University of Texas at Dallas
CS 6322 : Information Retrieval
Fall 2011
Instructor: Dr. Sanda Harabagiu
Grader: Bryan Rink
Issued: September 14th 2011
Due October 3rd 2011 before midnight

Problem (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 programming language (e.g. C/C++, lex/yacc, Java, etc). 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 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 database;
2. The number of unique words in the database;
3. The number of words that occur only once in the database;
4. The frequencies of the 30 most frequent words in the database; 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.