CE706 - Information Retrieval SU 2022

Assigment 1

Student ID: 2112102

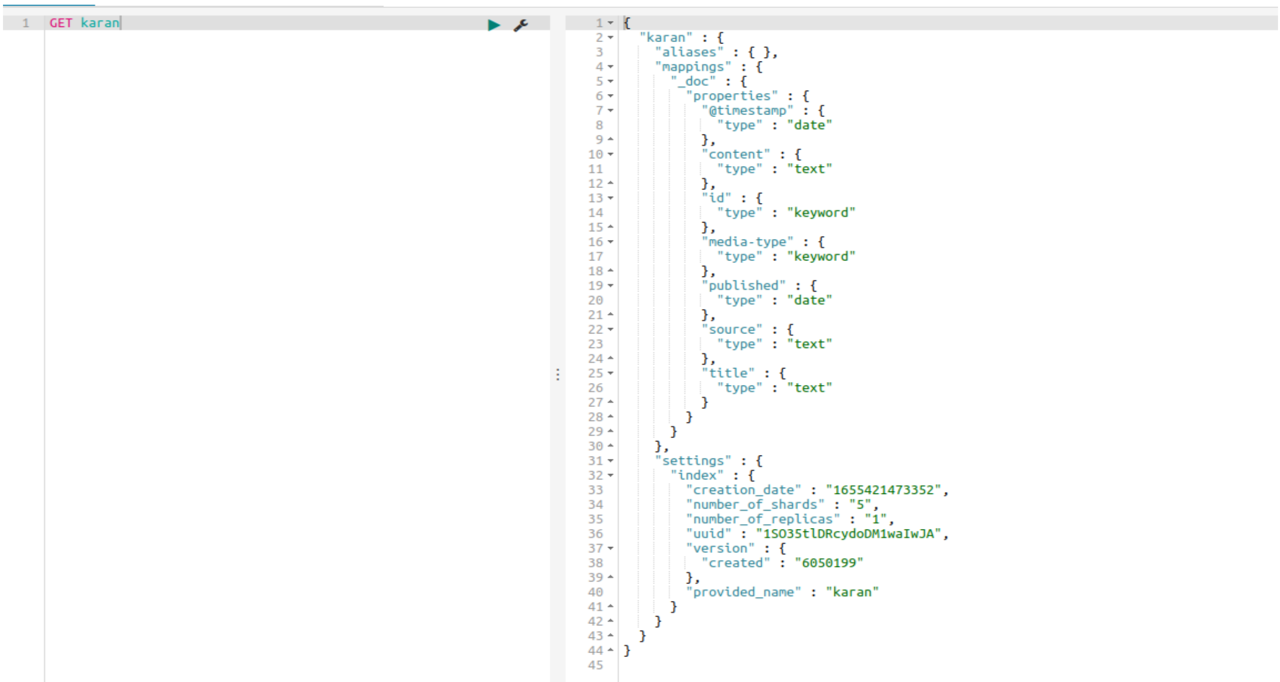
Instructions for running your system

I have Mentioned how I have performed various tasks of information retrivalby using online platforms elastic search and kibana, which are most commonly used softwares in industry. I have followed my lab work documents and had researched some to get better idea of these platforms. I have downloaded the dataset from the given site https://research.signal-ai.com/newsir16/signal-dataset.html. It is a signal media in which there are one million news articles gathered from different sources. After obtaining data I read that jsonl file in my code and then started performing the tasks. for that, I have splitted data into equal parts where each set had 1000 phrases only(for easy and fast analysing). After that I have started performing the tasks as asked eg. Data importing, Indexing, Tokenization, Stemming, Searching etc.

1. Indexing

With the help of indexing we can separate and retrieve the data from our dataset. Another concept is about replicas and shards. It is a mapping and it defines multiple types. Here I have got different indexing results using GET and PUT method.

