

University of Essex
Department of Computer Sciences

CE706-Information Retrieval
Information Retrieval Assignment 2: Elasticsearch and Evaluation

Submitted By
DILPA RAO: **1906319**
VIRAJ KUMAR DEWANGAN: **1901181**

Guided By
Dr Alba Garcia Seco Herrera

Date of Submission
9 April 2020

TABLE OF CONTENTS

Abstract

	Page
A. Introduction	3
B. Setting up Elastic Search	3
C. Setting up Kibana	4
D. Loading the Dataset	5
E. Indexing in Kibana	5
F. Description of Indexed Document	6
G. Searching in Kibana	6
H. Types of Queries in Kibana	7
1. Simple Query	7
2. Content Based Query	8
3. Regexp Query	8
4. Range Query	9
5. Boolean Query	9
I. Building a Test Collection	9
1. Politics	9
2. Technology	10
3. Health	11
4. Software giants	11
5. Sports	11
6. Rain	12
7. Movie	12
8. Book	13
9. Love	14
10. Holiday	15
J. Evaluation and Query Results (Recall and Precision)	16
K. Crowdsourcing Task and its Experience	17
L. Engineering a Complete System	17
M. Suggestions on Scope and Future Improvements possible	18

ABSTRACT

In this report, we will explain how we made a search engine of signalmedia-1m.json dataset using Elastic Search Engine and the GUI interface Kibana and loaded the data in Jupyter Notebook using Python. In Kibana we used the Kibana Query Language (KQL) and did simple searches and executed queries and built a Test Collection based on certain events. We used evaluation metrics like precision and recall determining the accuracy of our search results.

KEYWORDS:

Indexing, Searching, building a Test Collection, Evaluation, Engineering a Complete System

A. INTRODUCTION

The data was downloaded from the site <http://research.signalmedia.co/newsir16/signal-dataset.html> and in order to start Elastic search on windows the following programs were used Java, Elasticsearch and Kibana and Python using Anaconda . Out of the 1 million documents, 5000 documents were loaded into elastic search by writing a python script in Jupyter Notebook. The Kibana software tool was used for indexing, searching, building a test collection and evaluation.

Software Installation was done using the following sites:

1. Java : <https://www.java.com/en/download/win10.jsp>
2. Elasticsearch : <https://www.elastic.co/downloads/elasticsearch>
3. Kibana : <https://www.elastic.co/downloads/kibana>

B. SETTING UP ELASTICSEARCH

After downloading the Elasticsearch, it was in a zip format and we extracted the file. The Bin folder was opened, then clicked on the **elasticsearch.bat** and **'Run as administrator'**.

```
C:\WINDOWS\System32\cmd.exe

at org.elasticsearch.action.ActionListener$1.onFailure(ActionListener.java:71) [elasticsearch-7.6.2.jar:7.6.2]
at org.elasticsearch.action.ActionListener$1.onResponse(ActionListener.java:65) [elasticsearch-7.6.2.jar:7.6.2]
at org.elasticsearch.action.ActionRunnable.lambda$supply$0(ActionRunnable.java:58) [elasticsearch-7.6.2.jar:7.6.2]
2]
at org.elasticsearch.action.ActionRunnable$2.doRun(ActionRunnable.java:73) [elasticsearch-7.6.2.jar:7.6.2]
at org.elasticsearch.common.util.concurrent.AbstractRunnable.run(AbstractRunnable.java:37) [elasticsearch-7.6.2.jar:7.6.2]
at org.elasticsearch.common.util.concurrent.TimedRunnable.doRun(TimedRunnable.java:44) [elasticsearch-7.6.2.jar:7.6.2]
at org.elasticsearch.common.util.concurrent.ThreadContext$ContextPreservingAbstractRunnable.doRun(ThreadContext.java:692) [elasticsearch-7.6.2.jar:7.6.2]
at org.elasticsearch.common.util.concurrent.AbstractRunnable.run(AbstractRunnable.java:37) [elasticsearch-7.6.2.jar:7.6.2]
at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1128) [?:?]
at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:628) [?:?]
at java.lang.Thread.run(Thread.java:830) [?:?]
Caused by: org.elasticsearch.tasks.TaskCancelledException: cancelled
at org.elasticsearch.search.fetch.FetchPhase.execute(FetchPhase.java:150) ~[elasticsearch-7.6.2.jar:7.6.2]
at org.elasticsearch.search.SearchService.executeFetchPhase(SearchService.java:387) ~[elasticsearch-7.6.2.jar:7.6.2]
at org.elasticsearch.search.SearchService.executeQueryPhase(SearchService.java:367) ~[elasticsearch-7.6.2.jar:7.6.2]
at org.elasticsearch.search.SearchService.lambda$executeQueryPhase$1(SearchService.java:343) ~[elasticsearch-7.6.2.jar:7.6.2]
at org.elasticsearch.action.ActionListener.lambda$map$2(ActionListener.java:146) ~[elasticsearch-7.6.2.jar:7.6.2]
at org.elasticsearch.action.ActionListener$1.onResponse(ActionListener.java:63) ~[elasticsearch-7.6.2.jar:7.6.2]
... 9 more
```

