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	**Section:H, D(ECE),F(ECE)**	**Team Id:1**
SRN: PES1201701429	**Name: Muvazima Mansoor**	
SRN: PES1201700230	**Name: Kirthika Gurumurthy**	
SRN: PES1201700864	**Name: Nikunj Phutela**	
SRN: PES1201700088	**Name: Anantharam R U**	

ASSIGNMENT REPORT

Problem Statement:

To build a search engine using Environmental News dataset as the corpus and compare the performance with Elasticsearch**.**

Description

The dataset consists of 418 CSV files. Each row in these CSV files is considered as a document.

1. Building the Inverted Index

   The corpus of documents is parsed ,tokenized and stemmed by removing punctuation, splitting it on whitespace using the NLTK library. Then, a temporary hashtable is created that maps filenames to their list of tokens. This hashtable is repeatedly transformed until the final inverted index is created. To allow support for phrase (exact match) queries, the position of each word in the document is also stored.

   The structure of the inverted index is : {‘word’ : {‘doc_name’ : { ‘row’ : [list of positions’]}}}

2. Querying the Inverted Index

   The Query is also parsed, tokenized and stemmed before execution.

   a. Free text query
      The query is split into words and for each word, a list of documents that contain that word is obtained using the inverted index. Union operation is performed on the lists of documents to get the final result. If there are common snippets occurring as output for multiple words, these are removed to decrease redundancy.

   b. Phrase query
      Single word query is executed for every word in the input to obtain a list of documents for each word. Then, intersection of all the lists is taken that gives all the documents that contain all the words in the query. To ensure the correct ordering of the words in the document, the positional index(present in the inverted index) of each word is considered.

   c. Wild card query
      A permuterm index was generated for the purpose of wild cards. The wild card queries considered are X*, X*Y and *X*. The permuterm consists of all rotation points of the words in the inverted index and the query look up can be performed using the same.

3. Ranking Results

   To rank the results obtained from free text query and phrase query by their relevance to the query, tf-idf (term frequency- inverse document frequency) is used. A tf-idf vector of size N (length of the inverted index) is created for each document using the TfidfVectorizer module and stored as a sparse matrix.
   The tf-idf vector of size N is also created for the query.
   To calculate the similarity between the query and document vectors, cosine similarity is used which takes the dot product of the query and each document vector in the result set and divides it by the product of the magnitudes of these two vectors, which in turn returns the cosine of the angle between these vectors.
   The documents are then sorted in decreasing order of the cosine similarity score and top-k (in our case, k=10) documents are returned.

4. Performance Comparison
   To determine the efficiency of the engine, the top 10 results of 50 different queries were compared with the top 10 results of elasticsearch for the same query using mean average values of Precision, Recall and F-measure.

Output Screenshots

Displaying top-5 results

Free text query:

```
------------------------SEARCH ENGINE-----------------------------

Enter the Query : global warming
Similarity value: 0.770587587115239
Result # 1
Document Name: MSNBC.201607.csv
Row no:  18
Snippet:  'global warming.' what do you expect? trump says 'global warming is a hoax.'


Similarity value: 0.743528354212911
Result # 2
Document Name: BBCNEWS.201903.csv
Row no:  70
Snippet:  to tackle global warming.


Similarity value: 0.6962138002432555
Result # 3
Document Name: BBCNEWS.201908.csv
Row no:  10
Snippet:  tackle global warming.


Similarity value: 0.6904612633028764
Result # 4
Document Name: FOXNEWS.201409.csv
Row no:  42
Snippet:  global warming? not global warming. that's up ahead.


Similarity value: 0.6880266910235185
Result # 5
Document Name: CNN.201509.csv
Row no:  1
Snippet:  who believes in global warming?


Execution Time:  5.08513617515564
Do you want to search again? (y/n):
```

