CS 240

Data Structures and Algorithms

Fall 2013

1 Lab 10 (Project 1)

There are two goals for this assignment:

1. Practice implementation of a program that has several requirements. (It may not be easy to satisfy all requirements in one solution.)

2. Work together to accomplish the task. This requires effective communication/understanding of technical content.

2 Grading

- Five percent of this lab will be following submission instructions for all submissions throughout this assignment.

- Ten percent of this lab assignment will be the result of your evaluation of each other as a team – evaluation forms will be available when the assignment is due.

- Ten percent of this lab assignment will be the result of your implementation's performance relative to other students in the class. This means your code will be tested in a series of automated tests. This will be done using input and output redirection.

  - Top team – 10 points
  - Team 2 – 9 points
  - Team 3 – 8 points

  - Team 4 – 7 points
  - Teams 5 and 6 – 6 points
  - Teams 7 and 8 – 5 points
  - Teams 9, 10, and 11 – 4 points
  - Teams 12, 13, and 14 – 3 points
  - Teams 15, 16, and 17 – 2 points
  - Teams 18+ – 1 point

- The remainder will be based upon successful completion of the assigned tasks.

Note that because this is being graded competatively, neither of the TAs/CAs/Professor will be able to give you any numbers regarding how quickly this program should run.

3 Assignment

The web is made up of countless documents that often contain a lot of text. Search engines work by processing these files and aggregating the information contained therein.

You and your partner will write a Wikipedia processing tool that will take in a file (with a simplified format) that contains the text of various Wikipedia articles. You have several tasks to perform on the data contained within this data set.

The format of the text file is as is described, and as is illustrated in the small example `Wiki_small.txt` that is provided. Each article is defined by its title. The title of an article will always appear immediately before the text of the article, on its own line. The title will be surrounded with tags:
`<TITLE>` and `<\TITLE>`
The title will also always be plain text, may or may not contain stop words, and is guaranteed to be unique. The title may also have punctuation, and neither the words of the title, nor the title tags themselves are to be considered in the reports you are to provide.

Since this data set will contain real-world articles, you will need to strip off the punctuation (non-alpha-numeric characters) from the beginning and ends of words. This includes, but is not limited to, ‘.’,‘,’,‘!’,‘”’, etc. This however does mean that contractions and hyphenations within words should remain intact.

Many words in the English language are used repeatedly (and usually offer very little meaning), these properties make these words less relevant to most article, and are usually left out of a web crawler's calculations. You will be provided with a list of stop words to consider in a file named `stop.txt`. This file will be listed in mixed case (some words capital, some words lowercase), but your analysis is to consider any formation (regardless of case) as long as it matches the whole stop word itself. For example if the file contains `the`, the stop words you will need to eliminate are `the`, `The`, `THe`, and `THE`. Note that while you will be provided with a file named `stop.txt`, you are not guaranteed that the current version of this file will be the one used for testing (this includes attributes like the ordering of the stop words and the number of stop words included).

Tasks:

1. You are to retrieve the most frequently used words of the data set. When considering only non-stop words, you are to report the top 5% of words within the data set, and the title of the article in which this each such word appears most often. Report these results to a file named `report.txt` where each line contains:
   `x, y, z`
   Where `x` is the word in question, `y` is the number of occurrences of that word, and `z` is the title of article in which that word appears most. Report these values in order of non-ascending number of appearances.

2. You are to compute the percentage of the words in the document that are stop words.

3. You are to print out all of the words within the data set alphabetically (using dictionary ordering). With each word, you are to provide the number of occurrences within the data set. Note, this includes reporting of the number of occurrences of the stop words.

4 Due Date

This assignment will be due on April 23, 2014 before the start of your lecture period, which is at 4:00pm. Note that because this is a project, you are not able to use a late for this assignment. A lateness penalty of 10% per 24-hour period following the due date will be strictly enforced, and the latest possible submission is 7 days late. This due date does not excuse you from any labs held between when the assignment was given and when it was due. Attendance for those lab sessions is still required.