C. SETTING UP KIBANA

Similarly, like Elasticsearch we extracted the zip file and selected the Bin folder and clicked on Kibana.bat and 'Run as Administrator'. Once properly run, it will show an output like this:

```
C:\WINDOWS\system32\cmd.exe

log [16:54:56.625] [info][status][plugin:siem@7.6.2] Status changed from uninitialized to green - Ready
log [16:54:56.629] [info][status][plugin:remote_clusters@7.6.2] Status changed from uninitialized to green - Ready
log [16:54:56.631] [info][status][plugin:cross_cluster_replication@7.6.2] Status changed from uninitialized to green - Ready
- Ready
log [16:54:56.645] [info][status][plugin:upgrade_assistant@7.6.2] Status changed from uninitialized to green - Ready
log [16:54:56.662] [info][status][plugin:uptime@7.6.2] Status changed from uninitialized to green - Ready
log [16:54:56.665] [info][status][plugin:oss_telemetry@7.6.2] Status changed from uninitialized to green - Ready
log [16:54:56.667] [info][status][plugin:file_upload@7.6.2] Status changed from uninitialized to green - Ready
log [16:54:56.669] [info][status][plugin:data@7.6.2] Status changed from uninitialized to green - Ready
log [16:54:56.672] [info][status][plugin:lens@7.6.2] Status changed from uninitialized to green - Ready
log [16:54:56.687] [info][status][plugin:snapshot_restore@7.6.2] Status changed from uninitialized to green - Ready
log [16:54:56.692] [info][status][plugin:input_control_vis@7.6.2] Status changed from uninitialized to green - Ready
log [16:54:56.694] [info][status][plugin:kibana_react@7.6.2] Status changed from uninitialized to green - Ready
log [16:54:56.696] [info][status][plugin:management@7.6.2] Status changed from uninitialized to green - Ready
log [16:54:56.698] [info][status][plugin:navigation@7.6.2] Status changed from uninitialized to green - Ready
log [16:54:56.700] [info][status][plugin:region_map@7.6.2] Status changed from uninitialized to green - Ready
log [16:54:56.706] [info][status][plugin:telemetry@7.6.2] Status changed from uninitialized to green - Ready
log [16:54:56.708] [info][status][plugin:markdown_vis@7.6.2] Status changed from uninitialized to green - Ready
log [16:54:56.709] [info][status][plugin:ui_metric@7.6.2] Status changed from uninitialized to green - Ready
log [16:54:56.711] [info][status][plugin:table_vis@7.6.2] Status changed from uninitialized to green - Ready
log [16:54:57.301] [info][status][plugin:timelion@7.6.2] Status changed from uninitialized to green - Ready
log [16:54:57.303] [info][status][plugin:tagcloud@7.6.2] Status changed from uninitialized to green - Ready
log [16:54:57.305] [info][status][plugin:metric_vis@7.6.2] Status changed from uninitialized to green - Ready
log [16:54:57.307] [info][status][plugin:vega@7.6.2] Status changed from uninitialized to green - Ready
log [16:54:57.995] [warning][reporting] Generating a random key for xpack.reporting.encryptionKey. To prevent pending reports from failing on restart, please set xpack.reporting.encryptionKey in kibana.yml
log [16:54:58.003] [info][status][plugin:reporting@7.6.2] Status changed from uninitialized to green - Ready
log [16:54:58.034] [info][listening] Server running at http://localhost:5601
log [16:55:03.217] [info][server][Kibana][http] http server running at http://localhost:5601
```

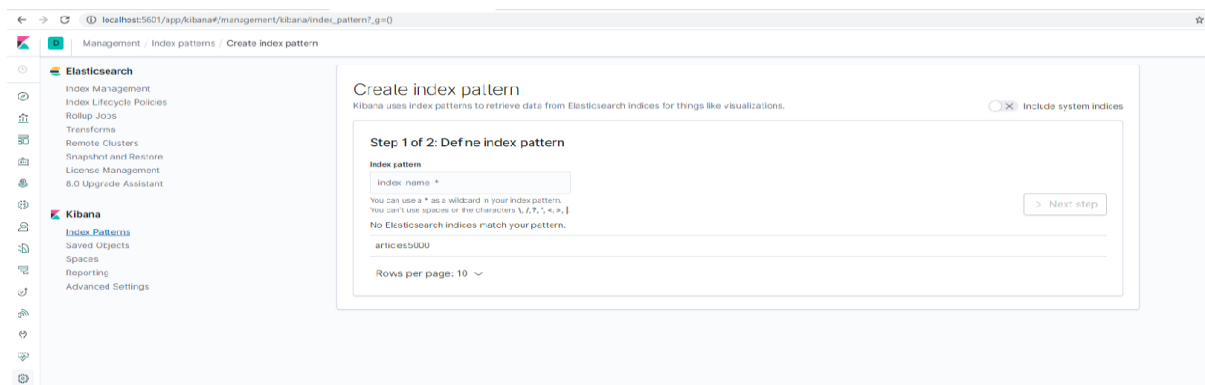
Both should remain working in the background. We can for proper installation by entering this URL for Elasticsearch and Kibana respectively - <http://localhost:9200/> and <http://localhost:5601/>.

D. LOADING THE DATASET

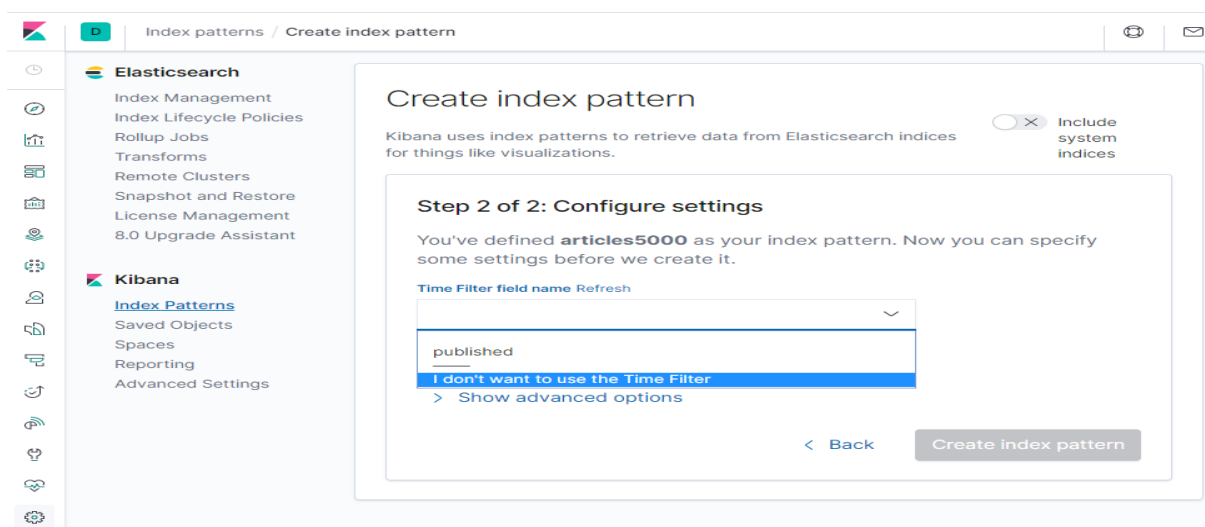
The dataset was loaded in Elasticsearch by running a script file in python. We selected only 5000 documents as loading more data was giving system errors and taking a long time, making the machine hang.

