**Indexing Report**

1.First we copy the files of 2009 to titles folder with the following command

**cp 2009/\*.txt titles**

2.Then we create the representation of all the files of 2009 and remove the stop words and store them with their frequency by using a shell script in “**/home/sm1264/varun/Indexing/others**” by the name unixscript.sh which is given as:

**y=0**

**for entry in /home/sm1264/varun/Indexing/titles/\***

**do**

**cat "$entry" | tr [A-Z] [a-z] | tr -cs "a-z\_" " " | tr " " "\n" | sort | uniq -c | sort -nr > /home/sm1264/varun/Indexing/out/$y.txt**

**./replacesw.sh /home/sm1264/varun/Indexing/others/sw.txt /home/sm1264/varun/Indexing/out/$y.txt | tr -s "\n" | grep [[:alpha:]] > $entry**

**y=$((y+1))**

**done**

This script stores the indexed files to the titles folder with the actual name of the file instead of 0.txt, 1.txt … as so on.

3.In the above script we used the “**replacesw.sh**” which compares the stop words file named as “**sw.txt**” and files in the **out** folder and removes all the stop words from the files in the **out** folder

replacesw.sh:

**#!/bin/bash**

**MYREGEX=\\b\(`perl -pe 's/\n/|/g' $1`\)\\b**

**perl -pe "s/$MYREGEX//g" $2**

4.**chmod 777 unixscript.sh**

**./unixscript.sh**

This executes the unixscript.sh and the Indexed files after removing the stop words are stored in the “**titles”** folder.