Bank Word Embeddings ptBR

Victor Takashi Hayashi

O presente documento destina-se à apresentação do Trabalho Final realizado pelo aluno Victor Takashi Hayashi para a disciplina PCS5029 (IMPLEMENTAÇÃO+RELATÓRIO).
O código fonte está disponível no GitHub: <https://github.com/vthayashi/BWE>.
O glossário financeiro do banco Central foi utilizado como corpus: <https://www.bcb.gov.br/content/cidadaniafinanceira/documentos_cidadania/biblioteca/glossario_cidadania_financeira.pdf>

A tarefa de Processamento de Linguagem Natural escolhida foi o cálculo de similaridade entre palavras do domínio específico bancário, e foi utilizada a arquitetura Continuous Bag Of Words como solução. O problema considerado foi a estimativa de similaridade semântica entre palavras de um domínio específico, o que pode ser interessante para processos de obtenção de conhecimento a partir de textos em linguagem natural (e.g., estruturando o conhecimento presente nestes dados não-estruturados). O corpus utilizado foi o glossário simplificado de termos financeiros do Banco Central (Bacen) (BACEN, 2020). A arquitetura Continuous Bag Of Words (CBOW) realiza a predição de uma palavra baseada nas palavras de seu contexto (MIKOLOV et al., 2013). A biblioteca Python gensim (REHUREK, 2020) foi utilizada para o treinamento da rede neural do CBOW. A biblioteca NLTK (NLTK, 2020) com suporte para português brasileiro foi utilizada como ferramenta auxiliar para o preprocessamento do texto de entrada, em conjunto com rotinas encontradas em trabalho de word embeddings da literatura (HARTMANN et al., 2017). Foram comparados dois word embeddings de escopo geral para o português, de modelos CBOW e Glove 50 dimensões com o modelo específico desenvolvido CBOW com 100 dimensões na tarefa de similaridade do cosseno entre os *word embeddings* da palavra 'banco' e outras relacionadas ao domínio bancário. Foram obtidos 10980 tokens e 1462 termos no vocabulário a partir do glossário do Banco Central, o que é um tamanho equivalente ao menor corpus (SARESP) utilizado em trabalho relacionado (HARTMANN et al., 2017). A Figura 1 é uma visualização 2D (obtida com Principal Component Analysis) dos word embeddings obtidos com o treinamento específico com o glossário do banco central. Os resultados estão sumarizados na Tabela 1. A similaridade de 5 palavras com a palavra 'banco' dentre 9 foram maiores com o word embedding específico, enquanto o modelo Glove do escopo geral apresentou melhores resultados de similaridade que o CBOW geral. O experimento indica que o treinamento específico tem potencial de melhorar os resultados do modelo CBOW para cálculo de similaridade entre algumas palavras do domínio específico bancário.

Palavras-chave: Word Embedding. CBOW. Domínio Bancário.

Palavras	CBOW geral	Glove	CBOW específico	espec.>geral
valor	0.571603	0.665027	0.695323	Sim
juros	0.248311	0.629176	0.637589	Sim
pagamento	0.522324	0.652546	0.593549	Não
poupança	0.220006	0.543432	0.632208	Sim
dinheiro	0.410055	0.621486	0.623654	Sim
cliente	0.472513	0.622152	0.564684	Não
comprar	0.014624	0.526575	0.580381	Sim
crédito	0.482977	0.811110	0.553456	Não
serviço	0.621220	0.625766	0.458480	Não

Tabela 1: Resultados de similaridade entre 'banco' e outras palavras.

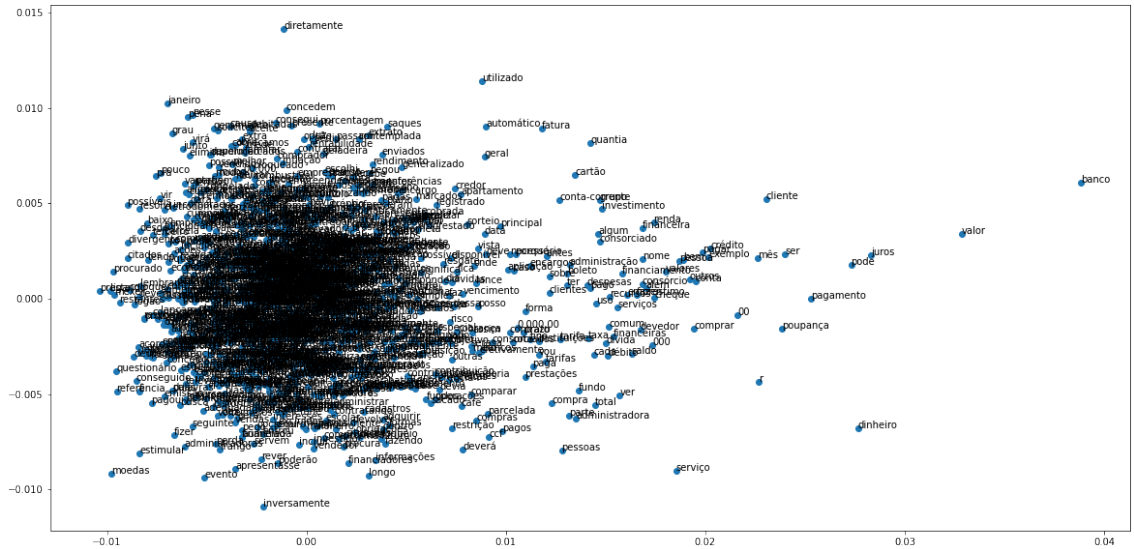

Figura 1: Visualização 2D obtida com PCA dos word embeddings obtidos.

Referências

BACEN. *Glossário Simplificado de Termos Financeiros do Banco Central (Bacen)*. 2020. <https://www.bcb.gov.br/content/cidadaniafinanceira/documentos_cidadania/biblioteca/glossario_cidadania_financeira.pdf>. Acessado em: 02/12/2020.

HARTMANN, N. et al. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. *arXiv preprint arXiv:1708.06025*, 2017.

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

NLTK. *Biblioteca Python NLTK*. 2020. <http://www.nltk.org/howto/portuguese_en.html>. Acessado em: 02/12/2020.

REHUREK, R. *Biblioteca Python Gensim*. 2020. <https://radimrehurek.com/gensim/models/word2vec.html#gensim.models.word2vec.Word2Vec>. Acessado em: 02/12/2020.