Nikhil Chakravarthy

Problem 3 – Word2Vec

Evaluate the efficacy of turning words into sparse vectors for the purposes of Deep Learning.

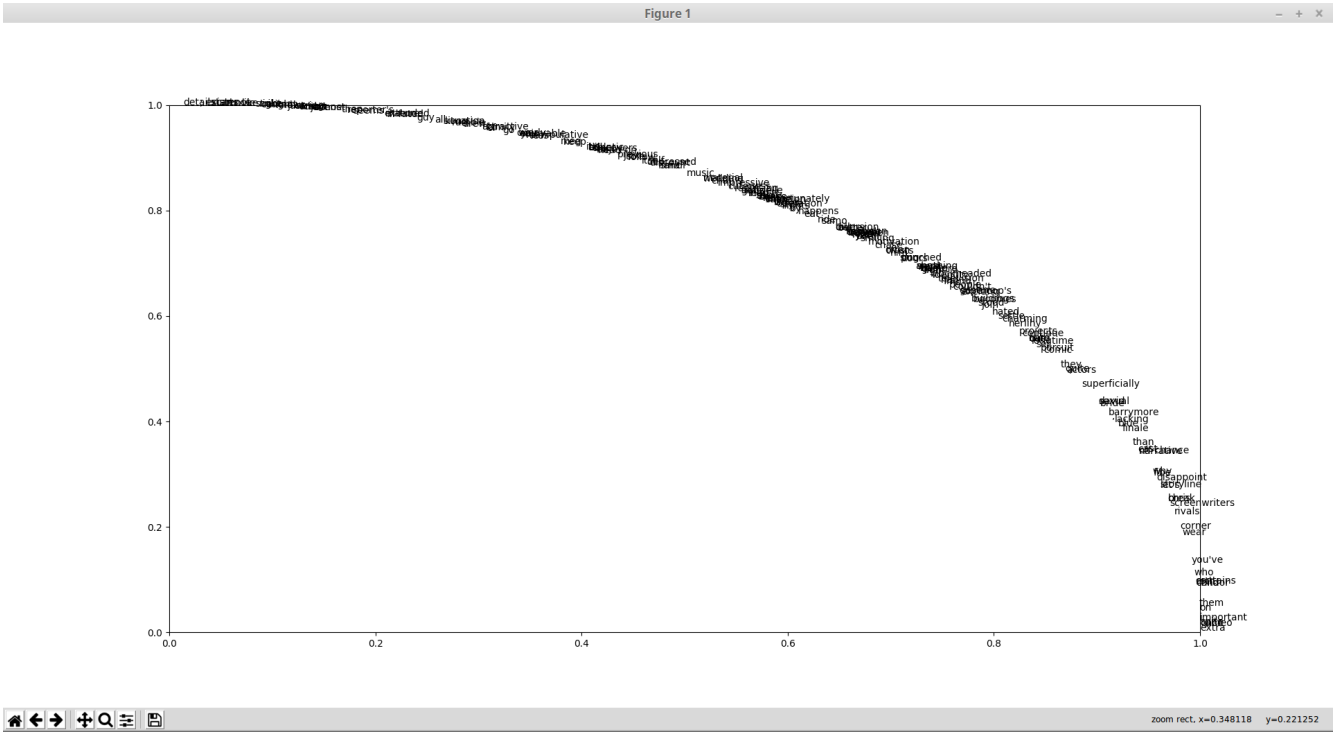
Problem description:

I have taken data from movie reviews (look at end of report for more information), and I have taken a few positive reviews and turned the words of the reviews into vectors, and tried to figure out if the algorithm correctly mapped related words close together. This is done by looking at a graph of words, where the dimensionality has been reduced to 2 dimensions for easy visualization, and seeing if words of similar relation are next to each other.