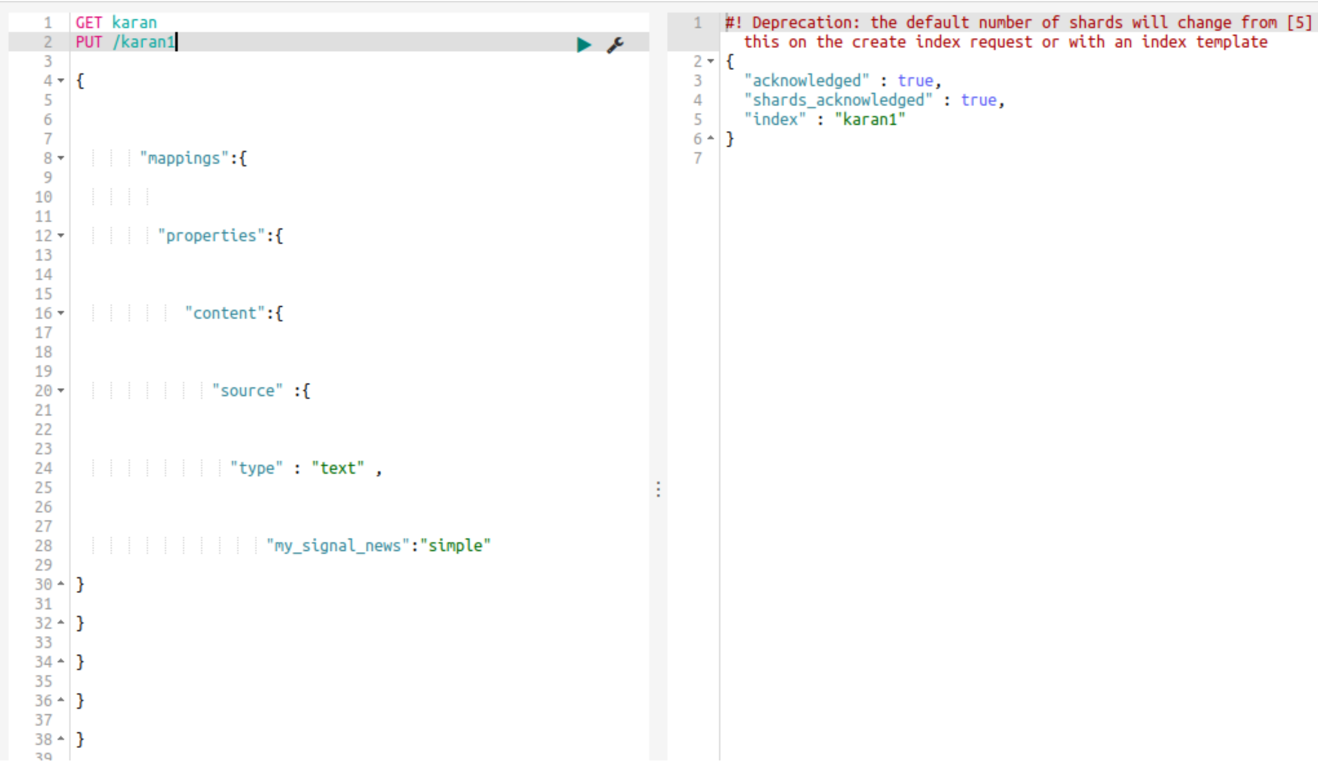
1.1 By using GET method:



```
GET karan
```

```
{
  "karan" : {
    "aliases" : { },
    "mappings" : {
      "_doc" : {
        "properties" : {
          "@timestamp" : {
            "type" : "date"
          },
          "content" : {
            "type" : "text"
          },
          "id" : {
            "type" : "keyword"
          },
          "media-type" : {
            "type" : "keyword"
          },
          "published" : {
            "type" : "date"
          },
          "source" : {
            "type" : "text"
          },
          "title" : {
            "type" : "text"
          }
        }
      }
    },
    "settings" : {
      "index" : {
        "creation_date" : "1655421473352",
        "number_of_shards" : "5",
        "number_of_replicas" : "1",
        "uuid" : "1SO35tlDRcydoDM1waIwJA",
        "version" : {
          "created" : "6050199"
        },
        "provided_name" : "karan"
      }
    }
  }
}
```

