CE706 - Information Retrieval SU 2022

Assigment 1

PG2111329

ELK stack : It is a combination of three different tools which are Elasticsearch, Kibana and logstash. These tools Logstash and Elasticsearch can work distinctly, the three useful tools are planned to be used as a combined answer, known as the Elastic Stack. In our report we are going to use only two Elasticsearch and Kibana .The summary of these tools are explained below

1)Elasticsearch API: Elasticsearch is a Java-based open source full-text search engine that is intended to be distributed, scalable, and near real-time. The Elasticsearch server is simple to set up, and the default configuration provided with it is suitable for standalone use without further alterations. The only parameter that has to be entered in the configuration file to start up an Elasticsearch cluster is the cluster name; Elasticsearch will take care of determining nodes on the web and tied them into a cluster (Kononenko, 2014).

2) Kibana: Kibana is a visualization platform available for elasticsearch. It delivers a web – based graphical user interface to visualize, search and many other queries on data stored in clusters. It can be used online and can be downloaded on your personal computers to visualize data . the important view of kibana has 4 components- dashboards, Management, Discover and Visualization (Bajer, 2017).

3)Logstash: It is a plugin event having many features. It takes data from different numbers of sources simultaneously, converts it and then transfers it to elasticsearch. In logstash to transfer data in elasticsearch we have to create a .conf file to send the whole dataset in elasticsearch (Bajer, 2017).

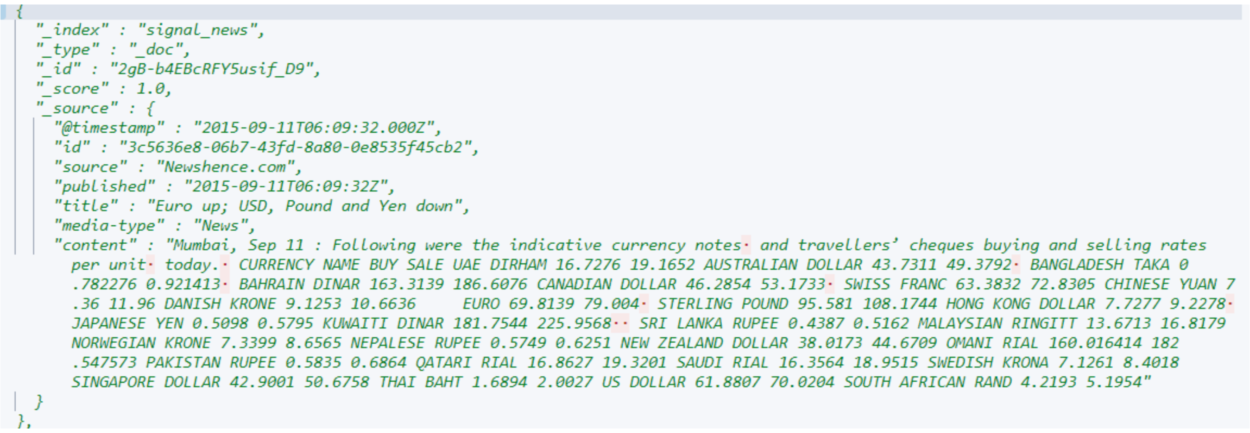
Dataset Description: The dataset named Signal Medai one_million news dataset by signal media is available to enable research on different news articles. It is envisioned to help the research community. The dataset is originally composed by Moreover Technologies which are signal providers. It contains a variety of news sources for a period of 1 month(1-30 sept 2015). It is a directory of 1 million articles that are mainly in English, Although they also articles in different languages. The main sources of these articles are Reuters, blogs and local news.

1) To download dataset: The dataset can be downloaded using this link. https://research.signal-ai.com/newsir16/signal-dataset.html

2) Format of Dataset: Once the dataset is downloaded you have a compressed zip file. You can extract files anywhere in your personal computer. The dataset provided is in json format, in which every line exemplifies an article as a JSON object. The size of dataset is 2.65 GB. The following fields are found in each article:

ID	Act as a unique identifier
Title	The main title of the dataset article
Content	Text which displsy as a conent in article
Source	The main sourec from where data is collected
Published	Date of publication
Type of Media	News

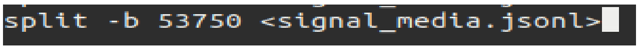
An Example of a Dataset:



```
{
  "_index" : "signal_news",
  "_type" : "_doc",
  "_id" : "2gB-b4EBcRFY5usif_D9",
  "_score" : 1.0,
  "_source" : {
    "@timestamp" : "2015-09-11T06:09:32.000Z",
    "id" : "3c5636e8-06b7-43fd-8a80-0e8535f45cb2",
    "source" : "Newshence.com",
    "published" : "2015-09-11T06:09:32Z",
    "title" : "Euro up; USD, Pound and Yen down",
    "media-type" : "News",
    "content" : "Mumbai, Sep 11 : Following were the indicative currency notes· and travellers’ cheques buying and selling rates
      per unit· today.· CURRENCY NAME BUY SALE UAE DIRHAM 16.7276 19.1652 AUSTRALIAN DOLLAR 43.7311 49.3792· BANGLADESH TAKA 0
      .782276 0.921413· BAHRAIN DINAR 163.3139 186.6076 CANADIAN DOLLAR 46.2854 53.1733· SWISS FRANC 63.3832 72.8305 CHINESE YUAN 7
      .36 11.96 DANISH KRONE 9.1253 10.6636     EURO 69.8139 79.004· STERLING POUND 95.581 108.1744 HONG KONG DOLLAR 7.7277 9.2278·
      JAPANESE YEN 0.5098 0.5795 KUWAITI DINAR 181.7544 225.9568·· SRI LANKA RUPEE 0.4387 0.5162 MALAYSIAN RINGITT 13.6713 16.8179
      NORWEGIAN KRONE 7.3399 8.6565 NEPALESE RUPEE 0.5749 0.6251 NEW ZEALAND DOLLAR 38.0173 44.6709 OMANI RIAL 160.016414 182
      .547573 PAKISTAN RUPEE 0.5835 0.6864 QATARI RIAL 16.8627 19.3201 SAUDI RIAL 16.3564 18.9515 SWEDISH KRONA 7.1261 8.4018
      SINGAPORE DOLLAR 42.9001 50.6758 THAI BAHT 1.6894 2.0027 US DOLLAR 61.8807 70.0204 SOUTH AFRICAN RAND 4.2193 5.1954"
  }
},
```

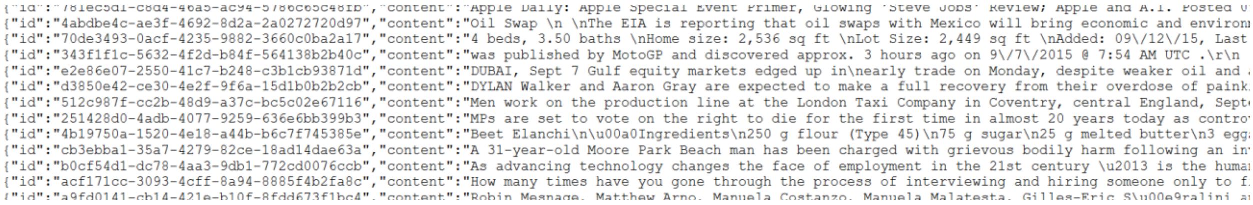
Obtain Dataset: First step I follow to extract data from signal_media.json file. To upload these data on kibana, first I try python “https://github.com/Tamanna1991/Elasticsearch” to upload. But it takes lots of time there. I also try to upload .json with logstash but unfortunaltely my .conf file not work well. That’s why I use Linux to divide the dataset into chunks.

Linux command to divide dataset into chunks.:



```
split -b 53750 <signal_media.jsonl>
```

Split will divide data, -b denotes size of chunks(can be changed according to requirements).Next signal_media is my document name. After running this command my signal_media.json file splits into small chunks named as xaa, xab,xac, etc. It consist of 50 separate document. Then I sort all these documents into VM editor and and upload on kibana to perform queries. Initially I took a sample of 1000 queries. In which dataset is related to trading for different countries. Basically it contains world wide local news. News related to sports , markets and technology. Here is a sample of dataset.



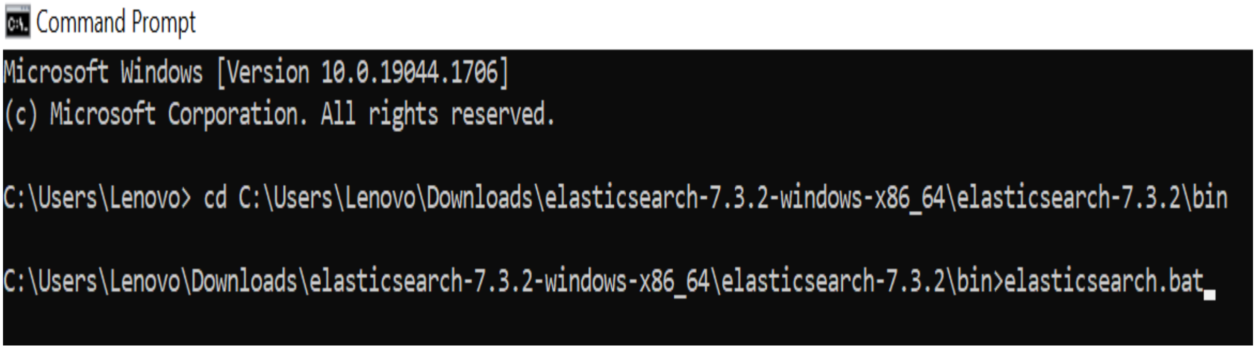
```
{"id":"781ec5d1-c8d4-46a5-ac94-5786c65c48fb","content":"Apple Daily: Apple Special Event Primer, Glowing 'Steve Jobs' Review; Apple and A.I. Posted o
{"id":"4abdbe4c-ae3f-4692-8d2a-2a0272720d97","content":"Oil Swap \n \nThe EIA is reporting that oil swaps with Mexico will bring economic and environ
{"id":"70de3493-0acf-4235-9882-3660c0ba2a17","content":"4 beds, 3.50 baths \nHome size: 2,536 sq ft \nLot Size: 2,449 sq ft \nAdded: 09\/12\/15, Last
{"id":"343f1f1c-5632-4f2d-b84f-564138b2b40c","content":"was published by MotoGP and discovered approx. 3 hours ago on 9\/7\/2015 @ 7:54 AM UTC .\r\n
{"id":"e2e86e07-2550-41c7-b248-c3b1cb93871d","content":"DUBAI, Sept 7 Gulf equity markets edged up in\nearly trade on Monday, despite weaker oil and
{"id":"d3850e42-ce30-4e2f-9f6a-15d1b0b2b2cb","content":"DYLAN Walker and Aaron Gray are expected to make a full recovery from their overdose of paink
{"id":"512c987f-cc2b-48d9-a37c-bc5c02e67116","content":"Men work on the production line at the London Taxi Company in Coventry, central England, Sept
{"id":"251428d0-4adb-4077-9259-636e6bb399b3","content":"MPs are set to vote on the right to die for the first time in almost 20 years today as contro
{"id":"4b19750a-1520-4e18-a44b-b6c7f745385e","content":"Beet Elanchi\n\u00a0Ingredients\n250 g flour (Type 45)\n75 g sugar\n25 g melted butter\n3 egg
{"id":"cb3ebba1-35a7-4279-82ce-18ad14dae63a","content":"A 31-year-old Moore Park Beach man has been charged with grievous bodily harm following an in
{"id":"b0cf54d1-dc78-4aa3-9db1-772cd0076ccb","content":"As advancing technology changes the face of employment in the 21st century \u2013 is the huma
{"id":"acf171cc-3093-4cff-8a94-8885f4b2fa8c","content":"How many times have you gone through the process of interviewing and hiring someone only to f
{"id":"a9fd0141-cb14-421e-b10f-8fdd673f1bc4","content":"Robin Mesnage, Matthew Arno, Manuela Costanzo, Manuela Malatesta, Gilles-Eric S\u00e9ralini a
```

1: Sample Data(1 to1000) set of file signal_media

Instructions for running your system : Once extraction is completed . Next step is to install Kibana and elasticsearch on your system. For my implementation of work I use Elasticsearch 7.3.2 and Kibana 7.3.2. Elasticsearch can be download using link https://www.elastic.co/downloads/elasticsearch .

