

## Homework #1 (due in 1 week)

### **Administrata (5 points)**

As soon as you can, please send me an email identifying yourself and telling me which email address you would like class-related mailings sent to. Any account is fine, but it should be one you can check regularly.

### **Corpus Statistics (65 points)**

Zipf's law (cf. Section 5.1 in the text) predicts that the number of times the  $i$ th most frequent word will be seen is about  $k/i$  times the frequency of the most common word, for *some*  $k$ . You can verify whether this is so by examining a collection of text and counting the number of occurrences for each word. For this assignment, two electronic texts have been placed on the course web page: the *Bible* (American Standard Version) and Jane Austen's *Sense and Sensibility*. Download these files, and write a program (or programs) to run on the texts individually (i.e., run separately on each text). The program should:

- Performs some normalization of the text. For example, removing punctuation and lower-casing words.
- Report the number of 'paragraphs' processed, the number of unique words observed, and the total number of words encountered. Be sure to describe how you determine what constitutes a word. For example, you should indicate if you ignore case-differences and how you treat punctuation symbols and digits.
- Keep track of both the total number of times each word is seen and the number of documents (paragraphs = documents) which the word occurs in.
- Identify the 30 most frequent words (by total count) and report the number of times each occurs (both in total and expressed as the number of documents it appears in).
- Also print the 100<sup>th</sup>, 500<sup>th</sup>, and 1000<sup>th</sup> most-frequent words and their frequencies of occurrence. (But please do not turn in a printout with the top 1000.)
- Compute and prints the *number* of words that occur in exactly one document. (For *Sense*, I believe *surplice* and *simpering* are such words.) What percentage of the dictionary terms occur in just one document?

You can (and should) build the required lexicon in one scan of the input text. After sorting terms by total frequency it should be fairly easy to extract the pieces of information that I am asking you to report. In the files paragraphs are indicated with <P ID=XXXX> tags indicating the start of each new paragraph. Some 'paragraphs' are short, some are longer; I think that none are empty, but I haven't verified this.

Lexicons are a key data structure for IR systems. In future assignments you will need a lexicon, so you might consider how to make your dictionary modular and reusable. In particular you will want it to fit in memory, to be storable on disk for subsequent reloading, and to enable efficient lookups. Some representations you might consider are in-memory hashtables, binary trees, or tries. If you use Java, using the built-in HashMap or a TreeMap classes is a very reasonable idea. Hand in your source code and the requested output described above.

The Bible text was obtained from <http://unbound.biola.edu/>; versions in numerous languages are available. Austen's novel was obtained from Project Gutenberg: <http://www.gutenberg.org/>. I attempted to process an English version of the Koran (several are available from PG); however, I found it difficult to automatically segment verses from those texts, so I gave up. The Biblical text is just a bit under 5MB in size; *Sense and Sensibility* is about 1/7th of the size.

I have coded a solution to the exercise above in ~ 240 lines of (verbose, commented) Java code and I have had former students write good solutions in about 2 pages of (good) Perl.

### **Questions (10 points apiece)**

[1] Problem 1.7 (pg. 13)

[2] Problem 1.9 (pg. 13) [note: consider 'optimal' to mean it is not possible to respond to a query with less processing effort]

[3] Problem 2.5 (pg. 35)

