CE-706 –Information Retrieval

Assignment 2: Elasticsearch and Evaluation

Rohini Subramaniam – 1908736

Srinidhi Karthikeyan -1900637

Introduction

The task assigned with this coursework is to design a model to search and retrieve documents from the **signal media one million news article**. We have used Elastic Search as the search engine, json file as the data set, Kibana for visual representation, search and retrieval of and the programming language python is used to access the Elasticsearch search engine. The components used in this assignment are briefly explained below.

Elasticsearch:

- Elasticsearch is a search engine.
- It is based on the Lucene library and is developed in JAVA.
- It gives the user a distributed, multitenant-capable full-text search engine which can be accessed using an HTTP web interface and schema-free JSON documents

Dataset:

- The dataset used is the **signal media one million news article.**
- The dataset contains one million articles mostly English articles but there are also non-English and multilingual articles.
- The articles are mainly from the major news organisations but other local news and blogs are also included.
- The data is stored in a JSONL format and each line is a JSON object representing one article.
- Fields:
    - **Id** - a unique identifier for the article
    - **Title** - the title of the article
    - **Content** - the textual content of the article
    - **Source** - the name of the article source
    - **Published -** the publication date of the article
    - **media-type -** either "News" or "Blog

Kibana:

- Kibana is an open source data visualization dashboard for Elasticsearch.
- It provides visualization methods like plots, charts, maps on huge volume of data of the indexed content on an elasticsearch cluster.
-

Software Requirements

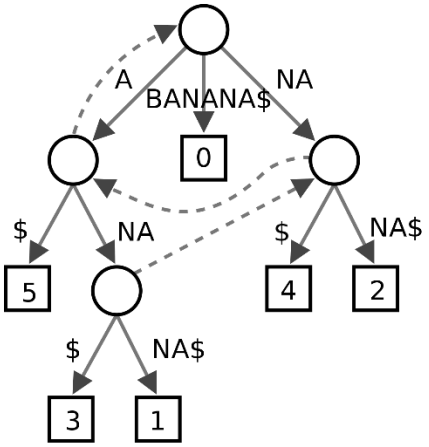
- **Java Runtime Environment:**
    - The basic requirement is to have Java Runtime Environment (java 7 or above) in the device.
    - System variable JAVA_HOME should be pointing to install location of JDK
- **Installing Elasticsearch:**
    - Download zip file from www.elastic.co for the required Operating System
    - Unzip downloaded file in your local file system

    - Run elasticsearch.bat file from the bin folder, set environment variable to the bin folder to directly run from the command prompt
    - To check if elastic search is running correctly, open web browser and launch the URL http://localhost:9200.
- **Installing Kibana:**
    - Download Kibana for required OS from same https://www.elastic.co/downloads/kibana page.
    - Unzip the downloaded file in your local file system
    - Run kibana.bat file (for windows) from the bin folder, set environment variable to the bin folder to directly run from the command prompt
    - Edit the kibana.yml file in config folder to point the elastic search to the URL http://localhost:9200 (if this entry is already available, just uncomment the line). This indicates kibana will connect to the elastic search during startup at the given URL
    - Run elastic search as indicated in the previous section
    - To check if kibana is installed correctly check the localhost:5601 on web browser
- **Run python code:**
    - Install elasticsearch by running the command Pip install elasticsearch in the command line.
    - Install Json by running the command pip install json in the command line.
    - Install Json_lines by running the command pip install Json-lines in the command line.
    - Specify the directory of the json file (dataset) in the part1 python file.

Indexing

- An index is a data structure for storing the mapping of fields to the corresponding documents.
- Index (indices) is / are defined on single / multiple fields, when a search query is run, the search engine executes a search based on the index (indices) instead of traversing through the entire content.
- The objective is to allow faster searches; however, this may potentially lead to the expense of increased memory usage and pre-processing time.
- But the index is stored in an additional storage which will increase the time taken to update but there is trade off with the time that is saved because of indexing.
- Following factors to be considered while designing the index / indices
    - Merge factors: Factors like how the data enters the index, how indexer traverses.
    - Storage Techniques: Whether the index data is compressed or filtered
    - Index Size: How much storage is used for storing the index.
    - Lookup Speed: How quickly the indexer can traverse through the index
    - Maintenance: How the index is maintained over time.
    - Fault Tolerance: How reliable the stored index are and how good they can handle faults

- Index Data Structures:
  - **Suffix tree:** It is type of compressed tree that contains all the suffixes of a text as their keys and positions in the text as their values.

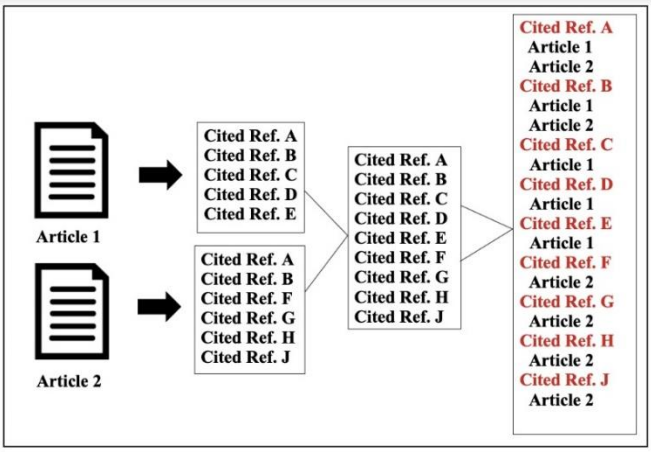


  - **Inverted Index:** The tokens are mapped to document ids in the form of hash table or binary tree.

