5th October 2020

UE17CS412 - Algorithms for Informational Retrieval (B.Tech 7th Sem Elective)

Assignment 1 : Build a search engine (15 Marks)

Guidelines:

a) Code has to be developed in **PYTHON.**

b) ***Assignment will have to be carried out in teams of size FOUR.***

c) Submission (Code, Readme files etc , Snapshot of results) will have to be done, on or before deadline in to the Google Drive shared folder.

d) Summary report of the assignment will ***have to be uploaded in to the Google Drive shared folder.***

e) Approx **4 *Weeks*** of time will be available before submission. Final submission will be on 17,18 & 19 November,2020.

f) Follow fair code of ethics and **develop your version** of code. You can discuss/consult with anyone, but write your version of the code. Plagiarism will get you zero marks !!

g) You will be called upon to Demo the assignment, to match with submission data you have provided in the Google Drive.

Problem Definition, Data Generation, Testing and Logging Stats

Problem:

a) Build a search engine for Environmental News NLP archieve.

b) Built a corpus for archieve with atleast 418 documents.

Data:

a) Use the following link for Environmental News NLP dataset.

https://www.kaggle.com/amritvirsinghx/environmental-news-nlp-dataset

Code:

Your Code should contain functionality to

- Search for the terms in the query
- Create Postings list
- Fill the Inverted Index
- Rank the pages
- Retrieve the data from the dictionary
- Query response time
- Measure the efficiency using precision,recall,F measure.

Demo:

- **Run your code and carryout the search with different queries.**
- **Retrieve the data, compile and compare metrics with any one of the search engine like Elasticsearch, Apache Solr,Apache Lucene, Google Cloud Search, Google Desktop Search for the same corpus.**
- **Measure the efficiency.**

Report:

- You should submit a hard copy, 4-page summary of your project
- Your report should include the code snippet/algorithm used, similarity check of retrieved data obtained with your search engine and any one search engine like Elasticsearch,Apache Solr, Apache Lucene, Google Cloud Search, Google Desktop Search.
- Interpretation of efficiency.

Last para of your report should contain your observations on the Learning Outcomes of this project.