Instruction: Please find your own text documents for this assignment. Text documents should be related your interests. Some of students will be randomly selected to present your work, so please prepare your assignment in preparation for presentation. You should submit both hard and soft copies until the due date. Late submission will not be accepted.

Due date: March 12th, 2018

Assignment I: Vector Space Model

Try to understand the basic idea of pre-processing, construct vector space representation for specific text documents, and compute similarity based on TF-IDF weighting among different text documents.

1. **Do Pre-Processing**

   1.1 Tokenization: tokenize each document into tokens.

   1.1. Normalization: normalize the tokens from step 1.

   1.2. Stemming: stem the tokens back to their root form.

   1.3. Stopword removal: remove stopwords.

   Please define your list of stopwords depending on context.

2. **Understand Zipf's Law**

   2.1. Show your implementation of text normalization module.

   2.2. Generate a curve for Zipf's law.

   Please specify your cutoffs.

3. **Construct a Vocabulary**

   3.1. List the resulting vocabulary.

   3.2. List top 50 and bottom 50 words according to TF in the resulting vocabulary, and represent their corresponding IDFs.

   Please define your TF.

4. **Compute Similarity between documents**

   4.1. Paste your implementation of similarity computation.

   4.2. For each document, list the top 3 most similar documents and the corresponding similarity.

Please define your similarity.