CPS842 – A1 doc

In this assignment I have imported Counter() for collections to easily get rid of term repeats. Importation of re is to do easy search and/or changes to a string variable with regex, nltk is for implementing PorterStemmer into the code, json is for easy write and read of the 2 created files by the invert() function, and the import of time is to use the time functions.

There are 2 global variables that will be used in the entire assignment. These variables just hold Boolean values that represent the users choice of either turning on/off stopwards and stemming.

Function removeField() takes on fields that are not necessary and remove them from a selected file and creates a new file with all elements removed. In my case it removes everything except for the title and abstract as it is stated in the assignment documentation that only those terms are used. The program will mainly use the new file to create the terms and add them into a dictionary or hashmap.

1. Invert() asks if the user wishes to turn on stemming or stopwords and then store those answers in global variables mentioned above. **This is what the program will ask before making the terms.**
2. Invert() will open the new file and go through it line by line while also performing actions on those lines to start creating a dictionary of terms and a postings
   1. Uses for loop to go through each line
   2. Use if statements to turn on and off variables that are Boolean
      1. Such as when to start adding in words to a dictionary
   3. As the program begins adding words it will also use regex to remove some unnecessary items
   4. If there is stemming then the use of join method to make a list back to a string
   5. If there is stopwards then there will be removal of words
   6. Use of Collection Counter to get rid of repeating words allowing easy counting
   7. Use of json to quickly write the dictionary.items() into a file so it can be easily accessed and extracted using json reading in test()
   8. There are also exceptions in case the files do not exist
3. The test() asks for user input and it doesn’t matter how the user writes the term as it is in a loop so if he types a non-existent word then it will ask for a term again. The input will also be made into lower case and striped of whitespace
   1. If stopwards were when performing #2, it will ask user to try again as it is a stopword. An eclipsed time also shows.
   2. If stemming is done, the users input will also be stemmed using the same package to match words in files
      1. Printed out words will the normal version
   3. When the program find a word it will print out
      1. Document ID: word frequency, how to read the positions and positions
      2. The Title is printed
      3. The document title
      4. The word “Summary” is printed
      5. The first occurrence of the word and 9 other words but if there are less than 10 words then it will print out what is there

HOW TO RUN:

1. Call invert()
2. Follow instructions on how to activate stemming or stopwards
3. Wait for files to be created
4. Run test()
5. Type term
6. Get results or type again
7. Repeat step 5 to 6 until user types zzend