

# MSCI 541 HW2 Report

Ramandeep Farmaha 20516974

## Problem 2

In order to demonstrate that Boolean AND retrieval is functioning properly, the following test document was created. The test document comprises of 4 individual articles in a format similar to the LATimes.gz file:

```
<DOC>
<DOCNO> RF020218-0001 </DOCNO>
<DOCID> 1 </DOCID>
<DATE>
<P>
February 2, 2018
</P>
</DATE>
<HEADLINE>
<P>
NEW TESLA VEHICLE SELLS 500 THOUSAND UNITS FIRST WEEK;
</P>
</HEADLINE>
<BYLINE>
<P>
By Ramandeep Farmaha
</P>
</BYLINE>
<TEXT>
<P>
Elon Musk car vehicle electricity Tesla green. Sales automaker Ford.
</P>
</TEXT>
<GRAPHIC>
<P>
New Tesla electric vehicle. All new touchscreen display.
</P>
</GRAPHIC>
<TYPE>
<P>
Article
</P>
</TYPE>
</DOC>
<DOC>
<DOCNO> RF020218-0002 </DOCNO>
<DOCID> 2 </DOCID>
<DATE>
<P>
February 2, 2018
</P>
</DATE>
<HEADLINE>
```

<P>  
NEW IPAD HAS EXPLODING ISSUES;  
</P>  
</HEADLINE>  
<BYLINE>  
<P>  
By Ramandeep Farmaha  
</P>  
</BYLINE>  
<TEXT>  
<P>  
In a strange twist of fate, Apple's newest iPad is experiencing explosive issues. Electricity and batteries and touchscreen technology.  
</P>  
</TEXT>  
<GRAPHIC>  
<P>  
iPad with shiny new display!  
</P>  
</GRAPHIC>  
<TYPE>  
<P>  
Article  
</P>  
</TYPE>  
</DOC>  
<DOC>  
<DOCNO> RF020218-0003 </DOCNO>  
<DOCID> 3 </DOCID>  
<DATE>  
<P>  
February 2, 2018  
</P>  
</DATE>  
<HEADLINE>  
<P>  
FORD CUTS 100,000 JOBS AMID POOR FIRST QUARTER REPORT  
</P>  
</HEADLINE>  
<BYLINE>  
<P>  
By Ramandeep Farmaha  
</P>  
</BYLINE>  
<TEXT>  
<P>  
Thanks to Tesla's new launch, Ford is experiencing a massive downturn. The beleaguered automaker reduce sales car by a lot.  
</P>  
</TEXT>  
<GRAPHIC>  
<P>  
Graph of Ford's plummeting sales.  
</P>  
</GRAPHIC>  
<TYPE>  
<P>

```

Article
</P>
</TYPE>
</DOC>
<DOC>
<DOCNO> RF020418-0004 </DOCNO>
<DOCID> 4 </DOCID>
<DATE>
<P>
February 2, 2018
</P>
</DATE>
<HEADLINE>
<P>
CANADIAN MAN ARRESTED FOR NOT APOLOGIZING ENOUGH IN TRAFFIC ACCIDENT
</P>
</HEADLINE>
<BYLINE>
<P>
By Ramandeep Farmaha
</P>
</BYLINE>
<TEXT>
<P>
Canadian man in rush to buy Tim Hortons coffee. Car hit snowplow. Couldn't get his
coffee. Didn't say sorry enough times.
</P>
</TEXT>
<GRAPHIC>
<P>
Canadian man showing no remorse.
</P>
</GRAPHIC>
<TYPE>
<P>
Article
</P>
</TYPE>
</DOC>

```

The queries below were then run using the Boolean AND retrieval script against the test document collection.

```

topics = {1: 'electric car', 2: 'touchscreen display', 3: 'car accident', 4:
'coffee', 5: 'car'}

```

After performing the Boolean AND retrieval script, the following was outputted:

```

1 q0 RF020218-0001 1 0 rfarmahaAND
2 q0 RF020218-0001 1 1 rfarmahaAND
2 q0 RF020218-0002 2 0 rfarmahaAND
3 q0 RF020418-0004 1 0 rfarmahaAND
4 q0 RF020418-0004 1 0 rfarmahaAND

```

```
5 q0 RF020218-0001 1 2 rfarmahaAND
5 q0 RF020218-0003 2 1 rfarmahaAND
5 q0 RF020418-0004 3 0 rfarmahaAND
```

### Topic 1: 'electric car'

Although the first, third and fourth articles in the document collection all contain the word 'car', only the first document contains both 'electric' and 'car'. Thus, the Boolean AND script outputted only the first article for the query.

### Topic 2: 'touchscreen display'

Both the first and second articles contain the words 'touchscreen' and 'display', while the other two articles don't contain either word.

### Topic 3: 'car accident'

Similar to the first topic, 3 out of the 4 articles contain the word 'car', but only the fourth article contains the word 'accident', thus it is the only result in the output.

### Topic 4: 'coffee'

Only the fourth article has the term 'coffee', and thus it is the only one outputted.

### Topic 5: 'car'

The first, second, and fourth articles all contain the word 'car', and thus they're outputted.

## Problem 3

### Topic 401: 'foreign minorities, Germany'

Rank	Docno	Relevance	Explanation
1	LA021890-0100	Non-relevant	Minorities refers to minorities of adults, not foreign minorities
2	LA090490-0093	Non-relevant	No match for 'foreign minorities'
3	LA050789-0068	Non-relevant	Document discusses countries other than Germany
4	LA122389-0060	Non-relevant	Article focuses on Romania
5	LA111289-0073	Non-relevant	Articles just mentions plight of minorities
6	LA121890-0117	Non-relevant	Examines many countries, not just Germany
7	LA040490-0003	Non-relevant	Focuses on Lithuanian politics
8	LA051390-0170	Non-relevant	Article concerns the USSR
9	LA052190-0065	Non-relevant	Article pertains Romania
10	LA050590-0114	Non-relevant	Article is about Latvian politics

Precision: 0.0

### Topic 403: 'osteoporosis'

Rank	Docno	Relevance	Explanation
1	LA120689-0083	Non-relevant	Relates to estrogen cream's effects on osteoporosis
2	LA111589-0004	Non-relevant	Article is on a study concerning menopause
3	LA051490-0120	Relevant	Document discusses taking an oral fluoride supplement to improve bone density
4	LA110490-0091	Non-relevant	Briefly mentions osteoporosis in passing; mostly about HGH
5	LA101890-0267	Relevant	Article discusses Didronel, an oral supplement for bone density loss
6	LA032489-0093	Non-relevant	Mentions osteoporosis in passing
7	LA071290-0133	Relevant	Focuses on Etidronate, which is a drug used to abate osteoporosis
8	LA092890-0067	Relevant	Article discusses calcium supplements for slowing bone loss
9	LA120390-0005	Non-relevant	Article briefly mentions osteoporosis
10	LA030689-0082	Non-relevant	Talks about osteoporosis, but does not discuss dietary intakes

Precision: 0.4