I Encontro de Iniciação Científica e Tecnológica
I EnICT
IFSP – Câmpus Araraquara
21 de Outubro de 2016

# UM BANCO DE *WORD EMBEDDINGS* PARA O PORTUGUÊS BRASILEIRO

DOUGLAS EDUARDO MODENA DOS SANTOS[1], MARCELO CRISCUOLO[2]

[1] Graduando em Análise e Desenvolvimento de Sistemas, IFSP Campus Araraquara, d.modena@live.com.
[1] Doutorando em Ciências de Computação e Matemática Computacional, ICMC-USP, São Carlos, mcrisc@icmc.usp.br.

**Área de conhecimento** (Tabela CNPq): Metodologia e Técnicas de Computação – 1.03.03.00-6

**RESUMO:** **Definição do Problema:** Em linguagens naturais, uma palavra possui significado além de um simples conjunto de letras organizados em certa ordem. Uma série de conceitos e alusões existem por trás de cada palavra na mente humana, porém o mesmo não é valido para sistemas computacionais, que reconhecem uma palavra simplesmente como um conjunto de caracteres. Modelos de língua que sejam adequados ao tratamento computacional são fundamentais para o sucesso dos métodos de processamento de línguas naturais. Nos últimos anos, a representação de palavras por meio de vetores tem gerado resultados bastante promissores. Tais vetores de palavras (*word embeddings*) são treinados por modelos neurais (Bengio et al., 2003; Mikolov et al., 2013) em grandes corpora por meio de técnicas de aprendizado de máquina não supervisionado. Esse processo de treinamento requer recursos computacionais e linguísticos. Sabe-se também que vetores treinados em diferentes domínios tendem a apresentar propriedades diferentes (Lai et al., 2016). **Objetivo:** Pretende-se criar um banco de vetores treinados em corpora do português brasileiro de pelo menos três domínios: conhecimento geral, texto jornalístico e domínio agropecuário. **Justificativa:** O banco de vetores será disponibilizado publicamente para uso em pesquisas e desenvolvimento na área de Processamento de Linguagens Naturais (PLN), poupando tempo e recursos que seriam necessários para sua geração. **Metodologia:** Diferentes corpora serão utilizados no treinamento, como o Wikipédia, corpus CETENFolha e textos públicos coletados da Internet. Para a geração dos vetores, serão utilizadas as ferramentas Word2Vec e GloVe. O banco de vetores será avaliado extrinsecamente por meio de tarefas de classificação textual, implementada com uso de ferramentas como Scikit-Learn e TensorFlow. Cada versão (domínio) do banco de vetores será avaliada separadamente e serão coletadas as métricas de precisão, cobertura e F1, referentes à tarefa de classificação. Os resultados dessas avaliações e outros metadados serão publicados juntamente com o banco de vetores. Este é um trabalho em desenvolvimento. A etapa de revisão bibliográfica encontra-se em estágio avançado, as ferramentas mencionadas já são conhecidas e os dados do corpus Wikipédia já foram coletados. Espera-se que o sucesso desse projeto beneficie outros trabalhos, dada a importância atual dos vetores de palavras para o PLN.

## REFERÊNCIAS BIBLIOGRÁFICAS

BENGIO, Yoshua et al. A neural probabilistic language model. **journal of machine learning research**, v. 3, n. Feb, p. 1137-1155, 2003.

GOTH, Gregory. Deep or shallow, NLP is breaking out. **Communications of the ACM**, v. 59, n. 3, p. 13-16, 2016.

LAI, Siwei et al. “How to Generate a Good Word Embedding?”. **IEEE Intelligent Systems**, v. PP, n.99, pp. 1-1, 2016.

MIKOLOV, Tomas et al. Efficient estimation of word representations in vector space. **arXiv preprint arXiv:1301.3781**, 2013.

PENNINGTON, Jeffrey; SOCHER, Richard; MANNING, Christopher D. Glove: Global Vectors for Word Representation. In: **EMNLP**. 2014. p. 1532-43.

VIEIRA, Renata; LIMA, Vera LS. Lingüística computacional: princípios e aplicações. In: **Anais do XXI Congresso da SBC. I Jornada de Atualização em Inteligência Artificial**. sn, 2001. p. 47-86.