Phrase query:

```
-------------------------SEARCH ENGINE-----------------------------

Enter the Query : "global warming is a hoax"
Similarity value: 0.8159752989085705
Result # 1
Document Name: MSNBC.201607.csv
Row no:  18
Snippet:  'global warming.' what do you expect? trump says 'global warming is a
hoax.'



Similarity value: 0.7124490442336275
Result # 2
Document Name: CNN.202001.csv
Row no:  33
Snippet:  global warming a as a hoax. what is your position on global warming? d
o you us think it's a hoax? no, not at all. nothing is a hoax about that. i want
 clean air and clean water.



Similarity value: 0.650954803059845
Result # 3
Document Name: CNN.202001.csv
Row no:  34
Snippet:  sees global warming as a hoax, a comment he made as his administration
 was rolling back environmental regulations. what is your position on global war
ming, do you think it's a hoax? not at all. nothing is a hoax about that, very s
erious subject.



Similarity value: 0.6471197199180826
Result # 4
Document Name: MSNBC.201701.csv
Row no:  5
Snippet:  let's play that. donald trump has called global warming a hoax caused
by the chinese. do you agree that global warming is a hoax? i do not, senator. s
o donald trump is wrong? i do not believe that climate change is a hoax.



Similarity value: 0.6105343586167118
Result # 5
Document Name: FOXNEWS.201701.csv
Row no:  34
Snippet:  do you agree that global warming is a hoax? i do not. so you disagree
with the president?



Execution Time:  0.16408967971801758
Do you want to search again? (y/n):
```

Wild card query (unranked)

1) x*

```
------------------------SEARCH ENGINE-----------------------------

Enter the Query : hell*
You have entered a wildcard query
Document Name: BBCNEWS.201701.csv
Row no:  205
Snippet:  no-one yet knows what he'll do. while he's recently met with clim
ate campaigner al gore, he's also just appointed several key cabinet member
s who've expressed sceptical views about climate change. and could we soon
see the land speed record broken? after funding setbacks



Document Name: BBCNEWS.201701.csv
Row no:  207
Snippet:  of the paris climate deal. no-one yet knows what he'll do. while
he's recently met with climate campaigner al gore, he's also just appointed
 several key cabinet members who've expressed sceptical views about climate
 change. and could we soon see the land speed record broken?



Document Name: BBCNEWS.201701.csv
Row no:  244
Snippet:  and during the election, he said he'd pull out of the paris clima
te deal. no-one yet knows what he'll do. while he's recently met with clima
te campaigner al gore, he's also just appointed several key cabinet members
 who've expressed sceptical views about climate change. and could we soon s
ee the land speed record broken?



Document Name: BBCNEWS.201702.csv
Row no:  107
Snippet:  it. the general idea is that he looks like he will be ruthless, t
hat things like climate change, foreign trade investment, and so on will go
 to hell and a high place. so even the fact that social security is
```

2) *x

```
-------------------------SEARCH ENGINE-----------------------------

Enter the Query : *llo
You have entered a wildcard query
Document Name: BBCNEWS.201907.csv
Row no:  744
Snippet:  this, experts say, it's down to climate change. this latest heatwave is likely
 to last the whole weekend, but there will be some quick relief as temperatures are set
to fall rapidly, early next week. stay with us on bbc world news, still to come: the was
hington monument turns into apollo eleven as america marks the 50th anniversary


Document Name: FOXNEWS.200907.csv
Row no:  382
Snippet:  are the people that want to do the job, here are the finances, here is the tec
hnology, go ahead and do it. greta: the apollo model might hold promise for solving prob
lems like energy and climate change. nasa is still facing a gap in


Document Name: FOXNEWS.201204.csv
Row no:  10
Snippet:  has nasa lowered its scientific standards to score political points? joining u
s from houston, former apollo 7 astronaut walter cunningham. he signed a letter to nasa'
s chief condemning the agency's position on climate change.


Document Name: FOXNEWS.201907.csv
Row no:  156
Snippet:  griff: pete's mentioned climate change. the 2020 campaign trail yesterday, of
course, historic moment. fifty years since one of the greatest achievements in our natio
n's history, the apollo 11 mission. elizabeth warren is comparing her plan, the green ne
w deal --  _


Document Name: FOXNEWS.201907.csv
Row no:  233
Snippet:  to fight climate change. what do you think about this, jesse watters? jesse: t
hey just set apollo program is racist and sexist last week so i don't know if that is a
comparison for the democrats. cory booker scares me. he looked hysterical -- tyrence: he
 was fired up.
```

