PES UNIVERSITY
(Established under Karnataka Act No. 16 of 2013)
100 Ft. Road, BSK III Stage, Bengaluru – 560 085
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SESSION: AUG-DEC 2020

Course Title: Algorithms for Information Retrieval		
Course code: UE17CS412		
Semester : VII sem	**Section: C**	**Team Id:19**
SRN: PES1201701868	**Name: Kritika Kapoor**	
SRN: PES1201700160	**Name: Shrutiya M**	
SRN: PES1201700740	**Name: Tejaswini A**	
SRN: PES1201701540	**Name: Shubha M**	

ASSIGNMENT REPORT

Problem Statement

- Build a search engine for Environmental News NLP archive.
- Build a corpus for archive with at least 418 documents.

Description

- A search engine was developed which uses the Environmental news dataset as a corpus.TF-IDF scores were used to score the terms and cosine similarity was used to rank the results.

- Firstly, the snippets were preprocessed through lemmatization, stop word removal, special character removal, case folding and equivalence classing.

- Inverted index was constructed using a hash table containing the terms ,the document the terms belonged to with their doc ID, frequency of the terms and the positional index of the terms.

- Permuterm index construction was also carried out to support wildcard queries.

- Types of queries handled:
    - Boolean Queries- Queries involving AND,OR and NOT were implemented using postfix stack evaluation having operand precedence order NOT, AND, OR from highest to lowest.
        - AND query was implemented by retrieving positional indices of both the left and right operands .Intersection of the results were returned.
        - OR query was implemented by retrieving positional indices of both left and right operands. Union of the results were returned.
        - NOT query was implemented by obtaining the positional indices of the input query and taking the set difference of the obtained results.

    - Phrase queries - Queries involving phrases (i.e 1 or more words)
        - The implementation involved doing an AND operation of all the words in the query and then obtaining the positional indices of the queries .Positions in the individual phrases were checked if they were one after the other and results were fetched accordingly.

    - Wildcard Queries - Queries where special characters are used to represent unknown characters in the text.
        - The queries were processed and based on the position and occurrence of the special characters representing the unknown text ,the queries were parsed. Accordingly, results were retrieved.

    - Mixed Queries - Boolean Queries can be given along with phrase and wildcard queries.

- Spelling error correction is taken care of using the edit-distance algorithm.

- A vector space model was constructed by considering the tf-idf scores of each of the tokens in the snippets .

- On passing input queries for retrieval, cosine similarity between the query vector and the results vector was used to obtain the most relevant results.

The application provides top most K relevant results (K given as input by the user) with details of document name, record number and the snippet.

Output Screenshots

- Mixed Query (involving phrase, boolean and wildcard query)

Mixed Queries

```
[82] ranked_results('(greenhouse gas) AND (NOT fuel AND pollu*n OR experimenting)',10)
```

```
168  results fetched in  0.48101115226745605 seconds. Top  10 are being shown below -

Document Name         Row number  Snippet
------------------  ------------  ------------------------------------------------------------------------------------------------------------------------
FOXNEWS.201803.csv             4  growers. cullen is experimenting with leds. the annual greenhouse gas pollution is equivalent to 3 million cars.
BBCNEWS.201901.csv            67  so as christmas trees rot, they comparatively give off huge amounts of greenhouse gases such as methane. but at th
FOXNEWS.201002.csv             0  reporter: he wants the agency to rethink the findings on green house gases announced in december. mainly a determi
FOXNEWS.201002.csv            18  epa administrator lisa jackson that human activity increases levels of greenhouse gas pollution. and that, 'greenh
BBCNEWS.201704.csv            25  we have 7 million people die every year because of pollution. and this is why it is so important that we go and re
BBCNEWS.201911.csv           200  and the second is about air pollution. and it so happens that air pollution is also adding greenhouse gases to the
BBCNEWS.201911.csv           241  so what we've identified are six steps that can be taken. the first one has to do with energy, and carbon dioxide
CNN.201906.csv                33  about a ton of greenhouse gases. while it can be true that incinerators produce fewer greenhouse gases, nationally
MSNBC.200909.csv              31  billions to capture carbon pollution so we can clean up our coal plants and just this week, we announced that for
FOXNEWS.200912.csv           436  i'm proud to announce epa has finalized the endangerment finding on greenhouse gas pollution and is now authorized
```

- Mixed Query with spelling error

Mixed Queries

```
[83] ranked_results('tom steyer AND wasingtonn',10)
```

```
Did you mean washington ?

34  results fetched in  0.29035353660583496 seconds. Top  10 are being shown below -

Document Name       Row number  Snippet
----------------  ------------  ------------------------------------------------------------------------------------------------------------------------
CNN.201912.csv             168  i'm tom steyer and i approve this message because the only way we get universal healthcare, address
CNN.201912.csv             159  i'm tom steyer and i approve this message because the only way we get universal healthcare, address
CNN.201912.csv             173  i'm tom steyer and i approve this message because the only way we get universal healthcare, address
CNN.201912.csv             172  i'm tom steyer and i approve this message because the only way we get universal healthcare, address
CNN.201912.csv             166  i'm tom steyer and i approve this message because the only way we get universal healthcare, address
CNN.201912.csv             152  i'm tom steyer and i approve this message because the only way we get universal healthcare, address
CNN.201912.csv             157  i'm tom steyer and i approve this message because the only way we get universal healthcare, address
CNN.201912.csv             167  i'm tom steyer and i approve this message because the only way we get universal healthcare, address
CNN.201912.csv             149  i'm tom steyer and i approve this message because the only way we get universal healthcare, address
CNN.201912.csv             156  i'm tom steyer and i approve this message because the only way we get universal healthcare, address
```

Interpretation of efficiency

The efficiency of the IR model was compared with Elasticsearch through the following unranked evaluation metrics. Elasticsearch was considered as the ground truth for evaluation of the performance of our search engine.

- Precision - The fraction of relevant instances among the retrieved instances, calculated to be **91.4%.**

- Recall - The fraction of the total amount of relevant instances that were actually retrieved, calculated to be **99.44%.**

- F1 score - The metric that combines recall and precision using the harmonic mean, calculated to be **94.25%.**

- Accuracy - The fraction of the retrieved results our IR model got right, calculated to be **99.40%.**

- Response Time - Elasticsearch took **4.33s** to retrieve a query, whereas our search engine took **1.46s**.

All the above measures were averaged over a bunch of 50 different queries of all kinds. Our search engine worked appreciably well in comparison to elastic search. Phrase query could be improved upon to generate appropriate results. The ranking algorithm could be improved by adding extra factors along with TF-IDF and cosine similarity.

Learning Outcome

- Thorough understanding of working of search engines.

- Hands on experience on building posting lists, different indexing, different types of queries and vector space model.

- Understanding about metrics used to evaluate the accuracy and efficiency of information retrieval systems.

Name and Signature of the Faculty