MABEL 400G”e ”BISC.PARATI MARIA PE 370GR”, sendo necessário expandir a busca por mais termos como ”BISC”e ”BISC.”. Além disto existem termos regionais ou similares, por exemplo as descrições ”BOLACHA PAPAGUARA MOTOR 400g”, ”COOKIE INT BAUDUCCO AVEIA PASSAS 40G”e ”COOKIES GRAN CASTANH PARA KOBBER 150G”levam novamente a expansão do problema para palavras como BOLACHA, COOKIE e COOKIES. Mas o problema ainda continua onde produtos contém as palavras acima

[1] ghdaru@ups.br
[2] philipedalmeida@usp.br
[3] pigato@usp.br
[4] castelo@icmc.usp.br

mas não são biscoitos, como por exemplo ”BISC PET CRACKER FORTAL 400G”, ”BISC PET DOG CROCK 500G”, ”BISC LACTA BIS FLOWP 126G, LAKA BCO”e ”CESTA SUPREME BISCOITO LIMAO 100G”que são biscoitos para cachorros, chocolate ou wafer e o último é um presente.

Desta forma este trabalho pretende, a partir da descrição de um produto, D, avaliar algoritmos de classificação para atribuir uma classe ou categoria C.

2 Trabalhos Relacionados

2.1 Definições

O problema de classificação, conforme [1], pode ser definido como um conjunto de instâncias D = {X1,...,XN}, tal que cada instância possui uma classe indexada pelos valores {1 ...k}, onde k é o número de classes. Um conjunto de treino é usado para construir um modelo de classificação. Uma nova instância desconhecida se utiliza do modelo construído para prever sua classe. O problema de classificação pode ser considerado *hard* quando se determina explicitamente uma classe ou *soft* quando se atribui probabilidades. Ainda em [1], este enumera algumas técnicas utilizadas tais como árvores de decisão, classificadores baseado em padrões, máquinas de vetores de suporte classificadoras, classificadores baseados em redes neurais, classificadores Bayesianos e meta algoritmos(combinação, empilhamento e sequenciamento), classificadores lineares e classificadores baseados em vizinhança sendo os mais importantes.

Já no artigo [4], este trata da avaliação a partir da similaridade entre vetores. Estas incluem modelos transdutivos, modelos baseados em variáveis latentes de Dirichlet, redução de dimensionalidade com decomposição em valores singulares, análise semântica latente, entre outros, que projetam a variável a ser predita nos vetores de cada classe e verificam sua similaridade.

É importante destacar que para todos os métodos descritos acima, faz-se necessário transformar o texto de entrada em um espaço vetorial. Algumas técnicas são sugeridas na literatura. A seguir destacam-se dois importantes métodos.

2.2 Vetorização

Sacola de palavras ou *bag of words* ou BoW, conforme [5] é um dos métodos de representação mais populares na categorização de objetos. A ideia principal é associar um número a cada palavra de uma descrição e representar a descrição pela presença ou não da palavra, com um valor 1 em sua posição.

Outra forma de representação vetorial é dado pelo artigo seminal de [3]. O termo *word2vec* surgiu a partir deste artigo e é uma técnica para processamento de linguagem natural que usa um modelo de rede neural para aprender associações de palavras de um grande corpus de texto o qual após treinado, pode detectar sinônimos ou sugerir palavras adicionais para uma frase incompleta. O nome *word2vec* indica ”palavra para vetor”, significando que cada palavra distinta é representada por uma lista particular de números, o vetor. Então uma função matemática simples (similaridade do cosseno entre os vetores) indica o nível de similaridade semântica entre as palavras representadas por esses vetores.

A definição do espaço vetorial pode influenciar no classificador escolhido.

2.3 Redução de Dimensionalidade

O problema de conversão de textos para vetores trazem o problema da dimensionalidade. Para este trabalho o tamanho do vocabulário gerado foi de mais de quarenta mil elementos. Ao se utilizar a técnica de sacola de palavras, qualquer palavra é representada por um vetor do tamanho do vocabulário. Para reduzir a dimensionalidade de um problema e manter a quantidade de informação disponível o máximo possível é utilizar técnicas projetivas. Uma desta técnicas é chamada de análise dos componentes principais, que em inglês é chamada de PCA.