Elasticsearch 7.3.2: Before installation of Elasticsarach and kibana install Java on your machine first then use steps to install Elasticsearch and Kibana. After Installation of Java : You will need to ‘restart’ your laptop in order to process after this Java installation, otherwise your system will not able to run elasticsearch. Now After Java installation follow steps below :

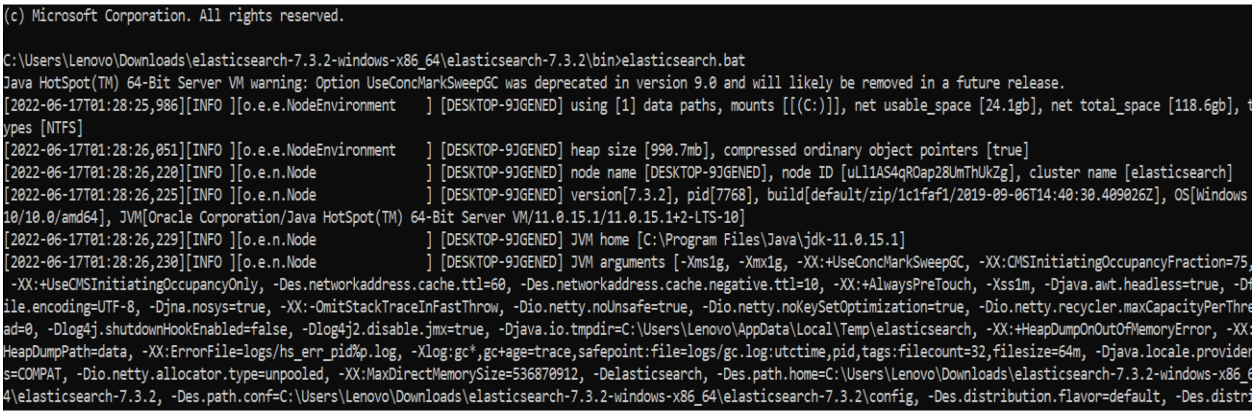
Step 1: After installation use below command to run elasticsearch.



```
Microsoft Windows [Version 10.0.19044.1706]
(c) Microsoft Corporation. All rights reserved.

C:\Users\Lenovo> cd C:\Users\Lenovo\Downloads\elasticsearch-7.3.2-windows-x86_64\elasticsearch-7.3.2\bin

C:\Users\Lenovo\Downloads\elasticsearch-7.3.2-windows-x86_64\elasticsearch-7.3.2\bin>elasticsearch.bat
```

2: Run Elasticsearch

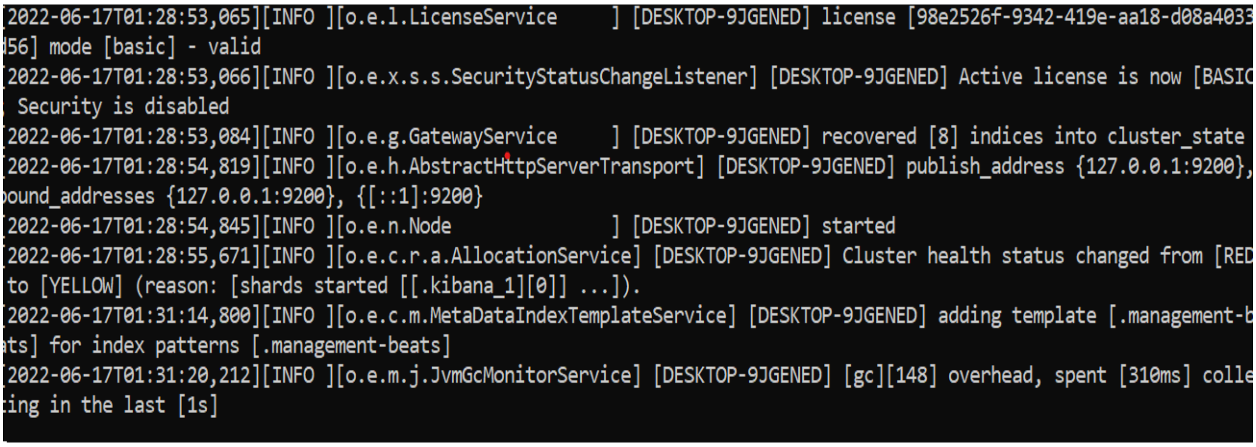
Step 2: Configuartion window will look like this: Once elasticsearch starts running screen display looks like below.



```
(c) Microsoft Corporation. All rights reserved.

C:\Users\Lenovo\Downloads\elasticsearch-7.3.2-windows-x86_64\elasticsearch-7.3.2\bin>elasticsearch.bat
Java HotSpot(TM) 64-Bit Server VM warning: Option UseConcMarkSweepGC was deprecated in version 9.0 and will likely be removed in a future release.
[2022-06-17T01:28:25,986][INFO ][o.e.e.NodeEnvironment    ] [DESKTOP-9JGENED] using [1] data paths, mounts [[(C:)]], net usable_space [24.1gb], net total_space [118.6gb], t
ypes [NTFS]
[2022-06-17T01:28:26,051][INFO ][o.e.e.NodeEnvironment    ] [DESKTOP-9JGENED] heap size [990.7mb], compressed ordinary object pointers [true]
[2022-06-17T01:28:26,220][INFO ][o.e.n.Node               ] [DESKTOP-9JGENED] node name [DESKTOP-9JGENED], node ID [uLl1AS4qROap28UmThUkZg], cluster name [elasticsearch]
[2022-06-17T01:28:26,225][INFO ][o.e.n.Node               ] [DESKTOP-9JGENED] version[7.3.2], pid[7768], build[default/zip/1c1faf1/2019-09-06T14:40:30.409026Z], OS[Windows
10/10.0/amd64], JVM[Oracle Corporation/Java HotSpot(TM) 64-Bit Server VM/11.0.15.1/11.0.15.1+2-LTS-10]
[2022-06-17T01:28:26,229][INFO ][o.e.n.Node               ] [DESKTOP-9JGENED] JVM home [C:\Program Files\Java\jdk-11.0.15.1]
[2022-06-17T01:28:26,230][INFO ][o.e.n.Node               ] [DESKTOP-9JGENED] JVM arguments [-Xms1g, -Xmx1g, -XX:+UseConcMarkSweepGC, -XX:CMSInitiatingOccupancyFraction=75,
 -XX:+UseCMSInitiatingOccupancyOnly, -Des.networkaddress.cache.ttl=60, -Des.networkaddress.cache.negative.ttl=10, -XX:+AlwaysPreTouch, -Xss1m, -Djava.awt.headless=true, -Df
ile.encoding=UTF-8, -Djna.nosys=true, -XX:-OmitStackTraceInFastThrow, -Dio.netty.noUnsafe=true, -Dio.netty.noKeySetOptimization=true, -Dio.netty.recycler.maxCapacityPerThre
ad=0, -Dlog4j.shutdownHookEnabled=false, -Dlog4j2.disable.jmx=true, -Djava.io.tmpdir=C:\Users\Lenovo\AppData\Local\Temp\elasticsearch, -XX:+HeapDumpOnOutOfMemoryError, -XX:
HeapDumpPath=data, -XX:ErrorFile=logs/hs_err_pid%p.log, -Xlog:gc*,gc+age=trace,safepoint:file=logs/gc.log:utctime,pid,tags:filecount=32,filesize=64m, -Djava.locale.provider
s=COMPAT, -Dio.netty.allocator.type=unpooled, -XX:MaxDirectMemorySize=536870912, -Delasticsearch, -Des.path.home=C:\Users\Lenovo\Downloads\elasticsearch-7.3.2-windows-x86_6
4\elasticsearch-7.3.2, -Des.path.conf=C:\Users\Lenovo\Downloads\elasticsearch-7.3.2-windows-x86_64\elasticsearch-7.3.2\config, -Des.distribution.flavor=default, -Des.distri
```

3: Configuartion window

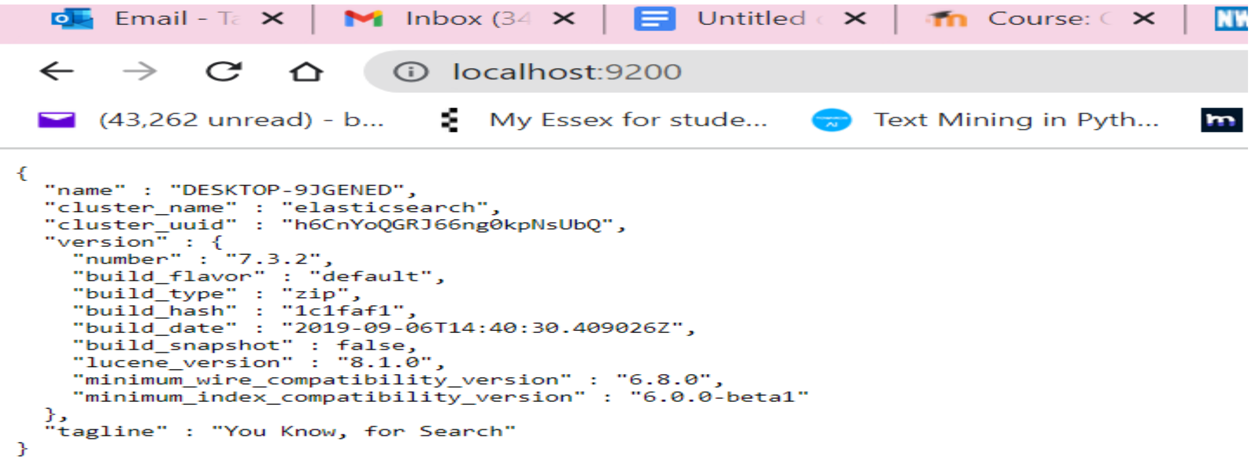
Step 3: Once configuration complete, In the end of screen, you will get bound address{127.0.01.9200}, which is a sign of successfully running your elasticserach engine on browser. You can also check it from browser by using {http://Localhost:9200}



```
[2022-06-17T01:28:53,065][INFO ][o.e.l.LicenseService     ] [DESKTOP-9JGENED] license [98e2526f-9342-419e-aa18-d08a4033
56] mode [basic] - valid
[2022-06-17T01:28:53,066][INFO ][o.e.x.s.s.SecurityStatusChangeListener] [DESKTOP-9JGENED] Active license is now [BASIC
 Security is disabled
[2022-06-17T01:28:53,084][INFO ][o.e.g.GatewayService     ] [DESKTOP-9JGENED] recovered [8] indices into cluster_state
[2022-06-17T01:28:54,819][INFO ][o.e.h.AbstractHttpServerTransport] [DESKTOP-9JGENED] publish_address {127.0.0.1:9200},
bound_addresses {127.0.0.1:9200}, {[::1]:9200}
[2022-06-17T01:28:54,845][INFO ][o.e.n.Node               ] [DESKTOP-9JGENED] started
[2022-06-17T01:28:55,671][INFO ][o.e.c.r.a.AllocationService] [DESKTOP-9JGENED] Cluster health status changed from [RED
 to [YELLOW] (reason: [shards started [[.kibana_1][0]] ...]).
[2022-06-17T01:31:14,800][INFO ][o.e.c.m.MetaDataIndexTemplateService] [DESKTOP-9JGENED] adding template [.management-b
ts] for index patterns [.management-beats]
[2022-06-17T01:31:20,212][INFO ][o.e.m.j.JvmGcMonitorService] [DESKTOP-9JGENED] [gc][148] overhead, spent [310ms] colle
ing in the last [1s]
```

4: Sign of successfully installation of Elasticsearch

Step 4: Run "http: localhost:9200" on any browser.



