

Programming Project

Preprocessing, indexing, counting, removing stop-words & stemming

Programming Language: Java

Team: Three students in group

Due date: 15/6/1433 – 6/5/2012

Gained points = 15

Your program is going to read a text file from the input stream and from file as well. Then, it will do the following tasks:

1. **Preprocessing:** in this task you are going to clean the text file from punctuations, special characters and numeric values.
2. **Tokenization:** split the file into words (you can choose to save tokens in an array, vector or list).
3. **Normalization:** convert all words into lowercase characters.
4. **Sorting:** sort the words in ascending order (a-z).
5. **Removing Duplicates:** remove all duplicate words. *Keep in mind to save the frequency before removal process*
6. **Counting:** count the number of words in the array (*not their frequency*), list, etc. Call this number $C1$.
7. **Printing:** for each word, print its frequency as shown in Table 1.

Table 1: Terms and their frequencies

Word	Frequency
apple	5
...	...
Total words	$C1$

8. **Stop-Words Removal:** compile a list of stop-words (high frequent words) and use this list to remove all stop-words from list. This link below will help you to define stop word list.
<http://www.ranks.nl/resources/stopwords.html>
9. **Printing and Computing Reduction:** count all remaining words in the list. Call this number $C2$. Print all words in the list as shown in Table 2.

Table 2: Terms and stop-words reduction

Word	Frequency
apple	5
...	...
Total words	$C2$
Reduction	$C2/C1 \times 100\%$

Porter Stemmer: named after its author Martin Porter. It is an algorithm used for removing the most common morphological and inflectional affixes from words in English. Its main use is as part of a term normalization process that is usually done when designing Information Retrieval systems.

The stemmer is available in many common programming languages including: Java, C#, VB 6, Lisp, Perl, Python and others. The Porter stemmer code can be downloaded from: <http://tartarus.org/~martin/PorterStemmer/>

10. **Using Porter Stemmer:** use Porter stemmer to stem the words and print their stems as shown in Table 3.
11. **Printing and Computing Reduction:** Computing the percentage of reduction after stemming.

Table 3: Terms and their stems

Word	Stem
apples	apple
car	car
cars	car
...	...

12. Finally, Construct a read me file containing all installation and runs instructions, it also include group's e-mail incase of problems during assessment of the program.

Enjoy