# Word Embeddings

Advanced Topics in Data Mining and Machine Learning | Jad Habouch | January 31, 2018

## 1. READINGS

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *CoRR*, abs/1301.3781: https://arxiv.org/pdf/1301.3781.pdf

Pennington, J., Socher, R. & Manning, C. D. (2014). Glove: Global Vectors for Word Representation. EMNLP (p./pp. 1532--1543), https://nlp.stanford.edu/pubs/glove.pdf

## 2. ABSTRACT

The complexity of biological and biomedical data, along with the wide range of use cases, questions, and data types, are compelling researchers to pursue machine learning and deep learning methods that were originally developed for other domains such as Linguistics and Natural Language Processing. Word Embeddings has the goal of quantifying and categorizing semantic resemblances between language elements. The basic premise started with a notion introduced by Firth, a British Linguist, that "a word is characterized by the company it keeps". The approach of characterizing words as vectors can be traced back to the development of vector space model for information retrieval. In this topic, we will introduce the latest in Word Embeddings from two prominent publications. The first paper we will explore is Word2Vec: Efficient Estimation of Word Representations in Vector Space by Mikolov et al. And the second paper is GloVe: Global Vectors for Word Representation by Pennington et.al. Lastly, we will discuss potential biological applications and cases where Word Embeddings methods can be applied.

### Word2Vec: Efficient Estimation of Word Representations in Vector Space

Word2Vec surveys the previous work conducted in word representation as a benchmark for the authors' effort to improve the vector operations and yet maintain the linear regularities among words. The paper demonstrates the heightened focus on building two new model architectures that maximize the accuracy and efficiency of learning high-quality word vectors from vast data sets with "billions of words, and with millions of words in the vocabulary". In this pursuit, the authors develop a comprehensive test set for measuring both syntactic and semantic regularities. Furthermore, the paper shows that various regularities can be learned with a high of accuracy. Additionally, Word2Vec shows that training time and accuracy are dependent on the dimensionality of the word vectors along with the quantity of the training data. The key output and contribution of this body of work is the development of model architectures that proved to surpass the top performing techniques developed previously, based on different types of neural networks. Word2Vec concludes that its techniques have resulted in the best performance for measuring syntactic and semantic word similarities.

### GloVe: Global Vectors for Word Representation

GloVe examines the recent question that has surfaced to inspect whether distributional word representations are best learned from count-based methods or from prediction-based methods. While, prediction-based models have earned more support; GloVe aims to show that the two classes are fundamentally comparable since they both utilize co-occurrence statistics of the corpus. However, the count-based methods offer superior efficiency when capturing global statistics. GloVe developed a model that has leveraged the benefit of count data while concurrently preserving the essential linear substructures that prevail in log-bilinear prediction-based methods such as Word2Vec. "GloVe purportedly "*is a new global log-bilinear regression model for the unsupervised learning of word representations that outperforms other models on word analogy, word similarity, and named entity recognition tasks*".

3. **SPOTLIGHT QUESTION:** What advantages or disadvantages do you see when applying Word Embeddings models to questions and applications outside the linguistics field? Particularly in biological and biomedical domains.