IR Assignment: Cieran Almond_1604959, Drew Naylor_154692

Introduction

In this project we ill be using elasticsearch and kibana to query a collection of over one million documents in an index. We will be calculating the precision and recall of each query for a number of document types. This will then contribute to a mean accuracy precision score (mAP) . We will then alter the information retrieval pipeline and compare the mAP scores of our result.

Methodology

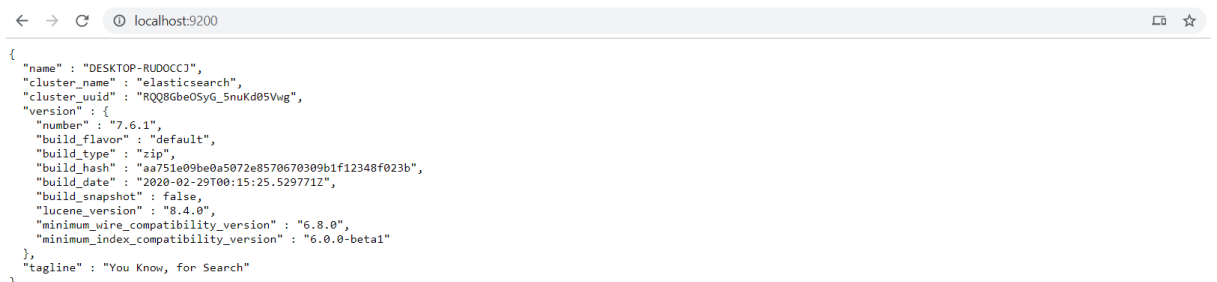We requested the dataset from here:

You can download the dataset directly using this link:
https://research.signal-ai.com/newsir16/signalmedia-1m.jsonl.gz

The articles of the dataset were originally collected by Moreover Technologies (one of Signal's content providers) from a variety of news sources for a period of 1 month (1-30 September 2015). It contains 1 million articles that are mainly English, but they also include non-English and multi-lingual articles. Sources of these articles include major ones, such as Reuters, in addition to local news sources and blogs.[1]

This data is provided in a JSON format, with each article containing different fields. We will load the JSON file into Kibana for this assignment.

1. Created a folder "ir_ass2" and extracted the dataset into that folder directory
2. Downloaded ElasticSearch for Windows from here:
   https://www.elastic.co/downloads/elasticsearch
3. Extracted the ElasticSearch folder into the "ir_ass2" directory
4. Opened command prompt. CD to the following path: \ir_ass2\elasticsearch-7.6.1\bin
5. In command line ran "elasticsearch.bat", let that run.
6. Open a web browser and navigate to http://localhost:9200/. If elasticsearch is working you'll see this screen:

7. Now we can see the data is loaded, we want to create an index to store the articles inside. This is done by running curl -X PUT 'http://localhost:9200/articles' [2]

```
C:\Users\My Computer>curl -X PUT "http://localhost:9200/articles"
{"acknowledged":true,"shards_acknowledged":true,"index":"articles"}
C:\Users\My Computer>
```

8. Now we need to install requests – this can be done by opening another command terminal and navigating to the python scripts directory.

```
D:\Python\Scripts>pip install requests
Collecting requests
  Downloading https://files.pythonhosted.org/packages/1a/70/1935c770cb3be6e3a8b78ced23d7e0f3b187f5cbfab4749523ed65d7c9b1
/requests-2.23.0-py2.py3-none-any.whl (58kB)
    |                              | 61kB 1.3MB/s
Collecting certifi>=2017.4.17 (from requests)
  Downloading https://files.pythonhosted.org/packages/b9/63/df50cac98ea0d5b006c55a399c3bf1db9da7b5a24de7890bc9cfd5dd9e99
/certifi-2019.11.28-py2.py3-none-any.whl (156kB)
    |                              | 163kB 1.3MB/s
Collecting chardet<4,>=3.0.2 (from requests)
  Downloading https://files.pythonhosted.org/packages/bc/a9/01ffebfb562e4274b6487b4bb1ddec7ca55ec7510b22e4c51f14098443b8
/chardet-3.0.4-py2.py3-none-any.whl (133kB)
    |                              | 143kB 1.3MB/s
Collecting idna<3,>=2.5 (from requests)
  Downloading https://files.pythonhosted.org/packages/89/e3/afebe61c546d18fb1709a61bee788254b40e736cff7271c7de5de2dc4128
/idna-2.9-py2.py3-none-any.whl (58kB)
    |                              | 61kB ...
Collecting urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1 (from requests)
  Downloading https://files.pythonhosted.org/packages/e8/74/6e4f91745020f967d09332bb2b8b9b10090957334692eb88ea4afe91b77f
/urllib3-1.25.8-py2.py3-none-any.whl (125kB)
    |                              | 133kB 595kB/s
Installing collected packages: certifi, chardet, idna, urllib3, requests
  WARNING: The script chardetect.exe is installed in 'd:\python\Scripts' which is not on PATH.
  Consider adding this directory to PATH or, if you prefer to suppress this warning, use --no-warn-script-location.
Successfully installed certifi-2019.11.28 chardet-3.0.4 idna-2.9 requests-2.23.0 urllib3-1.25.8
WARNING: You are using pip version 19.2.3, however version 20.0.2 is available.
You should consider upgrading via the 'python -m pip install --upgrade pip' command.
```

9. Once installed, navigate back to the directory the json file is located in. In my case: D:\ir_ass2, then run "index.py http://localhost:9200 ./sample-1M.jsonl"
10. Wait for indexing to complete, this may take some time.

```
Command Prompt - index.py  http
Added document: 16545
Added document: 16546
Added document: 16547
Added document: 16548
Added document: 16549
Added document: 16550
Added document: 16551
Added document: 16552
Added document: 16553
Added document: 16554
Added document: 16555
Added document: 16556
Added document: 16557
Added document: 16558
Added document: 16559
Added document: 16560
Added document: 16561
Added document: 16562
Added document: 16563
Added document: 16564
Added document: 16565
Added document: 16566
Added document: 16567
Added document: 16568
Added document: 16569
Added document: 16570
Added document: 16571
Added document: 16572
Added document: 16573
```