E. INDEXING IN KIBANA

After ensuring that the dataset is loaded, then the next step was indexing. Index was created by the following way:



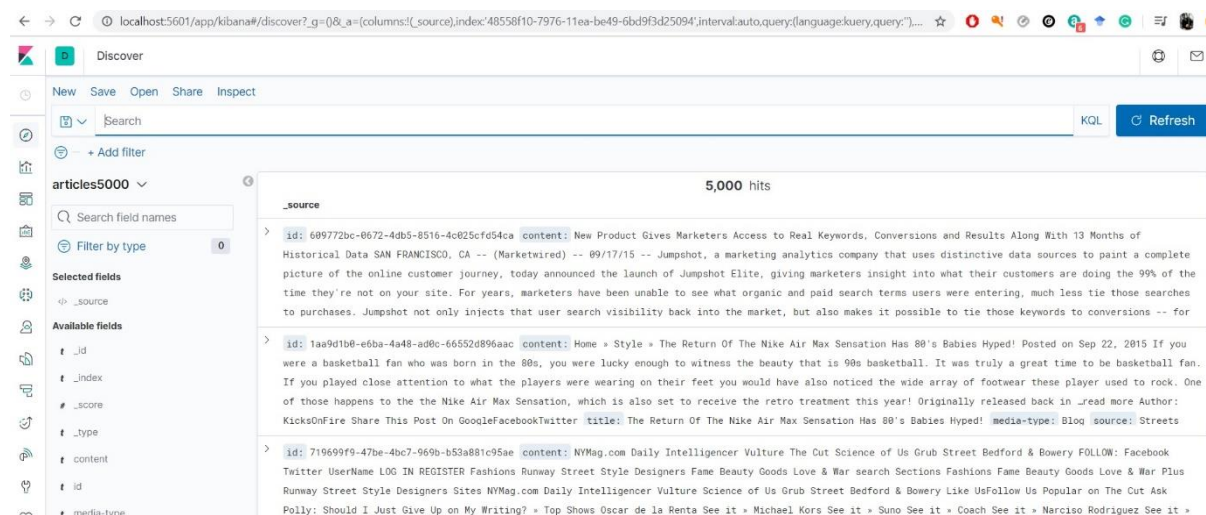
- In Kibana, open Management, then click Index Patterns.
- Click on Create index pattern and type 'articles5000' in the textbox.
- Click on the next step button.
- In Step 2 of 2: Configure Settings click on the down button arrow and select 'I don't want to use the Time Filter'.
- Finally click on the 'Create Index pattern button'.



F. DESCRIPTION OF INDEXED DOCUMENT

So, now the index pattern is defined on our 5000 documents. It is to be noted that the pattern must exist in the Elasticsearch and it must contain data for an index to be created. We can check for the indices in the management icon the last tool on the left menu of Kibana and click on it to get to Elasticsearch – Index management and Kibana – Index Patterns screen page.

The screenshot of our indexed document is displayed below-



The indexed document contains various fields like source, id, index, type, content, media type, date of publication on which the data is available and can be used to define custom searches and results in Kibana.


G. SEARCHING IN KIBANA

We can do simple search and more complex queries in Kibana. The Discover icon which is on the top of the left menu allows us to search, write complex queries and to save it. We can alternatively write the queries using the Dev Tools icon on the left menu.



Discover

New Save Open Share Inspect

 **Search** KQL Refresh

+ Add filter

articles5000

Search field names

Filter by type

Selected fields

_source

Available fields

5,000 hits

_source


- > **id:** db2be30b-630e-478c-a13f-06e404a81844 **content:** It is wisteria blooming time in Australia. I love wisteria and at one house we used to have it trained along trellis sitting on top of the back fence but it is hard beast to keep under control. I seem to recall the blooms appear on two year old wood, so great care is needed when you go at it with an axe and chainsaw. Please don't bother bringing white wisteria into the post. Who cares about white wisteria. I am talking mauve wisteria, proper like init. The Japanese like to torture control plants to a very fine degree and
- > **id:** d72baf68-1cd2-4152-86cb-7202e2780b29 **content:** ShareTwitter Facebook Google+ Email Listen Listening... / Originally published on September 23, 2015 7:53 am Copyright 2015 NPR. To see more, visit http://www.npr.org/. North Carolina Public Radio - WUNC is created in partnership with: © 2015 WUNC Bringing The World Home To You 120 Friday Center Dr | Chapel Hill, NC 27517 91.5 Chapel Hill 88.9 Manteo 90.9 Rocky Mount **title:** Lawsuits Will Be Next Battle In Sage Grouse Conservation Saga **media-type:** News **source:** WUNC **published:** Sep 23, 2015 @ 13:11:58.000
- > **id:** 6f4ed3e4-2c81-41a8-ba79-fd06ad86cd1a **content:** Listen NEW YORK Microsoft has given longtime executive Brad Smith the title of president, as the company continues its transition to a new generation of leadership. Smith's full title will be president and chief legal officer. He has been Microsoft's general counsel since 2002 and became an executive vice president in 2011.CEO Satya Nadella announced the appointment in an email to employees. Nadella said he wants Smith to play a bigger role in strengthening Microsoft's relationships and representing the company publicly. Smith

