CSCI 345: Information Storage and Retrieval — Assigned: Feb. 10
Project 1: Index Construction — Due: Feb. 28

This project deals with building an inverted index for a given collection of $2,916$ emails. Each email was modeled as a "Bag of Words": a set of words present in the email along with its frequency of occurrence. Some neutral words such as 'the', 'a', 'an' have been removed. The compressed archive consists of $2,916$ files. Each of these files consists of a list of words (one per line) along with their frequency count. The compressed archive is available at `http://www.cs.olemiss.edu/∼ychen/data/csci345_pj1.zip`.

You need to:

1. Build a dictionary using all the documents. The terms should be sorted in alphabatical order and stored in a text file (one term per line).
2. Construct postings for each term in the dictionary.
3. In practice, the dictionary is commonly kept in memory, with pointers to each posting list, which is stored on disk. In this project, you may store each posting list as a text file.
4. Include in your postings files the term frequency along with each docID.
5. Include in your dictionary the number of documents for each term.
6. Write a function/method named `getpostings()`, which can return the postings for a given input term.
7. Prepare a report. The report should include a cover page, a detailed description of data preprocessing, data structure, and the format of posting files (if needed). Include in your report the first 50 DocID's for the following queries:

   (a) *file*

   (b) *performance*

   (c) *read*

   (d) *window*

   (e) *subject*

   Please **DO NOT** include a hard copy of the source code in the report. **The report is due Friday, February 28.** You should **submit your source code files along with your report via Blackboard system.**