11. Indexing is completed after 8 hours:

Command Prompt

```
Added document: 999954
Added document: 999955
Added document: 999956
Added document: 999957
Added document: 999958
Added document: 999959
Added document: 999960
Added document: 999961
Added document: 999962
Added document: 999963
Added document: 999964
Added document: 999965
Added document: 999966
Added document: 999967
Added document: 999968
Added document: 999969
Added document: 999970
Added document: 999971
Added document: 999972
Added document: 999973
Added document: 999974
Added document: 999975
Added document: 999976
Added document: 999977
Added document: 999978
Added document: 999979
Added document: 999980
Added document: 999981
Added document: 999982
Added document: 999983
Added document: 999984
Added document: 999985
Added document: 999986
Added document: 999987
Added document: 999988
Added document: 999989
Added document: 999990
Added document: 999991
Added document: 999992
Added document: 999993
Added document: 999994
Added document: 999995
Added document: 999996
Added document: 999997
Added document: 999998
Added document: 999999
Added document: 1000000

D:\ir_ass2>
```
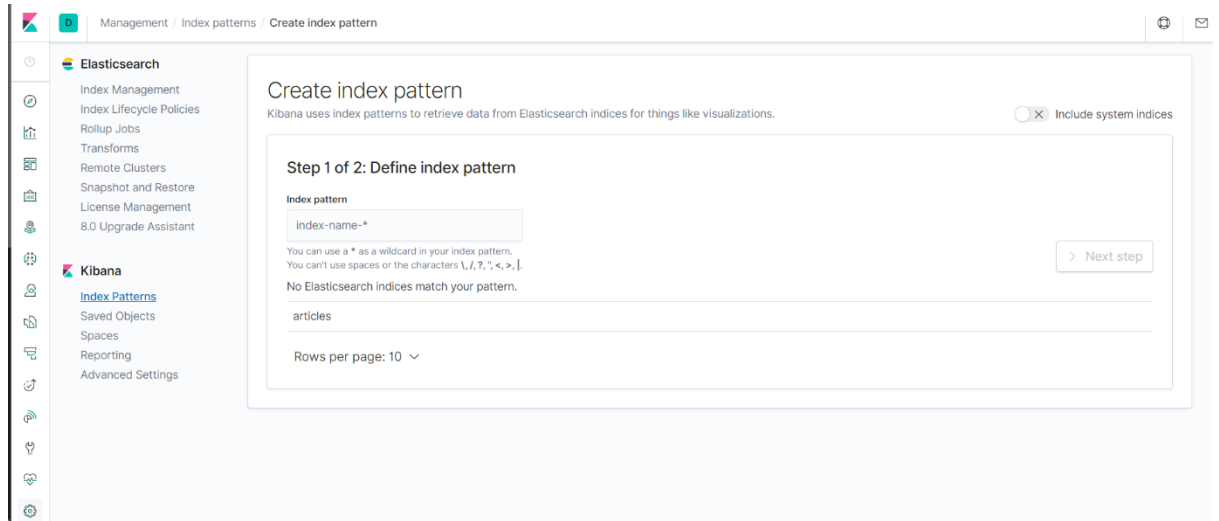
12. Once indexing is is completed, we can download Kirbana from here: https://www.elastic.co/downloads/kibana and extract it to "ir_ass2" directory.

13. Open another terminal window and cd to the directory kibana was installed, and then the bin folder. Now run "kibana.bat" in the command window.

```
D:\>cd D:\ir_ass2\kibana-7.6.1-windows-x86_64\bin

D:\ir_ass2\kibana-7.6.1-windows-x86_64\bin>kibana
  log   [15:53:27.487] [info][plugins-service] Plugin "case" is disabled.
  log   [15:54:12.923] [info][plugins-system] Setting up [37] plugins: [taskManager,siem,code,licensing,infra,encryptedS
avedObjects,usageCollection,metrics,canvas,timelion,features,security,apm_oss,translations,reporting,uiActions,data,navi
gation,share,status_page,newsfeed,kibana_legacy,management,dev_tools,home,spaces,cloud,apm,graph,inspector,expressions,v
isualizations,embeddable,advancedUiActions,dashboard_embeddable_container,eui_utils,bfetch]
  log   [15:54:12.924] [info][plugins][taskManager] Setting up plugin
  log   [15:54:12.940] [info][plugins][siem] Setting up plugin
  log   [15:54:12.942] [info][code][plugins] Setting up plugin
  log   [15:54:12.943] [info][licensing][plugins] Setting up plugin
  log   [15:54:12.945] [info][infra][plugins] Setting up plugin
  log   [15:54:12.946] [info][encryptedSavedObjects][plugins] Setting up plugin
  log   [15:54:12.948] [warning][config][encryptedSavedObjects][plugins] Generating a random key for xpack.encryptedSave
dObjects.encryptionKey. To be able to decrypt encrypted saved objects attributes after restart, please set xpack.encrypt
edSavedObjects.encryptionKey in kibana.yml
  log   [15:54:12.953] [info][plugins][usageCollection] Setting up plugin
  log   [15:54:12.955] [info][metrics][plugins] Setting up plugin
  log   [15:54:12.956] [info][canvas][plugins] Setting up plugin
  log   [15:54:12.968] [info][plugins][timelion] Setting up plugin
  log   [15:54:12.969] [info][features][plugins] Setting up plugin
  log   [15:54:12.970] [info][plugins][security] Setting up plugin
  log   [15:54:12.971] [warning][config][plugins][security] Generating a random key for xpack.security.encryptionKey. To
 prevent sessions from being invalidated on restart, please set xpack.security.encryptionKey in kibana.yml
```

14. Once this has completed running setup, the kibana server will be hosted on http://localhost:5601, open a browser and head to that URL.

15. Once in Kibana, we need to create an index pattern. For my index pattern, I used "articles" as the keyword.

16. Now we have to define the time filter field, this will allow you to filter data by time. I opted out of this option.

Create index pattern

Kibana uses index patterns to retrieve data from Elasticsearch indices for things like visualizations.

⊗ ☒ Include system indices

Step 2 of 2: Configure settings

You've defined **articles** as your index pattern. Now you can specify some settings before we create it.