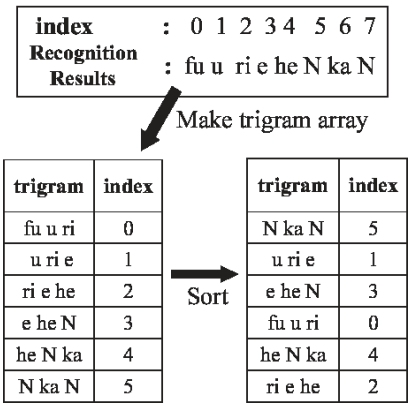
Token	Document Id
Harry	1, 2
Potter	1, 2
And	1, 2
The	1, 2
Half	1
Blood	1
Prince	1
Deathly	2
Hallows	2

Inverted index

  - **Citation index:** Stores citations or hyperlinks between documents.



  - **Ngram index:** Stores sequences of length of data to support other types of retrieval



- **Document-term matrix:** stores occurrences of words as a matrix.

Document Term Matrix

	intelligent	applications	creates	business	processes	bots	are	i	do	intelligence
Doc 1	2	1	1	1	1	0	0	0	0	0
Doc 2	1	1	0	0	0	1	1	0	0	0
Doc 3	0	0	0	1	0	0	0	1	1	1

Approach

The dataset used for this assignment is huge in size (approximately 3 GB) with millions of news articles. As it's a significantly large dataset, we considered to upload and index only 5000 articles in elastic search. The curtailed dataset is named as news_article. As indicated in the introduction, Elastic Search is used as the search engine, Kibana for visual representation, search and retrieval of and the programming language python is used to access the Elasticsearch search engine.

The dataset is loaded into kibana by indexing the pattern developed while setting up the elasticsearch connection. Here we index the pattern news_article.