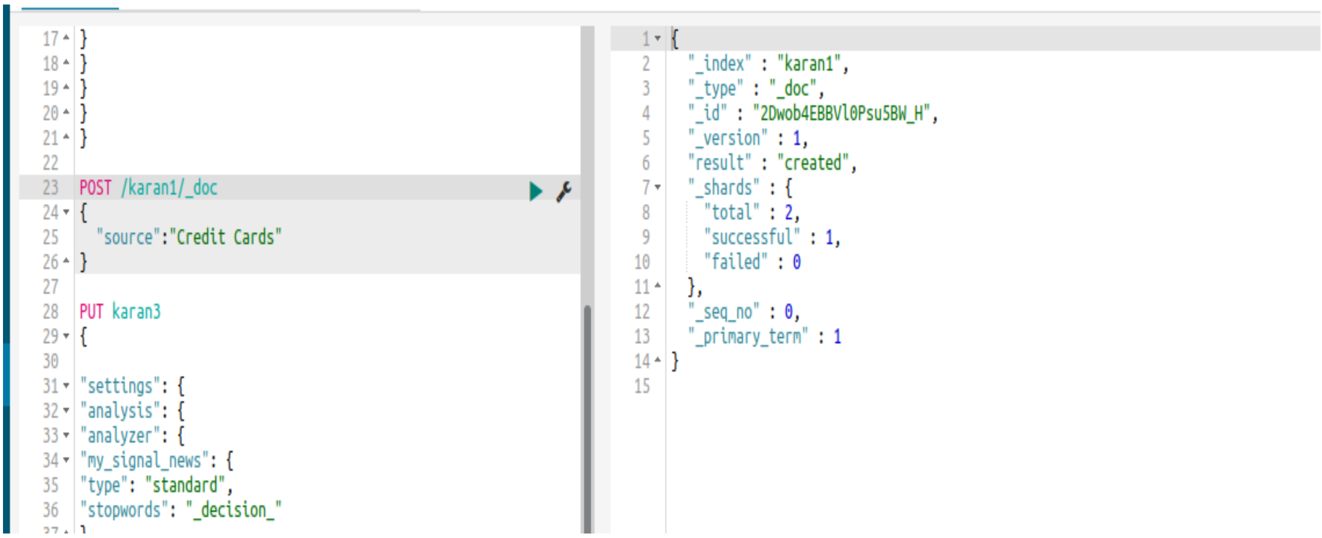
1.2 By using PUT:



```
GET karan
PUT /karan1

{


      "mappings":{


        "properties":{



          "content":{



            "source" :{




              "type" : "text" ,




                  "my_signal_news":"simple"

}

}

}

}

}
```

```
#! Deprecation: the default number of shards will change from [5]
  this on the create index request or with an index template

{
  "acknowledged" : true,
  "shards_acknowledged" : true,
  "index" : "karan1"
}
```

1.3 By using POST method



```
}
}
}
}
}

POST /karan1/_doc
{
  "source":"Credit Cards"
}

PUT karan3
{

"settings": {
"analysis": {
"analyzer": {
"my_signal_news": {
"type": "standard",
"stopwords": "_decision_"
}
```

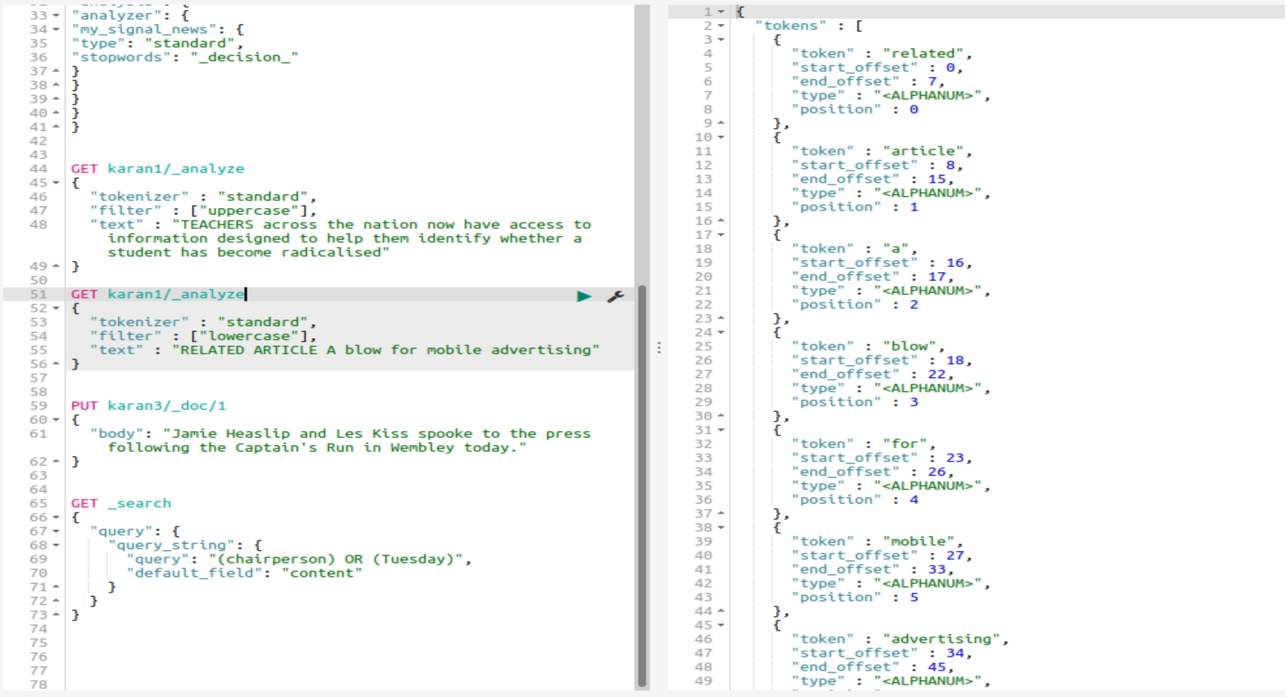
```
{
  "_index" : "karan1",
  "_type" : "_doc",
  "_id" : "2Dwob4EBBVl0Psu5BW_H",
  "_version" : 1,
  "result" : "created",
  "_shards" : {
    "total" : 2,
    "successful" : 1,
    "failed" : 0
  },
  "_seq_no" : 0,
  "_primary_term" : 1
}
```