Time Filter field name Refresh

[⌄]

The Time Filter will use this field to filter your data by time.
You can choose not to have a time field, but you will not be able to
narrow down your data by a time range.

❯ Show advanced options

❮ Back   Create index pattern

17. Once this is completed, you can head to the "Discover" tab, to check that all of your data has been properly indexed and loaded into Kibana. As you can see, I have over a million hits which implies the data has been correctly loaded.

Searching:

Firstly, some changes were made on the discover page before querying can begin.
The syntax used for searching is defaulted on Kibana Query Language, this needs to be turned off so Lucene can be used.

We can now add "_score" into the "Selected Fields" tab to be returned for our query. This will assign documents a score based of the query entered in the search field. Ensure that is sorting by ascending by hovering over the little blue arrow under score.

Add content and title to the selected fields tab. We should now return the Title, Content, and Score of each document we query now. The content will help us determine whether the document is relevant to our query.

The system should look something like this:

The system with a query having being entered into the search bar using Lucene.

Each of the documents are sorted in ascended order and returned highest to lowest.

Prerequisits before querying:

MAP is the mean average precision of a set of queries, where the number of queries in the set is divided by the average precision.

Precision and recall are metrics used to judge performance of a given classification model.

Recall can be calculated by identifying the "true positives" or "relevant" documents, divided by the "true positives" plus the "false negatives", which will be the non relevant documents in our case.[3]

Going forward, what we deem as relevant will be determined as matching our key search criteria in our query, for example if our query was "the American civil war", as long as the results related to this event, such that say a particular battle in this timeline was referenced, this would be deemed as relevant. The contents or "body" of the article is inspected and expected match what the query is about, either directly or through a related event.

Design Decisions

We decided to use BM25 as our default information retrieval method since it's a popular ranking function used by search engines to estimate the relevance of documents to a given query. Due to the nature of the assignment, we thought this would give us the best results, since we want to calculate precision, recall and MAP values based on the queries we give to our collection.

Test Collection

1st Event: Documents relating to **Work/Jobs**

MAP: The mean average precision for this document type is: (1.0+0.1+0.6)/3 = 0.566

Query: What's the best place to work

Results:

Title	Score	Relevant
Australia's best place to work	22.16	Y
Widen Named "Best Place to Work 2015"	20.01	Y
Atlassian is Australia's best place to work, again	19.09	Y
Facebook beats Google as the best place to work in the UK	18.50	Y
Expats Vote Singapore Best Place to Live and Work	18.25	Y
Singapore named best place for expats to live and work	17.48	Y
Best Places to Work Baltimore midsize first-place winner: IntelliGenesis	17.48	Y
Meltwater Australia named a	17.48	Y

Best Place to Work in 2015		
Best Places to Work, small, third-place: Applied Technology Group	17.48	Y
Commissions Inc Named as a Best Place to Work in Atlanta	16.77	Y
Precision:		10 / (10+0) = 1.0
Notes:		I would argue that all of the documents are relevant to the query, as they all include details of companies and places that claim to be among the best places to work.

Query: What's the worst place to work

Results:

Title	Score	Relevant
Australia's best place to work	16.32	N
Switzerland best place to grow old, India among the worst	16.08	Y
BRABUS Gives the Gigantic Mercedes G500 4x4² a Powerful Update	15.77	N
Whats the secret to better networking	15.70	N
A 'great place to be and a great place to work'	15.27	N
Club Library Contents List Now Available On-Line	15.10	N
Senior staff have the worst work-life balance	15.09	N
New Music: Dane Jalee – Whats Love	14.88	N
Widen Named "Best Place to Work 2015"	14.74	N
7th Veil (Kool Keith and H-Bomb) - Whats Up Now EP [2005]	14.74	N
Precision:		1 / (1+9) = 0.1
Notes:		This query seemed to struggle identifying any documents that detailed worst places to work, and again focused on the term "work" to return a list similar to the first query.

Query: How to get a job

Results:

Title	Score	Relevant
Interview Insider: How to Get a Job at Dropbox	19.64	Y
How to Get a Job During and After a Federal Clerkship	19.02	Y
How To Ace A Job Interview	17.01	Y
How to get a Deathgasm	16.33	N
How to Make Teaching a Dream Job	16.19	N
Opinion: How to nail a job interview	16.19	Y
How to Use Facebook to Land a		Y

Job	16.00		
How to get through a divorce	15.50	N	
How to Get a Cash Advance	15.50	N	
How did cop stripped of badge get job? | WIVB.com	15.47	N	
Precision:		6 / (6+4) = 0.6	
Notes:		This query managed to produce articles that used techniques and skills that would assist in landing a job.	

2nd Event: Documents relating to **Foods**

MAP: The mean average precision for this document type is: (0.3+0.6+0.7)/3 = 0.533

Query: Best food recepies

Results:

Title	Score	Relevant
Grandparents Day celebrations in Elmira	20.71	N
30 Ways in 30 Days SHARE A MEMORY - Recipies	13.17	Y
A Sweet Surprise Awaits You	12.73	N
Boston's Best Middle Eastern Food	12.68	N
Best Singapore Food | Singapore Food Reviews: Bottura Restaurant @ Suntec City	12.08	Y
The Buzz: Finding best food trucks	12.04	N
7 best raw food diet retreats	12.04	N
Best Thai Food In New York City	11.46	N
Indian Recipes, Cooking Videos, Healthy Recipes, Vegetarian Recipes, North and South Indian Recipies	11.20	Y
Is rice cereal the best food for baby?	10.93	N
Precision:	3 / (3+7) = 0.3	
Notes:	This query managed to produce articles that had the keyword food inside, though these weren't specifically finding recepies which was the main focus of the query.	