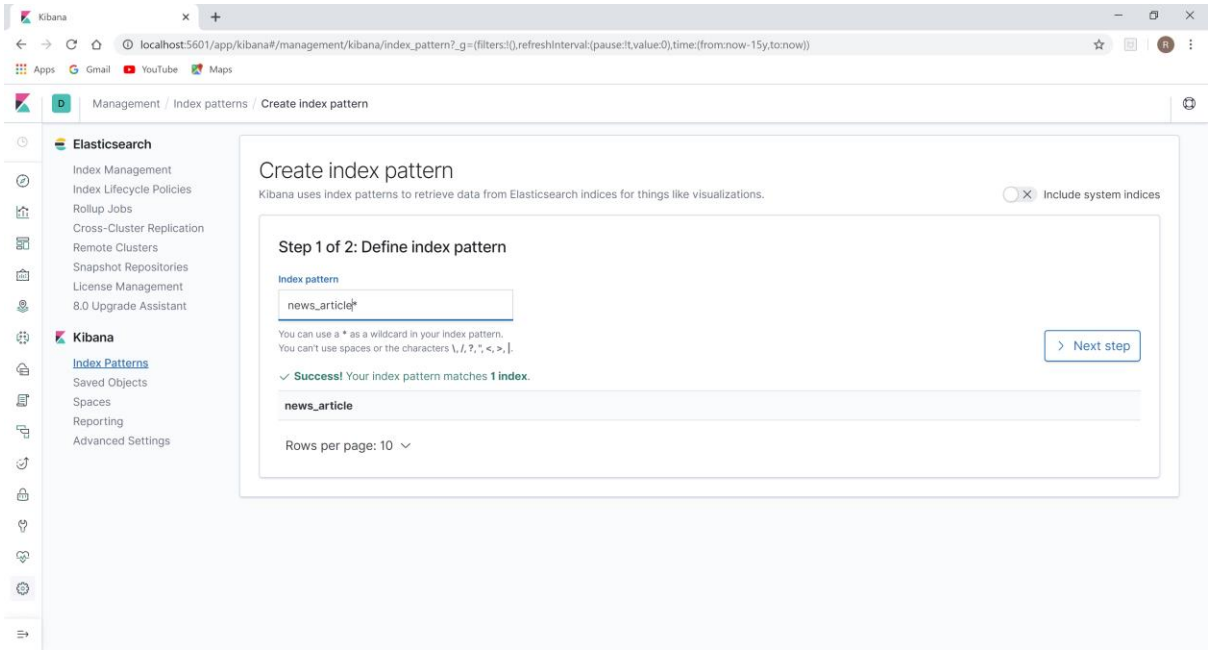


Figure 1 : Creating Index pattern

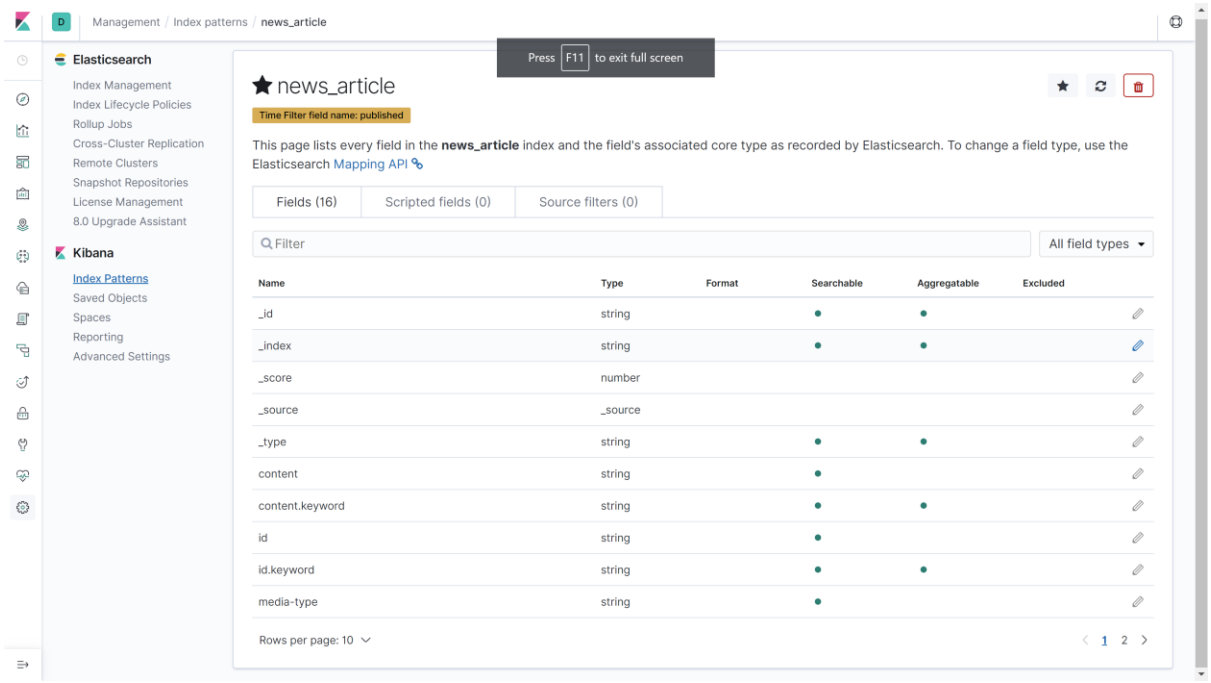


Figure 2: Fields in news_article

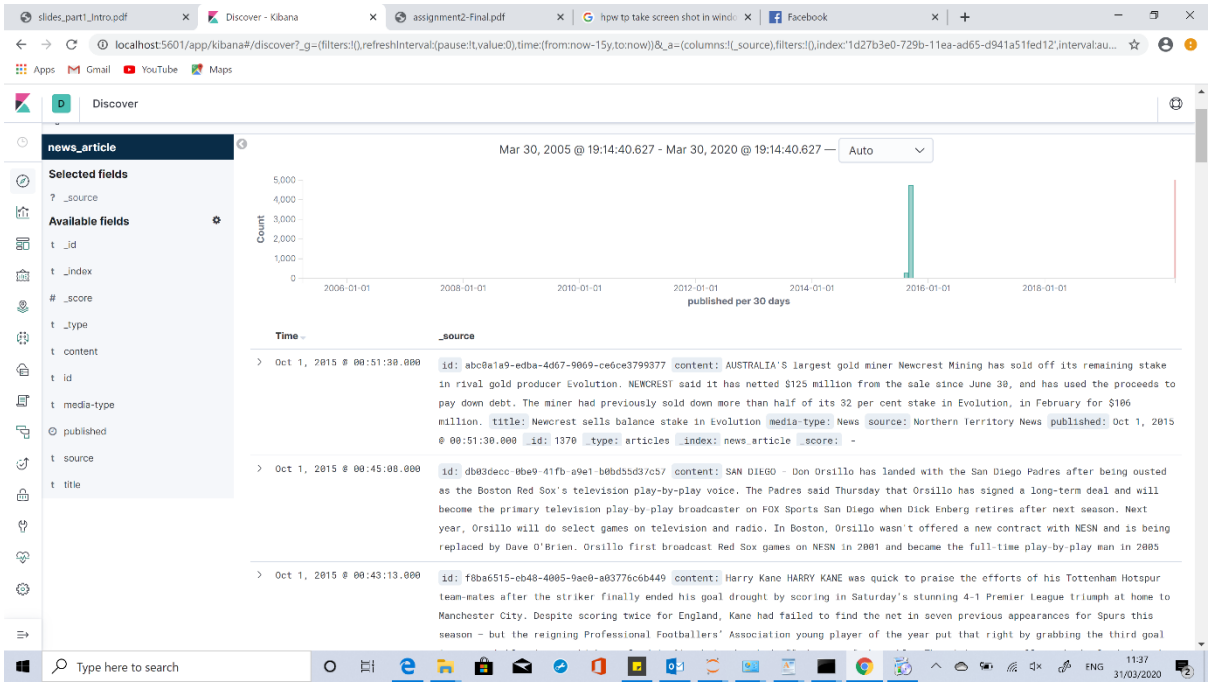
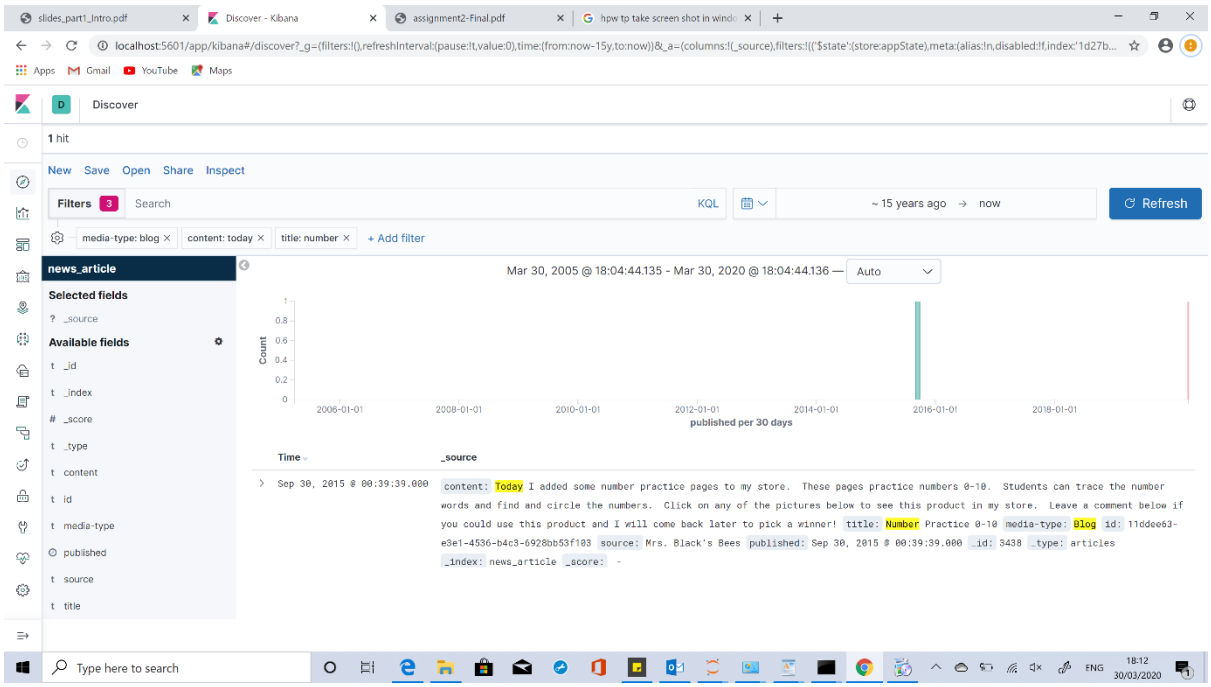


