INFORMATION RETRIEVAL: ASSIGNMENT 1

Group - 8

VIpul kesari(2018118), Nipun Jain(2018058), Aman Kumar(2018216),Harman Singh(2018284),Abhinava Phukan(2018004)

Link for Github Repository - nipun5/IR2021_A1_1-for-Group-No-8 (github.com)

PreProcessing

First of all, we have to download the dataset from a website which is (http://archives.textfiles.com/stories.zip) containing 467 files and 15 MB.

Let's START!
Pre-requisites:
install Python
install NLTK
1)First part of the assignment is to carry out the preprocessing steps on the dataset
start with pre-processing of text as it is important while cleaning the text helps you get quality output by removing all irrelevant text and getting the forms of the words etc.
There are various preprocessing steps ->

Convert text to lower case.
word tokenize
Lemmatize
Remove Number
Remove punctuation
Stop words removal

- Then we give the output as fully preprocessed text.

.

Methodology

2)Second part of the assignment is to implement the unigram inverted index.

We will Use a document ID for all the documents (usually the index during iteration)
Then we Preprocess the text by many methods.
Generate tokens from preprocessing the text
For every token, we have to add the document id, and then we have to repeat this process for all the documents. Then generate the new dictionary token. After that, we generate the posting list for all the tokens from the inverted index function. For the merging algorithms, we make appropriate functions for the boolean query and then we have to do all the work.

Assumptions

3) Third part of the assignment is to against some queries like x OR y,x AND y,x AND NOT y,x OR NOT y.
In this, we have to do that. After preprocessing the text and generating the posting list, we have to do make a merging algorithm, and the queries are considered from left to right, so no optimization has to be done.

Steps of use

All the code is written python file for this, we have to only run the files in the command prompt, and then we have to give the input as per the specialization we have to run the inverted index file also.