Query: italian restaurants

Results:

Title	Score	Relevant
Restaurants - Best Italian restaurants in Melbourne	20.71	Y
25 best Italian restaurants	19.54	Y
The Italian Job: KL's 10 best Italian restaurants	18.64	N
The 40 best restaurants in Swansea: The best places to eat in the city	15.65	N
Luciano's named one of Nevada's top 10 Italian restaurants	15.23	Y
NEW restaurants at Carnival City!	15.10	N
Marco's Pizza Announces Aggressive Growth Plan for Colorado, Wyoming	14.50	Y
Hidden Gem Spots Around Downtown	14.47	N
Meal Ticket: Dunne and Crescenzi, Dublin 2	14.45	Y
Neiman's Highlights Precious Jewelry, Italian Style	14.18	Y
Precision:	6 / (6+4) = 0.6	
Notes:		

Query: online grocery shopping

Results:

Title	Score	Relevant
Walmart Expands Online Grocery Shopping	25.48	Y
Woodman's launches online grocery shopping, delivery	25.48	Y
What are the Tips for Secure Online Grocery Shopping?	24.19	Y
Is online grocery shopping in Singapore cheaper than going to a supermarket?	23.60	Y
Wal-Mart grocery-pickup test grows	22.64	Y
SPECIAL REPORT: A New Era in Grocery Shopping	22.60	N
Walmart Doubles Down On Online Grocery Shopping With Curbside Pickup	22.49	Y
Packaged Facts: 30% of U.S. Consumers Order Online for Same Day Delivery	22.43	Y
Walmart expands e-grocery in eight cities	22.19	Y

Innovations in FMCG Digital Marketing: Whisk.com & Schwartz To Lead Session at IGD Online & Digital Summit 2015	`21.21`		N
Precision:		7 / (7+3) = 0.7	
Notes:			

3rd Event: Documents relating to **Sports**

MAP: The mean average precision for this document type is: (0.0+1.0+0.8)/3 = 0.56

Query: most popular sports in the world

Results:

Title	Score	Relevant
The 15 most popular YouTubers in the world	22.33	N
Meet the 30 most popular Vine stars in the world	20.52	N
The 10 most valuable sports teams in the world	20.33	N
India is one of the most popular countries in the world	19.71	N
These Are Canada's Most Popular Songs Around The World	18.86	N
Cowboys, Patriots NFL's most valuable sports franchises in the world	18.80	N
Cowboys most valuable sports team in world - Forbes	18.54	N
The Month in Review - The Most Popular Posts	17.68	N
Manchester United plc Announces Fourth Quarter and Fiscal 2015 Earnings Report Date	17.58	N
The Most Popular Biblical Names	17.50	N

Precision:		0 / (0+10) = 0.0
Notes:		This query didn't manage to produce any relevant search results due to it focusing on the term "in the world" rather than the keyword "sport" which is what the user is interested in.

Query: sports news

Results:

Title	Score	Relevant
Sports News | Latest Sports Headlines, Scores, Stats Analysis & News	14.96	Y
Sports Radio News	14.89	Y
German Summaries-Sports news	14.05	Y
Scottish Standings-Sports news	14.05	Y
Sports News Roundup September 30	13.30	Y
Sports news roundup — Sept. 11	13.30	Y
Sports news roundup — compiled Sept. 16	12.62	Y
CTV News sports segment: Sept. 2	12.62	Y
Western Morning News: deputy sports editor	12.6	Y
Sports news in brief, Sept. 25, 2015	12.01	Y
Precision:		10 / (10+0) = 1.0
Notes:		This query didn't manage to produce all relevant search results, as each document contained information detailing sports news events in both a certain timeframe and also a general overview.

Query: football world cup

Results:

Title	Score	Relevant
Matao ties with Oman in World Cup Qualifier 11 hours ago	16.03	Y
Russia Marks 2018 World Cup Milestone	15.74	Y
KJ slams FAM over shameful performance	15.55	N
Halfpenny ruled out of World Cup	15.52	Y
FOOTBALL: Russia starts 1,000-day countdown to World Cup	15.50	Y
Football: Cahill lifts Australia in World Cup qualifier	15.42	Y
Rugby and football enter cosy World Cup relationship	15.42	Y
Yokohama to host 2019 Rugby World Cup final	15.41	Y
PELE AUTOGRAPHED BRAZIL JERSEY Signed by Football Soccer Legend Memorabilia	15.41	N
1,000 days to Russia 2018: footballers and fans celebrate in Red Square	15.41	Y
Precision:		8 / (8+2) = 0.8

Notes:		

4th Event: Documents relating to **History**

MAP: The mean average precision for this document type is: (0.8+1.0+0.5)/3 = 0.76

Query: the american civil war

Results:

Title	Score	Relevant
Battles of the American Civil War	21.50	Y
Democracy and the American Civil War	21.50	Y
CULTURE Irish involvement in American Civil War explored	19.51	Y
California's statehood debate foreshadowed the American Civil War	19.51	Y
The "civil war" over Captain America: Civil War	18.57	Y
Civil War #4	17.93	N
New and Noteworthy: Drilling For Battle and Drilling for Reenactments	17.31	Y
Woodward Library host tribute to Lincoln	17.15	Y
Ready, aim, fire: American Civil War re-enactment comes to Cayuga	17.14	Y
Amputation- Reality Of War- Private Parmenter- Company G	16.99	Y
Precision:		8 / (8+2) = 0.8
Notes:		

Query: the Iraq war

Results:

Title	Score	Relevant
Iraq: The Other, More Important, War	17.73	Y
Labour leadership favourite Jeremy Corbyn 'attended Cairo conference which backed Iraqi attacks on British troops'	15.46	Y
The Latest: Trump links Iraq war to Jeb Bush	15.38	Y
The Saudi war in Iraq and	15.38	Yis

lessons for Pakistan		
Those in public life who made the catastrophic decision to support the Iraq war have prospered.	15.38	
Former US vice president defends invasion of Iraq	15.36	Y
Donald Trump: I'm the Only Candidate Who Opposed the Iraq War. Rand Paul: Uhhh	15.18	Y
US invasion of Iraq was 'genocidal war'	15.04	Y
New Film Investigates Bush, Cheney & Iraq War Deceit	14.99	Y
Sanders' Iran deal defense cites Iraq war	14.84	Y
Precision:		10 / (10+0) = 1.0
Notes:		

Query: the cold war era

Results:

