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

ASSIGNMENT REPORT

Course Title: Algorithms for Information Retrieval		
Course code: UE17CS412		
Semester : VII sem	**Section:C**	**Team Id:36**
SRN:PES1201701136	**Name: Kavya Khatter**	
SRN: PES1201700395	**Name: Shrinidhi Choragi**	
SRN:PES1201700847	**Name: Daksh Singhal**	

PROBLEM STATEMENT:

Building a search engine for Environmental News NLP Archive and perform the comparison of metrics with available search engines.

DESCRIPTION

The aim is to implement a search engine that answers the queries on Environmental News NLP Archive. In the first part, the corpus is created, by indexing the documents and snippets and secondly the search queries are answered using the index created.

To which features like, spell-check, ranking are added. To grade the implementation, similarity checks with the existing search engine- Elastic search.

Dataset

Environmental News NLP Archive : Contains 418 documents from various news stations containing the news article URL, Match DateTime, station, source, IAshowId, IA preview Thumb, snippet.

Inverted Index

Inverted Index is a data structure that is built to parse the documents that the queries are answered on. Given a query, the index is used to return the list of documents and snippets relevant for the query. The inverted index contains mappings from terms to the documents that those terms appear in. Wherein, each term is a key in the index whose value is its postings list. A posting list of a term is the list of documents that the term appears in.

Query Types

The search engine implemented answers the query types namely:

1. **One-word Queries**: Queries that consist of a single word.
2. **Free test Queries**: Queries that consist of sequences of words separated by space, where the result will be the implicit logical ‘OR’ of all the terms present in the query.
3. **Phrase Queries**: Queries that again consist of sequences of words separated by space, and inserted with the double quotes so that the documents to contain the terms in the query exactly in the specified order are to be fetched.
4. **Wild-card Queries**: Queries that are uncertain about the spelling of a term or when multiple spelling variants of a term exist.
5. **Proximity Queries**: Queries that require the term in the query to be occurring in the given proximity within a snippet.

Spell Check

If the query word exists in the vocabulary then we assume that it is correct.If this word does not exist in the vocabulary we try to find the most similar words.The similar words are sorted based on Jaccard Distance by computing the 2Q grams of the words and returned the 3 most similar words order by Similarity and Probability.