Figure 3: indexed news_article

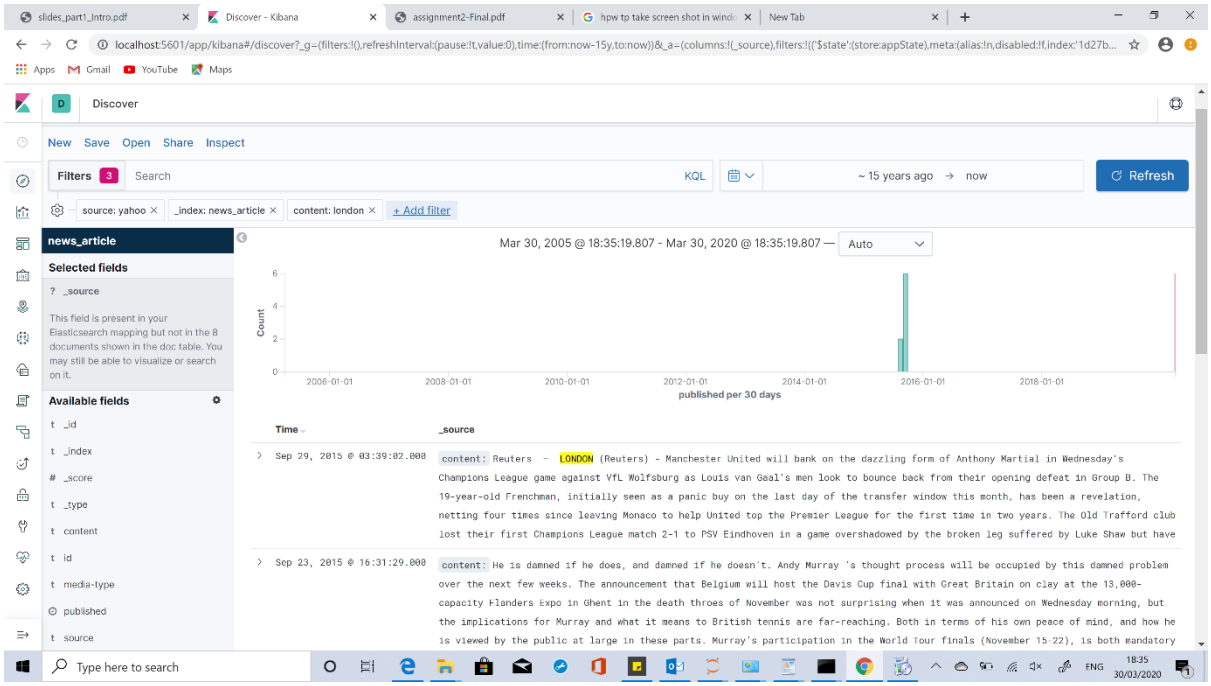
SEARCHING:

Once the dataset is loaded in Kibana, user can perform many interactive functions to search the data, with various filter / sort options provided in Kibana.
Following are the examples for retrieving data using various fields / indices.

