Universidade Federal Fluminense
Pós-Graduação em Ciência da Computação
Disciplina: Tópicos Avançados em Aprendizado de Máquina (2018.2)
Aluno: Henrique Bueno Rodrigues (D022.118.009)

Avaliando técnicas de "*Word Embedding*" em apresentações do TED

1. Introdução

"*Word embedding*" é o nome de um conjunto de técnicas em Processamento de Linguagem Natural (PLN) onde palavras de um vocabulário são mapeadas em vetores de números reais.

Em linguística, "*word embedding*" é estudado na área conhecida como Semântica Distribucional que tenta categorizar e quantificar similaridades entre as palavras considerando a distribuição de ocorrências das palavras em grandes conjuntos de dados.

Existem diferentes estratégias para essas técnicas, por exemplo, modelos baseados em predição e outros baseados em contagem [5]. Algumas das técnicas mais populares atualmente são Word2Vec [2], GloVe [3] e FastText [4].

O algoritmo Word2Vec é o mais popular para a computação de *embeddings*. Ele representa cada palavra como um vetor onde as relações entre as palavras podem ser obtidas a partir de operações entre esses vetores.

A técnica GloVe, que é baseada em contagem, trata cada palavra de um corpus como uma entidade atômica e gera um vetor para cada palavra. Nesse sentido, GloVe é parecido com Word2Vec, uma vez que ambos tratam palavras como a menor unidade para treinar.

Por fim, a técnica FastText trata cada palavra como uma composição de ngrams. Assim, o vetor para uma palavra é feito a partir da soma desses ngrams.

2. Proposta

A proposta desse trabalho é aplicar as técnicas de "*Word Embedding*" Word2Vec, GloVe e FastText sobre uma base de dados de apresentações do TED [1] e realizar análises comparativas das diferentes estratégias através da execução de operações entre os vetores que representam as palavras. A base de dados que será utilizada possui 2550 registros e 17 atributos.

Segue a lista de atividades previstas:

- Setembro: análise do *dataset* de apresentações do TED.
- Outubro: estudo e aplicação da técnica Word2Vec.
- Novembro: estudo e aplicação das técnicas GloVe e FastText.
- Dezembro: elaboração de relatório e apresentação final.

3. Referências

[1] https://www.kaggle.com/rounakbanik/ted-talks - Acessado em 11/09/2018.

[2] Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781* (2013).

[3] Pennington, Jeffrey, Richard Socher, and Christopher Manning. "Glove: Global vectors for word representation." *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014.

[4] Bojanowski, Piotr, et al. "Enriching word vectors with subword information." *arXiv preprint arXiv:1607.04606* (2016).

[5] Baroni, Marco, Georgiana Dinu, and Germán Kruszewski. "Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors." *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vol. 1. 2014.