2. Tokenization and Normalisation

The tokenizer breaks the data in individual tokens and gives output of stream of tokens. We use tokenization to protect the sensitive datawhen we are making preservation of utilities. Tokenization process is different than the encryption process in which all the sensitive data is been modified & stored.

2.1 Uppercase:

Here I have used uppercase filter to convert all the lowercase words in the word to the uppercase words.

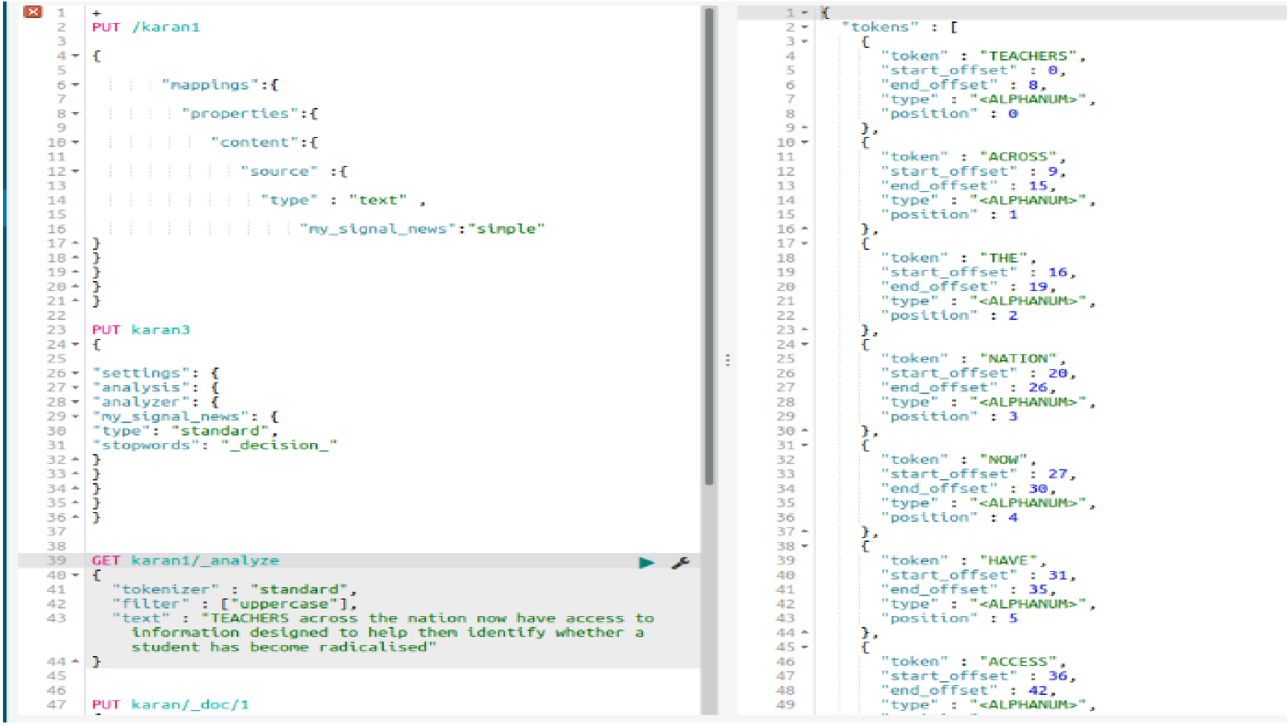


```
"analyzer": {
"my_signal_news": {
"type": "standard",
"stopwords": "_decision_"
}
}
}
}
}


GET karan1/_analyze
{
  "tokenizer" : "standard",
  "filter" : ["uppercase"],
  "text" : "TEACHERS across the nation now have access to
    information designed to help them identify whether a
    student has become radicalised"
}

GET karan1/_analyze
{
  "tokenizer" : "standard",
  "filter" : ["lowercase"],
  "text" : "RELATED ARTICLE A blow for mobile advertising"
}


PUT karan3/_doc/1
{
  "body": "Jamie Heaslip and Les Kiss spooke to the press
    following the Captain's Run in Wembley today."
}


GET _search
{
  "query": {
    "query_string": {
      "query": "(chairperson) OR (Tuesday)",
      "default_field": "content"
    }
  }
}
```

```
{
  "tokens" : [
    {
      "token" : "related",
      "start_offset" : 0,
      "end_offset" : 7,
      "type" : "<ALPHANUM>",
      "position" : 0
    },
    {
      "token" : "article",
      "start_offset" : 8,
      "end_offset" : 15,
      "type" : "<ALPHANUM>",
      "position" : 1
    },
    {
      "token" : "a",
      "start_offset" : 16,
      "end_offset" : 17,
      "type" : "<ALPHANUM>",
      "position" : 2
    },
    {
      "token" : "blow",
      "start_offset" : 18,
      "end_offset" : 22,
      "type" : "<ALPHANUM>",
      "position" : 3
    },
    {
      "token" : "for",
      "start_offset" : 23,
      "end_offset" : 26,
      "type" : "<ALPHANUM>",
      "position" : 4
    },
    {
      "token" : "mobile",
      "start_offset" : 27,
      "end_offset" : 33,
      "type" : "<ALPHANUM>",
      "position" : 5
    },
    {
      "token" : "advertising",
      "start_offset" : 34,
      "end_offset" : 45,
      "type" : "<ALPHANUM>",
```

