Keeyon Ebrahimi
NLP
Final Project

Correlation Station

Demo: Go to correlationstation.me on your web browser

Type 1 word in the search bar and click High or Low Variance
This Project explanation will make more sense after playing with the demo

Summary

Wanted to study the correlation between words in News Articles. Wanted to be able to search any word and find the most interesting related news topics. Studied the long term and short term changes in the co-occurrence of different words.

Technique being used here is word embedding. On a high level, if we have these two sentences

Angry dogs bite

Angry cats bite

We see that we have a structure of **Angry *blank* bite** . Cat and dog have both filled this structure and are made to be similar.

Data

Scraped **reuters.com** from 2008 - 2014. Total of 370k articles that span global, economic, and market news mostly.

Word Embedding

Used word embedding to map words to a 300 dimension space. Maps similar words close together while semantically different words are far apart. This is a very simple model, but it needs a large amount of data to train on. Word embedding can also be used for POS tagging, semantic analysis, NERs.

Trained on news articles that span from 2008 - 2014, so now we can study trends over seven years

Visualizing Data

Steps:

1. Built 7 different word2vec models for each of the 7 years.
2. Grabbed the top 10 similar words for a given word from each of the 7 models.
3. Calculated similarity with Cosine Similarity.
4. For each unique word in the grabbed 10 most similar words for each year, found the similarity of the original searched word with unique word in each year.
5. Found the 4 words that had the most amount of variance with similarity of each year and also the least amount of variance with the 7 years of similarity data.
6. Visualized data with some very serious D3 visualization, which again, can be seen at **correlationstation.me**

Technologies Used

Beautiful Soup (Web Scraping)

Gensim/NumPy/SciPy (Word Embedding)

Flask (Webpage backend)

D3 (Webpage frontend)

JQuery/Ajax (Webage communication between backend and frontend)

References

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop at ICLR, 2013.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. In Proceedings of NIPS, 2013.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic Regularities in Continuous Space Word Representations. In Proceedings of NAACL HLT, 2013.