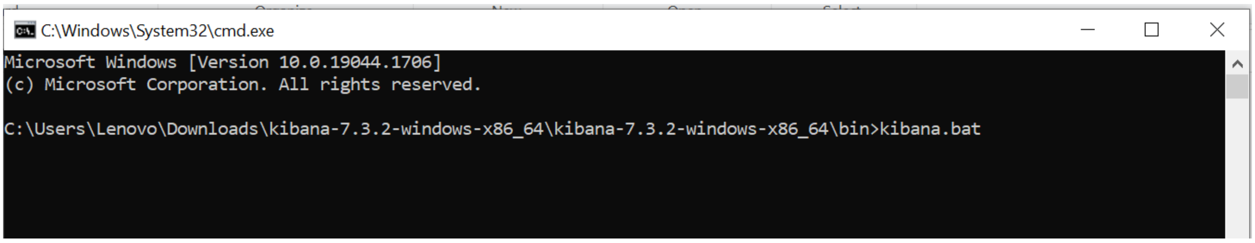
```
{
  "name" : "DESKTOP-9JGENED",
  "cluster_name" : "elasticsearch",
  "cluster_uuid" : "h6CnYoQGRJ66ng0kpNsUbQ",
  "version" : {
    "number" : "7.3.2",
    "build_flavor" : "default",
    "build_type" : "zip",
    "build_hash" : "1c1faf1",
    "build_date" : "2019-09-06T14:40:30.409026Z",
    "build_snapshot" : false,
    "lucene_version" : "8.1.0",
    "minimum_wire_compatibility_version" : "6.8.0",
    "minimum_index_compatibility_version" : "6.0.0-beta1"
  },
  "tagline" : "You Know, for Search"
}
```

5: Browser display after successfully installation

NOTE: Donot Close configuration window of elasticsearch.

Kibana 7.3.2: Elasticsearch is successfully installed on your wwindow , Next step is to install Kibana. Link to install kibana on window. https://www.elastic.co/downloads/kibana.

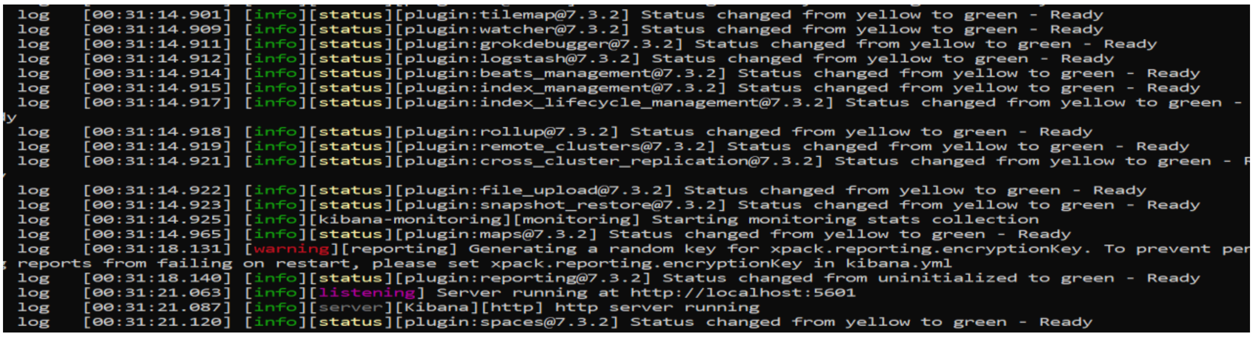
Step 1: After extraction of file. Run .bat file on command prompt. Window will appear like below.



```
Microsoft Windows [Version 10.0.19044.1706]
(c) Microsoft Corporation. All rights reserved.

C:\Users\Lenovo\Downloads\kibana-7.3.2-windows-x86_64\kibana-7.3.2-windows-x86_64\bin>kibana.bat
```

6: Run Kibana.bat

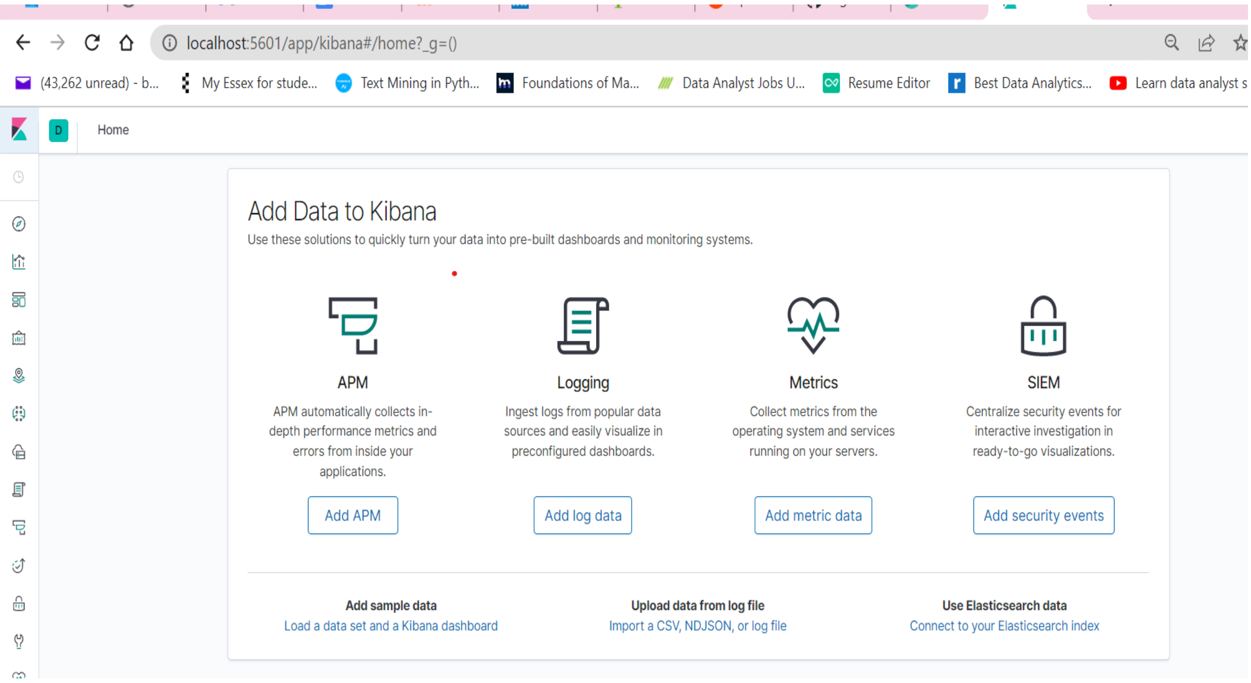
Step 2: Installation process will appaear on screen. Status will change from green to ready. In the last of screen http://localhost:5601. It can be checked on browser.



```
 log   [00:31:14.901] [info][status][plugin:tilemap@7.3.2] Status changed from yellow to green - Ready
 log   [00:31:14.909] [info][status][plugin:watcher@7.3.2] Status changed from yellow to green - Ready
 log   [00:31:14.911] [info][status][plugin:grokdebugger@7.3.2] Status changed from yellow to green - Ready
 log   [00:31:14.912] [info][status][plugin:logstash@7.3.2] Status changed from yellow to green - Ready
 log   [00:31:14.914] [info][status][plugin:beats_management@7.3.2] Status changed from yellow to green - Ready
 log   [00:31:14.915] [info][status][plugin:index_management@7.3.2] Status changed from yellow to green - Ready
 log   [00:31:14.917] [info][status][plugin:index_lifecycle_management@7.3.2] Status changed from yellow to green -
y
 log   [00:31:14.918] [info][status][plugin:rollup@7.3.2] Status changed from yellow to green - Ready
 log   [00:31:14.919] [info][status][plugin:remote_clusters@7.3.2] Status changed from yellow to green - Ready
 log   [00:31:14.921] [info][status][plugin:cross_cluster_replication@7.3.2] Status changed from yellow to green - R

 log   [00:31:14.922] [info][status][plugin:file_upload@7.3.2] Status changed from yellow to green - Ready
 log   [00:31:14.923] [info][status][plugin:snapshot_restore@7.3.2] Status changed from yellow to green - Ready
 log   [00:31:14.925] [info][kibana-monitoring][monitoring] Starting monitoring stats collection
 log   [00:31:14.965] [info][status][plugin:maps@7.3.2] Status changed from yellow to green - Ready
 log   [00:31:18.131] [warning][reporting] Generating a random key for xpack.reporting.encryptionKey. To prevent per
 reports from failing on restart, please set xpack.reporting.encryptionKey in kibana.yml
 log   [00:31:18.140] [info][status][plugin:reporting@7.3.2] Status changed from uninitialized to green - Ready
 log   [00:31:21.063] [info][listening] Server running at http://localhost:5601
 log   [00:31:21.087] [info][server][Kibana][http] http server running
 log   [00:31:21.120] [info][status][plugin:spaces@7.3.2] Status changed from yellow to green - Ready
```

7: local host kibana is ready to use

Step 3: Go to link http://localhost:5601. Kibana winodow will appear like this on your window.



Add Data to Kibana

Use these solutions to quickly turn your data into pre-built dashboards and monitoring systems.

APM

APM automatically collects in-depth performance metrics and errors from inside your applications.

Logging

Ingest logs from popular data sources and easily visualize in preconfigured dashboards.

Metrics

Collect metrics from the operating system and services running on your servers.

SIEM

Centralize security events for interactive investigation in ready-to-go visualizations.

Add sample data

Load a data set and a Kibana dashboard

Upload data from log file

Import a CSV, NDJSON, or log file

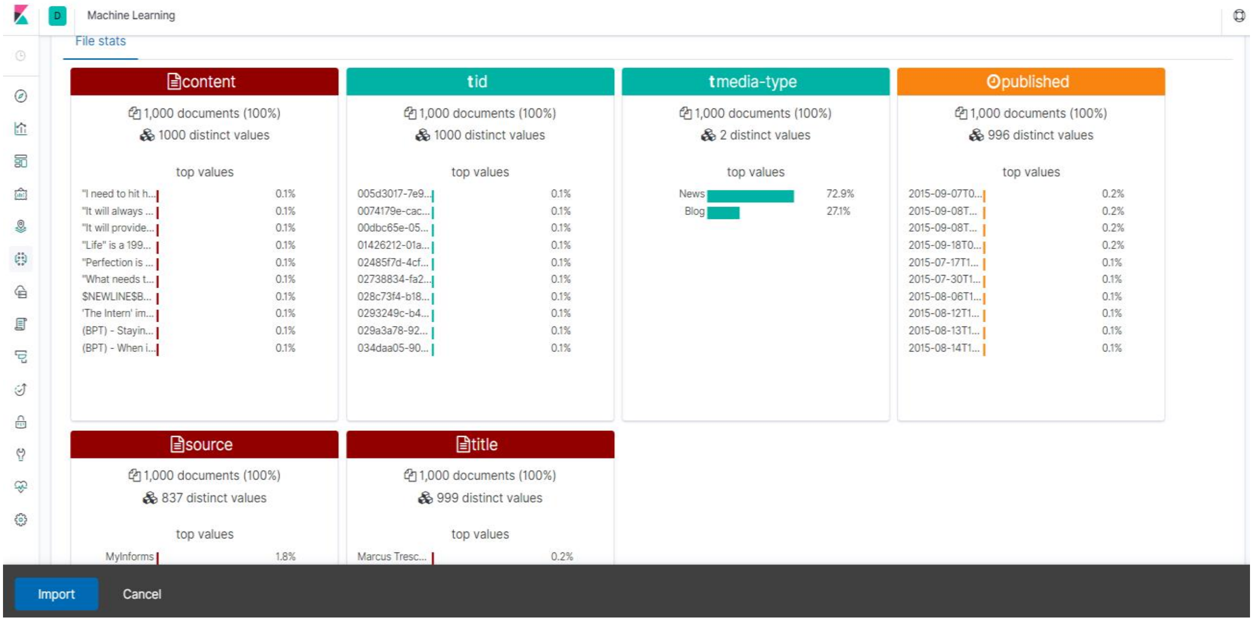
Use Elasticsearch data

Connect to your Elasticsearch index

8: Kibana window

INDEXING: An index can be believed as an elevated collection of documents whereas each document is collection of different fields, known as key-pairs that have your data.

