_____
University of Texas at Dallas
CS 6322 : Information Retrieval
Fall 2014
Instructor: Dr. Sanda Harabagiu
Grader: Travis Goodwin
Issued: September 15th 2014
Due October 6th 2014 before midnight
Homework 1
_____

Problem 1 (100 points)
Tokenization
_____

A copy of the publicly available Cranfield collection is located on the UTD
Apache machine at:
/people/cs/s/sanda/cs6322/Cranfield

Write a program to gather information about tokens in the Cranfield database. You
may use any of the following programming languages : C/C++, lex/yacc, perl, or
Java.

In the Cranfield collection, document and field boundaries are indicated with SGML
tags ("document markup"). SGML tags are not considered words, so they should not
be included in any of the information your program gathers. The SGML tags in this
data follow the conventional style:

<[/]?tag> | >[/]?tag (attr[=value])+>

The attributes and the values from the SGML conventional style are optional and
appear rarely or not at all in this data collection.

 Use your program to generate the following information.

1. The number of tokens in the Cranfield text collection;
2. The number of unique (e.g. distinct) tokens in the Cranfield text collection ;
3. The number of tokens that occur only once in the Cranfield text collection;
4. The 30 most frequent word tokens in the Cranfield text collection – list them and
their respective frequency information; and
5. The average number of word tokens per document.

Turn in this information with your program description. Also make sure that you
upload a separate README file describing the way your program should be run and
all the additional software you attach. Your program should run on any UTD Unix
machine.

HINT: Program Description

Describe the operation of your program and design decisions. Include the following information.

1. How long the program took to acquire the text characteristics.
2. How the program handles:
   A. Upper and lower case words (e.g. "People", "people", "Apple", "apple");
   B. Words with dashes (e.g. "1996-97", "middle-class", "30-year", "tean-ager")
   C. Possessives (e.g. "sheriff's", "university's")
   D. Acronyms (e.g., "U.S.", "U.N.")
3. Major algorithms and data structures.

Problem 2 (100 points)
Stemming

_____

After your parsed the documents from the publicly available Cranfield collection, which is located on the UTD Apache machine at:
/people/cs/s/sanda/cs6322/Cranfield

You are asked to apply stemming to the tokens that you have recognized. For this reasons, you need to run the Porter stemmer implementation of your choice. You can use any implementation of the Porter stemmer available in open-source (e.g. http://chianti.ucsd.edu/svn/csplugins/trunk/soc/layla/WordCloudPlugin/trunk/WordCloud/src/cytoscape/csplugins/wordcloud/Stemmer.java).

You will need to also report:
1. The number of distinct stems in the Cranfield text collection;
2. The number of stems that occur only once in the Cranfield text collection;
4. The 30 most frequent stems in the Cranfield text collection – list them and their respective frequency information; and
5. The average number of word stems per document.