**CS 6322.501**

**Information Retrieval**

**Program Description:**

**Algorithm:**

1. All the CarnField is sent as input and it goes to the readingFiles() function where files are processed in the following steps:
   * 1. If the File is a Directory then this function is called recursively with the Directory as the new input.
     2. If this is a File then it goes to step 2,
2. One File is processed at a time using the following steps:
   * 1. A line of text is read.
     2. Each Line goes through a set of processing where digits, Special characters, dashes and possessives are handled(it is removed).
     3. When the line processing is done, the line of text is split using the space as a delimiter and stored in a String array.
     4. In the String array each word(this is the finalized token) is processed at a time, it is stored in a “Token”(of HashMap type) data structure using the following conditions:

- If the token word is encountered the first time, insert into the “Token” with frequency 1.

- If the token word is encountered previously, increment the frequency.

v) Repeat the step until all the lines are processed.

1. Repeat Step 2 until all the Files have been processed.
2. From Step 1-3 are performed all the token are identified and stored in the “Token”.
3. Call the Stem function on each token and store returned stems into Stem(of type HashMap)

- If the token word is encountered the first time, insert into the “Stem” with frequency 1.

- If the token word is encountered previously, increment the frequency.

6. Sort the two maps and retrieve the top 30 frequent words and distinct terms.

**Token Processing done in 2 ii)**

A line of input is taken and processed using the following conditions and done in the following way and then they are stored in a String array where each word is separated using space as a delimiter.

1. The special characters of the word (ex:”/car//”) are removed.

2. If the word is of abbreviated form (ex: U.S.A), the dots are removed (USA) and it is converted to Lowercase.

3. If the word is of possession form (ex: University’s), the Possessives is removed (University).

4. If the word contains hyphens in the middle (ex: multi-layer), hyphens are removed and they are taken as two separate words.

5. If the word contains ‘- -‘ in the middle, then ‘- -‘ are removed.

6. If the word contains ‘,’ in the middle(“schidler,j”), then it is removed to form a single sentence (“schindlerj”).

7. If the word contains only ‘ such as in the case of their’middle’class then in this case the ‘ is removed and they all separated and taken as different words.

**Program-Overview:**

**1)Running Time**

The Program has taken around 3000 seconds to acquire the text characteristics during the test run.

**2)Handling special cases while tokenization:**

a. Upper and lower case words: All the tokens are converted to lower case words, so that words “People”, “people” and “pEople” go into the same entry(“people”).

b. Words with dashes: words having dash in the middle (“middle-class”) are separated and taken as two words. So, “middle” and “class” are taken as separate tokens.

c. Possessives: Possessive words (“schindler’s”) are transformed to non-possessive words by truncating‘s at the end(“schindler”).

d. Acronyms: Acronyms (“U.S.A”) are transformed by removing dots (“USA”) and then these Acronyms are transformed to Lowercase and are then processed.

**3) Major Algorithms and data structures:** algorithms is described in program description above. HashMap is used to store tokens and stems as insertion, retrieval and comparison takes linear time.

**Final Statistics of Tokens**

**Tokenization**

Total Number of the Documents: 1400

Total number of tokens: 229228

Number of unique words: 9123

Number of words that occur only once: 3853

Average number of tokens per document: 163

Top 30 Frequent Words:

S.No Word Frequency

1 the 19453

2 of 12714

3 and 6669

4 a 5937

5 in 4650

6 to 4560

7 is 4113

8 for 3491

9 are 2428

10 with 2263

11 on 1943

12 flow 1848

13 at 1834

14 by 1756

15 that 1570

16 an 1389

17 be 1271

18 pressure 1207

19 boundary 1156

20 from 1116

21 as 1113

22 this 1081

23 layer 1002

24 which 975

25 number 973

26 results 885

27 it 855

28 mach 824

29 theory 789

30 shock 712

Time taken to processes the token is: 1340

**Stemming**

Total Number of the Documents: 1400

Total number of tokens: 9123

Average number of tokens per document: 6

Number of unique words: 6416

Number of words that occur only once: 4924

Top 30 Frequent Words:

S.No Word Frequency

1 gener 15

2 observ 11

3 oper 10

4 determin 9

5 comput 9

6 deriv 9

7 integr 9

8 continu 8

9 simul 8

10 separ 8

11 indic 8

12 investig 8

13 predict 8

14 express 7

15 illustr 7

16 design 7

17 correl 7

18 origin 7

19 approxim 7

20 depend 7

21 develop 7

22 acceler 7

23 diffus 7

24 compar 7

25 differ 7

26 conduct 7

27 correct 7

28 us 7

29 stabil 7

30 interact 6