CSC2611 Exercise: Meaning construction from text

This exercise serves as a self-assessment of the course basics and is relevant to the next lab. It should take you no more than 4 hours to complete (excluding data transcription).

Step 1. Import `NLTK` in Python: `http://www.nltk.org/`. Download the Brown Corpus `http://www.nltk.org/book/ch02.html` for analyses below.

Step 2. Extract the 5000 most common English words (denoted by W) based on unigram frequencies in the Brown corpus. Report the 5 most and least common words you have found in the 5000 words. Update W by adding n words where n is the set of words in Table 1 of RG65 that were not included in the top 5000 words from the Brown corpus. Denote the total number of words in W as $|W|$.

Step 3. Construct a word-context vector model (denoted by $M1$) by collecting bigram counts for words in W. The output should be a $|W| \times |W|$ matrix (consider using `sparse matrices` for better efficiency), where each row is a word in W, and each column is a context in W that precedes row words in sentences. For example, if the phrase *taxi driver* appears 5 times in the entire corpus, then row *taxi* and column *driver* should have a value of 5 in the matrix.

Step 4. Compute *positive pointwise mutual information* on $M1$. Denote this model as $M1+$.

Step 5. Construct a latent semantic model (denoted by $M2$) by applying principal components analysis to $M1+$. The output should return 3 matrices, with different truncated dimenions at 10 (or a $|W| \times 10$ matrix, denoted by $M2_{10}$), 100 ($M2_{100}$), and 300 ($M2_{300}$).

Step 6. Find all pairs of words in Table 1 of RG65 that are also available in W. Denote these pairs as P. Record the human-judged similarities of these word pairs from the table and denote similarity values as S.

Step 7. Perform the following calculations on each of these models $M1$, $M1+$, $M2_{10}$, $M2_{100}$, $M2_{300}$, separately: Calculate cosine similarity between each pair of words in P, based on the constructed word vectors. Record model-predicted similarities: S_{M1}, $S_{M2_{10}}$, $S_{M2_{100}}$, $S_{M2_{300}}$.

Step 8. Report Pearson correlation between S and each of the model-predicted similarities. Create a GitHub repository that implements all of your analyses; you will need this repo for the next lab.