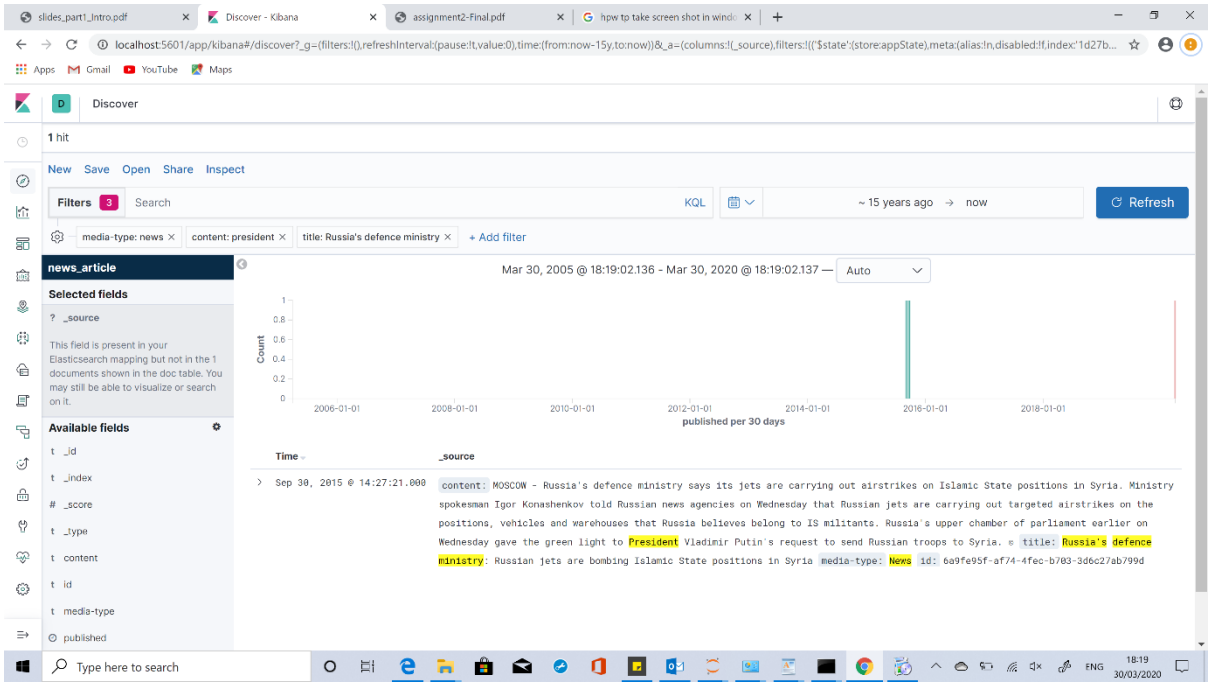
1. We have retrieved the articles where the **Media-type** is Blog, **content** is today, **title** is number



2. The following figure demonstrates the data retrieved while searching with the search criteria - **source**:yahoo, **_index**:News_article, **content**:London.



3. The below figure illustrates the results retrieved for the search criteria **media-type:** news, **content:** president, **title:** Russia's defense ministry.



TEST COLLECTION AND EVALUATION

Based on the news articles we considered from the dataset, we selected 10 events that the user may be interested to search and explain with examples on how the test results can be obtained.

Evaluation method

Precision and recall is used for evaluation metric. Precision is the number of relevant documents retrieved divided by the total number of documents retrieved by the query, whereas, recall is the number of relevant document retrieved divided but the total number of relevant documents in the data.

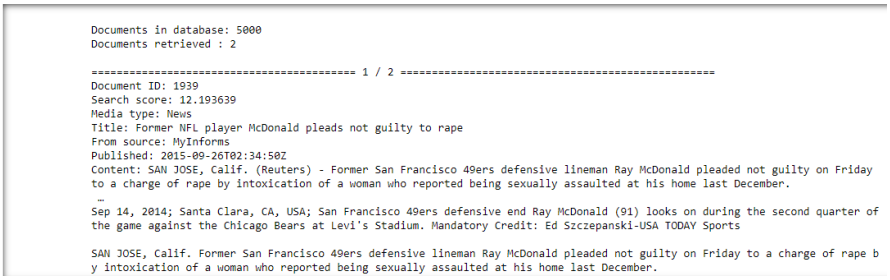
Event 1:
The user may want to look for sexual encounter from the content and that can be retrieved using the following query.

Query:

```
"query":{
   "match":{
      "content":{
         "query": "sexual encounter",
         "operator": "and"
      }
   }
}
```

Result:

There are two relevant documents in the collection and the query correctly bring them.



```
Documents in database: 5000
Documents retrieved : 2

========================================= 1 / 2 ================================================
Document ID: 1939
Search score: 12.193639
Media type: News
Title: Former NFL player McDonald pleads not guilty to rape
From source: MyInforms
Published: 2015-09-26T02:34:50Z
Content: SAN JOSE, Calif. (Reuters) - Former San Francisco 49ers defensive lineman Ray McDonald pleaded not guilty on Friday
to a charge of rape by intoxication of a woman who reported being sexually assaulted at his home last December.
...
Sep 14, 2014; Santa Clara, CA, USA; San Francisco 49ers defensive end Ray McDonald (91) looks on during the second quarter of
the game against the Chicago Bears at Levi's Stadium. Mandatory Credit: Ed Szczepanski-USA TODAY Sports

SAN JOSE, Calif. Former San Francisco 49ers defensive lineman Ray McDonald pleaded not guilty on Friday to a charge of rape b
y intoxication of a woman who reported being sexually assaulted at his home last December.
```

Evaluation:

We have different precision and recall for different document retrieval.

```
@K 1939 | P= 0.0005157 | R= 0.0002 Document id: 1939

@K 2504 | P= 0.0007987 | R= 0.0004 Document id: 2504
```

Event 2:

The user may want to look for Dating Apocalypse from the content and that can be retrieved using the following query.

Query:

```
"query":{
  "match":{
    "content":{
      "query":"sexual encounter",
      "operator":"and"
    }
  }
}
```

Result:

There is only one relevant document in the collection and the query correctly bring them.

```
Documents in database: 5000
Documents retrieved : 1

======================================== 1 / 1 ================================================
Document ID: 4292
Search score: 2.5976334
Media type: Blog
Title: London Film Festival 2015: Film Line-up: STEVE JOBS, SUFFRAGETTE, CAROL
From source: FilmBook
Published: 2015-09-01T16:25:25Z
Content: Steve Jobs, Suffragette, Carol, and the other gala and competition films for 2015 London Film Festival have been ann
ounced. The 59th annual London Film Festival, run by the British Film Institute, “is the UK’s largest public film event,… scr
eening a total of 238 fiction and documentary features, including 16 World Premieres, 8 International Premieres, 40 European
Premieres and 11 Archive films including 5 Restoration World Premieres…There will also be screenings of 182 live action and a
nimated shorts…The Festival showcases the best of world cinema to champion creativity, originality, vision and imagination, a
nd presents the finest contemporary international cinema from both established and emerging film-makers…the festival hosts hi
gh profile awards contenders, screens recently restored archive films, champions new discoveries and combines curatorial stre
ngth with red carpet glamor. It also provides an extensive program of industry events, public forums, education events, lectu
```

Evaluation:

We got a good precision but the recall is not that good because we are restricting the results using the “and” operator.