OUTPUT SCREENSHOTS

```
Enter the query:usa elections
Query Type: Free Text Query
['usa', 'elections']
Response_time:  0.011539936065673828
Document : 1
Similarity :  1.0
URL  :  https://archive.org/details/MSNBC_20091211_160000_MSNBC_News_Live#start/2291/end/2326
MatchDateTime  :  12/11/2009 16:38:26
Station  :  MSNBC
Show  :  MSNBC News Live
IAShowID  :  MSNBC_20091211_160000_MSNBC_News_Live
IAPreviewThumb  :  https://archive.org/download/MSNBC_20091211_160000_MSNBC_News_Live/MSNBC_20091211_160000_MSNBC_News_Live.thumbs/MSNBC_20091
Snippet  :  for a new interview with 'usa today.' palin talked about climate change, obama's nobel win and her involvement with the 2010 midte
Document : 2
Similarity :  0.7552233121809941
URL  :  https://archive.org/details/BBCNEWS_20170120_130000_BBC_News_at_One#start/851/end/886
MatchDateTime  :  1/20/2017 13:14:26
Station  :  BBCNEWS
Show  :  BBC News at One
IAShowID  :  BBCNEWS_20170120_130000_BBC_News_at_One
IAPreviewThumb  :  https://archive.org/download/BBCNEWS_20170120_130000_BBC_News_at_One/BBCNEWS_20170120_130000_BBC_News_at_One.thumbs/BBCNEWS
Snippet  :  president obama's plan to cut greenhouse gas emissions in the usa. our environment analyst roger harrabin assesses what he may do,
```

```
Enter the query:elect*
['electric', 'election', 'elect', 'elected', 'elections', 'electricity', 'electing', 'electorate', 'electoral', 'electrification', 'electorial
Response_time:  0.8735971450805664
Document : 1
Similarity :  0.25302692958134054
URL  :  https://archive.org/details/MSNBCW_20190908_220000_Meet_the_Press#start/1092/end/1127
MatchDateTime  :  9/8/2019 22:18:27
Station  :  MSNBC
Show  :  Meet the Press
IAShowID  :  MSNBCW_20190908_220000_Meet_the_Press
IAPreviewThumb  :  https://archive.org/download/MSNBCW_20190908_220000_Meet_the_Press/MSNBCW_20190908_220000_Meet_the_Press.thumbs/MSNBCW_2019
Snippet  :  and win big and win back the u.s. senate. that's the recipe to getting all of these things done on climate change and immigration
Document : 2
Similarity :  0.22684022171208082
URL  :  https://archive.org/details/MSNBCW_20140515_030000_All_In_With_Chris_Hayes#start/892/end/927
MatchDateTime  :  5/15/2014 3:15:07
Station  :  MSNBC
Show  :  All In With Chris Hayes
IAShowID  :  MSNBCW_20140515_030000_All_In_With_Chris_Hayes
IAPreviewThumb  :  https://archive.org/download/MSNBCW_20140515_030000_All_In_With_Chris_Hayes/MSNBCW_20140515_030000_All_In_With_Chris_Hayes.
Snippet  :  to deal with climate change, you're not taking money and putting it in the ground. you're giving it to other people, you're creati
Document : 3
```

```
Enter the query:fusa eelections
         word      Prob  Similarity
296        usa  0.001248    0.666667
7173   refusal  0.000295    0.500000
27464     usaa  0.000268    0.500000
Give the suggested word from the listusa
['usa', 'eelections']
            word      Prob  Similarity
1871    elections  0.003557    0.888889
20470  reelection  0.000309    0.800000
33295  selections  0.000027    0.800000
Give the suggested word from the listelections
['usa', 'elections']
Query Type: Free Text Query
['usa', 'elections']
Response_time:  11.218514919281006
Document : 1
Similarity :  1.0
URL  :  https://archive.org/details/MSNBC_20091211_160000_MSNBC_News_Live#start/2291/end/2326
MatchDateTime  :  12/11/2009 16:38:26
Station  :  MSNBC
Show  :  MSNBC News Live
IAShowID  :  MSNBC_20091211_160000_MSNBC_News_Live
IAPreviewThumb  :  https://archive.org/download/MSNBC_20091211_160000_MSNBC_News_Live/MSNBC_20091211_160000_MSNBC_News_Live.thumbs/MSNBC_20091
Snippet  :  for a new interview with 'usa today.' palin talked about climate change, obama's nobel win and her involvement with the 2010 midte
```

```
Enter the query:government /8 plan
[[0, 13], [0, 17], [0, 74], [26, 398], [26, 542], [28, 1138], [44, 8], [44, 35], [44, 43], [44, 44], [44, 45], [44, 46], [44, 47], [44, 56], [44, 58], [44, 15
Response_time:  0.006306648254394531
Document : /content/drive/My Drive/AIR/TelevisionNews/BBCNEWS.201701.csv
URL  :  https://archive.org/details/BBCNEWS_20170110_023000_Monday_in_Parliament#start/19/end/54
MatchDateTime  :  1/10/2017 2:30:34
Station  :  BBCNEWS
Show  :  Monday in Parliament
IAShowID  :  BBCNEWS_20170110_023000_Monday_in_Parliament
IAPreviewThumb  :  https://archive.org/download/BBCNEWS_20170110_023000_Monday_in_Parliament/BBCNEWS_20170110_023000_Monday_in_Parliament.thumbs/BBCNEWS_20170
Snippet  :  brazil's government is defending its plan to build dozens of huge hydro-electric dams. it argues the project will boost the economy and provide cle
Document : /content/drive/My Drive/AIR/TelevisionNews/BBCNEWS.201701.csv
URL  :  https://archive.org/details/BBCNEWS_20170110_040000_BBC_News#start/932/end/967
MatchDateTime  :  1/10/2017 4:15:47
Station  :  BBCNEWS
Show  :  BBC News
IAShowID  :  BBCNEWS_20170110_040000_BBC_News
IAPreviewThumb  :  https://archive.org/download/BBCNEWS_20170110_040000_BBC_News/BBCNEWS_20170110_040000_BBC_News.thumbs/BBCNEWS_20170110_040000_BBC_News_0009
Snippet  :  brazil's government is defending its plan to build dozens of huge hydro-electric dams in the amazon. it argues the project will boost the economy,
Document : /content/drive/My Drive/AIR/TelevisionNews/BBCNEWS.201701.csv
URL  :  https://archive.org/details/BBCNEWS_20170121_123000_Click#start/1703/end/1738
MatchDateTime  :  1/21/2017 12:58:38
Station  :  BBCNEWS
Show  :  Click
IAShowID  :  BBCNEWS_20170121_123000_Click
IAPreviewThumb  :  https://archive.org/download/BBCNEWS_20170121_123000_Click/BBCNEWS_20170121_123000_Click.thumbs/BBCNEWS_20170121_123000_Click_001678.jpg
Snippet  :  affordable healthca re plan. mr trump signed his first executive order forcing government offices to minimise the cost of the reforms until they ca
Document : /content/drive/My Drive/AIR/TelevisionNews/BBCNEWS.201910.csv
URL  :  https://archive.org/details/BBCNEWS_20191012_020000_BBC_News#start/1206/end/1241
```

```
Enter the query:"beena part"
Query Type: Phrase Query
Phrase Query: found in document 0, snippet 0
[[0, 0]]
Response_time:  0.001447439193725586
Document : /content/drive/My Drive/AIR/TelevisionNews/BBCNEWS.201701.csv
URL  :  https://archive.org/details/BBCNEWS_20170131_054500_BBC_News#start/493/end/528
MatchDateTime  :  1/31/2017 5:53:28
Station  :  BBCNEWS
Show  :  BBC News
IAShowID  :  BBCNEWS_20170131_054500_BBC_News
IAPreviewThumb  :  https://archive.org/download/BBCNEWS_20170131_054500_BBC_News/BBCNEWS_20170131_054500_BBC_News.thumbs/BBCNEWS_20170131_054500_BBC_News_000478.
Snippet  :  beena part to do. the airline industry has not been a part of this move to reduce carbon and teal last year. -- and teal. they agreed on a deal to cur
```

COMPARISON

We compare all queries types of our IR systems with elastic search engine and following are the response time:

Query	Our IR System	Elastic Search
Free Text Query	0.01 secs	0.063 secs
Wildcard Query	0.87 secs	0.052 secs
Proximity Query	0.0014 secs	NA
Phrase Query	0.012 secs	NA

INTERPRETATION OF EFFICIENCY

We have calculated the Precision and Recall of 5 queries for our IR system by measuring it against the elastic search results considering it as relevant documents.

For Free Text query Precision is 0.97 and Recall is 1. This means our IR system is retrieving all the relevant documents from the database plus a few false positives.

LEARNING OUTCOME

- Building IR systems for query searching.
- Building Posting listing and Dictionary using B-Trees
- Handling Different types of Queries
- Building Spelling Correction using Jaccard Coefficient
- Hands on of ElasticSearch

Name and Signature of the Faculty