H. TYPES OF QUERIES IN KIBANA

1. Simple Query

Discover

New Save Open Share Inspect

 **NASA** KQL Refresh

+ Add filter

articles5000

Search field names

Filter by type

Selected fields

_source

Available fields

18 hits

_source

- > **content:** (NEW YORK) - The red planet has space fans red hot with anticipation. **NASA** said it will announce Monday that a Mars mystery has been "solved," but did not say what the mystery was, only... The post 'Mars Mystery Solved': **NASA** to Announce 'Major' News appeared first on... (NEW YORK) - The red planet has space fans red hot with anticipation. **NASA** said it will announce Monday that a Mars mystery has been "solved," but did not say what the mystery was, only that it "detail a major science finding." "I suspect it's going to be something that will increase our interest in going to Mars," former astronaut Mike Massimino said. Speculation on social media about what the announcement might be included aliens and water on the planet. Massimino said that if the announcement is that there is water, that could have great significance. "That there is some sort of life form that might be discovered, and it's also the... **title:** 'Mars Mystery Solved': **NASA** to Announce 'Major' News **id:** 33ede757-4c7e-44ad-808e-c7401061af32
- > **content:** SOURCE California Science Center LOS ANGELES "The Journey to Space Exhibition and 3D Film explores both the history and bright future of human space travel, along with the risks and innovative solutions involved," said Jeffrey N. Rudolph Journey to Space: The Exhibition offers hands-on and multimedia experiences Guests will get a hands-on, class-board experience at what it takes to live and work in space. The exhibition examines the extraordinary environment of space, including the very real dangers astronauts face during their mission above-Earth and the adaptations that engineers have developed to help them survive. Visitors will learn about the vacuum of space, radiation, meteoroids, and temperature extremes, while getting a look at all of the ways in which the forbidding environment of space can challenge human exploration. Guests will find hands-on activities to explore the science of getting to space. This includes learning about "weightlessness" and how it affects the body during a long-term
- > **content:** Press Release From: Aerospace Industries Association Although the end of summer indicates the beginning of classes for students across the country, it also signals the kickoff of the world's largest annual student rocketry contest. Registration for the Team America Rocketry Challenge (TARC) is now open for teams of 7-12th grade students through December 4. TARC is the U.S. aerospace and defense industry's flagship program designed to encourage students to pursue study and careers in science, technology, engineering and math (STEM). Structured to emulate the aerospace industry's design, fabrication and testing process, TARC requires teams to build and fly a model rocket that meets challenging design requirements and precise targets for altitude and flight duration. Each year, TARC's rules and scoring parameters change to challenge the students' ingenuity and encourage a fresh approach to rocket design. This year's rules require teams to build and launch a rocket carrying two raw eggs to 850 feet and
- > **content:** **NASA** mistakes the moon for the sun in a Tweet, prompting a flurry of social media ridicule and eventually, a correction. **NASA** mistakes the moon for the sun. **title:** **NASA** can't tell the difference between the sun and the moon **id:** b043307a-2580-410e-a327-ea7fd2144668 **media-type:** news **source:** MyInfores **published:** Sep 10, 2015 @ 13:54:35.000 **id:** L8HMYE8E1EjYHNglyem **_type:** article **index:** articles5000 **score:** 0
- > **content:** The gold standard of measuring inequality is the Gini coefficient . It's been rising in the United States since the 1970s. But last year, when Gini was calculated for metro areas, Seattle didn't make the top 10. Those included Bridgeport, Conn.; Naples, Fla., New York; Miami, and Port St. Lucie, Fla. The Wall Street Journal crunched Census data and came up with slightly different rankings . Still, Seattle didn't make the top 30 metropolitan areas in income inequality (nor the 30 least unequal, although Olympia did). That doesn't mean all is well, of course. A new report from the federal Bureau of Labor Statistics looks at the so-called 90-10 ratio. In other words, the ratio of the 90th percentile in income to the 10th percentile. Thus, the best paid 10 percent of wage earners in the country (the 90th percentile) earned at least \$88,330 annually, while the lowest paid 10 percent (the 10th percentile) earned less than \$10,190. The data cover 2003 to 2013, and not surprisingly location, metro size and

2. Content Based Query

The screenshot shows the Elasticsearch Kibana interface with a query of `content:"Lung Cancer"`. The left sidebar shows the 'Selected fields' list with `_source` selected. The main panel displays 12 hits. The first hit is a document about the role of Memantine in neuroprotection during whole brain irradiation. The second hit is a document about Clovis Oncology's rating. The third hit is a document about Corey Seager's performance. The fourth hit is a document about radon exposure and lung cancer. The fifth hit is a document about oncology researchers.

3. Regexp Query

The screenshot shows the Elasticsearch Kibana interface with a query of `content:robot*`. The left sidebar shows the 'Selected fields' list with `_source` selected. The main panel displays 26 hits. The first hit is a document about the human workforce being made obsolete. The second hit is a document about the Japanese humanoid robot Pepper. The third hit is a document about the SOURCE California Science Center. The fourth hit is a document about the NEW YORK, NY -- (Marketsrod) --. The fifth hit is a document about neurosciences.

4. Range Query

The screenshot shows the Elasticsearch Discover interface. The search bar contains the query `content: robot*`. The left sidebar shows the 'Selected fields' as `_source` and 'Available fields' including `_id`, `_index`, `_score`, `_type`, `content`, `id`, `media-type`, `published`, `source`, and `title`. The main panel displays 26 hits. The first hit is a document from the `articles5000` index with a score of 0. The content of the first hit is: "As advancing technology changes the face of employment in the 21st century - is the human workforce being made obsolete? Martin Ford is the founder of a Silicon Valley software firm and the author of Rise of the Robots: Technology and the Threat of a Jobless Future. Geoff Colvin is senior editor at Large at Fortune magazine and author of Humans Are Underrated: What High Achievers Know That Brilliant Machines Never Will. (cont...) Source: Will Robots create more jobs than they destroy? | Technology | The Guardian [id: b6cf54d1-dc78-4aa3-9db1-772c08676ccb] title: 'Will robots create more jobs than they destroy?'... media-type: Blog source: whoar.co.nz published: Sep 6, 2015 8 21:41:26.000 _id: k5H9XREIEifYHkQpynj _type: article _index: articles5000 _score: 0".

