**CS 6322 Information Retrieval**
**Spring 2020**
**Homework 1**

**NetID – TXK190012**
**Name: Theja Shree Kunam**

This document provides description of programs Tokenizer and Stemmer

## Tokenizer Statistics :

****TOKEN STATISTICS****

| Sr. No. | Information | Frequency |
|---|---|---|
| 1. | No. of tokens | 232928 |
| 2. | No. of unique tokens | 8406 |
| 3. | No. of tokens that appear only once | 3184 |
| 4. | 30 most frequent tokens | 91116 |
| | - the | 19450 |
| | - of | 12717 |
| | - and | 6678 |
| | - a | 6273 |
| | - in | 4651 |
| | - to | 4563 |
| | - is | 4114 |
| | - for | 3493 |
| | - are | 2429 |
| | - with | 2265 |
| | - on | 1944 |
| | - flow | 1849 |
| | - at | 1835 |
| | - by | 1756 |
| | - that | 1570 |
| | - an | 1388 |
| | - be | 1271 |
| | - pressure | 1207 |
| | - boundary | 1156 |
| | - from | 1116 |
| | - as | 1113 |
| | - this | 1081 |
| | - layer | 1002 |

- which 975
- number 973
- j 894
- results 885
- it 856
- mach 824
- theory 788

5. Average word tokens per document 166.3771

Total time elapsed = 2.78 seconds

To acquire tokens from the text:

- All SGML tags were replaced by blank spaces
- All numbers were replaced by blank spaces
- All special characters like [+^:,?;=%#&!@*_)(}/{\.] were replaced by blank spaces
- All occurrences of the possessive were replaced by null characters i.e. Peter's becomes Peter
- All occurrences of the apostrophe were replaced by null characters i.e. churches' replaced by churches
- All occurrences of the hyphen were replaced by blank spaces i.e. indo-american replaced by indo american
- All extra blank spaces were removed and all words were converted to lowercase

**Data Structures used:**

- HashMap token_count used to store tokens and their frequencies of occurrence
- ArrayList frequents used to store 30 most frequent tokens.

## Algorithms:

### 1. Update_Counts

**Description** : This algorithm takes as input a list of strings containing data from the Cranfield Collection and outputs a HashMap containing individual tokens and their frequencies.
**Input**: A list of strings T containing text data from Cranfield Collection
**Output**: A HashMap H containing Tokens and their frequencies

**Initialize** : an empty hashmap H

```
For string 'x' in T do:
|    A = x.split( )     //retrieve individual tokens from text and save into a list
|    for word in A :
|    |   if word not in H.keys( ):
|    |    |  H[word] = 1   //new token gets added to the HashMap
|    |    else:
|    |    |  H[word] +=1    // update token count for every occurrence in T
|    |    end if
|     end for
End for
```

Return : HashMap H

### 2. Get_Frequent

**Description** : This algorithm takes as input hashmap containing individual tokens and their counts and returns 30 most frequent tokens.
**Input**: A Hashmap H containing Tokens and their frequencies sorted in the decreasing order
**Output**: An ArrayList L containg 30 most frequent tokens

**Initialization** : An empty ArrayList L, integer count = 0

```
For x in H.keys( ):
|   while(count < 30)
|   |   L.add(x)     // add tokens to list L
|   |  count = count+1
|    end while
End for
```

Return : ArrayList L

## Stemmer Statistics:

The open source implementation provided on https://tartarus.org/martin/PorterStemmer/ was modified to run against the corpus.

| Sr. No. | Information | Frequency |
|---|---|---|
| 1. | No. of unique stems | 5847 |
| 2. | No. of stems that appear only once | 2103 |
| 3. | 30 most frequent stems | 93207 |
| | - the | 19450 |
| | - of | 12725 |
| | - and | 6678 |
| | - a | 6304 |
| | - in | 4674 |
| | - to | 4563 |
| | - is | 4114 |
| | - for | 3493 |
| | - ar | 2458 |
| | - with | 2265 |
| | - on | 2262 |
| | - at | 2136 |
| | - flow | 2080 |
| | - by | 1756 |
| | - that | 1570 |
| | - an | 1393 |
| | - pressur | 1384 |
| | - numb | 1346 |
| | - be | 1271 |
| | - boundari | 1185 |
| | - late | 1145 |
| | - from | 1116 |
| | - as | 1113 |
| | - result | 1087 |
| | - thi | 1082 |
| | - which | 975 |
| | - effec | 920 |
| | - j | 894 |
| | - method | 887 |
| | - theori | 881 |
| 4. | Average number of stems per document | 166.3771 |

Total time elapsed = 1.36 seconds.