2.2 Lowercase:

Here I have used lowercase filter to cover all the uppercase words to the lowercase words.

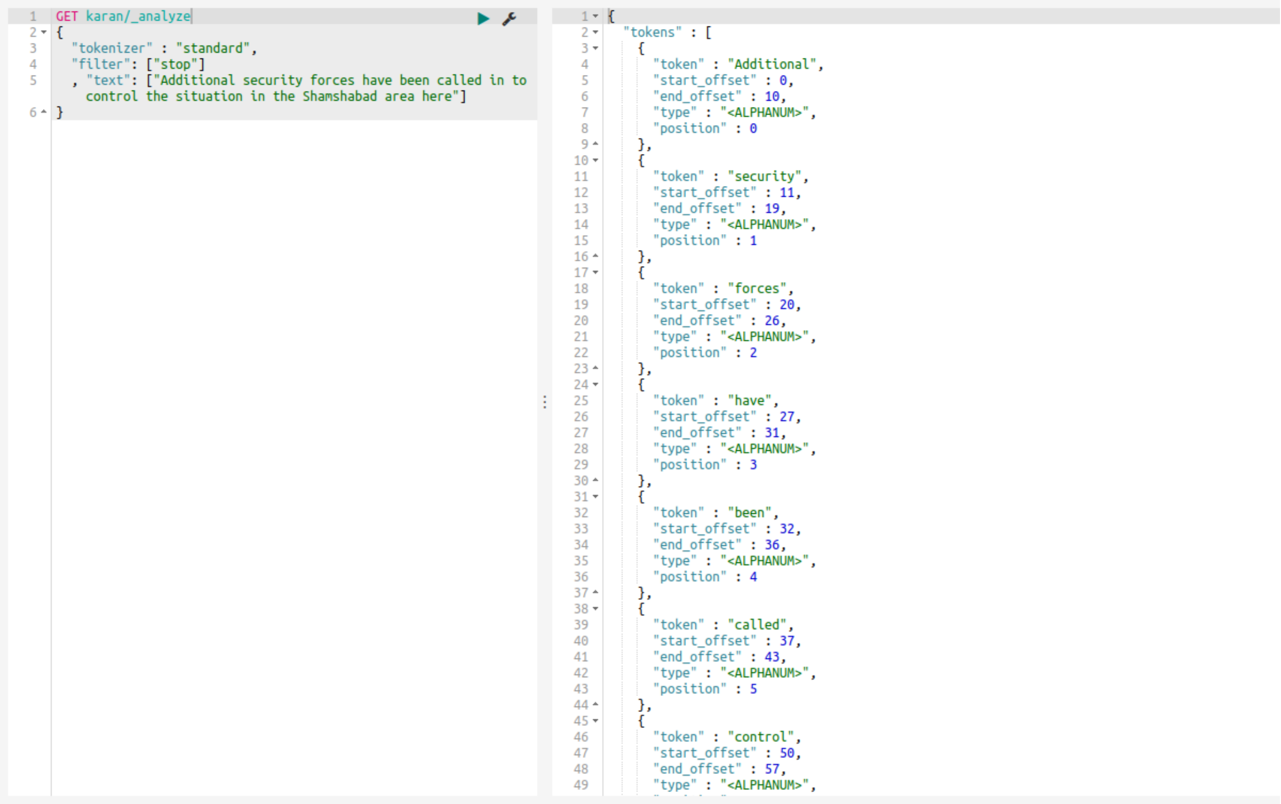


```
+
PUT /karan1

{

      "mappings":{

        "properties":{

          "content":{

            "source" :{

              "type" : "text" ,

                  "my_signal_news":"simple"
}
}
}
}
}

PUT karan3
{

"settings": {
"analysis": {
"analyzer": {
"my_signal_news": {
"type": "standard",
"stopwords": "_decision_"
}
}
}
}
}


GET karan1/_analyze
{
  "tokenizer" : "standard",
  "filter" : ["uppercase"],
  "text" : "TEACHERS across the nation now have access to
    information designed to help them identify whether a
    student has become radicalised"
}


PUT karan/_doc/1
```

```
{
  "tokens" : [
    {
      "token" : "TEACHERS",
      "start_offset" : 0,
      "end_offset" : 8,
      "type" : "<ALPHANUM>",
      "position" : 0
    },
    {
      "token" : "ACROSS",
      "start_offset" : 9,
      "end_offset" : 15,
      "type" : "<ALPHANUM>",
      "position" : 1
    },
    {
      "token" : "THE",
      "start_offset" : 16,
      "end_offset" : 19,
      "type" : "<ALPHANUM>",
      "position" : 2
    },
    {
      "token" : "NATION",
      "start_offset" : 20,
      "end_offset" : 26,
      "type" : "<ALPHANUM>",
      "position" : 3
    },
    {
      "token" : "NOW",
      "start_offset" : 27,
      "end_offset" : 30,
      "type" : "<ALPHANUM>",
      "position" : 4
    },
    {
      "token" : "HAVE",
      "start_offset" : 31,
      "end_offset" : 35,
      "type" : "<ALPHANUM>",
      "position" : 5
    },
    {
      "token" : "ACCESS",
      "start_offset" : 36,
      "end_offset" : 42,
      "type" : "<ALPHANUM>",
```

3. Selecting Keywords

Keywords are the words which are frequently used by the user while searching in the search engine. Here I have used stop filter, It is used to eliminate stop words. Here analyzer makes the token of the given input to give better results.