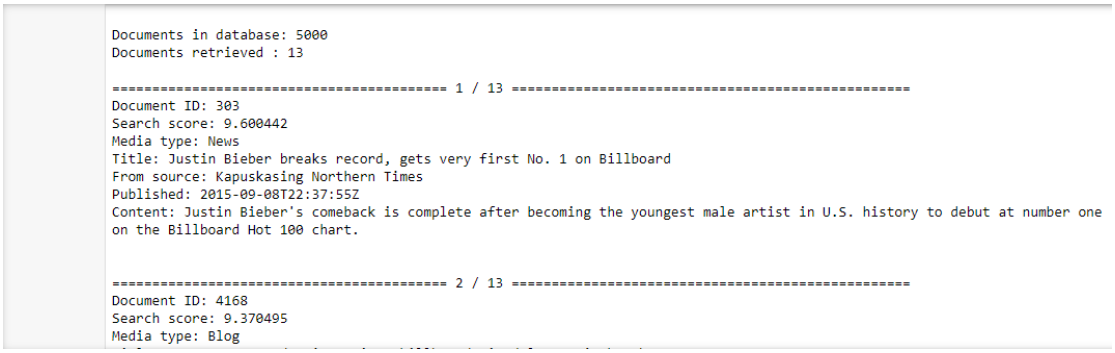
```
@K 4292 | P= 0.000233 | R= 0.0002 Document id: 4292
```

Event 3:
The user may want to know about the billboard charts from the content and that can be retrieved using the following query.

Query:

```
{
    "query": {
      "bool": {
        "must": [
          {
            "match": {
              "content": "billboard"
            }
          }
        ]
      }
    }
}
```

Result:
There are thirteen relevant documents in the collection and the query correctly bring them.



```
Documents in database: 5000
Documents retrieved : 13

========================================= 1 / 13 ==================================================
Document ID: 303
Search score: 9.600442
Media type: News
Title: Justin Bieber breaks record, gets very first No. 1 on Billboard
From source: Kapuskasing Northern Times
Published: 2015-09-08T22:37:55Z
Content: Justin Bieber's comeback is complete after becoming the youngest male artist in U.S. history to debut at number one
on the Billboard Hot 100 chart.


========================================= 2 / 13 ==================================================
Document ID: 4168
Search score: 9.370495
Media type: Blog
```

Evaluation:
We got a good precision but the recall is not that good because we are restricting the results because we are considering only a part of the total dataset.

```
@K 303 | P= 0.0066007 | R= 0.0004 Document id: 303
```

Event 4:
The user may want to know about the Business Review Albany from the content and that can be retrieved using the following query.

Query:

```
{
    "query": {
      "bool": {
        "must": [
          {
            "match_phrase_prefix": {
              "source": "Business Review Albany"
            }
          }
        ]
      }
    }
  }
```

Results:
There is only one relevant document in the collection because the match_phrase query type is used which is a strict method and the query correctly bring them.

```
Documents in database: 5000
Documents retrieved : 1

======================================== 1 / 1 ================================================
Document ID: 3340
Search score: 15.8747635
Media type: News
Title: Apple unveils large-screen iPad Pro, new iPhones (Video)
From source: Business Review Albany
Published: 2015-09-09T18:46:02Z
Content: Apple announced a variety of new products at its event Wednesday.

The company unveiled the iPad Pro, an update to its tablet that can display a full-sized virtual keyboard and sports a nearly
13-inch screen with more pixels than a Retina Macbook. The Pro boasts 10 hours of battery life, displays 5.6 million pixels,
weighs 1.57 pounds, is just 6.9 mm thick, contains a four-speaker sound system, and uses a new chip that is nearly twice as f
ast as the iPad Air 2.
```

Evaluation:

We got a good precision but the recall is not that good because we are restricting the results using the match_phrase_prefix type.

```
@K 303 | P= 0.0066007 | R= 0.0004 Document id: 303
```

Event 5:
The user may want to know about the iPad pro from the content of media type News and that can be retrieved using the following query.

Query:

```
"query": {
        "bool": {
          "must": [
            {
              "match_phrase_prefix": {
                "content": "iPad Pro"
              }
            },
            {
              "match_phrase_prefix": {
                "media-type": "News"
              }
            },
          ]
        }
    }
```

Result:

There is only one relevant document in the collection because the "match_phrase_prefix" and the search is specifically about the media type News query type is used which is a strict method and the query correctly bring them.

```
Media-type:News
Content keyword:iPad Pro

Documents in database: 5000
Documents retrieved : 5

======================================== 1 / 5 ========================================
Document ID: 3340
Search score: 341.52222
Media type: News
Title: Apple unveils large-screen iPad Pro, new iPhones (Video)
From source: Business Review Albany
Published: 2015-09-09T18:46:02Z
Content: Apple announced a variety of new products at its event Wednesday.

The company unveiled the iPad Pro, an update to its tablet that can display a full-sized virtual keyboard and sports a nearly
13-inch screen with more pixels than a Retina Macbook. The Pro boasts 10 hours of battery life, displays 5.6 million pixels,
```

Evaluation:

We got a good precision but the recall is not that good because we are restricting the results using the match_phrase_prefix type.

```
@K 7 | P= 0.1428571 | R= 0.0002 Document id: 7
```