Title	Score	Relevant
Cold War-era weaponry in pictures	24.12	Y
CIA declassifies trove of Cold War-era intelligence memosAUSTIN, Texas — As the U.S. and Russia reached the brink of nuclear war in 1962, President John F. Kennedy received top-secret intelligence from the CIA that a new warhead	23.07	Y
CIA declassifies Cold War-era intel memos	22.95	Y
CIA declassifies trove of Cold War-era intelligence memos	21.87	Y
The Cold War Horse	20.93	N
Lindsay Smith	18.73	N
Library Lecture Sept. 24: German Rocketeers in the Heart of Dixie	18.36	N
Cold War to Icy War	18.16	Y
Is NASA running out of plutonium?	17.40	N
Nordic Games acquires remaining property for Codename: Panzers	16.96	N
Precision:		5 / (5+5) = 0.5
Notes:		

5th Event: Documents relating to **Motorsport/Cars**

MAP: The mean average precision for this document type is: (0.9+0.9+0.6)/3 = 0.8

Query: mercedes-benz

Results:

Title	Score	Relevant
Brabus Maybach Rocket with 900HP! (Full Tour)	24.40	N
Mercedes-Benz of Kansas City offers special Certified Pre-Owned Mercedes-Benz deals	24.33	Y
Loeber Motors welcomes 2016 Mercedes-Benz C300 with lease specials	24.28	Y
The car transformed only with a single touch	24.21	Y
Loeber Motors welcomes 2016 Mercedes-Benz C300 with lease specials	24.17	Y
Mansory's Mercedes-Benz S63 AMG Coupe – 1,000 hp!	24.12	Y
Mansory releases Mercedes-AMG S63 Coupe Black Edition with 1,000 PS	24.01	Y
Certified Pre-Owned Mercedes-Benz models shine at Loeber Motors	23.89	Y
Mercedes to invest Rs 1,000 crore on luxury car production in Maharashtra	23.85	Y
Mercedes Benz of Westwood to Take Delivery of New Metris Mid-Sized Van in October	23.82	Y
Precision:		9 / (9+1) = 0.9
Notes:		

Query: formula 1

Results:

Title	Score	Relevant
In pictures - the best of Singapore	19.38	Y
Italy preview quotes - Williams, Toro Rosso, McLaren & more	19.38	Y
Hamilton: Matching Senna's 41 wins would be 'emotional'	19.38	Y
Exclusive Sergio Perez Q&A: Marina Bay causing Force India headaches	19.38	Y
McLaren: Button and Alonso the preferred line-up for 2016	19.38	Y
Toto Wolff Q&A: Tyre investigation exposed rules confusion, not wrongdoing	19.38	Y
Do you remember... F1's first ever night race	19.38	Y
IndyCar – Three Up, Three Down: Sonoma	16.36	N
Lauda: Red Bull could have had Mercedes engines	16.36	Y
Google Alert - formula 1	15.77	Y
Precision:	9 / (9+1) = 0.9	
Notes:		

Query: driver Kimi Räikkönen

Results:

Title	Score	Relevant
Stock Update: United Parcel Service Inc (NYSE:UPS) – UPS Invites Scuderia Ferrari Driver Kimi Räikkönen to Deliver UPS City Center Race Experience in Singapore	37.85	Y
UPS Invites Scuderia Ferrari Driver Kimi Räikkönen to Deliver UPS City Center Race Experience in Singapore	32.62	Y
F1, Hamilton wins in Suzuka ...	32.62	N

	Score	Relevant
Championship standings after Italy	29.41	Y
Vettel takes Singapore pole as Mercedes slump	29.20	Y
Hamilton on pole, Ferrari 2nd and 3rd	27.78	N
Monza Qualifying: Hamilton takes pole with Ferrari in the hunt	26.26	N
Mercedes faces tyre breach investigation	24.27	N
Monza FP3: Hamilton tops but stalked by Vettel	23.96	N
Ericsson docked three places for impeding	23.60	N
Precision:	4 / (4+6) = 0.6	
Notes:		

6th Event: Documents relating to **Science**

MAP: The mean average precision for this document type is: (0.6+0.6+0.2)/3 = 0.466

Query: elements in the periodic table

Results:

Title	Score	Relevant
Periodic Table of the Elements 2015 (Education)	31.87	Y
MOVED: Periodic Table of the Elements Display Case	30.40	Y
The Periodic Table of Confetti	24.54	N
C 29, September 11, 2015	23.78	Y
For the Record: Why did Studham leave?	23.49	N
Nature Watch: PBS series brings the elements to life	23.16	N
Table of the Elements	23.13	Y
Where In Our Solar System Are We Most Likely To Find Life	22.44	N
My teenage sons are the smartest people I know. And sometimes...	22.34	N
Monday, September 8 - A Day	21.07	N
Precision:	4 / (4+6) = 0.6	

Notes:		

Query: the atomic bomb

Results:

Title	Score	Relevant
Enc: LDS Youth acabou de enviar um vídeo	21.06	N
Great News! No Need To Worry About An Iranian Nuke	20.18	Y
The Handheld Analog Computer That Made the Atomic Bomb	20.18	Y
North Korea Restarts Atomic Bomb Plants	20.18	Y
The Truth About Atom Bombs and Defense Against Them	20.18	Y
Replica of most powerful nuclear bomb ever goes on display in Moscow	20.18	Y
Nuclear spy Mordechai Vanunu arrested for giving TV interview	20.18	N
Replica of most powerful nuclear bomb ever goes on display in Moscow	19.32	Y
China, U.S., Iran reach broad consensus over nuclear reactor in Tehran	19.27	N
70 years after Hiroshima: Faculty of Arts to host conference, Hiroshima atom bomb survivor on Sept. 25 to 26	19.19	Y
Precision:	6 / (6+4) = 0.6	
Notes:		

Query: planets in our solar system

Results:

Title	Score	Relevant
Our Solar System: You Are Now Leaving Earth	30.69	Y
Listen To The Solar System Sing With This Fantastic Simulation	29.17	N
Planets Transit The Desert In 7-Mile Scale Model Of The Solar System	28.76	N