3) x*y

```
-------------------------SEARCH ENGINE-----------------------------

Enter the Query : h*llo
You have entered a wildcard query
Document Name: BBCNEWS.201703.csv
Row no:  39
Snippet:  hello, i'm ros atkins, this is outside source. let's look through some of the main
 stories here in the bbc newsroom. america is expanding its presence in syria. it's sending
400 extra marines to support local militia trying to drive the islamic state group out of ra
qqa. the new head of the us environmental protection agency claims carbon dioxide emissions
are not a major factor in climate change.



Document Name: BBCNEWS.201703.csv
Row no:  183
Snippet:  hello, i'm ros atkins, this is outside source. in the past few minutes, donald tru
mp has signed a new executive order. rolling back a raft of obama's climate change policies
including restrictions on coal-fired power stations. the today's executive action i take



Document Name: BBCNEWS.201704.csv
Row no:  93
Snippet:  to industrial development. but on top of that, there is climate change. bringing h
igher temperatures, that makes bleaching more likely. time for a look at the weather now - h
ere's matt taylor. hello. fine end to monday across most parts



Document Name: BBCNEWS.201704.csv
Row no:  76
Snippet:  also tens of thousands of climate change protesters in several us cities across th
e country. that is it from me in the team. goodbye for now. hello, there. good evening.



Document Name: BBCNEWS.201705.csv
Row no:  268
Snippet:  hello. you are with bbc world news. the headlines: france's president-elect, the p
ro-european centrist emmanuel macron, has promised to unite the country, restore confidence
```

Interpretation of efficiency

Elasticsearch results are considered as the relevant results for a given query. To determine efficiency, the mean average values of Precision, Recall and F measure are computed for the top 10 results.

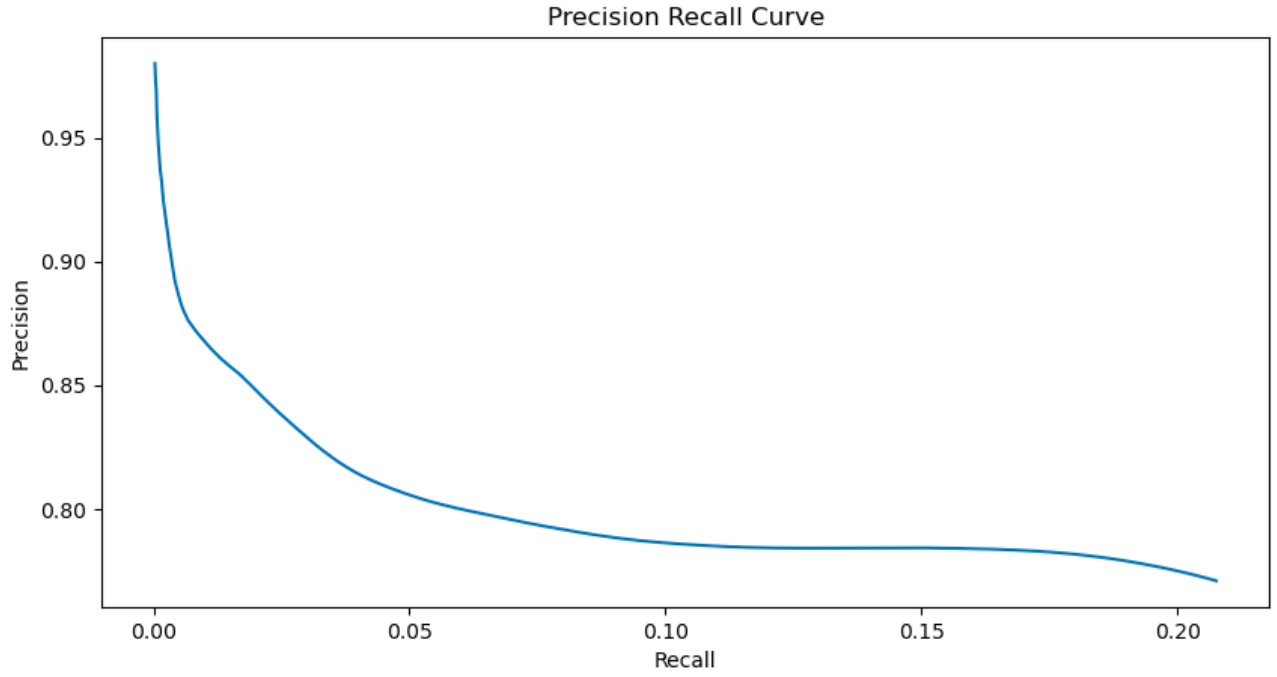
```
#Precision Recall F measure for free text
evaluationmeasure(retrievedfree)
```

```
Precision:  0.42947368421052634
Recall:  0.43683083511777304
F measure:  0.43312101910828027
```