Event 6 – iphone model comparison

Query 1

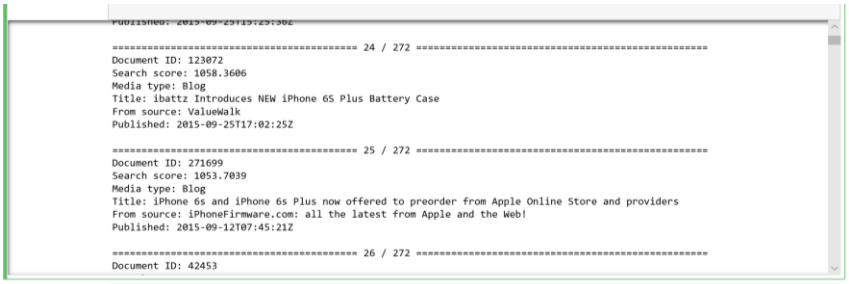
User would like to compare iphone 6 camera with iphone 6s

```
  {
     "query": {
      "match": {
          "content": {
           "query": "Compare iphone cameras",
           "operator": "and"
           }
        }
     }
  }
      , size=tdocument)
```

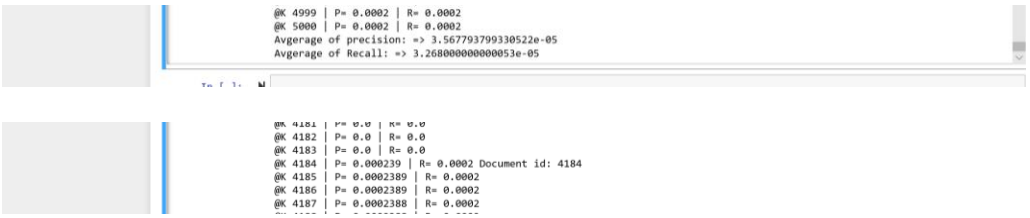
Query 2

User would like to retrieve information from blogs for iphone6s

```
  {
     "query": {
       "bool": {
          "must": [
             {
                "match_phrase_prefix": {
                   "content": "iphone 6s plus"
                }
             },
             {
                "match_phrase_prefix": {
                   "media-type": "Blog"

                }
             },

          ]
       }
     }
  }
             , size=tdocument)
```



Evaluation



Event 7 – Job Market

Query 1

User would like to check information on US job market between 2015/01/01 to 2016/31/12

```
{
      "query":{
        "bool":{
          "must":[
             {
                "match_phrase_prefix":{
                "content":
                "US Job Market"
             }
             },
             {
             "range":{
                "published":{
                "gte":
                "2010/01/01"
             ,
             "lte":
             "2016/31/12"
             ,
             "format":
             "yyyy/mm/dd||yyyy"
               }
              }
             }

          ]
```

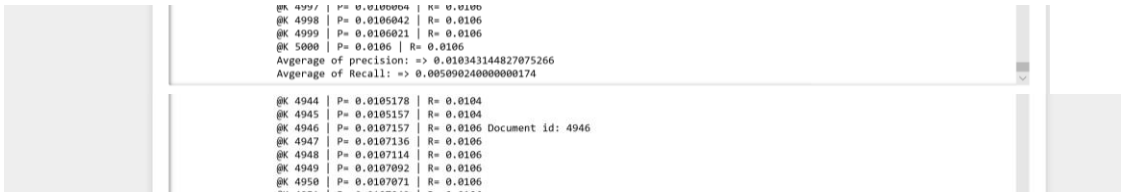
```
        }
      }
  }
            , size=tdocument)
```

Result

Documents were retrieved in the specified date range.

Evaluation

It can be observed precision is better than recall in this search



Query 2

If the user types in wrong spelling for” market”

```
    "query": {
      "fuzzy": {
        "content": {
          "value" : "mratek",
          "fuzziness": "AUTO"
        }

      }
    }
  }
            , size=tdocument)
```

Event 8 – Medical Tourism
User would like to access information on medical tourism.

Query 1

```
{
   "query": {
     "bool": {
       "must": [
         {
           "match_phrase_prefix": {
             "title": "Medical Tourism"
           }
         },

       ]
     }
   }
}
          , size=tdocument)
```

Result

11 documents were retrieved with title containing medical tourism.

Evaluation

Precision and recall is zero because only 11 document is retrieved out of 5000 documents. It will be better if whole dataset is considered.

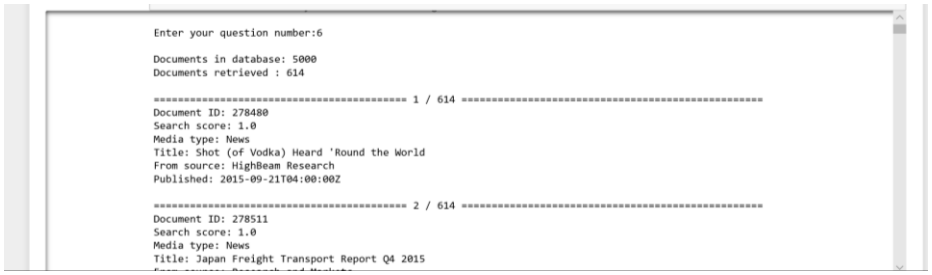
Query 2

User would like to find articles on research

```
{
   "query": {
      "wildcard": {
         "source":{
            "value":"research*",
            "boost": 1.0,
            "rewrite": "constant_score"

               }
      }

   }
}
```

Result

Documents with value “research” in source are retrieved.



EVALUATION
It cab boserved precision is better than recall.