Step 1: For indexing of dataset, I choose social network new related documents **.**



content

1,000 documents (100%)

1000 distinct values

top values

"I need to hit h...	0.1%
"It will always ...	0.1%
"It will provide...	0.1%
"Life" is a 199...	0.1%
"Perfection is ...	0.1%
"What needs t...	0.1%
$NEWLINE$B...	0.1%
'The Intern' im...	0.1%
(BPT) - Stayin...	0.1%
(BPT) - When i...	0.1%

tid

1,000 documents (100%)

1000 distinct values

top values

005d3017-7e9...	0.1%
0074179e-cac...	0.1%
00dbc65e-05...	0.1%
01426212-01a...	0.1%
02485f7d-4cf...	0.1%
02738834-fa2...	0.1%
028c73f4-b18...	0.1%
0293249c-b4...	0.1%
029a3a78-92...	0.1%
034daa05-90...	0.1%

media-type

1,000 documents (100%)

2 distinct values

top values

News	72.9%
Blog	27.1%

published

1,000 documents (100%)

996 distinct values

top values

2015-09-07T0...	0.2%
2015-09-08T...	0.2%
2015-09-08T...	0.2%
2015-09-18T0...	0.2%
2015-07-17T1...	0.1%
2015-07-30T1...	0.1%
2015-08-06T1...	0.1%
2015-08-12T1...	0.1%
2015-08-13T1...	0.1%
2015-08-14T1...	0.1%

source

1,000 documents (100%)

837 distinct values

top values

MyInforms	1.8%

title

1,000 documents (100%)

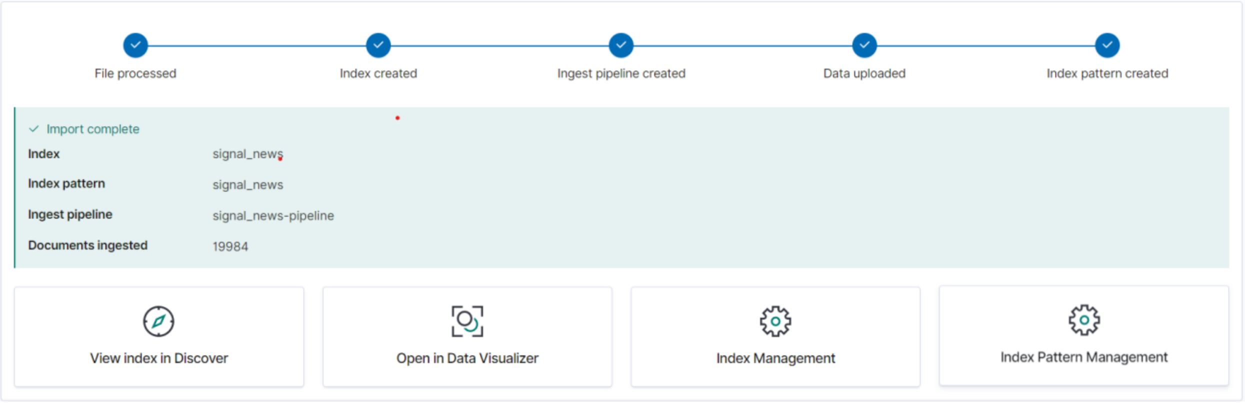
999 distinct values

top values

Marcus Tresc...	0.2%

9: Uploading of document

Step 2: Click on import, Next index-pattern window will appear. For index pattern I fill "signal_news", now which will be my index name. A pipeline is created after the index pattern.

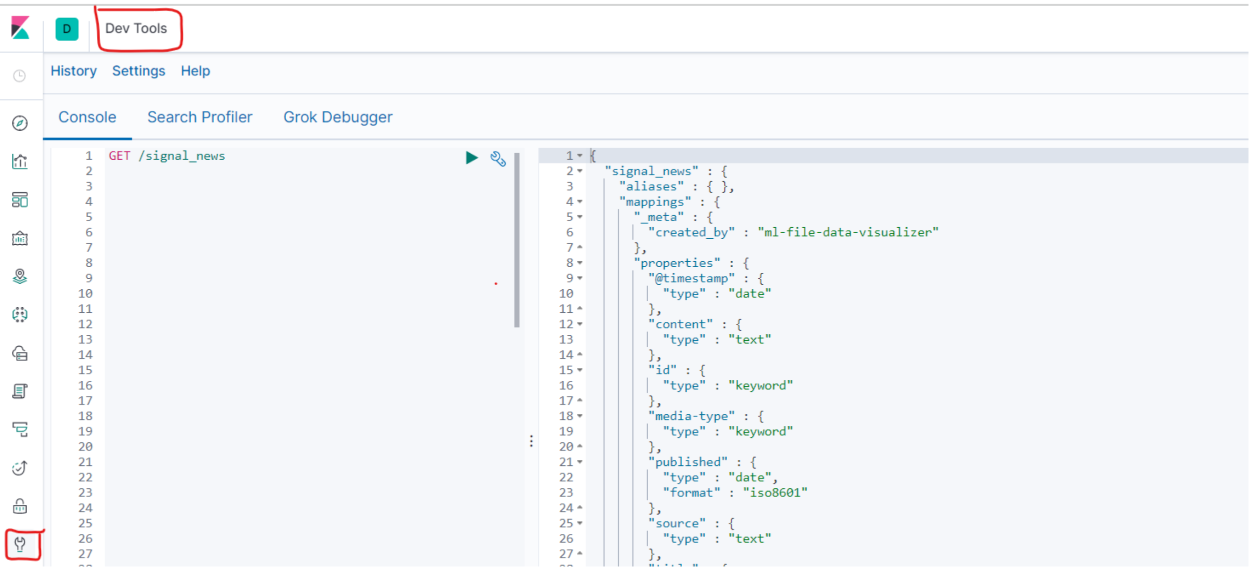


Import complete

Index	signal_news
Index pattern	signal_news
Ingest pipeline	signal_news-pipeline
Documents ingested	19984

10: **Import Data and Index created**

Step 3: After clicking on Dev tools as highlighted below , a console window of dev tools is appeared, click on Triangle ▶ mark to see index name and information related to document

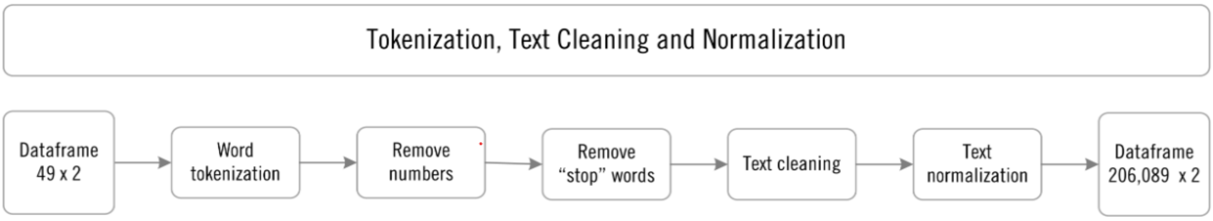


```
GET /signal_news
```

```
{
  "signal_news" : {
    "aliases" : { },
    "mappings" : {
      "_meta" : {
        "created_by" : "ml-file-data-visualizer"
      },
      "properties" : {
        "@timestamp" : {
          "type" : "date"
        },
        "content" : {
          "type" : "text"
        },
        "id" : {
          "type" : "keyword"
        },
        "media-type" : {
          "type" : "keyword"
        },
        "published" : {
          "type" : "date",
          "format" : "iso8601"
        },
        "source" : {
          "type" : "text"
        },
```

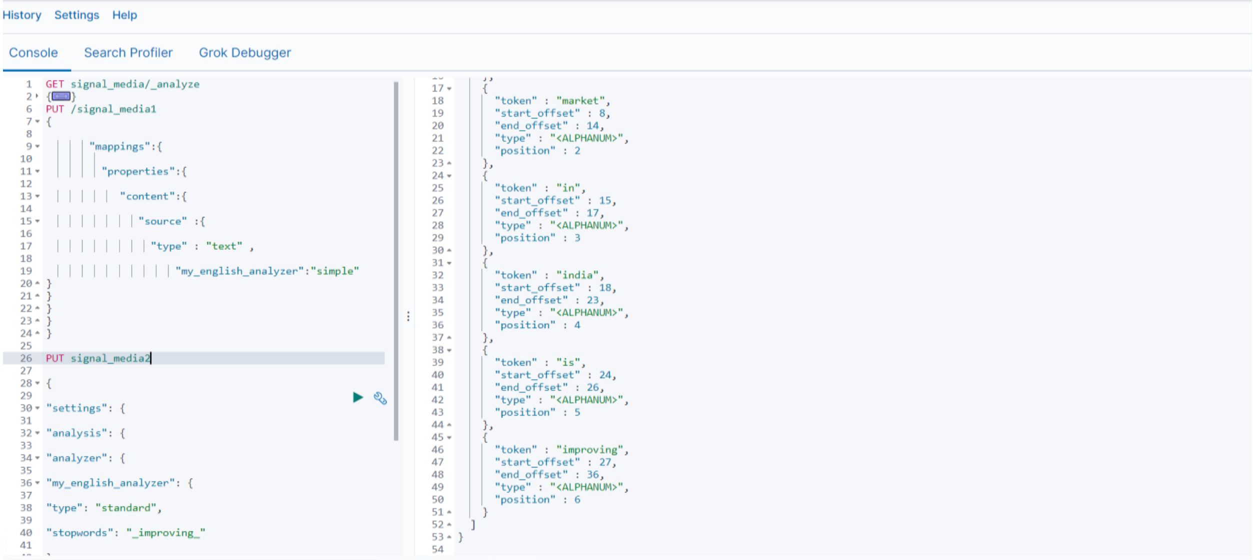
11: Kibana console and Dev tools

Tokenization and Normalisation: To study a raw string of text, it must first be tokenized. Complete Roadmap for Noramlization and tokenization is present below:



12: **Complete Roadmap For Tokenization**

Step 1: All steps are desrcribed with the help of screenshots. “Simple tokenizer” is created, lowercase filter is applied. For example : “ Market in india is _improving_.” After simple toenizer output will be(‘market”, “in”,”india”.”is”,”improving). For content and source type is text. Token name is my_english_anlalyzer.