5. Boolean Query

The screenshot shows the Elasticsearch Discover interface. The search bar contains the query `content: (*Microsoft Corp* or *IBM*) and media-type: "news"`. The left sidebar shows the 'Selected fields' as `_source` and 'Available fields' including `_id`, `_index`, `_score`, `_type`, `content`, `id`, `media-type`, `published`, `source`, and `title`. The main panel displays 17 hits. The first hit is a document from the `articles5000` index with a score of 0. The content of the first hit is: "Listen NEW YORK Microsoft has given longtime executive Brad Smith the title of president, as the company continues its transition to a new generation of leadership. Smith's full title will be president and chief legal officer. He has been Microsoft's general counsel since 2002 and became an executive vice president in 2011. CEO Satya Nadella announced the appointment in an email to employees. Nadella said he wants Smith to play a bigger role in strengthening Microsoft's relationships and representing the company publicly. Smith will help lead the company on issues like privacy, security and accessibility, he said. Smith, 56, joined Microsoft in 1993 and has held several other titles, including company secretary. Smith is a familiar face for investors because he has been a regular presence on the company's investor conference calls for almost a decade, said industry analyst Katherine Ebert of Piper Jaffray. 'It is interesting that they would put him as president of the company because he".

I. BUILDING A TEST COLLECTION

The following 10 queries were built, and their results are captured as follows-

1. Politics

First query retrieves 21 documents based on the content: **Prime Minister Narendra Modi** and adding additional condition of **media-type: "news"** and **published >= "2015-**

09-15", we retrieve 12 documents. So, the number of hits were reduced from 21 to 12. It can be clearly shown from the figures displayed below.

The image displays two screenshots of the Elasticsearch Kibana interface, illustrating the process of refining search results.

Top Screenshot: The search query is "content: 'Prime Minister Narendra Modi'". The left sidebar shows the "Selected Fields" section with the following fields: `_source`, `_id`, `_index`, `_score`, `_type`, `content`, `id`, `media-type`, `published`, `source`, and `title`. The main panel shows 21 hits. The first hit is from "Mumbai News.Net" dated Wednesday 2nd September, 2015, with the content: "Oppo eyes India opportunity to partner with Foxconn". The second hit is from "Mumbai News.Net" dated Tuesday 22nd September, 2015, with the content: "SC stays summons issued to Kejriwal for making inflammatory speech".

Bottom Screenshot: The search query is refined to "content: 'Prime Minister Narendra Modi' and media-type: 'news' and published >= '2015-09-15'". The left sidebar shows the same "Selected Fields" section. The main panel shows 12 hits. The first hit is from "Mumbai News.Net" dated Tuesday 22nd September, 2015, with the content: "SC stays summons issued to Kejriwal for making inflammatory speech". The second hit is from "Mumbai News.Net" dated Tuesday 22nd September, 2015, with the content: "President Obama opened a three-day series of meetings at the United Nations on Sunday by calling on all countries to 'step up' efforts to eradicate poverty".

2. Technology

First query retrieves 14 documents based on the content: **"data science" or "AI" or "Robotics"** and adding additional condition of **media-type: "news"**, we retrieve 10 documents. So, the number of hits were reduced from 14 to 10. It can be clearly shown from the figures displayed below.

content: "Data science" or "AI" or "Robotics"

articles5000

Search field names

Filter by type

Selected fields

Available fields

14 hits

Content: Chinese giant Baidu is getting in on the phone personal assistant game with the launch of Duer, marking a major improvement on the previous system launched on the Baidu app three years ago. The post appeared first on Silicon. Chinese giant Baidu is getting in on the phone personal assistant game with the launch of Duer, marking a major improvement on the previous system launched on the Baidu app three years ago. Baidu's Duer, which effectively translates to "Du Secretary" was given a major demonstration with aims of using artificial intelligence (AI) and machine learning to challenge the systems developed by Apple and Google. Much has been made of its development considering that one of the driving forces behind its machine learning technology is world-renowned expert in the field, Andrew Ng, who joined the company back in 2014. Much like its competitors, Baidu's Duer is working towards allowing the user to place orders online with services like buying cinema tickets or ordering takeaway food.

Content: The Japanese humanoid robot Pepper, which sold out of its first 1,000 units in one minute in Japan this June, will get a personality makeover for the US market: it'll go from cute and bubbly to snarky and sarcastic, MIT Technology Review reports. Editor Will Knight met a Pepper unit in Boston this week, and reported back some very distinct changes in the robot's personality designed to make it more appealing to Americans: high fives instead of bows; sarcastic seizes instead of songs. In the MIT report, Knight said he asked an Americanized Pepper if it's like Terminator, to which it responded: "Do I really have to answer that? Pepper is a robot that's designed to recognize human emotions, and is supposed to read social situations so it can interact with you like a person can. And since people already ask Siri questions that get sassy comebacks, Pepper's gotta be ready. Advertisement "In the U.S., we have this kind of C-3PO idea, where he's kind of snarky and kind of smart," Alla Pyros.

Content: A change in prime minister could boost the Australian economy, analysts say. Source: AAP AUSTRALIAN businesses are anticipating a shot in the arm from the change in prime minister before Malcolm Turnbull has even spelled out policy. INDUSTRY bodies welcomed the arrival of a PM with a track record in business and warned to the idea that the former investment banker can succeed where Tony Abbott failed, using a more collegiate and co-operative style of government to push through economic reforms. Analysts say he could lift consumer confidence through his popularity with voters and help business by articulating the need for reform. Business Council of Australia president Catherine Livingstone said companies would respond to leadership that "explains our national challenges ... while respecting the intelligence of the community to embrace change for the better". "A stronger economy for all will require a 10-year plan including ambitious tax reform that supports growth and fairness, a new framework for

