Joe Puhalla

<https://arxiv.org/pdf/1301.3781.pdf>

**Efficient Estimation of Word Representations in Vector Space**, Tomas Mikolov, 7 Sep 2013

This paper proposes two unique approaches to represent words in vector form. The proposed methodology of word2vec with two possible approaches, bag of words and skip gram model. Word2vec is a group of related models that are used to produce word embeddings. These models are shallow, two-layer neural networks that are trained to reconstruct linguistic contexts of words. Its input is a text corpus and its output is a set of vectors: feature vectors that represent words in that corpus. In the continuous bag of words approach, the model predicts the current word from a group of surrounding context words. The skip gram architecture gives a heavier weight to nearby context words rather than distant ones. The quality of word embedding increases with increasing dimension up to a point where increases become marginal. Patterns such as “Man is to Woman as Brother is to Sister” can be generated through algebraic operations on the vector representations of these words such that the vector representation of “Brother” - ”Man” + ”Woman” produces a result which is closest to the vector representation of “Sister” in the model.

<https://arxiv.org/pdf/1809.08390.pdf>

**Constructing Financial Sentimental Factors in Chinese Market Using Natural Language Processing**, Junfeng Jiang, Jiahao Li, Likelihood Technology, Sun Yat-sen University, 22 Sep 2018

This paper proposes an algorithm to identify the sentiment within the Chinese market, and then find a correlation between that sentiment and the performance of the market. The paper uses methodology including tokenization and word2vec, as presented in the above paper, in order to determine sentiment for the market. The paper shows there is significant correlation between sentiment and the market, especially during the crash of 2015. The methodology proposed using the word2vec approach combined with a wordnet and each word will have 200 dimensions. Words with a positive meaning are given a positive value between 0 and 1 and words with a negative meaning are given a negative value between 0 and -1. These sentiment scores are then summed up to give a sentiment score for an entire news headline. Furthermore, the model is adjusted so that headlines do not only affect sentiment values for that day, but also for the following days since large headlines are known to have lingering effects. Lastly a correlation coefficient was used to determine the relationship between the sentiment of headlines and the market movement.

Alternative Papers:

<https://arxiv.org/pdf/1811.06173.pdf>

**Leveraging Financial News for Stock Trend Prediction with Attention-Based Recurrent Neural Network**, Huicheng Liu, Department of Electrical and Computer Engineering, Queen’s University, Canada, 15 Nov. 2018

<https://arxiv.org/pdf/1902.08684.pdf>

**Discovering Language of the Stocks,** Marko Poženel and Dejan Lavbič,13 Feb. 2019