README

(VAIBHAV VARSHNEY, MT17065)

Preprocessing

In the preprocessing phase, I have created a dictionary considering each word as a key to the dictionary and corresponding to each key, there is a list of document id of the documents containing the key.

Preprocessing is divided in the following phases:

(1) Removal of headers: For removing the headers, I considered the following pattern: Each line of header has a word followed by a colon. Keeping this in mind, I removed such lines until I found a newline.
(2) Next punctuation is removed followed by tokenizing the document.
(3) Finally stemming is done. (Steps 2 and 3 are implemented using nltk library).

Assumption

(1) In the case of applying skip list merge algorithm, it is assumed that the best length for skipping is sqrt(l) where l is length of list. (According to the notes).
(2) Query is assumed to be provided in the format without punctuation.

Files included:

(1) Preprocess.py - used for creating the inverted index dictionary
(2) Function.py - all the function or, and etc are defined in here.
(3) Run.py - used for executing the queries.
(4) All_names.json - it is the json file containing names of all files.
(5) dataDictionary_modified_name.json contains the inverted index of the words.