Content: SAN MATEO, Calif., and NEW YORK, Sept. 18, 2015 /PRNewswire/ -- Feedzai, a DATA SCIENCE company that uses real-time, machine-based learning to analyze big data and minimize risk in the financial services industry, today announced a partnership with Secure, the industry leader in real-time online identity verification solutions. Working together, the two companies will provide their mutual clients with a "plug-and-play" solution that combines online and social data-based authentication with fraud risk scoring. Feedzai will incorporate the Secure Social Biometrics Platform as part of its Fraud Prevention that Learns software to provide ultra-fast processing of Big Data to minimize fraudulent transactions, chargebacks and reduce manual reviews costs for financial institutions. Click here to share this news on Twitter: http://t.t.ec/8aC9a Feedzai also announced updates to its software platform that provide improved risk management and fraud prevention solutions. Its powerful machine learning software

Content: Share. New games, new modes. By Sega and Sports Interactive have announced three new Football Manager games that are scheduled for release this year. As expected, Football Manager 2016 will be coming out on PC and Mac, but Football Manager Touch will also be coming out for PC and tablets. Previously known as Football Manager Classic, Touch is the quick play mode which has been within the full game before (and for tablets since March). This is the first time it will be available separately through Steam and on tablets. Football Manager Mobile will also be coming to all iOS and Android devices. "For some time now it's been clear to us that there is no 'one size fits all' football management experience," says Sports Interactive head Miles Jacobson. "The introduction of Football Manager Touch as a standalone offering – playable across computer and tablet – means that we now offer something for everyone." As for new features, FM2016 now includes a create-a-club mode, and a fantasy draft mode where you

content: ("data science" or "AI" or "Robotics") and media-type: "news"

articles5000

Search field names

Filter by type

Selected fields

Available fields

10 hits

Content: Chinese giant Baidu is getting in on the phone personal assistant game with the launch of Duer, marking a major improvement on the previous system launched on the Baidu app three years ago. The post appeared first on Silicon. Chinese giant Baidu is getting in on the phone personal assistant game with the launch of Duer, marking a major improvement on the previous system launched on the Baidu app three years ago. Baidu's Duer, which effectively translates to "Du Secretary" was given a major demonstration with aims of using artificial intelligence (AI) and machine learning to challenge the systems developed by Apple and Google. Much has been made of its development considering that one of the driving forces behind its machine learning technology is world-renowned expert in the field, Andrew Ng, who joined the company back in 2014. Much like its competitors, Baidu's Duer is working towards allowing the user to place orders online with services like buying cinema tickets or ordering takeaway food.

Content: A change in prime minister could boost the Australian economy, analysts say. Source: AAP AUSTRALIAN businesses are anticipating a shot in the arm from the change in prime minister before Malcolm Turnbull has even spelled out policy. INDUSTRY bodies welcomed the arrival of a PM with a track record in business and warned to the idea that the former investment banker can succeed where Tony Abbott failed, using a more collegiate and co-operative style of government to push through economic reforms. Analysts say he could lift consumer confidence through his popularity with voters and help business by articulating the need for reform. Business Council of Australia president Catherine Livingstone said companies would respond to leadership that "explains our national challenges ... while respecting the intelligence of the community to embrace change for the better". "A stronger economy for all will require a 10-year plan including ambitious tax reform that supports growth and fairness, a new framework for

Content: SAN MATEO, Calif., and NEW YORK, Sept. 18, 2015 /PRNewswire/ -- Feedzai, a DATA SCIENCE company that uses real-time, machine-based learning to analyze big data and minimize risk in the financial services industry, today announced a partnership with Secure, the industry leader in real-time online identity verification solutions. Working together, the two companies will provide their mutual clients with a "plug-and-play" solution that combines online and social data-based authentication with fraud risk scoring. Feedzai will incorporate the Secure Social Biometrics Platform as part of its Fraud Prevention that Learns software to provide ultra-fast processing of Big Data to minimize fraudulent transactions, chargebacks and reduce manual reviews costs for financial institutions. Click here to share this news on Twitter: http://t.t.ec/8aC9a Feedzai also announced updates to its software platform that provide improved risk management and fraud prevention solutions. Its powerful machine learning software

Content: Share. New games, new modes. By Sega and Sports Interactive have announced three new Football Manager games that are scheduled for release this year. As expected, Football Manager 2016 will be coming out on PC and Mac, but Football Manager Touch will also be coming out for PC and tablets. Previously known as Football Manager Classic, Touch is the quick play mode which has been within the full game before (and for tablets since March). This is the first time it will be available separately through Steam and on tablets. Football Manager Mobile will also be coming to all iOS and Android devices. "For some time now it's been clear to us that there is no 'one size fits all' football management experience," says Sports Interactive head Miles Jacobson. "The introduction of Football Manager Touch as a standalone offering – playable across computer and tablet – means that we now offer something for everyone." As for new features, FM2016 now includes a create-a-club mode, and a fantasy draft mode where you

Content: Hong Kong plans to rejuvenate its flagship industrial estates to attract more modern and specialized types of manufacturing, said Allen Ma Kam-sing, chief executive of Hong Kong Science and Technology Parks Corporation. The first phase is likely to see the building of more multi-story factory buildings by 2020. He said, ahead of the 2015 International Association of Science Parks and Areas of Innovation World Conference to be held in Beijing from Tuesday to Friday. The plan will prioritize high value-added tenants involved in industries such as advanced robotics, pharmaceuticals and biomedical devices, said Ma. Media-type: news id: 4802ec0-3ec7-4890-a046-fa10f97feas title: HK plans to rejuvenate industrial estates source: CHNdaily.com.cn published: Sep 12, 2015 @ 00:55:10.800 id: 0q9WVEE5EYwQvrytu _type: article _index: articles5000 _score: 0

3. Health

First query retrieves 12 documents based on the content: **"Lung Cancer"** and adding additional condition of **media-type: "news"**, we retrieve 9 documents. So, the number of hits were reduced from 12 to 9.

