UNIVERSIDADE FEDERAL DO AMAZONAS

INSTITUTO DE COMPUTAÇÃO

BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

Classificação de Produtos através de Modelos de Linguagem

Rúben Jozafá Silva Belém

Manaus - AM

Setembro de 2018

Sumário

1 Introdução

1.1 Contextualização

1.2 Objetivos

1.3 Organização do Trabalho

2 Revisão Bibliográfica

Muitos trabalhos encontrados na literatura abordam o problema de classificação de produtos. Alguns dos métodos clássicos de aprendizado de máquina podem ser utilizados para esse problema, como o *SVM - Support Vector Machine* (Máquina de Vetores de Suporte) (JOACHIMS, 1998), *K-NN - K Nearest Neighbours* (K vizinhos mais próximos) (CHAKRABARTI, 2003), e *Naive Bayes* (RISH, 2001). No entanto, as conclusões desses trabalhos apontam que utilizar desses métodos não produz um resultado tão satisfatório. Portanto, verificou-se trabalhos especificamente aplicados à classificação de produtos.

A *Deep Categorization Network* (HA; PYO; KIM, 2016), ou *DeepCN*, uma rede profunda de ponta a ponta formada por múltiplas redes neurais recorrentes (RNNs) é alimentada com metadados provindos de produtos de comércio eletrônico. A rede possui camadas totalmente conectadas e uma camada *softmax*. Para cada atributo de item como nome, marca ou fabricante há uma RNN dedicada. Elas geram vetores de valor real que caracterizam a semântica das palavras, descartando assim um processo de treinamento prévio como o *word2vec* (MIKOLOV et al., 2013).

Outra abordagem é a de uma Rede Profunda de Fusão de Níveis de Decisão (ZAHAVY et al., 2016), para classificação de produtos multimodais. Um produto multimodal no contexto desse trabalho é composto por texto e imagem, os quais são utilizados como entrada. A rede multimodal melhora a precisão *Top-1%* em uma base coletada do *Walmart*. São utilizadas duas Redes Neurais Convolutivas (CNN), uma para texto e outra para imagens. A CNN de texto tem precisão maior que a de imagem quando há poucos produtos. Quando há muitos, a CNN de imagem tem melhor desempenho. Isso serviu de incentivo para combinar as saídas das duas CNNs em uma terceira rede neural de "policiamento", que aprende a escolher entre o resultado da CNN de imagem ou de texto.

Uma vez que a abordagem deste trabalho é baseada em *modelos de linguagem*, também foi realizada uma busca a respeito dessa forma de classificação. Um dos trabalhos mais recentes e que se destaca é um estudo sobre algoritmos de suavização em modelos de linguagem para categorização de itens de um comércio eletrônico (SHEN et al., 2012). Esses algoritmos são úteis para ajustar o estimador de máxima verossimilhança. Em outras palavras, regulam alguns parâmetros do modelo para a que a escassez de dados não afete tanto o desempenho. Os modelos de linguagem geralmente designam probabilidades aos termos de uma consulta, com base em suas ocorrências nos documentos. No entanto, algumas palavras da consulta podem não existir em nenhum dos documentos. O algoritmo de suavização então determina quanto se deve descontar da massa de probabilidade das palavras da consulta que aparecem na coleção, para atribuir tal desconto às palavras "estranhas".

Analisou-se o desempenho do modelo de linguagem com cinco algoritmos de suavização: Suavização de *Laplace*, de *Jelinek–Mercer*, de Distribuição de *Dirichlet*, de Encolhimento, e por último, o Desconto Absoluto. Os testes foram realizados com uma base de dados do *eBay* que possui uma hierarquia de categorias de 7 níveis, com 34 categorias no primeiro nível, e um total de 19000 categorias "folha" (que não são "pai" de nenhuma outra categoria), com mais de 18 milhões de produtos. Dos cinco algoritmos, o que teve melhores resultados foi o da Suavização de Distribuição de *Dirichlet*. Então, passou-se a considerar somente este último algoritmo no resto do trabalho. Foram analisadas também 3 tipos de influência nos resultados: a do tamanho da base de treino, do tamanho dos documentos das categorias, e da especificidade de palavras entre categorias (se as palavras entre duas categorias são pouco distintas, naturalmente a classificação se torna mais difícil).

Referências

CHAKRABARTI, S. *Mining the Web: Discovering Knowledge from Hypertext Data*. Amsterdam: Morgan Kaufmann, 2003. 3

HA, J.; PYO, H.; KIM, J. Large-scale item categorization in e-commerce using multiple recurrent neural networks. In: KRISHNAPURAM, B. et al. (Ed.). *KDD*. ACM, 2016. p. 107–115. ISBN 978-1-4503-4232-2. Disponível em: <http://dblp.uni-trier.de/db/conf/kdd/kdd2016.html#HaPK16>. 3

JOACHIMS, T. Text categorization with support vector machines: Learning with many relevant features. In: *Proceedings of the 10th European Conference on Machine Learning*. Berlin, Heidelberg: Springer-Verlag, 1998. (ECML'98), p. 137–142. ISBN 3-540-64417-2, 978-3-540-64417-0. Disponível em: <https://doi.org/10.1007/BFb0026683>. 3

MIKOLOV, T. et al. Distributed representations of words and phrases and their compositionality. In: *NIPS*. Curran Associates, Inc., 2013. p. 3111–3119. Disponível em: <http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf>. 3

RISH, I. An empirical study of the naive bayes classifier. In: IBM NEW YORK. *IJCAI 2001 workshop on empirical methods in artificial intelligence*. [S.l.], 2001. v. 3, n. 22, p. 41–46. 3

SHEN, D. et al. A study of smoothing algorithms for item categorization on e-commerce sites. *Neurocomputing*, v. 92, p. 54–60, 2012. Disponível em: <https://doi.org/10.1016/j.neucom.2011.08.035>. 4

ZAHAVY, T. et al. Is a picture worth a thousand words? A deep multi-modal fusion architecture for product classification in e-commerce. *CoRR*, abs/1611.09534, 2016. Disponível em: <http://arxiv.org/abs/1611.09534>. 3