Event 9- Social Tourism

User would like to get articles on new delhi published on 01/01/2015

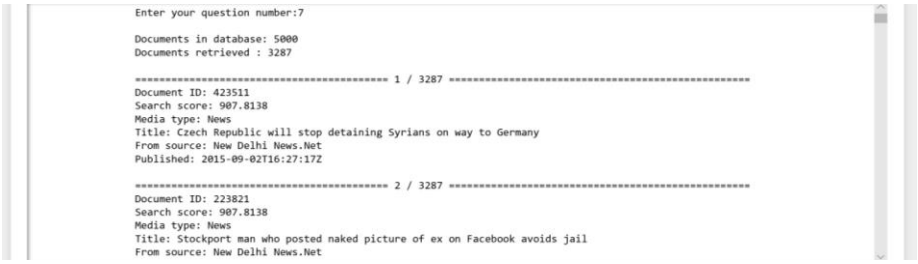
Query 1

```
{
   "query":{
      "bool":{
         "must":[
            {
               "match_phrase_prefix":{
               "content":
               "new delhi"
            }
```

```
            },
            {
            "range":{
               "published":{
               "gte":
               "2015/01/01"
            ,
            "format":
            "yyyy/mm/dd||yyyy"
              }
             }
            }

         ]
      }
   }
}
```

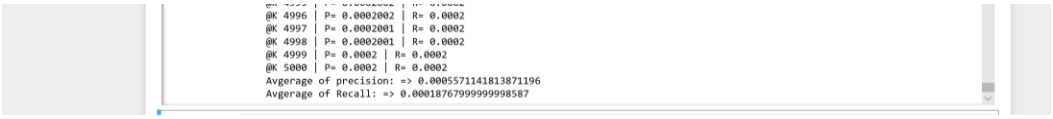
Result

Documents related to newdelhi published after 2015/01/01 were retrieved.



Evaluation

In this case precision is much better than recall



Query 2

User would like to get information on Tajmahal from blogs

```
{
    "query": {
```

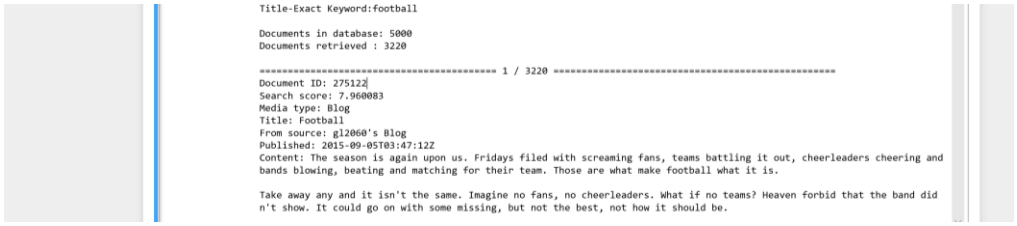
```
      "bool": {
        "must": [
          {
            "match_phrase_prefix": {
              "content": "Tajmahal"
            }
          },
          {
            "match_phrase_prefix": {
              "media-type": "Blog"

            }
          },

        ]
      }
    }
  }
          , size=tdocument)
```

Event 10 – Sports
User would like to get information on Foot ball

Query 1

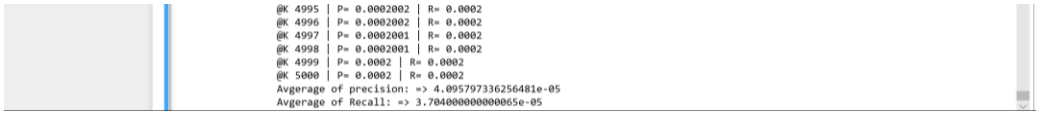
```
  {
    "query": {
      "term": {
        "title": “football”
      }
    }
  }
          , size=tdocument
```

Result
3220 documents related to football



Evaluation

It can be observed precision is better than recall



Query 2

User would like to get information on sports published between specific dates, say 2010/01/01 and 2016/31/12

{

“query”:{

“bool”:{

“must”:[

{

“match_phrase_prefix”:

{

“Title”:“Sports”

}

},

{

“range”:{

“publishes”:{

“gte”:

“2010/01/01”

,

“lte”:

“2016/31/12,

“format”:

“yyyy/mm/dd||yyyy”

}

}

}

}

]

}

}

}

Crowdsourcing:

The task was random videos were played continuously and we should press space bar whenever we see the video that was played previously. The second task was done after 24 hrs to check if we remember the videos, we watched the previous day. We were able to tell them correctly. But we also guessed few videos wrongly and those were the videos that were similar to the previous day videos. For example, a football video was played the previous day. But a similar but different video was played but we were confused and pressed the space bar. The task was useful to understand how much our brain can remember things.

Conclusion

The search engine, elasticsearch is built using python. The dataset provided was huge due to the unavailability of the university labs a collection of 5000 articles which is a mix of news and blog articles is chosen for the convenience of working in personal laptop. The 5000 articles was uploaded in elastic search. Kibana GUI was used to access the articles indexed in elastic search.A set of queries for different events were used to retrieve data depending on users requirement.The Evaluation metric Precision and recall was used to assess the efficiency of the query. It can be observed precision and recall can be increased with the increase in data size.

References:

Images:

1) https://en.wikipedia.org/wiki/Suffix_tree

2) https://stackoverflow.com/questions/33929377/what-exactly-does-the-data-structure-of-the-inverted-index-in-solr-looks-like

3) https://www.isko.org/cyclo/citation

4) https://www.darrinbishop.com/blog/2017/10/text-analytics-document-term-matrix/

5) https://www.semanticscholar.org/paper/Spoken-Term-Detection-by-N-gram-Index-with-Exact-Sakamoto-Nakagawa/0225141bf52373d2a6a852083e0db468e7bf9b35/figure/0

Content:

1) https://en.wikipedia.org/wiki/Search_engine_indexing
2) https://research.signal-ai.com/newsir16/signal-dataset.html