4. Software giants

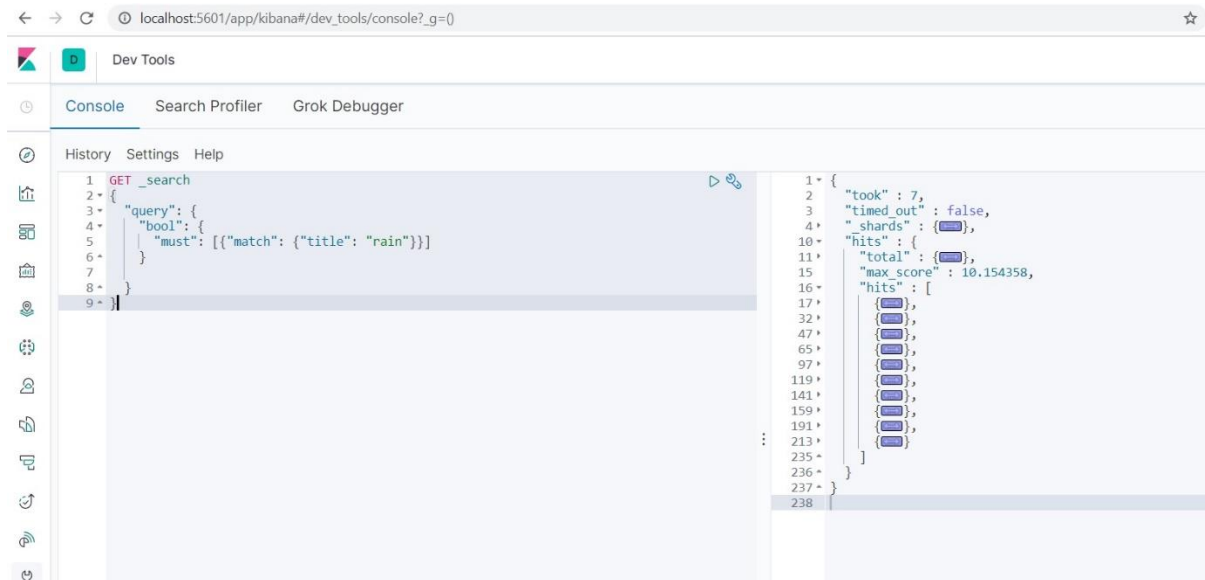
First query retrieves 21 documents based on the content: **"Microsoft Corp" or "IBM"** and adding additional condition of **media-type: "blog"**, we retrieve 4 documents. So, the number of hits were reduced from 21 to 4.

5. Sports

First query retrieves 48 documents based on the content: **"Rugby"** and adding additional condition of **"World Cup 2015"**, we retrieve 4 documents. So, the number of hits were reduced from 48 to 4. It can be clearly shown from the figures displayed below.

6. Rain

Rain yields the following search with hit number 10. This is obvious by counting the number of hits in blue rectangles in the figure.



7. Movie

Similarly, the search on movie gives 119 hits which is shown below –



8. Book

Initial query on the word: “book” gives 211 hits which reduces to 2 hits when publishing date filter of Sep 19 2015 is added. This is displayed below-

The screenshot shows the Kibana Discover interface. The search bar contains the word "book". The results list shows four entries, each with a "content" field. The first entry discusses fashion staples and the "No white after Labor Day" rule. The second entry discusses Iron Maiden's album "The Book of Souls". The third entry discusses the Iran Deal. The fourth entry discusses Leigh's performance in the Super League.

content
These Fashion Staples Encourage One to Wear White After Labour Day By: - Published: Sep 4, 2015 The origins of the "No white after Labor Day" rule are shrouded in mystery and confusion. It's possible that the rule was conceived simply as a reminder not to wear lightweight, easily stained white clothing outside of the summer season, which was book -ended by Memorial Day and Labor Day. It's also thought that in the late 1800s and early 1900s, the white after Labor Day rule was part of a vast and complex set of fashion and etiquette rules meant to separate older land-owning families from the nouveau riche population in the U.S. However, since the 1950s and 60s, when the white after Labor Day custom reached its peak and then began to decline, it's become entirely acceptable to wear white
Heavy metal rockers Iron Maiden have landed the top spot with their latest album, five years after The Final Frontier in 2010. The London-based band, led by frontman Bruce Dickinson, released their 16th studio record, The Book Of Souls, which went straight into Number One in the UK's Official Albums Chart with combined sales of just over 60,000. Iron Maiden are back at the top of the album chart after a five-year absence. The group's first double album and their longest to date, which was delayed while Dickinson recovered from tongue cancer, also topped the vinyl albums charts. DJ Sigala, originally from Norfolk, toppled Justin Bieber from the number one place in the singles chart with his track Easy Love. The track, which samples Jackson 5's 1970 Top 10 hit ABC, has racked up a combined
Influential Emirati Business Leader And Commentator: "Iran Deal Goes From Risky To Farical" HT: Memri. At First, "I Shrugged Off The News" Of The Secret Agreement "As A Figment Of Someone's Heated Imagination" In his August 25 article, Al-Habtoor wrote: "When I first learned from the news that the International Atomic Energy Agency (IAEA) had signed a secret agreement permitting Iran to self-monitor at least one of its major nuclear sites, I shrugged off the news as a figment of someone's heated imagination. "It is inconceivable that the world's nuclear watchdog, known for its professionalism and stringent monitoring, would sign off on something so bizarre - or so I initially believed. "Iraq, whose nuclear activities, both civilian and military, were dismantled
LEIGH'S hopes of promotion to the Super League were dented with a 34-12 defeat to Halifax. Paul Rowley's team was stunned by their second tier rivals, upsetting the form book to register their opening qualifying victory. Leigh had beaten their Shay hosts on both previous meetings this year but found themselves 28-0 down at half-time. Facing their most important 40 minutes of the year, the visitors made an encouraging start to the second period. Adam Higson and Fulful Moimoi touched down within nine minutes of the re-start to reduce their arrears to just eight points. But that was close as Leigh came to following up last week's encouraging victory over Sheffield Eagles. Halifax steadied their nerves to put the game out of reach with tries from Scott Murrell and Ste Tyrer. Share article

The screenshot shows the Kibana Discover interface with a filter applied: "book published on Sep 19 2015". The results list shows two entries, each with a "published" field. The first entry is "articles5000" with a published date of "Sep 19, 2015 @ 04:18:47.000". The second entry is "articles5000" with a published date of "Sep 19, 2015 @ 04:18:47.000".

