Progress Fall 2015

January 26, 2016

Done by Linan Qiu github.com/linanqiu for Apoorv Agarwal and Owen Rambow in partial satisfaction of the requirements of the COMS W3998 Undergraduate Research Course.

As opposed to the research log, this is more about what I learned personally and the various detours I took during this course.

1 Codeword Detection

The codeword detection project was the mainstay of this semester.

1.1 Replicating Previous Work and Word Similarity Model

1. (1 Week) I took over the codeword detection project, replicating the word done by Mariyam
2. (1 Week) A naive comparison of cosine distances between words proved to be too noisy to be feasible. Started thinking of alternatives to the cosine distance comparison method. At the same time, talked to Karl Stratos regarding an alternative to Word2Vec.
3. (1 Week) Turns out a simple comparison of word similarity (which theoretically is similar to cosine distance, but is more noise tolerant) was able to reveal codewords (with high false positives) on the Enron corpus.

At this point, Apoorv and I realized that a synthetic dataset need to be created for efficient further testing, since actually testing codewords from the Enron dataset required a very large reference corpus (so that the words like `raptor` appeared frequently enough).

1.2 Synthetic Dataset Creation

1. (1 Week) Successfully parsed the WSJ corpus using undocumented code in NLTK. Created a synthetic dataset using the TF-IDF method documented in the research log. Found this to produce rather nonsensical words.
2. (1 Week) Improved on the TF-IDF method by using a Gamma distribution to sample words from the dataset (effectively awarding words that have high TF-IDF, but penalizing words that appear way too infrequently). From the WSJ dataset, we were able to obtain words such as `yen`, `credit`, `stake` etc.
3. (1 Week) Wrote a routine to properly perform substitution, run Word2Vec on an original and substituted corpus, save the results accordingly and retrieve them systematically.
4. (1 Week) Performed analysis on the results of the synthetic codeword dataset, showing high accuracy in codeword detection.
5. (1 Week) Retrieved matching between codewords and original words to a high degree.
6. (1 Week) Compiled results into the research log.

2 Literature

I have not had time last semester to digest much of the literature that Apoorv has sent me. This is largely my fault – I have been overly focused on the engineering aspects of this project and neglected the theoretical portions. This will be a focus of this semester.

3 GPU

1. (1 Week) Along the way, I tried to use the GPU accelerated version of Word2Vec. Turns out the ones that were celebrated produced nonsensical results. This wasted a week of time.
2. (Currently low priority) Reproduce Word2Vec using TensorFlow.

4 Challenges

- Python: I was new to Python, and that resulted in a rather slow pace of progress in code. This has been largely resolved and I'm now more than comfortable in the language.
- Literature: I will need to read more papers this semester if I want to get this progress submitted to ACL2016.