Esta técnica, conforme [2], permite aumentar a interpretabilidade ao mesmo tempo que minimiza a quantidade perdida de informação. Consiste em criar novas variáveis não correlacionadas. Para realizar a mudança, reduz-se a quantidade de bases, mantendo a quantidade de variabilidade dos dados. Esta variabilidade é dada pela equação abaixo.

$$\tilde{\lambda}_i = \frac{\lambda_i}{\sum_j \lambda_j}$$

Sendo λ_i o i-ésimo auto-valor ordenado do maior para o menor da matriz de covariância e $\tilde{\lambda}_i$ a variabilidade explicada pelo i-ésimo auto-valor.

3 Resultados

Foi aplicado o Word2Vec para todos os elementos do vocabulário, utilizando-se uma janela de tamanho 5 sobre todas as descrições tokenizadas do corpus. Isto gerou uma matriz com Vxn, onde V é o tamanho do vocabulário e n é o tamanho do vetor da palavra obtido pela técnica Word2Vec. Esta matriz foi reduzida para uma matriz V'x2, utilizando-se as duas componentes principais. Isto permite visualizar as palavras em duas dimensões. Após isto, agrupa-se as descrições de cada categoria e gera-se seu vocabulário V1. O mesmo procedimento é realizado para uma segunda categoria de interesse V2. Após gera-se o conjunto V1 - V2, V2 - V1 e V1 $\bigcap$ V2 e plota-se seus vetores bidimensionais. Para permitir uma dinâmica de análise foi construído um aplicativo em python utilizando-se a biblioteca streamlit.

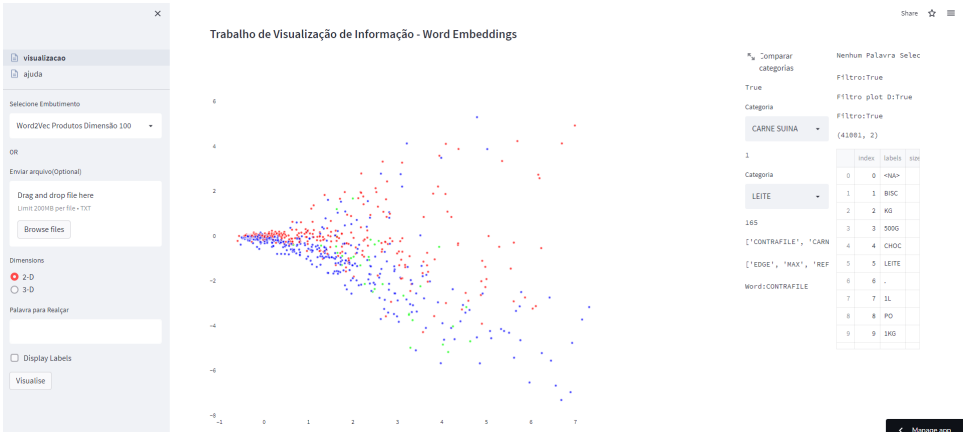

Figura 1: Aplicativo gerado

https://ghdaru-category-embeddings-comparison-visualizacao-9kraei.streamlitapp.com/

4 Conclusões e sugestão de trabalhos futuros

A técnica permitiu rapidamente identificar categorias com grande superposição de informação. Categorias ”independentes”possuem uma nuvem em direções opostas, sendo que as palavras da intersecção se distribuindo em torno do eixo horizonte. O que não ocorre com categorias similares, tais como SABONETE e SABONETE INFANTIL. A partir da visualização é possível direcionar trabalhos futuros tais como avaliar o vetor projeção de cada categoria para avaliar suas similaridades, utilizando-se os conjuntos diferenças.

Referências

[1] Charu C Aggarwal e ChengXiang Zhai. “A survey of text classification algorithms”. Em: *Mining text data*. Springer, 2012, pp. 163–222.

[2] Ian T Jolliffe e Jorge Cadima. “Principal component analysis: a review and recent developments”. Em: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374.2065 (2016), p. 20150202.

[3] Tomas Mikolov et al. *Efficient Estimation of Word Representations in Vector Space*. 2013. DOI: 10.48550/ARXIV.1301.3781. URL: https://arxiv.org/abs/1301.3781.

[4] Ge Song et al. “Short text classification: a survey.” Em: *Journal of multimedia* 9.5 (2014).

[5] Yin Zhang, Rong Jin e Zhi-Hua Zhou. “Understanding bag-of-words model: a statistical framework”. Em: *International journal of machine learning and cybernetics* 1.1 (2010), pp. 43–52.