[tt] NS 3036: More than 100 billion billion Earth-like planets might exist	28.76	N
This is a Scale Model of the Solar System Like You've Never Seen Before	28.52	N
What are asteroids?	27.88	N
Filmmakers Create a Seven-Mile-Wide Scale Model of the Solar System With an Earth the Size of a Marble	26.88	N
Friends create scale model of solar system in Nevada desert	26.73	N
Next Professor Brian Cox? Brainy toddler lists all the planets in our solar system	26.12	Y
What will be hottest space research in next ten years?	26.08	N
Precision:		2 / (2+8) = 0.2
Notes:		

6th Event: Documents relating to **Music**

MAP: The mean average precision for this document type is: (0.8+0.3+0.7)/3 = 0.6

Query: drake albums

Results:

Title	Score	Relevant
Type A to Z: Listen Download What a Time to Be Alive Album by Future and Drake	22.99	Y
Drake and Future's What A Time To Be Alive Debuts At No. 1	22.24	Y
The World Rates Drake and Future's Mixtape 'What A Time To Be Alive'	19.20	Y
ALBUM SALES (week 39): Drake, Future, Mac Miller & more!	19.06	Y
Drake and Future's Collaborative Album Debuts at No. 1 on Billboard 200 Chart	18.77	Y
Drake & Future reveal overhyped 'What A Time To Be Alive' mixtape	18.11	Y
Fetty Wap, T.I And More To Perform At The 2015 BET Hip-Hop Awards	18.03	N
Buffy Sainte-marie Wins Polaris Music Prize	18.01	N
Drake and Future's 'What a Time to Be Alive' Debuts at No. 1	17.29	Y

Title	Score	Relevant
Drake & Future: The Evolution of Rap's Latest (Potential) Power Duo	17.04	Y
Precision:		8 / (8+2) = 0.8
Notes:		

Query: now thats what i call music songs

Results:

Title	Score	Relevant
NOW That's What I Call Music! Presents Today's Biggest Hits on NOW That's What I Call Music! Vol. 56	30.99	Y
Now That's What I Call Music- 3	30.61	Y
Now That's What I Call Music- 5	27.93	Y
Ozzy Osbourne Song Meanings - Mr. Tinkertrain	27.93	N
WIB Interviews Rah Digga (Exclusive)	24.95	N
Maybe thats why I do what I do...	24.11	N
Reggae and the dole: an interview with UB40	23.52	N
Collin Raye	22	N
Now That's What I Call Marketing Vol. 1: the 14 most infuriatingly catchy ad songs	21.88	N
What We Will Remember	21.75	N
Precision:		3 / (2+7) = 0.3
Notes:		

Query: Tyler, the creator

Results:

Title	Score	Relevant
Tyler the Creator and Danny Brown GTA Online Livestream Today Tues Sept 22	19.05	Y
Tyler The Creator Speaks Out On UK Ban	19.00	Y
Tyler, The Creator On Recent Bans: "I'm Getting Treated Like	18.09	Y

A Terrorist"		
'Teen Wolf' Creator: Tyler Posey Vacate Lead Role Soon?	18.03	N
Odd Future to Reform at Tyler, the Creator's Camp Flog Gnaw	17.86	Y
'Teen Wolf' Creator: Tyler Posey Vacate Lead Role Soon?	17.58	N
ASAP Rocky and ScHoolboy Q perform "Electric Body" on Kimmel — watch	17.34	Y
Freestyles & Laughs :: Watch Tyler, the Creator x ASAP Rocky's Tour Trailer	17.34	Y
Chris Brown Faces Australia Ban, 'One Hell of a Nite Tour' Dates In Peril	17.06	N
Tyler, The Creator On U.K. Ban: "I'm Being Treated Like a Terrorist"	16.65	Y
Precision:		7 / (7+3) = 0.7
Notes:		

7th Event: Documents relating to **Film**

MAP: The mean average precision for this document type is: (1.0+0.9+1.0)/3 = 0.96

Query: actor - Leonardo DiCaprio

Results:

Title	Score	Relevant
Leonardo DiCaprio Joins $2.6 Trillion Fossil Fuel Divestment Movement	35.42	Y
Fox News Flubs 'Mona Lisa' Artist: Leonardo ... DiCaprio	34.10	Y
News Anchor Says Leonardo DiCaprio Painted Mona Lisa, Is So Close to Being Right	33.64	Y
Leonardo DiCaprio joins fossil fuel divestment movement as it reaches $2.6 trillion mark	33.62	Y
Leonardo Dicaprio to be awarded with Screen Actors Guild's Actors Inspiration Award	31.71	Y
Leonardo DiCaprio Is Unrecognisable In 'The Revenant', A Film Which Must Surely Bring Him A Gong	31.37	Y

Title	Score	Relevant
Watch: Leonardo DiCaprio out for revenge in new 'The Revenant' trailer	31.22	Y
Leonardo DiCaprio urges action 'to save the planet and help the underprivileged' at Global Citizen Festival in NYC	31.22	Y
Leonardo DiCaprio's Mona Lisa? Fox News Presenter Goofs Up During Broadcast, Sends Twitter Into A Frenzy [Video]	30.83	Y
Leonardo DiCaprio (Finally) Ditches the Bushy Beard During Bike Ride--Check it Out!	30.68	Y
Precision:		10 / (10+0) = 1.0
Notes:		

Query: the avengers

Results:

Title	Score	Relevant
Marvel Releases Preview Images of All-New, All-Different Avengers for "Avengers #0"	14.84	Y
AVENGERS #0 Assembles Earth's Mightiest This October!	14.67	Y
AVENGERS #0 ASSEMBLES EARTH'S MIGHTIEST HEROES	14.67	Y
Avengers Grimm (2015) Gratis	14.67	Y
DAME DIANA RIGG GUESTS AT BFI SCREENING OF CUSHING AVENGERS CYBERNAUT SCREENING	14.61	N
Thor & The Vision Kiss In 'Avengers: Age Of Ultron' Blooper Reel	14.41	Y
This Avengers: Age of Ultron Poster Perfectly Represents That Manic Marvel Movie	14.28	Y
Time Runs Out for Sokovia in Avengers Clip	14.20	Y
WATCH: "Avengers: Age of Ultron" Honest Trailer Expresses Concerns for Joss Whedon	14.18	Y
The UnPopular Opinion: Avengers: Age of Ultron	14.14	Y
Precision:		9 / (9+1) = 0.9
Notes:		