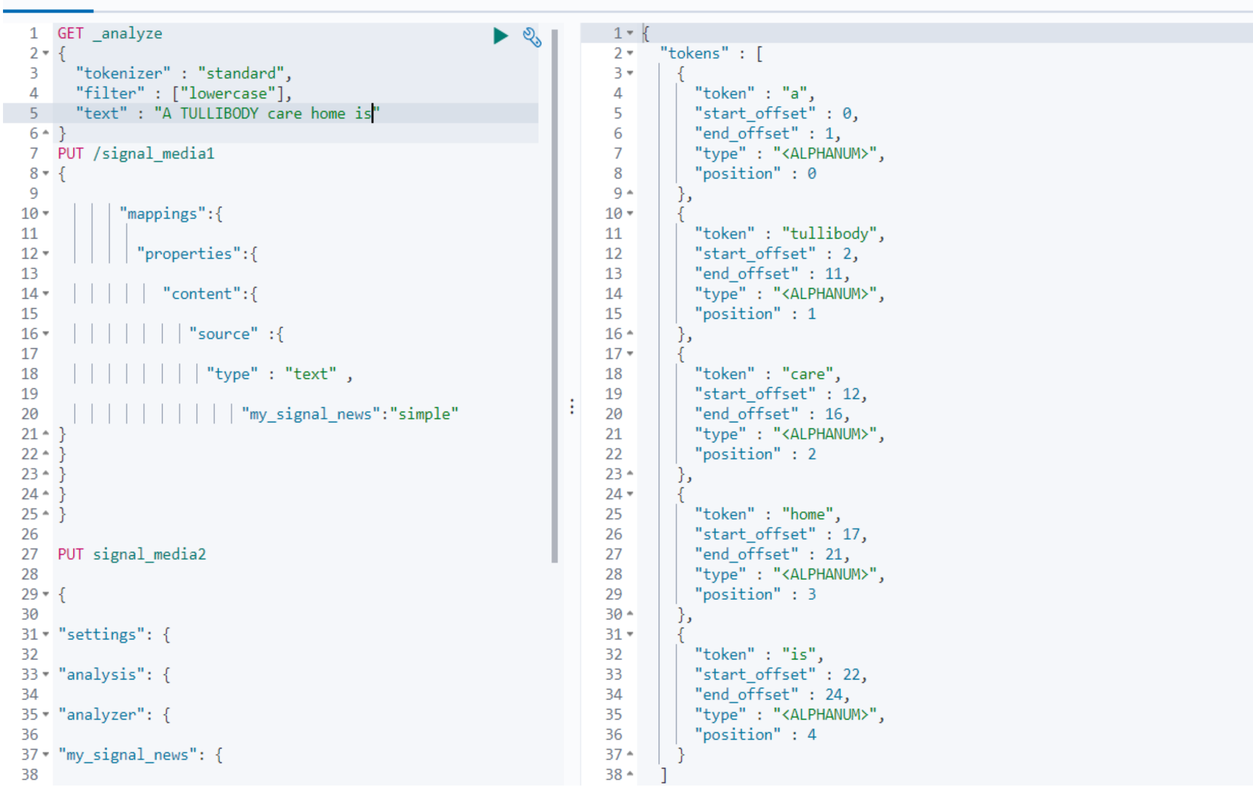


```
GET signal_media/_analyze
{}
PUT /signal_media1
{

          "mappings":{

            "properties":{

              "content":{

                  "source" :{

                    "type" : "text" ,

                        "my_english_analyzer":"simple"
}
}
}
}
}

PUT signal_media2

{

"settings": {

"analysis": {

"analyzer": {

"my_english_analyzer": {

"type": "standard",

"stopwords": "_improving_"

```

```
    },
    {
      "token" : "market",
      "start_offset" : 8,
      "end_offset" : 14,
      "type" : "<ALPHANUM>",
      "position" : 2
    },
    {
      "token" : "in",
      "start_offset" : 15,
      "end_offset" : 17,
      "type" : "<ALPHANUM>",
      "position" : 3
    },
    {
      "token" : "india",
      "start_offset" : 18,
      "end_offset" : 23,
      "type" : "<ALPHANUM>",
      "position" : 4
    },
    {
      "token" : "is",
      "start_offset" : 24,
      "end_offset" : 26,
      "type" : "<ALPHANUM>",
      "position" : 5
    },
    {
      "token" : "improving",
      "start_offset" : 27,
      "end_offset" : 36,
      "type" : "<ALPHANUM>",
      "position" : 6
    }
  ]
}
```

13: Simple tokenization

Step2: Removal of Stop words and Tokenization



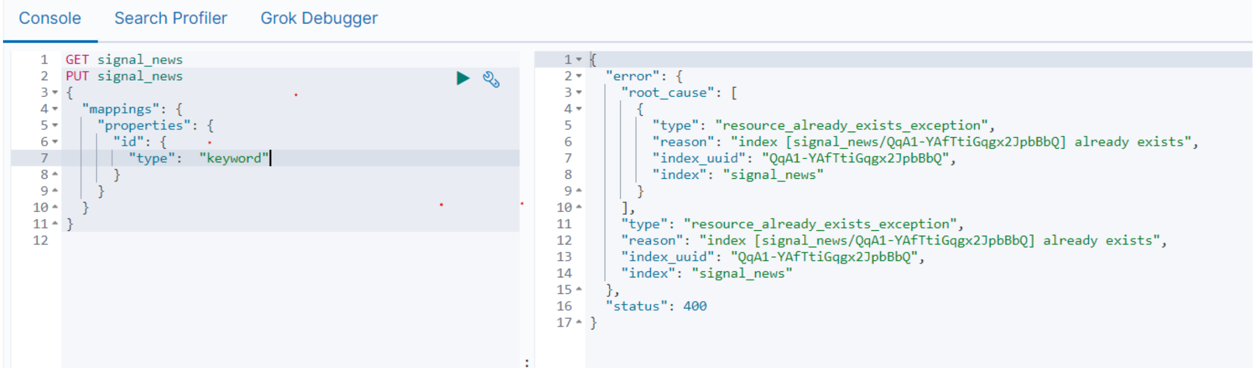
```
GET _analyze
{
  "tokenizer" : "standard",
  "filter" : ["lowercase"],
  "text" : "A TULLIBODY care home is"
}
PUT /signal_media1
{

          "mappings":{

            "properties":{

              "content":{

                  "source" :{

                    "type" : "text" ,

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

PUT signal_media2

{

"settings": {

"analysis": {

"analyzer": {

"my_signal_news": {

```

```
{
  "tokens" : [
    {
      "token" : "a",
      "start_offset" : 0,
      "end_offset" : 1,
      "type" : "<ALPHANUM>",
      "position" : 0
    },
    {
      "token" : "tullibody",
      "start_offset" : 2,
      "end_offset" : 11,
      "type" : "<ALPHANUM>",
      "position" : 1
    },
    {
      "token" : "care",
      "start_offset" : 12,
      "end_offset" : 16,
      "type" : "<ALPHANUM>",
      "position" : 2
    },
    {
      "token" : "home",
      "start_offset" : 17,
      "end_offset" : 21,
      "type" : "<ALPHANUM>",
      "position" : 3
    },
    {
      "token" : "is",
      "start_offset" : 22,
      "end_offset" : 24,
      "type" : "<ALPHANUM>",
      "position" : 4
    }
  ]
```

14: **Lowercase filter for Text normalization**

Selecting Keywords: This will be used when you want to select id, phone number, email-id and address.

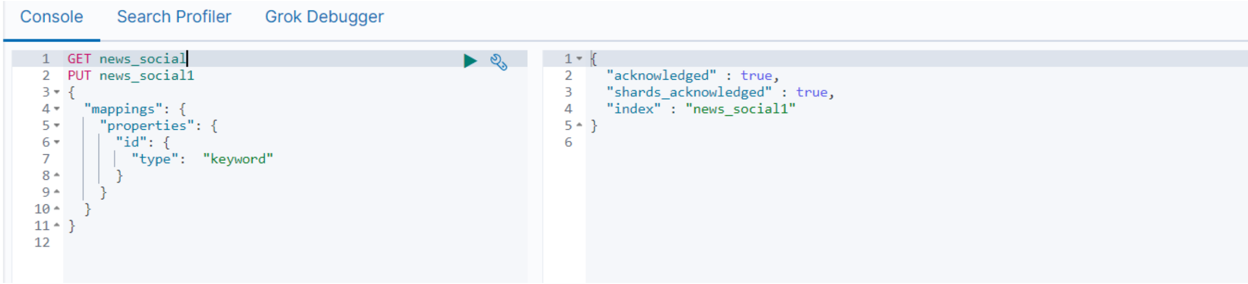
Step1: Error: While performing keyword selection step , an error accured due same name mapping. I remove this error with changing the name of mapping, and then I got results.



```
GET signal_news
PUT signal_news
{
  "mappings": {
    "properties": {
      "id": {
        "type":  "keyword"
      }
    }
  }
}
```

```
{
  "error": {
    "root_cause": [
      {
        "type": "resource_already_exists_exception",
        "reason": "index [signal_news/QqA1-YAfTtiGqgx2JpbBbQ] already exists",
        "index_uuid": "QqA1-YAfTtiGqgx2JpbBbQ",
        "index": "signal_news"
      }
    ],
    "type": "resource_already_exists_exception",
    "reason": "index [signal_news/QqA1-YAfTtiGqgx2JpbBbQ] already exists",
    "index_uuid": "QqA1-YAfTtiGqgx2JpbBbQ",
    "index": "signal_news"
  },
  "status": 400
}
```

15: Keyword selection(Error 1)

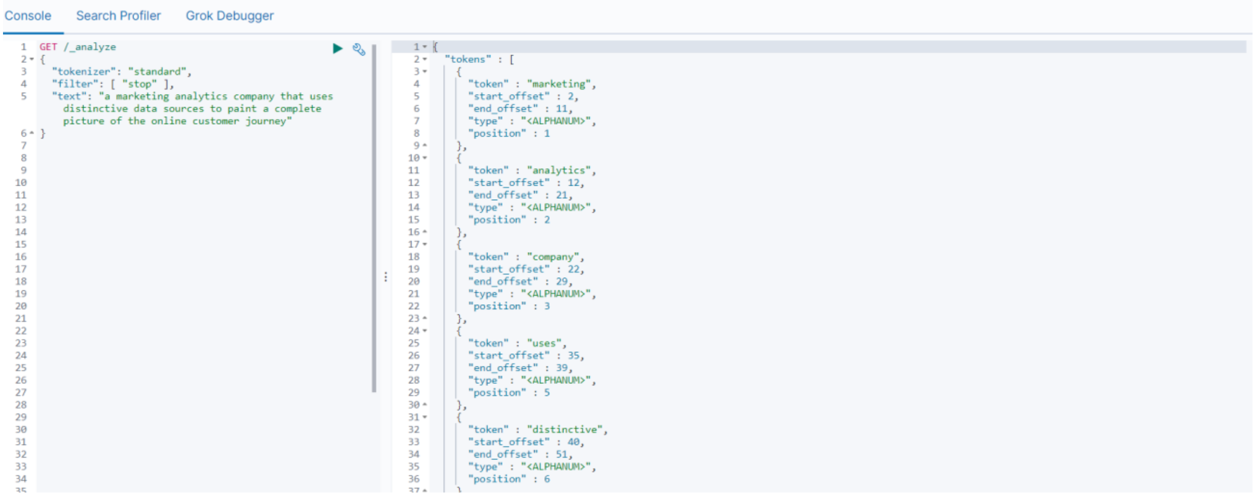


```
GET news_social
PUT news_social1
{
  "mappings": {
    "properties": {
      "id": {
        "type":  "keyword"
      }
    }
  }
}
```

```
{
  "acknowledged" : true,
  "shards_acknowledged" : true,
  "index" : "news_social1"
}
```

16: **Get Results(Error resolved by changing mapping name)**

Step2: Stop word removal: It will remove stop words from the token phrase. For example These words can be “in”,”of”,”and”,”not”,”such”,”that” . Created a standard toeknizer and filter name stop to remove stop words from phrase.