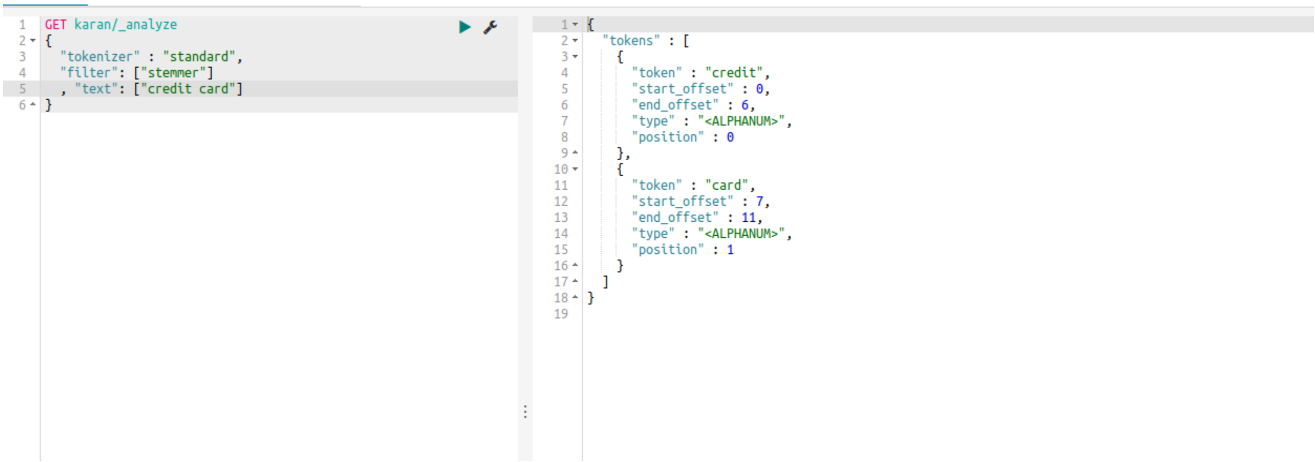


```
GET karan/_analyze
{
  "tokenizer" : "standard",
  "filter": ["stop"]
  , "text": ["Additional security forces have been called in to
    control the situation in the Shamshabad area here"]
}
```

```
{
  "tokens" : [
    {
      "token" : "Additional",
      "start_offset" : 0,
      "end_offset" : 10,
      "type" : "<ALPHANUM>",
      "position" : 0
    },
    {
      "token" : "security",
      "start_offset" : 11,
      "end_offset" : 19,
      "type" : "<ALPHANUM>",
      "position" : 1
    },
    {
      "token" : "forces",
      "start_offset" : 20,
      "end_offset" : 26,
      "type" : "<ALPHANUM>",
      "position" : 2
    },
    {
      "token" : "have",
      "start_offset" : 27,
      "end_offset" : 31,
      "type" : "<ALPHANUM>",
      "position" : 3
    },
    {
      "token" : "been",
      "start_offset" : 32,
      "end_offset" : 36,
      "type" : "<ALPHANUM>",
      "position" : 4
    },
    {