Query: Black Mirror on Netflix

Results:

Title	Score	Relevant
Netflix renews 'Black Mirror'	27.71	Y
Sci-fi series 'Black Mirror' finds a new home on Netflix	26.91	Y
'Black Mirror': 12 new episodes on Netflix	26.47	Y
Netflix Confirms 'Black Mirror' For 12 New Episodes (But Not Because of David Cameron or Pigs)	26.45	Y
'Black Mirror': Netflix orders 12 new episodes TV just now	26.44	Y
We're Getting Even More Black Mirror on Netflix	25.99	Y
'Black Mirror' Anthology Series Gets New Season On Netflix	25.99	Y
Netflix Books 12 New 'Black Mirror' Episodes 3 days ago	25.84	Y
Netflix Orders Original 'Black Mirror' Episodes Even as It Plans New Self-Sufficiency	25.78	Y
Netflix 'to fund new Black Mirror'	25.49	Y
Precision:		10 / (10+0) = 1.0
Notes:		

8th Event: Documents relating to **Education**

MAP: The mean average precision for this document type is: (0.2+0.8+0.6)/3 = 0.533

Query: top universities in the UK

Results:

Title	Score	Relevant
UK ranked second in the world in universities league table	19.44	Y
QS World University Rankings 2015/16	18.82	Y
British universities slip down in global rankings	18.47	N
Two UAE Universities Ranked in The Top 800 Best Universities Worldwide	18.21	N
British universities slip downing global rankings	17.94	N
Australian unis climb in global rankings	17.77	N
South Africa home to top African research institutions		N

	17.76	
Australian universities climb in Times Higher Education world rankings	17.71	N
Top UK Universities | Part Three – East Midlands & East of England	17.59	N
Australian universities climb in Times Higher Education world rankings	17.45	N
Precision:		2 / (2+8) = 0.2
Notes:		

Query: Cambridge university

Results:

Title	Score	Relevant
Exploring the university town of Cambridge	16.40	Y
Cambridge News published Cambridge named UK's best university in Sunday Times Good...	15.88	Y
Infinitus and Cambridge Jointly Establish International Research Centre	15.86	Y
Good new! Cambridge University mulling to reintroduce universal entrance exam	15.76	Y
Three University of Auckland students awarded top scholarships	15.66	N
"The Cambridge Companion to Christian Political Theology" (Hovey & Phillips, eds.)	15.38	Y
CambridgeNews published Cambridge University and Cambridge to go head to head in Town...	15.04	Y
British universities slip downing global rankings	14.90	Y
'We cannot stand alone' Cambridge University vice-chancellor...	14.86	Y
Monsignor Doyle graduate nets scholarship	14.68	N
Precision:		8 / (8+2) = 0.8
Notes:		

Query: technology used in learning

Results:

Title	Score	Relevant
Technology hinders learning	16.62	Y
How E-learning Is Used in Various Industries – Some Examples - e-Learning Feeds	15.68	Y
Discovering what's new in Learning & Development technology	15.42	Y
Fujitsu Develops Technology that Uses Machine Learning to Quickly Generate Predictive Models from Massive Datasets	15.23	Y
Social media and apps can benefit medical learning, academics say	15.00	N
Classrooms Today Are Digital, Connected	14.37	Y
Technology doesn't make school pupils smarter: study	14.13	Y
Transform your PowerPoint into a PowerCast using Camtasia!	14.12	N
Dato Announces New Machine Learning Toolkits and Coursera Training	14.12	N
Our Learning Story	14.09	N
Precision:	6 / (6+4) = 0.6	
Notes:		

9th Event: Documents relating to **Geography**

MAP: The mean average precision for this document type is: (0.7+1.0+0.1)/3 = 0.6

Query: climbing Mt Everest

Results:

Title	Score	Relevant
Everest (2015) New In Theaters (HD TS) Thriller	Jake Gyllenhaal	29.31	Y
Everest	27.80	Y
Mark Inglis says Mt Everest limits should be based on competence	27.29	Y
'Doesn't matter how many limbs	26.20	Y

you've got'		
Everest Bans Climbers: Nepalese Government To Issue Climbing Restrictions On Mount Everest	26.11	Y
Nepal to Place Climbing Restrictions on Everest?	24.79	Y
'Everest' is an edge-of-your-seat thriller that's true to life	24.47	N
Everest & The Vast Unknown	24.22	Y
'Everest' expert Nick Heil says...	24.05	N
West Coast Outdoor Weather	23.87	N
Precision:	7/ (7+3) = 0.7	
Notes:		

Query: is climate change affecting polar bears

Results:

Title	Score	Relevant
Starving Polar Bear Photo: Don't Blame (Just) Climate Change	40.78	Y
International Agreement Reached On Polar Bear Protection	38.7	Y
Polar Bears May Be Starving in the Barents Sea Area, But Thanks to Russia We Don't Know If It's Climate Change	33.99	Y
Startling image of emaciated polar bear: Sign of climate change?	33.99	Y
Hungry polar bears trap Arctic researchers	32.65	Y
Photo of Sickly Polar Bear Signals Tragedy Ahead in Arctic	32.15	Y
Polar Bears Might Survive Ice Melting By Hunting New Prey	31.61	Y
International agreement reached on polar bear protection	31.48	Y
discoverynews: Photo of Sickly Polar Bear Signals Tragedy Ahead...	29.87	Y
Biologists and Alaskan Locals Free Polar Bear Trapped By Fishnet	29.66	N
Precision:	10 / (10+0) = 1.0	
Notes:		

Query: what are the most obese countries

Results:

Title	Score	Relevant
Which U.S. states are most obese?	22.45	N
Here are the countries where people eat the most meat	18.76	N
Diabetes Should Be a Factor in Weight-Loss Surgery Decision: Study	18.34	N
Mapping America's most obese states	18.34	N
WHO: Europeans 'alarmingly' obese	18.34	Y
Bloomberg: These Are the 10 Most Competitive Countries in the World	18.32	N
Diabetes Should Be a Factor in Weight-Loss Surgery Decision: Study	18.13	N
### best cure for diabetes Reviews	18.03	N
CDC: Alabama 5th most obese state	17.77	N
These Are the 10 Most Competitive Countries in the World - Bloomberg Business	17.47	N
Precision:	1 / (1+9) = 0.1	
Notes:		