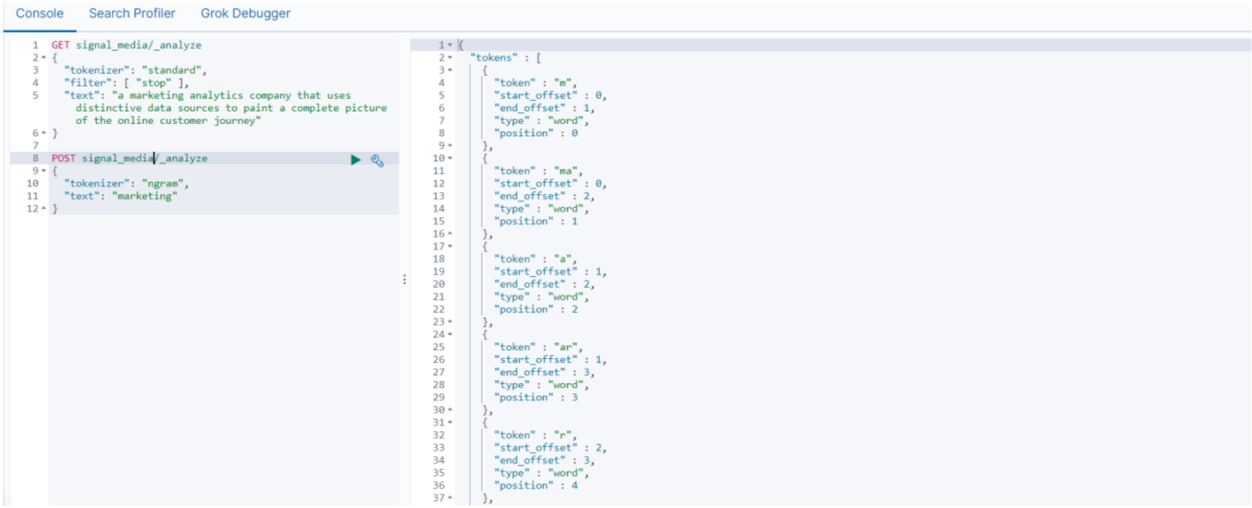


```
GET /_analyze
{
  "tokenizer": "standard",
  "filter": [ "stop" ],
  "text": "a marketing analytics company that uses
    distinctive data sources to paint a complete
    picture of the online customer journey"
}
```

```
{
  "tokens" : [
    {
      "token" : "marketing",
      "start_offset" : 2,
      "end_offset" : 11,
      "type" : "<ALPHANUM>",
      "position" : 1
    },
    {
      "token" : "analytics",
      "start_offset" : 12,
      "end_offset" : 21,
      "type" : "<ALPHANUM>",
      "position" : 2
    },
    {
      "token" : "company",
      "start_offset" : 22,
      "end_offset" : 29,
      "type" : "<ALPHANUM>",
      "position" : 3
    },
    {
      "token" : "uses",
      "start_offset" : 35,
      "end_offset" : 39,
      "type" : "<ALPHANUM>",
      "position" : 5
    },
    {
      "token" : "distinctive",
      "start_offset" : 40,
      "end_offset" : 51,
      "type" : "<ALPHANUM>",
      "position" : 6
    }
```

17:Stop word Removal

Step 3: N-gram tokenizer will break text in to words , whenever it discover a list of detailed characters, then it divide N-grams of each word of the stated length. It can be 2, 3, etc[Lab2]. For example I perform tokenizer on word marketing, anlyzer gave us output “m”,”a”,”r”,”k”,”e”,”t”,”i”,”i”,”n”,”g”.



```
GET signal_media/_analyze
{
  "tokenizer": "standard",
  "filter": [ "stop" ],
  "text": "a marketing analytics company that uses
    distinctive data sources to paint a complete picture
    of the online customer journey"
}

POST signal_media/_analyze
{
  "tokenizer": "ngram",
  "text": "marketing"
}
```

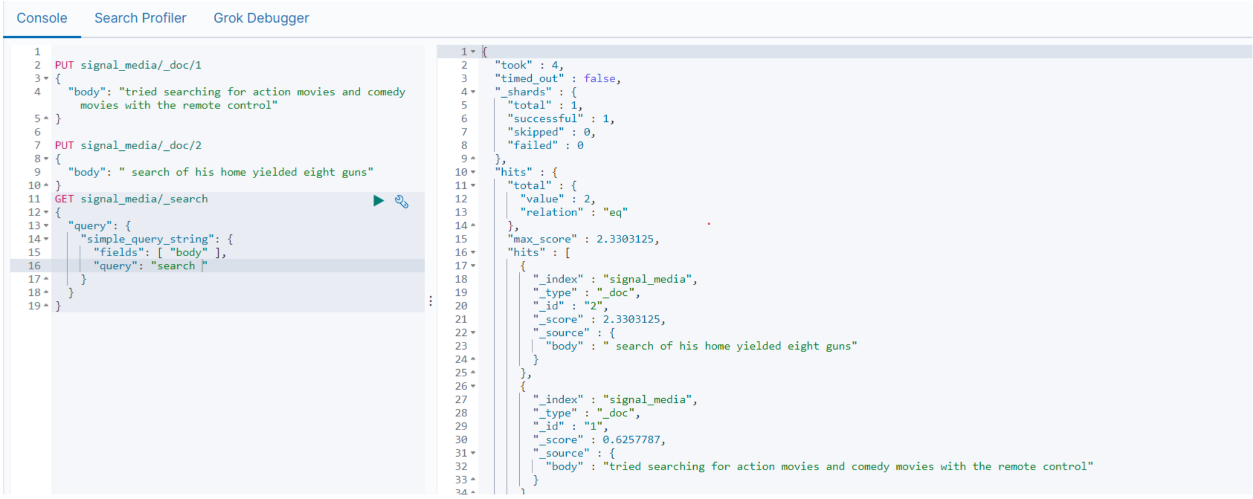
```
{
  "tokens" : [
    {
      "token" : "m",
      "start_offset" : 0,
      "end_offset" : 1,
      "type" : "word",
      "position" : 0
    },
    {
      "token" : "ma",
      "start_offset" : 0,
      "end_offset" : 2,
      "type" : "word",
      "position" : 1
    },
    {
      "token" : "a",
      "start_offset" : 1,
      "end_offset" : 2,
      "type" : "word",
      "position" : 2
    },
    {
      "token" : "ar",
      "start_offset" : 1,
      "end_offset" : 3,
      "type" : "word",
      "position" : 3
    },
    {
      "token" : "r",
      "start_offset" : 2,
      "end_offset" : 3,
      "type" : "word",
      "position" : 4
    },
```

18: N- Gram(Use of N-Gram on token"marketing"

Stemming: Basically stemming is the way of reducing a word to its stem by removing suffixes. For example : In both cases stem is a ”study” after reomoving suffix and stem[Lab2]

Form	**Suffix**	**Stem**
Studies	-es	Studi
Studying	-ing	Study

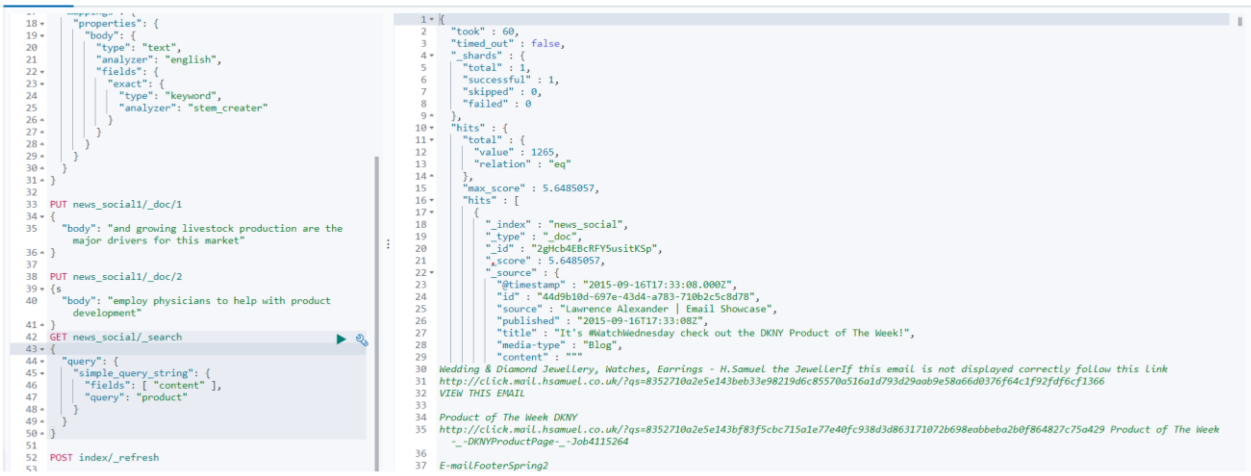
Step1: For stemming we put two documents docs, doc1 and doc2, then apply search within “body” for on form , In output two documents with specific body text are retrieved .



```

PUT signal_media/_doc/1
{
  "body": "tried searching for action movies and comedy
    movies with the remote control"
}

PUT signal_media/_doc/2
{
  "body": " search of his home yielded eight guns"
}
GET signal_media/_search
{
  "query": {
    "simple_query_string": {
      "fields": [ "body" ],
      "query": "search "
    }
  }
}
```

```
{
  "took" : 4,
  "timed_out" : false,
  "_shards" : {
    "total" : 1,
    "successful" : 1,
    "skipped" : 0,
    "failed" : 0
  },
  "hits" : {
    "total" : {
      "value" : 2,
      "relation" : "eq"
    },
    "max_score" : 2.3303125,
    "hits" : [
      {
        "_index" : "signal_media",
        "_type" : "_doc",
        "_id" : "2",
        "_score" : 2.3303125,
        "_source" : {
          "body" : " search of his home yielded eight guns"
        }
      },
      {
        "_index" : "signal_media",
        "_type" : "_doc",
        "_id" : "1",
        "_score" : 0.6257787,
        "_source" : {
          "body" : "tried searching for action movies and comedy movies with the remote control"
        }
      }
```

