CSC2611 Lab: Word embedding and semantic change

In this lab, you will explore the word embedding model *word2vec* and extend the analyses in the earlier exercise to both synchronic and diachronic settings. **Deliverable:** Submit a single PDF report (with your name on the first page) that addresses all the questions [15 points] and include link to the GitHub repository; the repository should include an executable Python or Jupyter Notebook file that replicates all of the findings that you report in the PDF write-up [5 points].

1 Synchronic word embedding [7 points]

Step 1. Download the pre-trained word2vec embeddings from https://code.google.com/archive/p/word2vec/, specifically, the file "GoogleNews-vectors-negative300.bin.gz".

Step 2. Using gensim, extract embeddings of words in Table 1 of RG65 that also appeared in the set W from the earlier exericse, i.e., the pairs of words should be identical in all analyses. An example use of `gensim` is provided below.

```
> from gensim.models import KeyedVectors
> model = KeyedVectors.load_word2vec_format('GoogleNews-vectors-negative300.bin',
binary=True)
> model['dog']
```

Step 3. Calculate cosine distance between each pair of word embeddings you have extracted, and report the Pearson correlation between word2vec-based and human similarities. [1 point] Comment on this value in comparison to those from LSA and word-context vectors from analyses in the earlier exercise. [1 point]

Step 4. Perform the analogy test based on data here (or as provided) with the pre-trained word2vec embeddings. Report the accuracy on the semantic analogy test and the syntactic analogy test (see *Note* below). [2 points] Repeat the analysis with LSA vectors (300 dimensions) from the earlier exercise, and commment on the results in comparison to those from word2vec. [1 point] *Note*: It is expected that the number of entries you could test with LSA would be smaller than that based on word2vec. For a fair comparison, you should consider reporting model accuracies based on the small test set, for both word2vec and LSA.

Step 5. Suggest a way to improve the existing set of vector-based models in capturing word similarities in general, and provide justifications for your suggestion. [2 points]

2 Diachronic word embedding [8 points]

Step 1. Download the diachronic word2vec embeddings from the course syllabus page. These embeddings capture historical usage of a small subset of English words over the past century.

Step 2. Propose three different methods for measuring degree of semantic change for individual words and report the top 20 most and least changing words in table(s) from each measure. Measure the intercorrelations (of semantic change in all words, given the embeddings from Step 1) among the three methods you have proposed and summarize the Pearson correlations in a 3-by-3 table. [3 points]

Step 3. Propose and justify a procedure for evaluating the accuracy of the methods you have proposed in Step 2, and then evaluate the three methods following this proposed procedure and report Pearson correlations or relevant test statistics. [2 points]

Step 4. Extract the top 3 changing words using the best method from Steps 2 and 3. Propose and implement a simple way of detecting the point(s) of semantic change in each word based on its diachronic embedding time course—visualize the time course and the detected change point(s). [3 points]