Experiment:

Movie reviews were taken off the web from (http://www.cs.cornell.edu/people/pabo/movie-review-data/), and a small subsection of 3 positive reviews are taken and used for the experiment (only 3 because of run time issues). Then, sparse vectors are created from the words in order to represent each word as a vector, and a

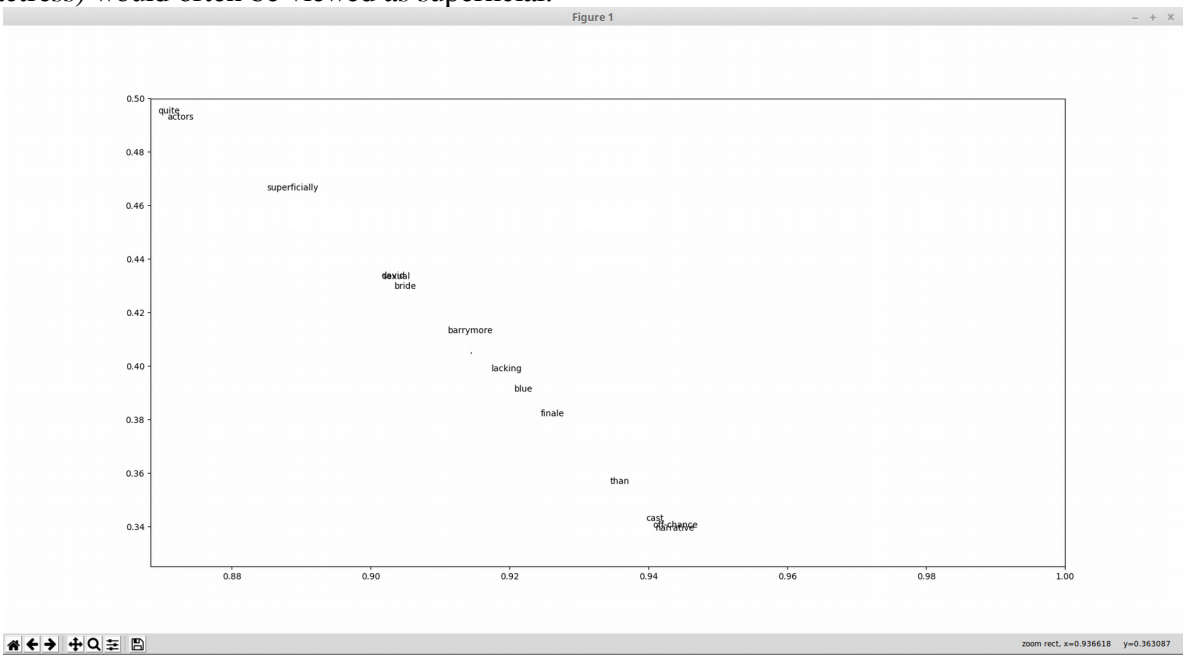
Results (found in prob3/ as images):



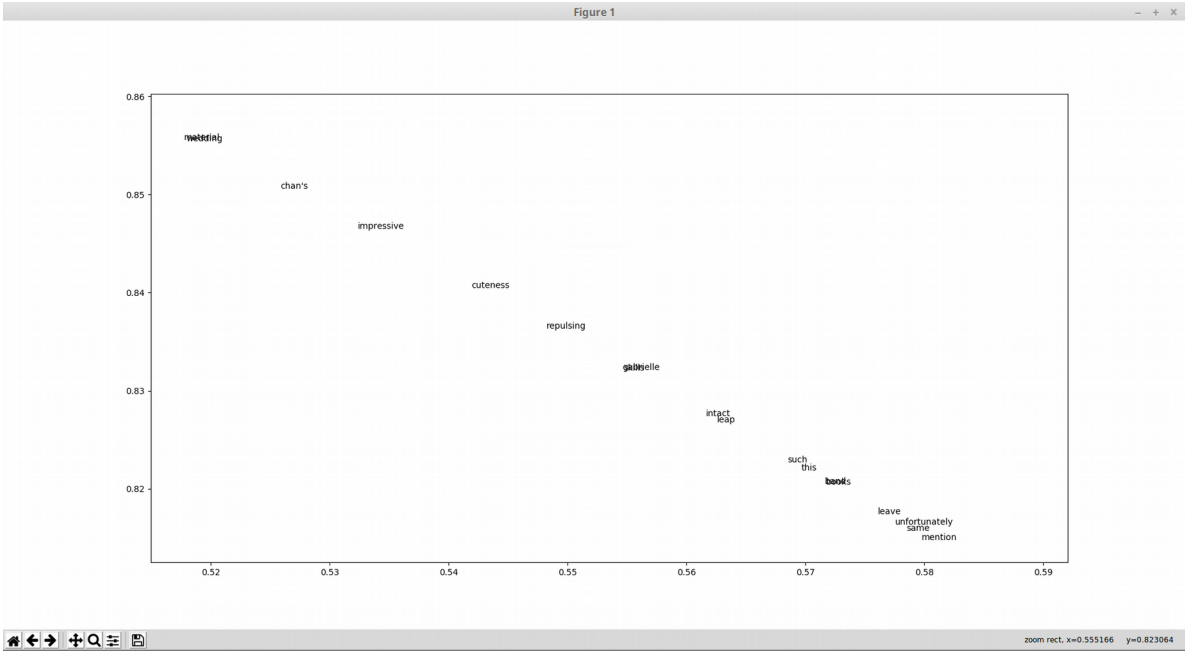
Above is the resulting 2-D word map that my Word2Vec algorithm generated. However, it is very hard to read. I zoomed into a few sections where I found interesting relations between words that I feel show that