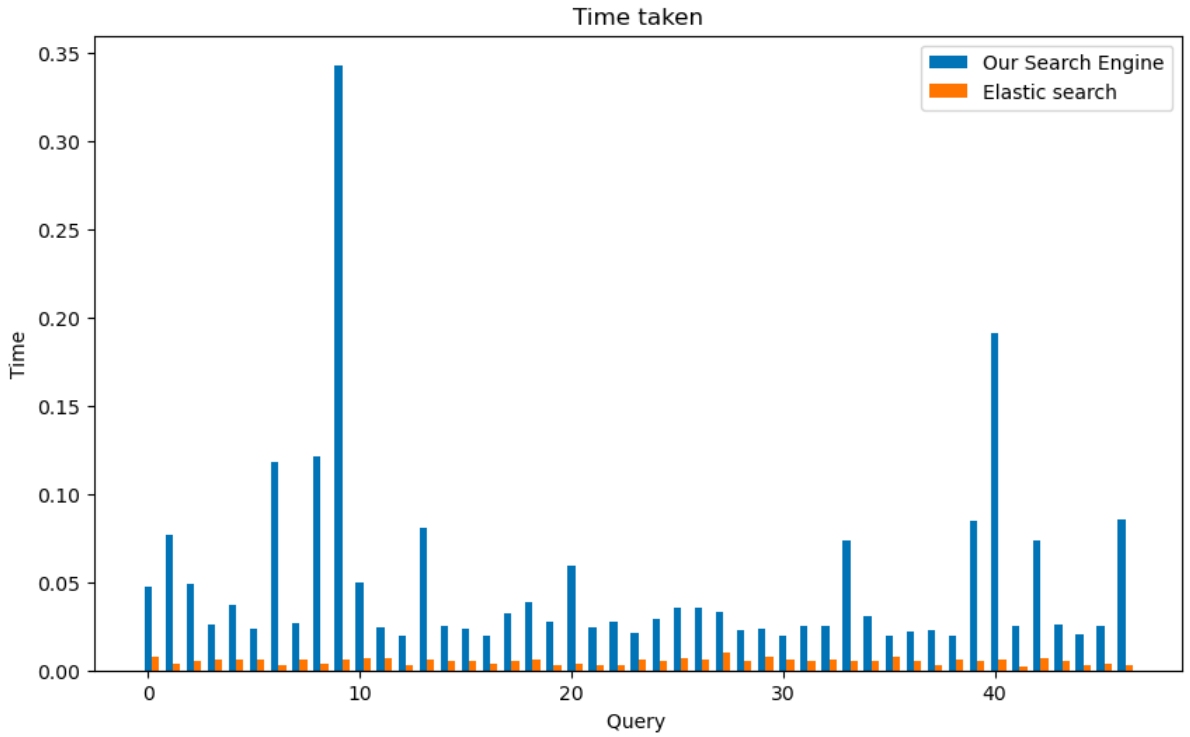
```
#Precision Recall F measure for Phrase text
evaluationmeasure(retrievedphrase)
```

```
Precision:  0.5876288659793815
Recall:  0.24411134903640258
F measure:  0.3449319213313162
```

Precision Recall Curve for 5000 results:



Comparison of time taken by our search engine and elasticsearch for the top ten results of 50 queries.



Learning Outcome

Our method uses hash tables for the creation of the inverted index. The index has been saved in a pickle file for faster retrieval by the engine. We performed three types of queries - free text , phrase and wildcard. For free text, we ensured that while calculating the union the end results were unique and no snippets were repeated to decrease redundancy. We experimented with different ranking techniques such as BM25 and TF IDF with cosine similarity and chose TF IDF with cosine similarity for our search engine with the sparse vectors for each snippet precomputed and stored as a pickle for increasing the retrieval time. Our search engine generates an output which has a precision of 0.58(for phrase queries, with elasticsearch considered as the benchmark) for a set of chosen 50 queries , ranking the top 10 results. As future work , we could experiment with different methods of ranking and compare the time and accuracy for the same. Gaining a hands on insight into the way search engines build search engines build indices, rank and query documents has helped us gain a deeper insight into the working of a search engine, and is just a step forward into understanding the underlying complexity in popular search engines such as Google, Elasticsearch, etc.

Name and Signature of the Faculty