Group Members

- Saad Iqbal 32903
- Usman Iqbal 32902

Indexing

- Assigned unique ID's to every document
  - Whole numbers i.e. 0, 1, 2, 3, …
- "dirent.h" library
- CSV file for recording DocIDs
- Less than 5 minutes

Parsing

- Ensures tokenization of words.
- Treated all special characters except the “alphabets” and “numbers” as delimiters.

HitList

- Read every document and made a CSV file for each one
  - 0.csv, 1.csv, 2.csv, …
- Recorded every word and its occurring frequency in that file
- Around 2 hours

Inverted Indexing

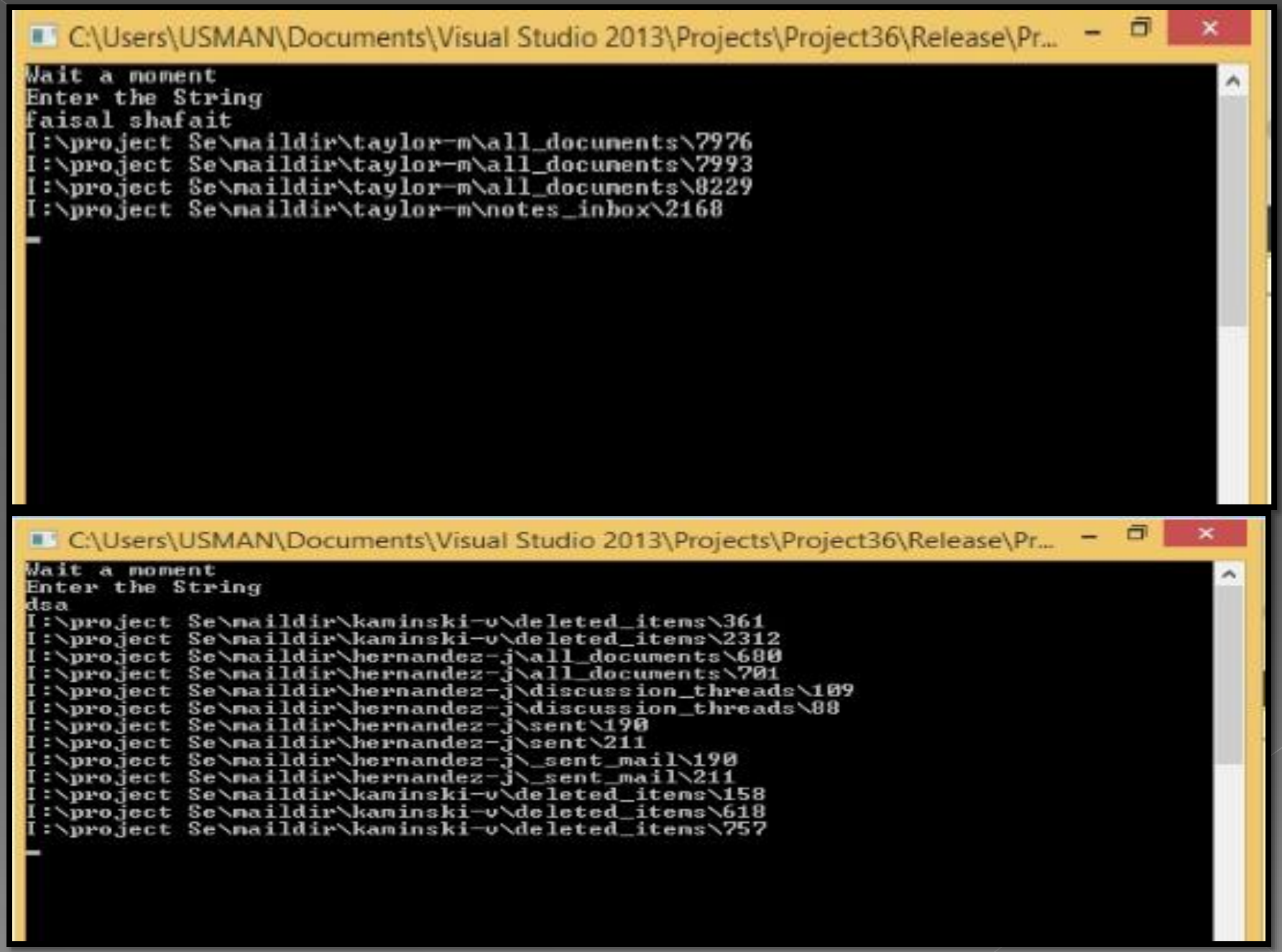
- Map of words to vector of nodes containing DocIDs and occuring frequency
- Made CSV file for every word and recorded that mapping in it
  - WordName_hashvalue.csv
- Partially on RAM and partially on disk
- Almost 4 hours

Sorting

- Sorted CSV file of each word on the basis of frequencies in different documents
- 1-2 hours

Searching

- Single-word query
  - Simply checking that word exists or not and display the results
- Multi-words query
  - Taking top 1000 DocIDs from the sorted file of each word
  - Making a map on the basis of priority
    - Files containing most of the query words will have highest priority
    - Words which are being occurred less in the whole dataset will have high priority against those who are occurring many times e.g. the, a, an etc
- That map will be sorted on the basis of highest priority and top results are shown


C:\Users\USMAN\Documents\Visual Studio 2013\Projects\Project36\Release\Pr...
Wait a moment
Enter the String
faisal shafait
I:\project Se\maildir\taylor-m\all_documents\7976
I:\project Se\maildir\taylor-m\all_documents\7993
I:\project Se\maildir\taylor-m\all_documents\8229
I:\project Se\maildir\taylor-m\notes_inbox\2168
C:\Users\USMAN\Documents\Visual Studio 2013\Projects\Project36\Release\Pr...
Wait a moment
Enter the String
dsa
I:\project Se\maildir\kaminski-v\deleted_items\361
I:\project Se\maildir\kaminski-v\deleted_items\2312
I:\project Se\maildir\hernandez-j\all_documents\680
I:\project Se\maildir\hernandez-j\all_documents\701
I:\project Se\maildir\hernandez-j\discussion_threads\109
I:\project Se\maildir\hernandez-j\discussion_threads\88
I:\project Se\maildir\hernandez-j\sent\190
I:\project Se\maildir\hernandez-j\sent\211
I:\project Se\maildir\hernandez-j_sent_mail\190
I:\project Se\maildir\hernandez-j_sent_mail\211
I:\project Se\maildir\kaminski-v\deleted_items\158
I:\project Se\maildir\kaminski-v\deleted_items\618
I:\project Se\maildir\kaminski-v\deleted_items\757