Grading Method, Fall 2017

Homework 1: Corpus Statistics

1. (10 pts) Choose or Collect Appropriate Data: the two documents should be sufficiently different to yield good questions and of sufficient length for the word frequency and bigram lists to be useful.

2. (30 pts) Process each document and produce the frequency, bigram frequency and bigram PMI score lists, with processing steps chosen to produce lists suitable for analysis of your question.
a. (10 pts) Description of processing steps: tokenization, lower case, stopwords or lemmatization, word frequencies, bigram frequencies and bigram PMI with frequency filter of 5 or greater, and state why you chose those options.
b. (10 pts) Discuss any issues with the lists and describe how the bigrams scored by frequency are different that the bigrams scored by PMI.

3a. (10 pts) Define a comparison question between the two documents.
b. (20 pts) Answer the question by picking examples from the lists and discussing how they show that the documents are different, not just reporting numbers. Discussion may include collection steps if significant, which will count towards discussion of differences.

Additional merit (10 pts): Choose some aspect of the processing or analysis that requires additional thought or work. Some options are

- If you collect your own data, describe that work or steps necessary to obtain documents ready for processing (from part 1)
- During processing, describe additional steps, for example, if you define or modify stopword lists to suit your documents or analysis question (from part 2a)
- Make trigram lists and include in your discussion
- Expand the question or analysis (from part 3).

Interpretation of numeric grades as letter grades:

90 – 100	A
85 – 89.9	A-
80 – 84.9	B+
75 – 79.9	B
70 – 74.9	B-

below 70 has similar interpretation in the C and lower range.

Late assignment submissions will be accepted, but will be penalized:
1 Week late- 7 points taken off (1/2 letter grade)
More than 1 week late- 14 points taken off (1 letter grade)