Assignment 2

Submitted by

Swati Verma

MT19073

Question 1

Assumptions:

1. No metadata has been removed from any of the documents.

Preprocessing steps:

1. Words in all the documents are converted to lower case.
2. Punctuations are removed from all the documents.
3. Expansion such as don't to do not has not been handled because the library was not working on my system.
4. Stopwords have been removed from all the documents using nltk library.
5. To convert surface words to root word lemmatization is used instead of stemming because lemmatization gives meaningful words.
6. All the above steps are done both for documents and query which is given by the user.

Methodology:

1. After tokenization, while lemmatizing each word two dictionaries are created. The first dictionary contains each document name as key and a list of lemmatized tokens corresponding to that document as value. The second dictionary contains each unique word as keys and the number of documents in which they occur as values. The second dictionary will help in computing idf values.
2. For Jaccard's coefficient based document retrieval, Jaccard coefficient value is calculated between query given by user and each document using the formula jc=(document intersection query)/(document union query) and the values are stored in a dictionary

where keys are document names and values are the Jaccard coefficient corresponding to that document and query. After that, the dictionary is sorted and top k documents are extracted from the dictionary.

3. For tf-idf based document retrieval, user is asked to enter the variation of tf and idf which they want to calculate. Based on the user input, tf-idf values for each word are calculated and are summed up and stored in a dictionary where keys are document names and values are the tf-idf value corresponding to that word. After that, the dictionary is sorted and top k documents are extracted from the dictionary.
4. One title dictionary is created which contains document names as keys and a list containing title tokens corresponding to that document as values. Now while calculating tf-idf values, each query term is seen, if that term is present in the title then tf-idf value is kept the same otherwise tf-idf value is multiplied by 0.7. Basically, 0.3 weightage is given to title and 0.7 weightage is given to the body of the document and k top documents are retrieved in the same way as done in the above step.
5. For cosine similarity based document retrieval, with and without including title, the cosine similarity between query and document is calculated and stored in dictionary where keys are document names and values are the cosine similarity corresponding to that document and query. After that, the dictionary is sorted and top k documents are extracted from the dictionary.

Question 2

Assumptions:

1. No metadata has been removed from any of the documents.

Preprocessing steps:

1. Words in all the documents are converted to lower case.
2. Punctuations are removed from all the documents.
3. Expansion such as don’t to do not has not been handled because the library was not working on my system.

4. Stopwords have been removed from all the documents using nltk library.
5. To convert surface words to root word lemmatization is used instead of stemming because lemmatization gives meaningful words.
7. All the above steps are done both for documents and query which is given by the user.

Methodology:

1. A sentence is taken from the user as input. Each term in that sentence is checked and if the term is present in the dictionary then it is not taken otherwise edit distance is taken between each term and all the dictionary terms and stored in a new dictionary and finally top k words with minimum edit distance are shown for each word.