**Reverse Indexing Report**

1. First, I’ve copied the **titles** folder into the **ReverseIndexing** folder. Then I’ve swapped the contents of the files to the form of **word : freq**

i.e 125 information 🡪 information 125

Using the “**swap.sh”** in the **/home/sm1264/varun/ReverseIndexing** directory, I’ve swapped the contents of the files and stored them back in the **titles** folder

swap.sh:

**y=0**

**for entry in titles/\*txt**

**do**

**awk '{ print $2 " " $1 }' $entry > temp/$y.txt**

**mv temp/$y.txt $entry**

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

**done**

This stores all the words in the **titles** folder as word:freq

1. Then I merged all the words in titles folder to one file using the following commands. These following commands only store the words without frequency to **mergefinal.txt**

**cat titles/\*txt > inimerge.txt**

**cat inimerge.txt |uniq > finmerge.txt**

**awk '{ print $2 }' finmerge.txt > mergefinal.txt**

1. **Reverse Index**
2. Firstly, I’ve made a backup of the **titles** folder to **titlesbak** folder and **mergefinal.txt** to **mergesorted.txt** to make sure I won’t mess up with the data.
3. Then I wrote a shell script that would do the reverse index of the files that I’ve got in the **titlesbak** folder using the **“revindexmainscript.sh”** stored in the **titlesbak** folder which is given as

**for i in $(</home/sm1264/varun/ReverseIndexing/mergesorted.txt)**

**do**

**echo "$i :" >> /home/sm1264/varun/ReverseIndexing/reverseindexwithtitlename.txt**

**grep -rw $i \*| tr ":" " "|awk '{print $1 ":" $3}'|xargs|tr " " ":" >> /home/sm1264/varun/ReverseIndexing/reverseindexwithtitlename.txt**

**done**

This script takes the word from the **mergesorted.txt** and checks for that word through the files in the **titlesbak** folder and stores the **final output** which is of the format as shown below to the **reverseindexwithtitlename.txt**

aa :

abc.txt:2:bcd.txt:1:xyz.txt:1…………….

**Other trial**

Then I made the sorted the merged file that we have and stored it in **mergesortedx.txt**

Then I executed the shell script with the mergesortedx.txt and stored the result in **reverseindexwithtitle.txt**. I’ve observed that the execution time with the sorted file is much lesser and the output fie size is much less than the one we got earlier. But as output is concerned the **reverseindexwithtitlename.txt** looks better than **reverseindexwithtitle.txt**

**Additional Stuff**

I made the indexes for the files 0.txt … with the actual files before I stored them with the actual titles names in the titles folder.

So I used this script to rename the titles with a prefix of the files such as 0.txt,1.txt........

indexscript.sh

**n=0;**

**for f in titlesindex/\*.txt; do mv "$f" "$((n++)).txt\_INDEX\_OF\_$f";**

**done**

Then I copied all the file names of titlesindex folder to indexes.txt as below

[sm1264@ci-lis-3lb2lq2 titlesindex]$ ls -1 | grep .txt > Indexes.txt