_index	_type	published	_id
articles5000	article	Sep 19, 2015 @ 04:18:47.000	FtTVWHEBveVKN2BEyRYW
articles5000	article	Sep 19, 2015 @ 04:18:47.000	8tTWIHEBveVKN2BE21DS

9. Love

The plain vanilla query on Love yields 379 hits and when filtered with media type blog just gives 170 results. This is displayed below –

The screenshot shows the Kibana Discover interface. The search bar contains the query 'love'. The left sidebar shows the 'articles5000' dataset selected. The 'Selected fields' list includes '_id', '_index', '_type', 'published', and '_score'. The 'Available fields' list includes '_id', '_index', '_type', 'published', and '_score'. The main table displays 379 hits for the 'media-type' field, with values including 'Blog' and 'News'.

The screenshot shows the Kibana Discover interface with a filter applied: 'media-type: Blog'. The search bar still contains the query 'love'. The left sidebar shows the 'articles5000' dataset selected. The 'Selected fields' list includes '_id', '_index', '_type', 'published', and '_score'. The 'Available fields' list includes '_id', '_index', '_type', 'published', and '_score'. The main table displays 170 hits for the 'media-type' field, with all values being 'Blog'.

The query on holiday gives 81 hits and on further adding a filter content on iPhone gave 1 hit. The kibana screenshots are as under –

The screenshot displays the Kibana search interface. The top navigation bar shows the URL: `localhost:5601/app/kibana#/discover?_g=(filters:[])&_a=(columns:!(media-type),filters:[]).index:48558f10-7976-11ea-be49-6bd9f3d25094;interval:auto;query:(lan...`. The main interface is divided into several sections:

- Discover Panel:**
 - Filters:** A filter for `holiday` is applied.
 - Selected fields:** `media-type` is selected.
 - Available fields:** A list of fields including `_id`, `_index`, `_type`, `published`, `_score`, and `content`.
- Search Results:**
 - media-type:** A list of 81 hits, all of which are `News`.
 - content:** A single hit (1 hit) for the `content` field. The text of the hit is highlighted in yellow, indicating a match with the search query.

The bottom of the image shows a Windows taskbar with various application icons and a system clock indicating 12:42 PM on 4/8/2020.

J. EVALUATION AND QUERY RESULTS (RECALL AND PRECISION)

We evaluated our system using the parameters Recall and Precision.

Recall is defined by the following ratio –

Recall = Number of relevant documents returned / Total number of relevant documents

Precision is given as –

Precision = Number of relevant documents returned/ Total number of returned documents

The sample size of returned documents taken for evaluation in our assignment is K= 10.

The Test Collection consisted of ten queries relating to specific events together with their expected results which is tabulated as under –

EVALUATION RESULTS TABLE			
Index			
K = Number of documents taken as sample size for evaluation or calculation.K=10			
S.No	Query	Recall	Precision
1	Politics	8/12	7/10
2	Sports	3/4	3/10
3	Technology	5/10	5/10
4	Software Giants	2/4	2/10
5	Health	4/9	4/10
6	Love	3/7	3/10
7	Rain	2/6	2/10
8	Movie	1/8	1/10
9	Book	3/9	3/10
10	Holiday	5/7	5/10

K. CROWDSOURCING TASK AND ITS EXPERIENCE

There was a crowdsourcing task in the assignment 2 which consisted of 20 % marks of the total assignment and was independent in nature, that is, this crowdsourcing task could be done irrespective of implementing this Elastic Search system.

The task consisted of 2 phases. Each phase composed of small videos to be seen and the task was to hit spacebar if the video presented was a repeated one. Each video was of around 5 seconds length and the focus was on checking the recalling and memory of the participant.

We had to login using a link given in the mail using our University of Essex credentials and also in the settings of the google chrome browser Javascript was to be enabled in the site settings to view the videos of the task.

The videos were extremely funny and relaxing. We scored above 90 % in the first phase.

There was a gap of 24 – 72 hours between Phase 1 and 2. The password for those who completed the phase 1 was given as Registration number_DaY2. Again, the same task was to be performed. This phase was focussed on how much one retains after the gap of 1 or 2 days in recognising the videos are repeated or new.

This phase was also very nice with each one of us scoring above 80%. It shows that with time, the memory and recalling capacity shows a decrease.

L. ENGINEERING A COMPLETE SYSTEM

We successfully engineered a holistic elasticsearch system whose blueprint is displayed below –

1. Installing → JAVA environment
2. Installing → Elasticsearch using batch file on Command prompt(cmd). Then running elastic search on localhost 9200
3. Installing → Elasticsearch using batch file on Command prompt(cmd). Then running elastic search on localhost 5601
4. Installing → Jupyter and Anaconda for Python 3 environment required for the system
5. Signalmedia.jsonl Dataset → Loaded into elasticsearch localhost using Anaconda Jupyter Python Environment

6. Indexing → Implemented for 5000 articles from the dataset since the system was getting freeze on including complete 1 million documents of the dataset into the system. **[LIMITATION OF SYSTEM]**
7. Searching → Using Kibana Query Language (KQL) we executed various searches outlined in the test collection section of this report.
8. Evaluation → Using recall and precision parameters we evaluated the system individually for each query in the test collection and tabulated them properly.

M.SUGGESTIONS ON SCOPE AND FUTURE IMPROVEMENTS POSSIBLE –

Due to practical and runtime problems of too much system hanging and freezing, we limited our search and article size to 5000.

For more comprehensiveness, this can be increased further to 1 million documents and evaluation parameter K can be increased which was fixed in our study to 10 because huge manual calculations involved in recall and precision was making the evaluation very complex.

This can be solved using faster data processing Tensor Processing Units (TPUs).