my algorithm did find some patterns. Note that these views are extremely viewed in, so almost all words inside these views are closely related, according to the algorithm.

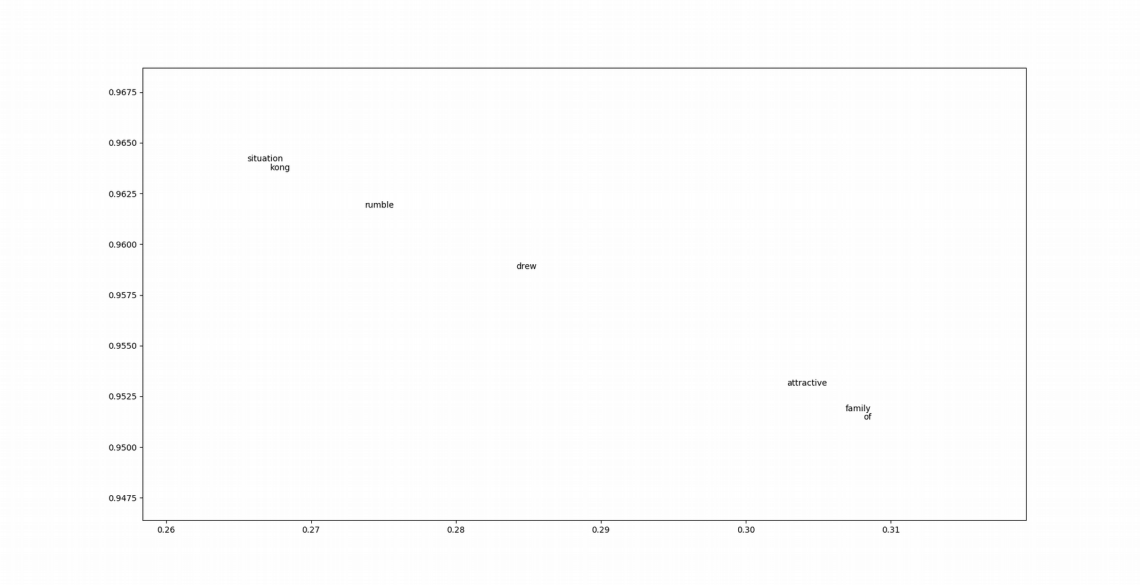
Barrymore and superficial: I found that the name “Barrymore” and “superficial” were close in distance, which is both hilarious and also makes sense, since movies with “Barrymore” (I believe a female rom-com type actress) would often be viewed as superficial.



Here is a view where opposing words “cuteness” and “repulsing” are close to each other. This again makes sense, as they would be used to describe similar things but in opposite directions.



Here is another set of related words with “kong” and “rumble”. King Kong would likely cause a rumble, relating these two words again.



Overall, I believe that much more data is needed for this algorithm to generate a more accurate map for some domain like movie reviews. I did not run more data because of time constraints, as I mentioned before. However, with some relations being found and correctly mapped out, I can reasonably say that word2vec can help generate some relations and add some order into text data.

This data was first used in Bo Pang and Lillian Lee,
``A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization
Based on Minimum Cuts'', Proceedings of the ACL, 2004.