19: Stemming from two documents with search in body

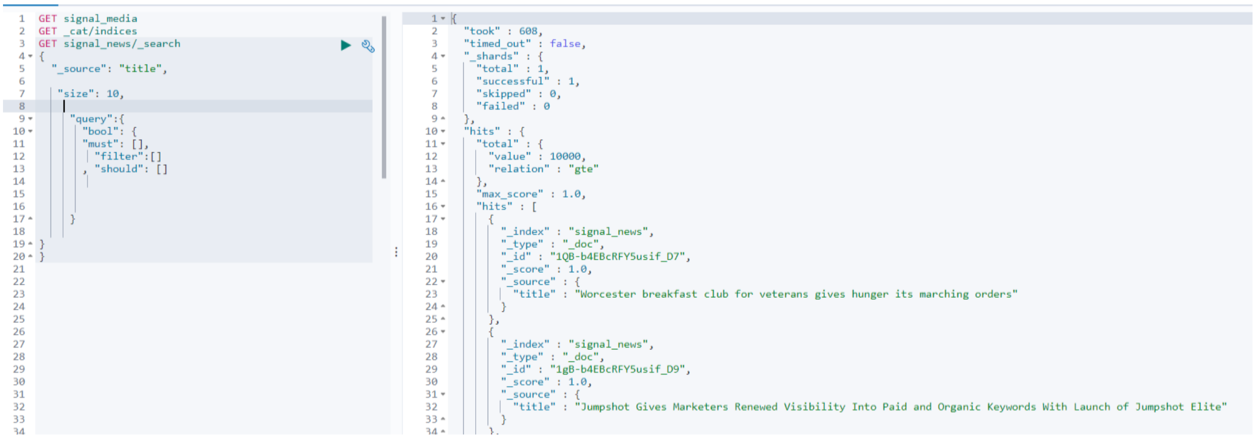


```
  "mappings" : {
    "properties": {
      "body": {
        "type": "text",
        "analyzer": "english",
        "fields": {
          "exact": {
            "type": "keyword",
            "analyzer": "stem_creater"
          }
        }
      }
    }
  }
}

PUT news_social1/_doc/1
{
  "body": "and growing livestock production are the
    major drivers for this market"
}

PUT news_social1/_doc/2
{s
  "body": "employ physicians to help with product
    development"
}
GET news_social/_search
{
  "query": {
    "simple_query_string": {
      "fields": [ "content" ],
      "query": "product"
    }
  }
}

POST index/_refresh
```

```
{
  "took" : 60,
  "timed_out" : false,
  "_shards" : {
    "total" : 1,
    "successful" : 1,
    "skipped" : 0,
    "failed" : 0
  },
  "hits" : {
    "total" : {
      "value" : 1265,
      "relation" : "eq"
    },
    "max_score" : 5.6485057,
    "hits" : [
      {
        "_index" : "news_social",
        "_type" : "_doc",
        "_id" : "2gHcb4EBcRFY5usitKSp",
        "_score" : 5.6485057,
        "_source" : {
          "@timestamp" : "2015-09-16T17:33:08.000Z",
          "id" : "44d9b10d-697e-43d4-a783-710b2c5c8d78",
          "source" : "Lawrence Alexander | Email Showcase",
          "published" : "2015-09-16T17:33:08Z",
          "title" : "It's #WatchWednesday check out the DKNY Product of The Week!",
          "media-type" : "Blog",
          "content" : """
Wedding & Diamond Jewellery, Watches, Earrings - H.Samuel the JewellerIf this email is not displayed correctly follow this link
http://click.mail.hsamuel.co.uk/?qs=8352710a2e5e143beb33e98219d6c85570a516a1d793d29aab9e58a66d0376f64c1f92fdf6cf1366
VIEW THIS EMAIL

Product of The Week DKNY
http://click.mail.hsamuel.co.uk/?qs=8352710a2e5e143bf83f5cbc715a1e77e40fc938d3d863171072b698eabbeba2b0f864827c75a429 Product of The Week
  -_-DKNYProductPage-_-Job4115264

E-mailFooterSpring2
```

20: Output of Stemming on word “product”

Searching: Searching of a query means retrieve specific information from documents. We can retrieves information from two documents like match phrases and filters.

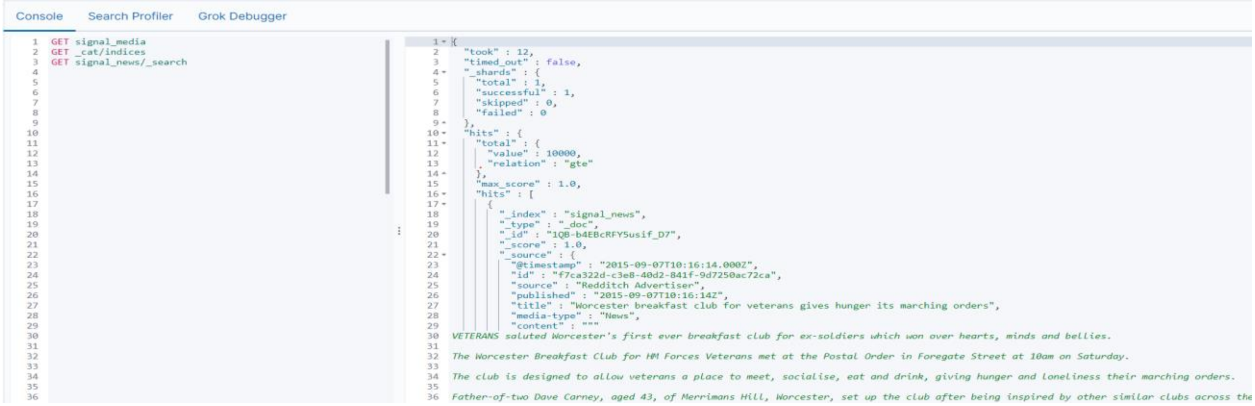


```
GET signal_media
GET _cat/indices
GET signal_news/_search
{
  "_source": "title",

   "size": 10,

     "query":{
       "bool": {
       "must": [],
         "filter":[]
       , "should": []



     }

}
}
```

```
{
  "took" : 608,
  "timed_out" : false,
  "_shards" : {
    "total" : 1,
    "successful" : 1,
    "skipped" : 0,
    "failed" : 0
  },
  "hits" : {
    "total" : {
      "value" : 10000,
      "relation" : "gte"
    },
    "max_score" : 1.0,
    "hits" : [
      {
        "_index" : "signal_news",
        "_type" : "_doc",
        "_id" : "1QB-b4EBcRFY5usif_D7",
        "_score" : 1.0,
        "_source" : {
          "title" : "Worcester breakfast club for veterans gives hunger its marching orders"
        }
      },
      {
        "_index" : "signal_news",
        "_type" : "_doc",
        "_id" : "1gB-b4EBcRFY5usif_D9",
        "_score" : 1.0,
        "_source" : {
          "title" : "Jumpshot Gives Marketers Renewed Visibility Into Paid and Organic Keywords With Launch of Jumpshot Elite"
        }
      },
```

Query: 1 Where in my document the published field have exact phrase.“ Jumpshot Gives Marketers Renewed Visibility Into Paid and Organic”
Query2: What pages on my document contain specific phrase?
Query 3: What pages on document contain a specific word from a title?

1) Step 1: search indices

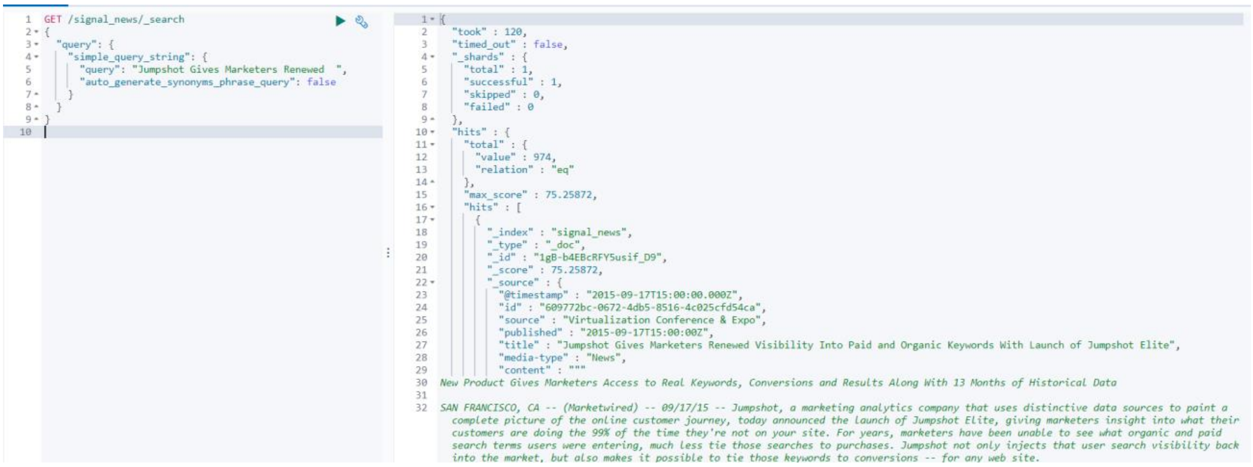


```
GET signal_media
GET _cat/indices
GET signal_news/_search
```

```
{
  "took" : 12,
  "timed_out" : false,
  "_shards" : {
    "total" : 1,
    "successful" : 1,
    "skipped" : 0,
    "failed" : 0
  },
  "hits" : {
    "total" : {
      "value" : 10000,
      "relation" : "gte"
    },
    "max_score" : 1.0,
    "hits" : [
      {
        "_index" : "signal_news",
        "_type" : "_doc",
        "_id" : "1QB-b4EBcRFY5usif_D7",
        "_score" : 1.0,
        "_source" : {
          "@timestamp" : "2015-09-07T10:16:14.000Z",
          "id" : "f7ca322d-c3e8-40d2-841f-9d7250ac72ca",
          "source" : "Redditch Advertiser",
          "published" : "2015-09-07T10:16:14Z",
          "title" : "Worcester breakfast club for veterans gives hunger its marching orders",
          "media-type" : "News",
          "content" : """
VETERANS saluted Worcester's first ever breakfast club for ex-soldiers which won over hearts, minds and bellies.

The Worcester Breakfast Club for HM Forces Veterans met at the Postal Order in Foregate Street at 10am on Saturday.

The club is designed to allow veterans a place to meet, socialise, eat and drink, giving hunger and loneliness their marching orders.

Father-of-two Dave Carney, aged 43, of Merrimans Hill, Worcester, set up the club after being inspired by other similar clubs across the
```

21: Search Indices

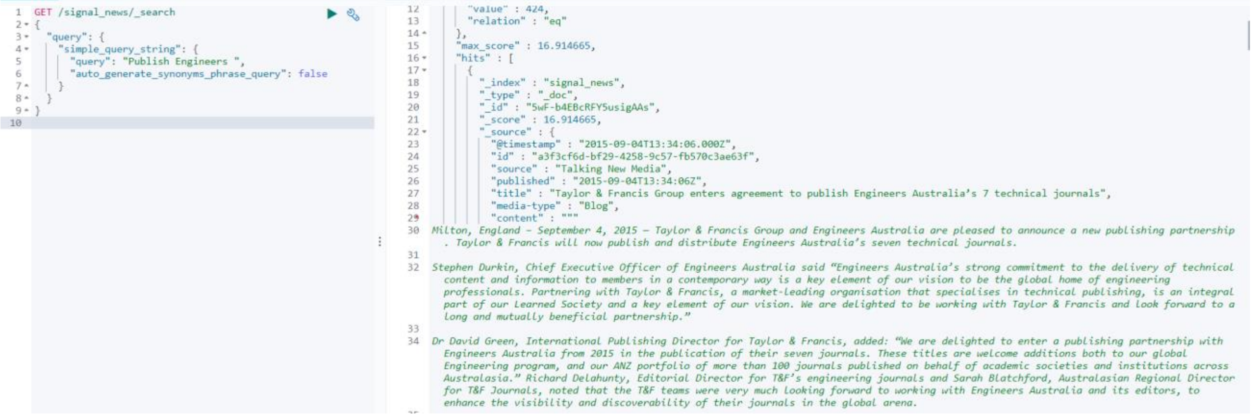
Step 2: Search Title exact phrase from document(Query1)



```
GET /signal_news/_search
{
  "query": {
    "simple_query_string": {
      "query": "Jumpshot Gives Marketers Renewed  ",
      "auto_generate_synonyms_phrase_query": false
    }
  }
}
```

```
{
  "took" : 120,
  "timed_out" : false,
  "_shards" : {
    "total" : 1,
    "successful" : 1,
    "skipped" : 0,
    "failed" : 0
  },
  "hits" : {
    "total" : {
      "value" : 974,
      "relation" : "eq"
    },
    "max_score" : 75.25872,
    "hits" : [
      {
        "_index" : "signal_news",
        "_type" : "_doc",
        "_id" : "1gB-b4EBcRFY5usif_D9",
        "_score" : 75.25872,
        "_source" : {
          "@timestamp" : "2015-09-17T15:00:00.000Z",
          "id" : "609772bc-0672-4db5-8516-4c025cfd54ca",
          "source" : "Virtualization Conference & Expo",
          "published" : "2015-09-17T15:00:00Z",
          "title" : "Jumpshot Gives Marketers Renewed Visibility Into Paid and Organic Keywords With Launch of Jumpshot Elite",
          "media-type" : "News",
          "content" : """
New Product Gives Marketers Access to Real Keywords, Conversions and Results Along With 13 Months of Historical Data

SAN FRANCISCO, CA -- (Marketwired) -- 09/17/15 -- Jumpshot, a marketing analytics company that uses distinctive data sources to paint a
  complete picture of the online customer journey, today announced the launch of Jumpshot Elite, giving marketers insight into what their
  customers are doing the 99% of the time they're not on your site. For years, marketers have been unable to see what organic and paid
  search terms users were entering, much less tie those searches to purchases. Jumpshot not only injects that user search visibility back
  into the market, but also makes it possible to tie those keywords to conversions -- for any web site.
```

Query 2: What pages on my document contain specific phrase?

