CSE 231 Fall 2009

Programming Project #6

Assignment Overview
This project focuses on Input / output file manipulation, as well as more experience on string processing. It is due Monday, October 26th before midnight. The application is to simulate a search engine that finds a query in a collection of files. This project is worth 40 points.

The Problem
The **search engine** is supposed to check a list of given files for a given query by the user. Search engines usually have complicated search algorithms but in this project we will take a slightly simpler approach. We will look for "exact" matches of a query in the documents we are searching. We say "exact" because we will do some pre-processing of both the documents and the queries before we run our algorithm

Preprocessing
Stop words are common words that are considered uninformative from a query point of view. A stop word is a connecting word that is mostly not informative such as articles, determinants, pronouns, some examples include: the, on, of, year,….. you are provided with a list of those words to exclude them from both your query and your file. Before you write your algorithm, you must remove stop words from both your files and your queries. You are required to write a function for this purpose.

Program Operation:

1. Prompt the user for a query to be examined
2. Remove stopwords from the query
3. For every file with the suffix .txt, do the following
    a. remove stopwords from the document
    b. look for an exact match (the exact non-stopword words in sequence) in the document
    c. If no such sequence is found, report so for that document
    d. If the sequence if found, report the line in the document where it was found

Example Interaction

```
Enter your query ...

justice of the ministry official

doc1.txt
query not found...

doc2.txt
found at rows :  [12, 7]

doc3.txt
```

```
query not found...

doc4.txt
found at rows :  [2]

doc5.txt
query not found...
```

Program Specifications
You project will do the following:

1. Create a function, call it removeStopWords(), that reads the "stoplist" file that we provide and return the final list of stop words.
2. You are required to have at least one other function in the code, your choice as to what it does.
3. Prompt the user to enter a query. Call the stopword function above to exclude all stop words from the query.
4. Read the first file and save it in a list of lists i.e., matrix data structure, where each line in the file is a list of words. Note that you need to remove punctuation from all the lines and words.
5. Start scanning each line/list of the document sequentially using a loop looking for the query components. Note that the size of the query is important at this point. You need to slide each row according to the size of the query. For example :
   The query : "justice of the ministry official"
   After stop word removal it will be [justice, ministry, official]
   Assume a line of the document is :
   [Iran elects a new justice at the ministry official on 12 June], but after stop removal it becomes:
   [Iran, elects, new, justice, ministry, official, June]
   you need to compare the query against:
   First slice: Iran, elects, new,
   Second slice: elects, new, justice
   Third slice : , new, justice, ministry
   Fourth slice: justice, ministry, official,
   there you found it!
   It is clear that this processing suggests a loop ….
6. report the line number of the docuent and continue.
7. move on to the second file …

Program Notes:

- Some example document files are provided for your testing.
- The lower() function is important when you read each word, because no match case is required in this assignment.
- The str() , list() functions are also useful here.
- The split() function is important to break a line into a list.

- The strip function is good to get rid of unnecessary characters attached to either end of a word, such as " , /)(." and all other punctuations. Remember, strip only works at the ends, not anywhere else!
- ''.join(mylist) is used to convert the list into a string.

Deliverables

proj06.py -- your source code solution (remember to include your section, the date, project number and comments).

1. Please be sure to use the specified file name, i.e. "proj06.py"
2. Save a copy of your file in your CS account disk space (H drive on CS computers).
3. Electronically submit a copy of the file.