10th Event: Documents relating to **Art**

MAP: The mean average precision for this document type is: (0.7+0.3+0.2)/3 = 0.4

Query: vincent van gogh paintings

Results:

Title	Score	Relevant
SVA Photo Exhibition in Southern France Pays Tribute to Van Gogh	42.15	Y
Van Gogh Museum Edition Collection on tour coming to Southgate Centre	41.45	Y
Takii creates sunflower display at Van Gogh Museum	41.18	Y
Van Gogh painting could fetch	39.04	Y

up to $70 million at auction		
Vincent van Gogh lays claim to the sunflower	38.45	N
Van Gogh Museum Edition Collection on tour coming to Southgate Centre	38.36	Y
Dutchman Van Gogh finally meets Norway's Munch	38.12	N
Van Gogh painting could bring up to $70m at auction	38.10	Y
Van Gogh estimated to bring $50M to $70M at NYC fall auction	38.05	Y
Jaeger-LeCoultre artistic tribute to Van Gogh	37.46	N
Precision:		7 / (7+3) = 0.7
Notes:		

Query: who won the turner prize

Results:

Title	Score	Relevant
Who won the Startup Prize	26.57	N
Profile: The Turner Prize	23.03	Y
Meet the scientist who won 'parody Nobel Prize' for unboiling egg	20.11	N
Liverpool workshop moved as Turner Prize exhibition heads north	19.79	Y
TV Choice Awards 2015: Ant and Dec win big, and Sheridan Smith's Cilla accolade	19.79	N
Conspiracy theorists and fur coats adorn Turner prize's Glasgow show	19.79	Y
TV Choice Awards 2015: Full list of winners and a moving tribute to Cilla Black by the actress who played her	19.67	N
Morley and Jones win 2015 Burghill Valley Golf Club Pairs Open	19.21	N
Sheridan Smith pays tribute to Cilla Black at TV Choice Awards	18.86	N
Joan enjoys a successful week at ladies' open tournament	18.43	N
Precision:		3 / (3+7) = 0.3
Notes:		

Query: words most expensive painting

Results:

Title	Score	Relevant
World's Most Expensive Painting Comes to the Phillips Collection	21.70	Y
Most Expensive Girl	18.61	N
Most expensive cities	18.61	N
World's Most Expensive Cities	17.56	N
America's Most Expensive Rental Markets	16.62	N
World's Most Expensive Vacuum Cleaner	16.62	N
This artist turned famous art masterpieces into 8-bit watercolors — see if you can identify them	16.24	N
Van Gogh painting 'Landscape Under a Stormy Sky' could fetch $70 million at Sotheby's auction	15.96	Y
Small Yet Comfortable Green House In Melbourne, Australia	15.95	N
Check out Merseyside's most expensive homes	15.77	N
Precision:		2 / (2+8) = 0.2
Notes:		

A Short Description and Motivation of Our Evaluation Methodology

We will be comparing the mean average precision (mAP) for each document type. We will be comparing the retrieval models of TD*IDF, as default on Elasticsearch, to boolean. A higher mAP indicates a more accurate retrieval methodology.

Evaluation

To change to Boolean retrieval model, we closed our articles repository, using command
- D:\ir_ass2>curl -XPOST "http://localhost:9200/articles/_close"
- {"acknowledged":true,"shards_acknowledged":true,"indices":{"articles": {"closed":true}}}

From here we used command curl –XPUT "http://localhost:9200/articles/_settings" to insert this line of JSON:

- {"index": {"similarity": {"default": {"type": "boolean"}}}}

We then proceeded to test this had worked, by opening articles again through this command:

- D:\ir_ass2>curl -XPOST "http://localhost:9200/articles/_open"
- {"acknowledged":true,"shards_acknowledged":true}

With this configured, we re-queried each event described and documented above, the table below shows the results of Boolean mAP compared to our BM-25 mAP results.

Events	BM-25 (TF*IDF) mAP	Boolean mAP
1	0.566	0.5
2	0.533	0.48
3	0.56	0.62
4	0.76	0.645
5	0.466	0.301
6	0.6	0.285
7	0.96	0.68
8	0.533	0.542
9	0.6	0.4
10	0.4	0.367

Given the results above it is clear that BM-25 out performs the boolean retrieval model with regards to mAP. This indicated that the BM-25 (tf*IDF variant) is a more precise retrieval model in this context. This is probably due to the size of the collection. TF * IDF performs better when ranking document relevancy in large collections like this. Boolean can out

perform TF *IDF in other scenario such as small collections and even when ranking the quality of writing in a document as the document is represented in a vector format.

Improvement to the Current Solution:

To improve the current system we could make changes to both the pipeline and retrieval models. Some example changes we could implement could be a stemming algorithm, for example Porter Stemming, this allows the user to match more documents to their search term. Another improvement could be to implement the BM-25 retrieval model, to have a more refined version of TF * IDF for a better relevancy score. This could result in more relevant documents being returned to our user.

Crowdsourcing Task:

Crowdsourcing is a type of participative online activity in which an individual, an institution, a non-profit organization, or company proposes to a group of individuals of varying knowledge, heterogeneity, and number, via a flexible open call, the voluntary undertaking of a task. For this experiment we worked separately. We were given the task of watching a series seven second video clips and determining if they had repeated in the sequence. If the video has repeated we had to press the space bar to indicated we had recognised the repetition. The experiment was then repeated twenty-four hours later to see if our responses had improved. Interestingly Cieran's responses had improved over the twenty-four hours with regards to recognising repeated clips, however his false positive rate increased. Drew's on the other hand, had a decreased rate of repetition recognition and an increase rate of false positives. The site had a couple of latency issues when displaying videos, but this might be due to the original video quality. Other than this slight issue, we both managed to complete the task very easily.

Sources:

https://research.signal-ai.com/newsir16/signal-dataset.html [1]
https://github.com/signal-ai/Signal-1M-Tools [2]

https://towardsdatascience.com/breaking-down-mean-average-precision-map-ae462f623a52 [3]