CS 1520 / CoE 1520: Programming Languages for Web Applications (Spring 2012)

Department of Computer Science, University of Pittsburgh

Assignment #1: Perl

Released: February 7th, 2012 **Due:** 11:59pm, Friday, February 17th, 2012

Goal

Gain familiarity with Perl.

Description

In this assignment, you will emulate an information retrieval engine. Given a set of plain-text documents, you will build a reverse index for them and then given user input in the form of one or two keywords, you will produce a list of documents that contain these keywords.

You will need to write two different Perl programs:

1. `index.pl` builds a reverse index from a collection of plain-text documents.
   Usage: `index.pl -dir <directory_name> -index <index_filename>`

   A reverse index helps quickly identify which documents have a specific word, by creating a list of keywords and associating with each keyword the list of all documents that contain it. The index file will have the following format:

   ```
   keyword1 filename11:cnt11 filename12:cnt12 fiename13:cnt13 ...
   keyword2 filename21:cnt21 filename22:cnt22 fiename23:cnt23 ...
   ...
   ```

   For example, filename11, filename12, filename13 are the names of the files that contain keyword1, and cnt11, cnt12, cnt13 are the number of occurrences of keyword1 in these files respectively. A file will not be listed in a reverse index entry for a keyword if that keyword does not appear in the file (i.e., no zero entries). You should order the filenames by decreasing counts (i.e., cnt11 $>=$ cnt12 $>=$ cnt13 etc.). Note that keywords, filenames, and counts have no spaces in them.

   All the files in the directory specified at command line should be used. When building the index, you should ignore stop words, such as ”the”, ”a”, ”him”, ”her”, etc. The complete list of stop words is provided at
   `http://db.cs.pitt.edu/courses/cs1520/spring2012/assign/a1.stopwords.txt`
   Also, we will provide sample files for you to test your programs.

2. `search.pl` reads in the reverse index and asks for one or two keywords from the user. It returns the list of documents that contain all the specified keywords, sorted by the number of occurrences (i.e., the document with the most occurrences of the keyword(s) should be listed first).
   Usage: `search.pl -index <index_filename> [-batch -key1 <word1> [-key2 <word2>]]`

   search.pl will work in two different modes: interactive (the default) and batch. In interactive mode, search.pl should read the user input from the standard input and immediately print the answers. In batch mode (specified by -batch), search.pl will only search once either for the one keyword (specified via -key1 and no -key2 given) or both keywords (specified via -key1 and -key2). The

interactive version is what you should primarily use for your testing & debugging, whereas the batch version is what we will use for grading. The core functionality should be the same, but in the batch mode, the index is build multiple on every run.

Output of search.pl will be in the following form:

```
KEYWORD(S): <keyword1> <keyword2>
cnt1 filename1
cnt2 filename2
cnt3 filename3
...
```

where cnt1 $>$ cnt2 $>$ cnt3, etc.

What to submit

Two Perl programs that perform the tasks listed above, along with any additional libraries that you have developed (and are shared by the programs). Name your programs `index.pl` and `search.pl`. There is no standard naming scheme for the libraries.

Academic Honesty

The work in this assignment is to be done *independently*, by you and only you. Discussions with other students on the assignment should be limited to understanding the statement of the problem. **Cheating in any way, including giving your work to someone else, will result in an F for the course and a report to the appropriate University authority for further disciplinary action.**

How to submit your assignment

We will use a Web-based assignment submission interface. To submit your assignment:

- If you have more than one file to submit, prepare your assignment for uploading, by generating a single zip file with all the files.
- Go to the class web page `http://db.cs.pitt.edu/courses/cs1520/spring2012` and click the Submit button.
- Use your pittID as the username and the password you specified at the contact information form for authentication. There is a reminder service via email if you forgot your password. You must have already submitted your contact information, if you have not yet you need to do so now.
- Upload your assignment file to the appropriate assignment (from the drop-down list).
- Check (through the web interface) to verify what is the file size that has been uploaded and make sure it has been submitted in full. **It is your responsibility to make sure the assignment was properly submitted.**

You must submit your assignment before the due date (11:59pm, Friday, February 17th, 2012) to avoid getting any late penalty. The timestamp of the electronic submission will determine if you have met the deadline. There will be no late submissions allowed after 11:59pm, Sunday, February 19th, 2012.

[Last updated on February 7, 2012 at 8:22pm EST]