      "token" : "called",
      "start_offset" : 37,
      "end_offset" : 43,
      "type" : "<ALPHANUM>",
      "position" : 5
    },
    {
      "token" : "control",
      "start_offset" : 50,
      "end_offset" : 57,
      "type" : "<ALPHANUM>",
```

4. Stemming or Morphological Analysis

Stemmer filter reduces the word and convert it to its root form. It takes care if the word variants matches during the search. It is a normalisation technique used in Neural Language Processing. Stemming helps to reduce the dimancity of data.

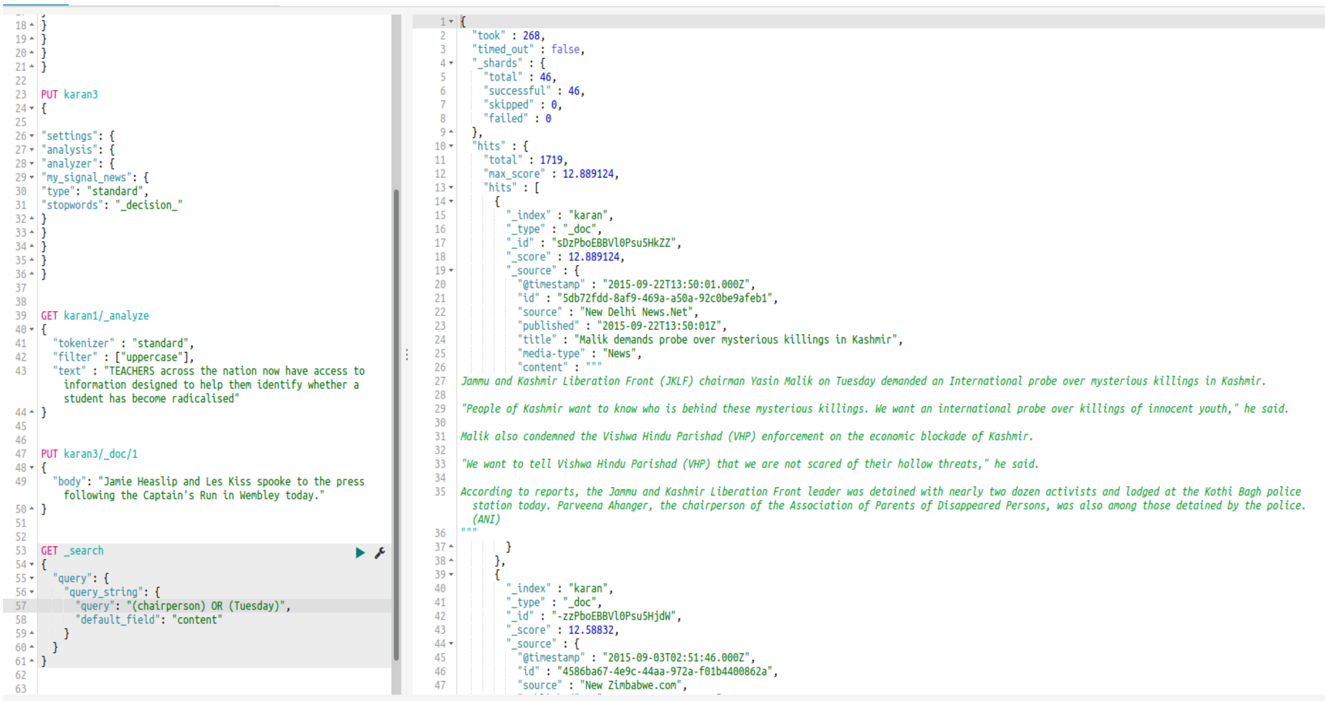


```
GET karan/_analyze
{
  "tokenizer" : "standard",
  "filter": ["stemmer"]
  , "text": ["credit card"]
}
```

```
{
  "tokens" : [
    {
      "token" : "credit",
      "start_offset" : 0,
      "end_offset" : 6,
      "type" : "<ALPHANUM>",
      "position" : 0
    },
    {
      "token" : "card",
      "start_offset" : 7,
      "end_offset" : 11,
      "type" : "<ALPHANUM>",
      "position" : 1
    }
  ]
}
```