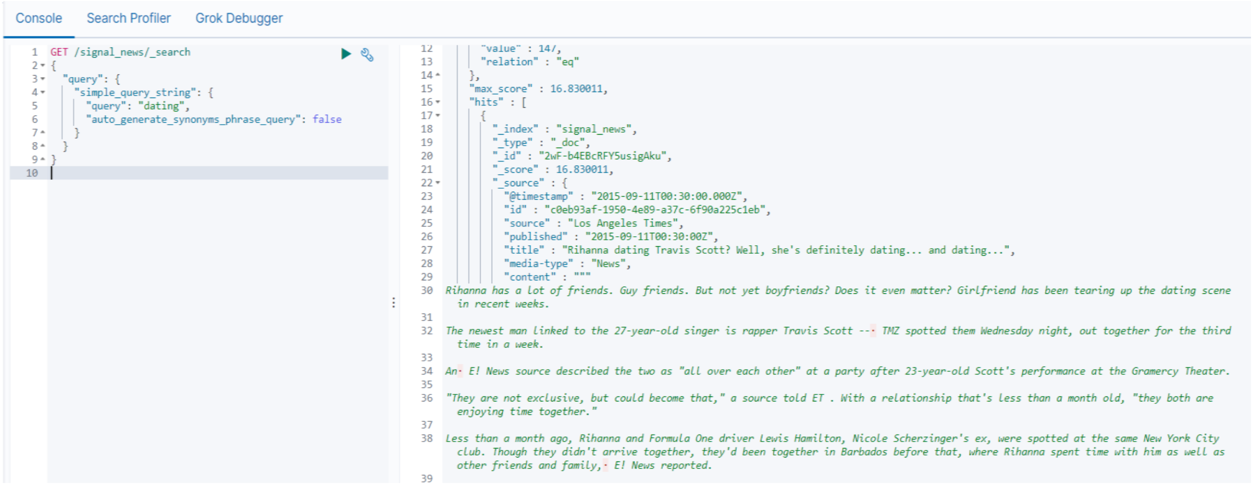


```
GET /signal_news/_search
{
  "query": {
    "simple_query_string": {
      "query": "Publish Engineers ",
      "auto_generate_synonyms_phrase_query": false
    }
  }
}
```

```
      "value" : 424,
      "relation" : "eq"
    },
    "max_score" : 16.914665,
    "hits" : [
      {
        "_index" : "signal_news",
        "_type" : "_doc",
        "_id" : "5wF-b4EBcRFY5usigAAs",
        "_score" : 16.914665,
        "_source" : {
          "@timestamp" : "2015-09-04T13:34:06.000Z",
          "id" : "a3f3cf6d-bf29-4258-9c57-fb570c3ae63f",
          "source" : "Talking New Media",
          "published" : "2015-09-04T13:34:06Z",
          "title" : "Taylor & Francis Group enters agreement to publish Engineers Australia's 7 technical journals",
          "media-type" : "Blog",
          "content" : """
Milton, England – September 4, 2015 – Taylor & Francis Group and Engineers Australia are pleased to announce a new publishing partnership
  . Taylor & Francis will now publish and distribute Engineers Australia's seven technical journals.

Stephen Durkin, Chief Executive Officer of Engineers Australia said "Engineers Australia's strong commitment to the delivery of technical
  content and information to members in a contemporary way is a key element of our vision to be the global home of engineering
  professionals. Partnering with Taylor & Francis, a market-leading organisation that specialises in technical publishing, is an integral
  part of our Learned Society and a key element of our vision. We are delighted to be working with Taylor & Francis and look forward to a
  long and mutually beneficial partnership."

Dr David Green, International Publishing Director for Taylor & Francis, added: "We are delighted to enter a publishing partnership with
  Engineers Australia from 2015 in the publication of their seven journals. These titles are welcome additions both to our global
  Engineering program, and our ANZ portfolio of more than 100 journals published on behalf of academic societies and institutions across
  Australasia." Richard Delahunty, Editorial Director for T&F's engineering journals and Sarah Blatchford, Australasian Regional Director
  for T&F Journals, noted that the T&F teams were very much looking forward to working with Engineers Australia and its editors, to
  enhance the visibility and discoverability of their journals in the global arena.
```

22: Output Specific word from Title

Query 3: What pages on document contain a specific word from a title?

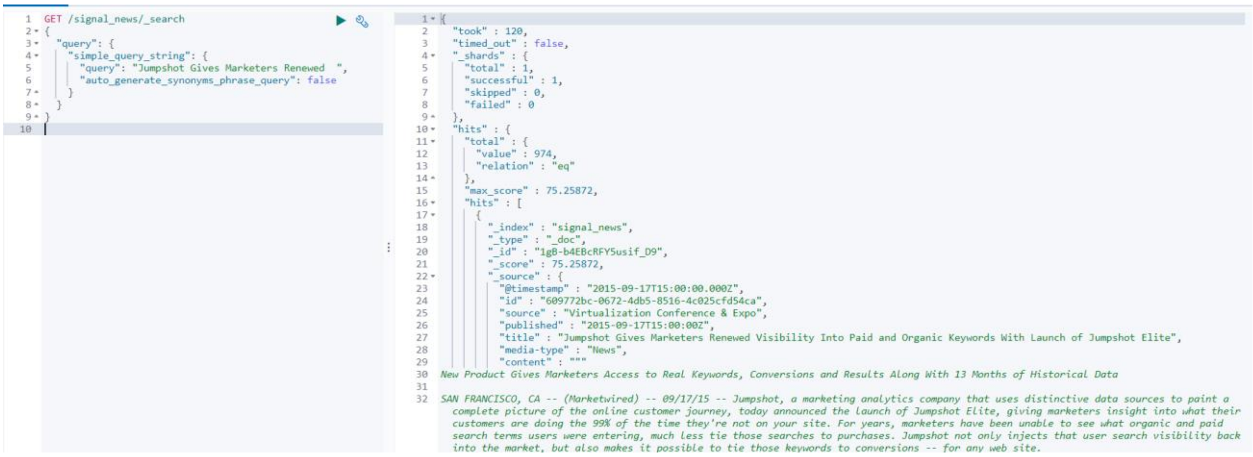


```
GET /signal_news/_search
{
  "query": {
    "simple_query_string": {
      "query": "dating",
      "auto_generate_synonyms_phrase_query": false
    }
  }
}
```

```
      "value" : 147,
      "relation" : "eq"
    },
    "max_score" : 16.830011,
    "hits" : [
      {
        "_index" : "signal_news",
        "_type" : "_doc",
        "_id" : "2wF-b4EBcRFY5usigAku",
        "_score" : 16.830011,
        "_source" : {
          "@timestamp" : "2015-09-11T00:30:00.000Z",
          "id" : "c0eb93af-1950-4e89-a37c-6f90a225c1eb",
          "source" : "Los Angeles Times",
          "published" : "2015-09-11T00:30:00Z",
          "title" : "Rihanna dating Travis Scott? Well, she's definitely dating... and dating...",
          "media-type" : "News",
          "content" : """
Rihanna has a lot of friends. Guy friends. But not yet boyfriends? Does it even matter? Girlfriend has been tearing up the dating scene
  in recent weeks.

The newest man linked to the 27-year-old singer is rapper Travis Scott --· TMZ spotted them Wednesday night, out together for the third
  time in a week.

An· E! News source described the two as "all over each other" at a party after 23-year-old Scott's performance at the Gramercy Theater.

"They are not exclusive, but could become that," a source told ET . With a relationship that's less than a month old, "they both are
  enjoying time together."

Less than a month ago, Rihanna and Formula One driver Lewis Hamilton, Nicole Scherzinger's ex, were spotted at the same New York City
  club. Though they didn't arrive together, they'd been together in Barbados before that, where Rihanna spent time with him as well as
  other friends and family,· E! News reported.

```

23: Search based on Query (Output)

Output: General Title search query with maximum score



```
GET /signal_news/_search
{
  "query": {
    "simple_query_string": {
      "query": "Jumpshot Gives Marketers Renewed  ",
      "auto_generate_synonyms_phrase_query": false
    }
  }
}
```

```
{
  "took" : 120,
  "timed_out" : false,
  "_shards" : {
    "total" : 1,
    "successful" : 1,
    "skipped" : 0,
    "failed" : 0
  },
  "hits" : {
    "total" : {
      "value" : 974,
      "relation" : "eq"
    },
    "max_score" : 75.25872,
    "hits" : [
      {
        "_index" : "signal_news",
        "_type" : "_doc",
        "_id" : "1gB-b4EBcRFY5usif_D9",
        "_score" : 75.25872,
        "_source" : {
          "@timestamp" : "2015-09-17T15:00:00.000Z",
          "id" : "609772bc-0672-4db5-8516-4c025cfd54ca",
          "source" : "Virtualization Conference & Expo",
          "published" : "2015-09-17T15:00:00Z",
          "title" : "Jumpshot Gives Marketers Renewed Visibility Into Paid and Organic Keywords With Launch of Jumpshot Elite",
          "media-type" : "News",
          "content" : """
New Product Gives Marketers Access to Real Keywords, Conversions and Results Along With 13 Months of Historical Data

SAN FRANCISCO, CA -- (Marketwired) -- 09/17/15 -- Jumpshot, a marketing analytics company that uses distinctive data sources to paint a
  complete picture of the online customer journey, today announced the launch of Jumpshot Elite, giving marketers insight into what their
  customers are doing the 99% of the time they're not on your site. For years, marketers have been unable to see what organic and paid
  search terms users were entering, much less tie those searches to purchases. Jumpshot not only injects that user search visibility back
  into the market, but also makes it possible to tie those keywords to conversions -- for any web site.
```

24: Title search with maximum score

References:

[1] Bajer, M. (2017). Building an IoT Data Hub with Elasticsearch, Logstash and Kibana. *2017 5th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW)*, (pp. 63-68).

[2]Kononenko, O. a. (2014). Mining Modern Repositories with Elasticsearch. *Proceedings of the 11th Working Conference on Mining Software Repositories* (pp. 328–331). New York, NY, USA: Association for Computing Machinery.

[3] https://www.elastic.co/guide/en/elasticsearch/reference/current/index.html

[4] https://www.kibana.co/guide/en/elasticsearch/reference/current

[5] https://research.signal-ai.com/newsir16/signal-dataset.html

[6] Lab 1 worsheet

[7] Lab 2 worksheet