MSCI 541: HW 2

Jonathan Chen
20722167
October 17th, 2021

General

All programs were created using Java (version 17, 2021-09-14 LTS, Java(TM) SE Runtime Environment (build 17+35-LTS-2724), using the IntelliJ IDE.

To run the programs, the pre-built jar files can be used. To run the IndexEngine, go to the *out/artifacts/IndexEngine_jar/* directory in the command line and then run *java -jar IndexEngine.jar [path_to_latimes.gz] [output_path]*. Similarly, to run the BooleanAND program, go to the *out/artifacts/BooleanAND_jar/* directory in the command line and then run *java -jar BooleanAND.jar [path to indexed data] [path to queries file] [output filename]*.

Question 2

The following test collection containing four documents was created:

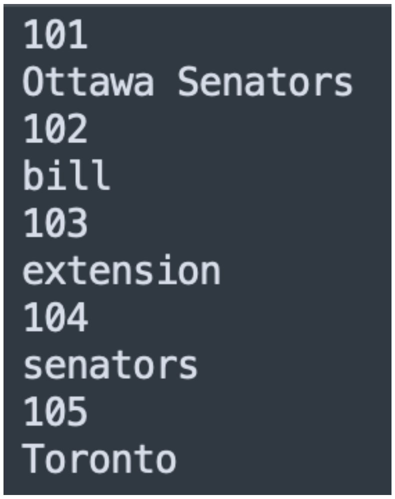
```
<DOC>
<DOCNO> JC101421-0001 </DOCNO>
<DOCID> 1 </DOCID>
<DATE>
<P>
October 14, 2021
</P>
</DATE>
<LENGTH>
<P>
25 words
</P>
</LENGTH>
<HEADLINE>
<P>
VICTORY: Senators Win Big over Maple Leafs
</P>
</HEADLINE>
<TEXT>
<P>
The Ottawa Senators defeated the Toronto Maple Leafs to start the new NHL season.
Brady Tkachuk scores three goals en route to a 5-2 win.
</P>
</TEXT>
```

```
<TYPE>
<P>
Article
</P>
</TYPE>
</DOC>
<DOC>
<DOCNO> JC101421-0002 </DOCNO>
<DOCID> 2 </DOCID>
<DATE>
<P>
October 14, 2021
</P>
</DATE>
<LENGTH>
<P>
34 words
</P>
</LENGTH>
<HEADLINE>
<P>
Senate Blocks Bill C-12345
</P>
</HEADLINE>
<TEXT>
<P>
Members of the Canadian Senate voted 154-153 against the passing of Bill C-12345. The
senators voted against this Bill for ethical reasons.
</P>
</TEXT>
<TYPE>
<P>
Article
</P>
</TYPE>
</DOC>
<DOC>
<DOCNO> JC101421-0003 </DOCNO>
<DOCID> 3 </DOCID>
<DATE>
```

```
<P>
October 14, 2021
</P>
</DATE>
<LENGTH>
<P>
30 words
</P>
</LENGTH>
<HEADLINE>
<P>
SIGNED: Brady Tkachuk Signs Extension with Ottawa Senators
</P>
</HEADLINE>
<TEXT>
<P>
Brady Tkachuk has signed a new 7 year extension with the Senators. Although he will
miss the game tonight, he is expected to be named captain within the coming days.
</P>
</TEXT>
<TYPE>
<P>
Article
</P>
</TYPE>
</DOC>
<DOC>
<DOCNO> JC101421-0004 </DOCNO>
<DOCID> 4 </DOCID>
<DATE>
<P>
October 14, 2021
</P>
</DATE>
<LENGTH>
<P>
30 words
</P>
</LENGTH>
<HEADLINE>
```

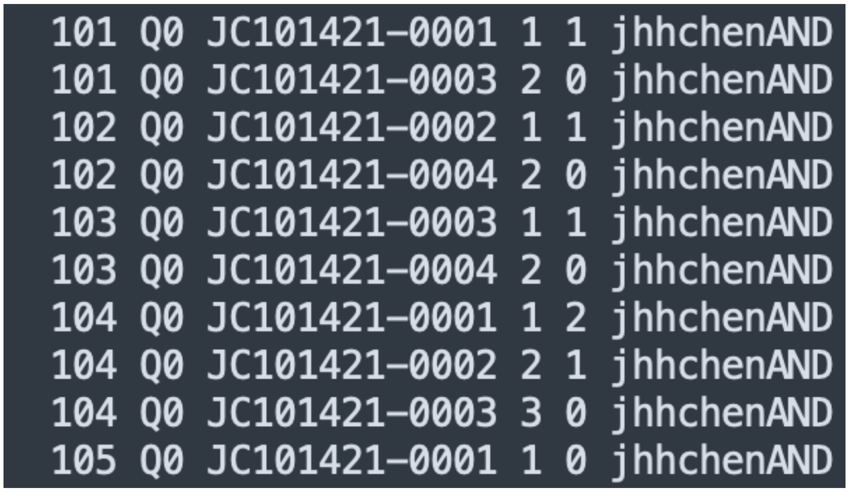
```
<P>
Lucky Duck?
</P>
</HEADLINE>
<TEXT>
<P>
A duck wandering the streets of Waterloo was gifted a lifetime supply of worms. This
duck was given this award after being judged to have the "longest bill" in Waterloo.
</P>
</TEXT>
<GRAPHIC>
<P>
Duck with the longest bill. Did it have a bill extension?
</P>
</GRAPHIC>
<TYPE>
<P>
Article
</P>
</TYPE>
</DOC>
```

