CSE508 : Information Retrieval

Assignment 2:

Q(i):
1): Jaccard Coefficient based document retrieval:

Preprocessing: In preprocessing i take all the files in one folder then using os liberaty i read all the documents one by one and read these files bu using open method and than using NLTK liberary i tokenize this string and store all the tokens in a list .Now i remove all the stopwords in this list. Same i did with query it stores in a string and than using same liberary i store tokenized words of a query and then i remove all the stopwords. And change these lists to the sets.

Methodology: Now i have two sets on which i applied jaccards coeficient formula and store all the documents and there score in a dictionary as key value pair. Now i sort this dictionary by its value.Now i return top k key elements of this dictionary.

2) Tf-Idf based document retrieval:

Preprocessing: In preprocessing i take all the files in one folder then using os liberaty i read all the documents one by one and read these files bu using open method and than using NLTK liberary i tokenize this string and store all the tokens in a list . Now i remove all the stopword in this list and change all the numbers to words using NUM2Words liberary. Now i preprocss all the index files and store the titles as the value of the key as document id. And store this dictionary in a pickle. Now i store all the distict words in dictionary as key of this dictionary and value as list of the documents term frequency corresponding to every word. I use 2 varient to store this term frequency for checking the relevency of the query output. Now i calulated tf-idf value in the and multiply this with the corresponding words term frequency. If this word occured in title i simply doubled the score to give the more weightage the document. Now i store this formed dictionary in a pickle file.Now i take query and preproces this query and make the tokens and change these tokens by num2words to words if there is any numerical value in the query.

Methodology: After reading the stored pickle file i apply td-idf varient which calculates all the documents tf-idf value and store it in a dictionary doc id as a key and total score as value now i sort this dictionary and sort by value and return top k documents.

3) Tf-Idf based vector space document retrieval:

Preprocessing: After reading above stored dictionary we length normalized the each document vector corresponding to each word and store this modified dictionary in pickle file. Now i take query and preproces this query and make the tokens and change these tokens by num2words to words if there is any numerical value in the query.

Methodology: Now i read this pickle file an take cosine score for each documents by dot the query vector with the each document vector corresponding to all the words. Now i store all the scores in a

dictionary as a value of the key as document id. Now i sort all the documents by its values and return top k documents.
Assumption: All queries are valid.

Q2: Minimum edit distance:

Preprocessing: I read the file and tokenize it by using NLTK liberary and store these words in a list. Now i take the query and tokenize with the same liberary and take the set of this list which are not available in the previous list.

Methodology: Now i apply edit distance for the above listed words and store all edit distances coresponding to every word and now i sort these words lists by its value and return top k elements.