5. Searching

This search API helps to seach the item in elastic search. Here I have fired two quries and it gives result according to how many tokens are made of those words. It can also gives data of past times like which things or network was ran or searched by the user.



```
}
}
}
}

PUT karan3
{

"settings": {
"analysis": {
"analyzer": {
"my_signal_news": {
"type": "standard",
"stopwords": "_decision_"
}
}
}
}
}


GET karan1/_analyze
{
  "tokenizer" : "standard",
  "filter" : ["uppercase"],
  "text" : "TEACHERS across the nation now have access to
    information designed to help them identify whether a
    student has become radicalised"
}


PUT karan3/_doc/1
{
  "body": "Jamie Heaslip and Les Kiss spooke to the press
    following the Captain's Run in Wembley today."
}


GET _search
{
  "query": {
    "query_string": {
      "query": "(chairperson) OR (Tuesday)",
      "default_field": "content"
    }
  }
}
```

```
{
  "took" : 268,
  "timed_out" : false,
  "_shards" : {
    "total" : 46,
    "successful" : 46,
    "skipped" : 0,
    "failed" : 0
  },
  "hits" : {
    "total" : 1719,
    "max_score" : 12.889124,
    "hits" : [
      {
        "_index" : "karan",
        "_type" : "_doc",
        "_id" : "sDzPboEBBVl0Psu5HkZZ",
        "_score" : 12.889124,
        "_source" : {
          "@timestamp" : "2015-09-22T13:50:01.000Z",
          "id" : "5db72fdd-8af9-469a-a50a-92c0be9afeb1",
          "source" : "New Delhi News.Net",
          "published" : "2015-09-22T13:50:01Z",
          "title" : "Malik demands probe over mysterious killings in Kashmir",
          "media-type" : "News",
          "content" : """
Jammu and Kashmir Liberation Front (JKLF) chairman Yasin Malik on Tuesday demanded an International probe over mysterious killings in Kashmir.

"People of Kashmir want to know who is behind these mysterious killings. We want an international probe over killings of innocent youth," he said.

Malik also condemned the Vishwa Hindu Parishad (VHP) enforcement on the economic blockade of Kashmir.

"We want to tell Vishwa Hindu Parishad (VHP) that we are not scared of their hollow threats," he said.

According to reports, the Jammu and Kashmir Liberation Front leader was detained with nearly two dozen activists and lodged at the Kothi Bagh police
  station today. Parveena Ahanger, the chairperson of the Association of Parents of Disappeared Persons, was also among those detained by the police.
  (ANI)
"""
        }
      },
      {
        "_index" : "karan",
        "_type" : "_doc",
        "_id" : "-zzPboEBBVl0Psu5HjdW",
        "_score" : 12.58832,
        "_source" : {
          "@timestamp" : "2015-09-03T02:51:46.000Z",
          "id" : "4586ba67-4e9c-44aa-972a-f01b4400862a",
          "source" : "New Zimbabwe.com",
```