Next, the queries file was created using the following queries:



```
101
Ottawa Senators
102
bill
103
extension
104
senators
105
Toronto
```

Finally, the BooleanAND retrieval output is:



```
101 Q0 JC101421-0001 1 1 jhhchenAND
101 Q0 JC101421-0003 2 0 jhhchenAND
102 Q0 JC101421-0002 1 1 jhhchenAND
102 Q0 JC101421-0004 2 0 jhhchenAND
103 Q0 JC101421-0003 1 1 jhhchenAND
103 Q0 JC101421-0004 2 0 jhhchenAND
104 Q0 JC101421-0001 1 2 jhhchenAND
104 Q0 JC101421-0002 2 1 jhhchenAND
104 Q0 JC101421-0003 3 0 jhhchenAND
105 Q0 JC101421-0001 1 0 jhhchenAND
```

Topic 101: "Ottawa Senators"
The term "Ottawa Senators," is mentioned in two documents: JC101421-0001 and JC101421-0003. The BooleanAND method correctly identified these two documents from text contained within the "TEXT" or "HEADLINE" sections. Moreover, the BooleanAND retrieval correctly omitted the JC101421-0002 document which contains the word "senators," but not the entire term.

Topic 102: "bill"
The term "bill" is mentioned in two documents: JC101421-0002 (with a capitalized first letter) and JC101421-0004. The BooleanAND method correctly identified these two documents by ignoring the letter casing, and also omitted the other two documents.

Topic 103: "extension"
The term "extension" is mentioned in two documents: JC101421-0003 and JC101421-0004. The BooleanAND method correctly identified these two documents from text contained within the "TEXT" or "GRAPHIC" sections.

Topic 104: "senators"
The term "senators" is mentioned in three documents: JC101421-0001, JC101421-0002 and JC101421-0003. The BooleanAND method correctly identified all three documents, ignored letter casing, and omitted the last document.

Topic 105: "Toronto"
The term "Toronto" is only mentioned in JC101421-0001. The BooleanAND method correctly returns this one document.

Question 3

Topic 401: foreign minorities, Germany

Rank	Docno	Relevance	Reasoning
1	LA122990-0070	Not-relevant	Discusses Soviet Jewish immigrants immigrating to Israel
2	LA040389-0047	Not-relevant	Discusses NATO's need to develop new policies
3	LA040490-0003	Not-relevant	Discusses countries wishing to secede from the Soviet Union
4	LA021890-0100	Not-relevant	Discusses the reunification of Germany
5	LA052190-0065	Not-relevant	Discusses elections in Romania
6	LA100889-0019	Not-relevant	Discusses the 54th International Pen Congress
7	LA082690-0052	Not-relevant	Discusses a group of motorcyclists travelling along the Silk Road
8	LA090490-0093	Not-relevant	Discusses the potential impacts of a Gulf War
9	LA050590-0114	Not-relevant	Discusses Latvian independence
10	LA050789-0068	Not-relevant	Does not discuss language and cultural differences impeding the integration in Germany in a significant way

Precision = 0.0

Topic 403: osteoporosis

Rank	Docno	Relevance	Reasoning
1	LA033089-0013	Relevant	Discusses eating vegetables to prevent osteoporosis
2	LA082890-0074	Not-relevant	Discusses the impacts of hormone therapy
3	LA111589-0004	Not-relevant	Discusses postmonopausal hormone therapies and heart disease
4	LA051490-0120	Not-relevant	Discusses impact of fluoride on cancer and osteoporosis
5	LA033089-0019	Relevant	Diet may contribute to osteoporosis
6	LA042189-0027	Not-relevant	Discusses clothing
7	LA110490-0091	Not-relevant	Discusses the impact of an age reversing hormone
8	LA111989-0048	Not-relevant	Discusses human health across different periods in time
9	LA101890-0267	Not-relevant	Does not discuss nutrition and preventing osteoporosis
10	LA050290-0050	Not-relevant	Discusses research funding for diseases related to aging

Precision = 0.2