Information Retrieval Assignment 1

Group-98

Answer 1:

Methodology

A file is used to maintain a list of all file names. After that, each file's data is preprocessed. The "utf-8" codec was used to decode the majority of the files, but "Unicode escape" was utilized for some.
Finally, utilizing all of the terms, an inverted index was created.

Preprocessing Steps

1. Lowercase text conversion
2. Using nltk for tokenization.
3. nltk is used to remove stop words.
4. Characters that aren't alphanumeric are eliminated.
5. All characters that appear just once are eliminated.
6. Finally, all of the words are compiled into a collection.

Assumptions

Input Query does not care about the lower or upper case.
We get the results from the query terms that aren't stemmed. It may also be shown for stemmed terms in the demo.

Question 2:

Methodology:

Created a dictionary of unique words present in all files and stored the number of occurrences and positions in the posting list. For searching the phase query in the dictionary it searches word by word in the dictionary and retrieves the document ids and indexes of the word in the document. Based on the indexes and doc ids gives the documents that have the phrase query.

Assumption-

Queries are case insensitive.