MSCI 541 Homework 2

Thomas Kleinknecht

Problem 2 - BooleanAND Retrieval

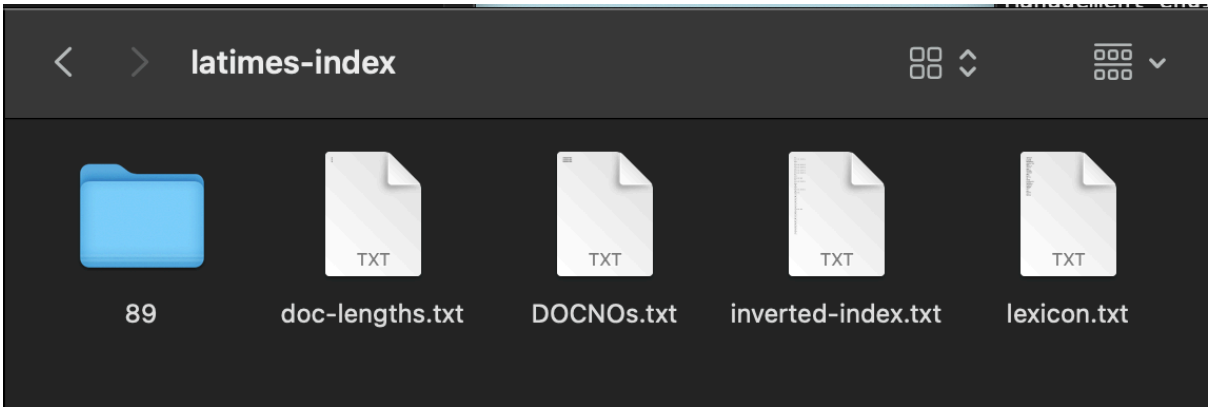
Search Functionality:

Test Documents (4):

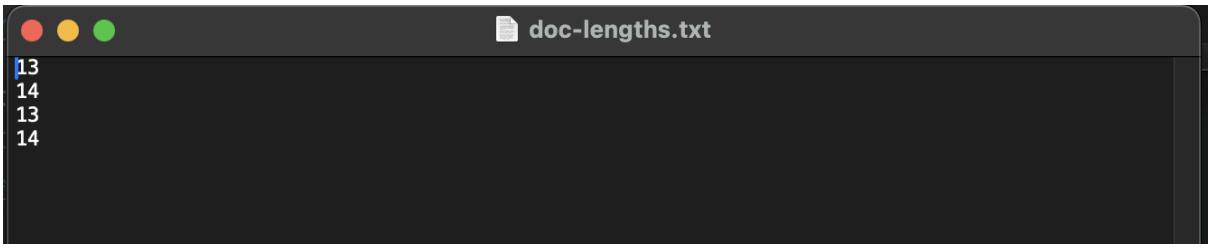
testdocs.txt

```
<DOC>
<DOCNO> LA010189-0007 </DOCNO>
<HEADLINE>
<P>
Vacation: Florida
Program: Management engineering
Home: Oakville, Ontario
Name: Matthew Erxleben
Interest: Data
</P>
</HEADLINE>
</DOC>
<DOC>
<DOCNO> LA010189-0008 </DOCNO>
<HEADLINE>
<P>
Vacation: Banff
Program: Management engineering
Home: Halifax, NS
Name: Thomas Kleinknecht
Interest: Product, Data
</P>
</HEADLINE>
</DOC>
<DOC>
<DOCNO> LA010189-0009 </DOCNO>
<HEADLINE>
<P>
Vacation: Vancouver
Program: Management engineering
Home: Brampton, Ontario
Name: Abhinav Sondhi
Interest: Product
</P>
</HEADLINE>
</DOC>
<DOC>
<DOCNO> LA010189-0010 </DOCNO>
<HEADLINE>
<P>
Vacation: New York
Program: Management engineering
Home: Toronto, Ontario
Name: John Kachura
Interest: Product
</P>
</HEADLINE>
</DOC>
```

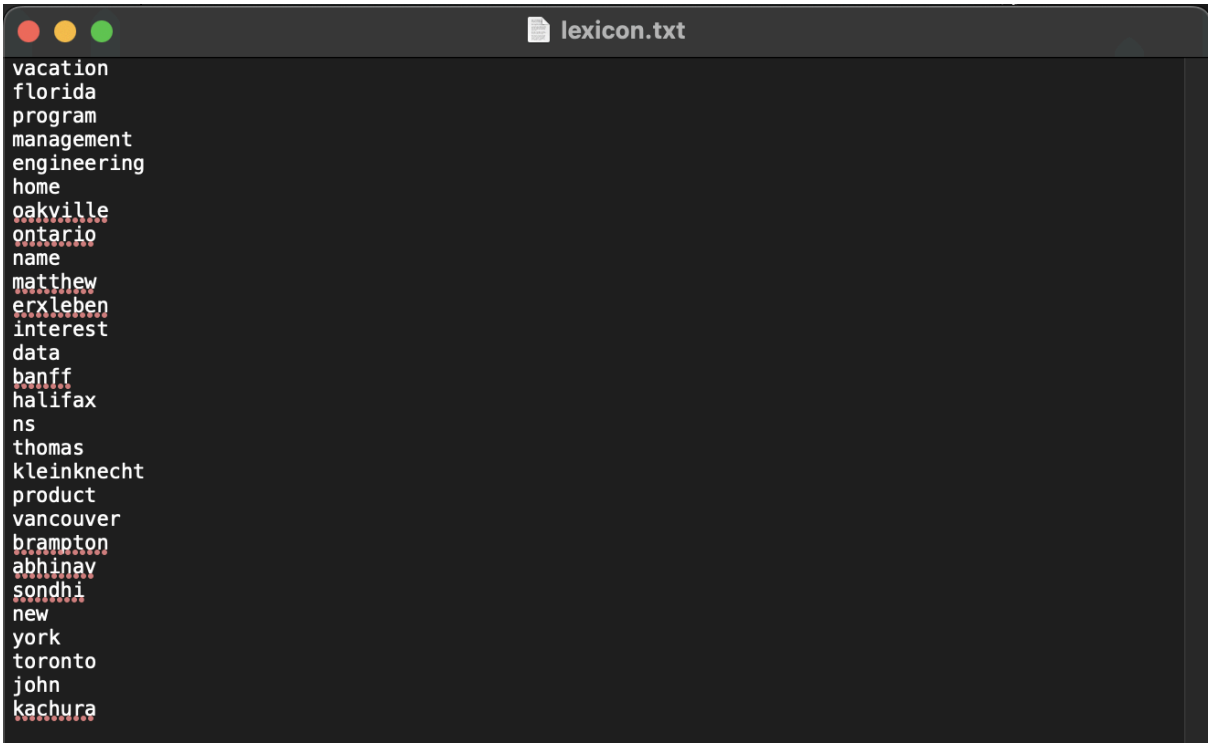
These were then gzipped and read into IndexEngine. This resulted in the lexicon, inverted index, and doc length files being created:



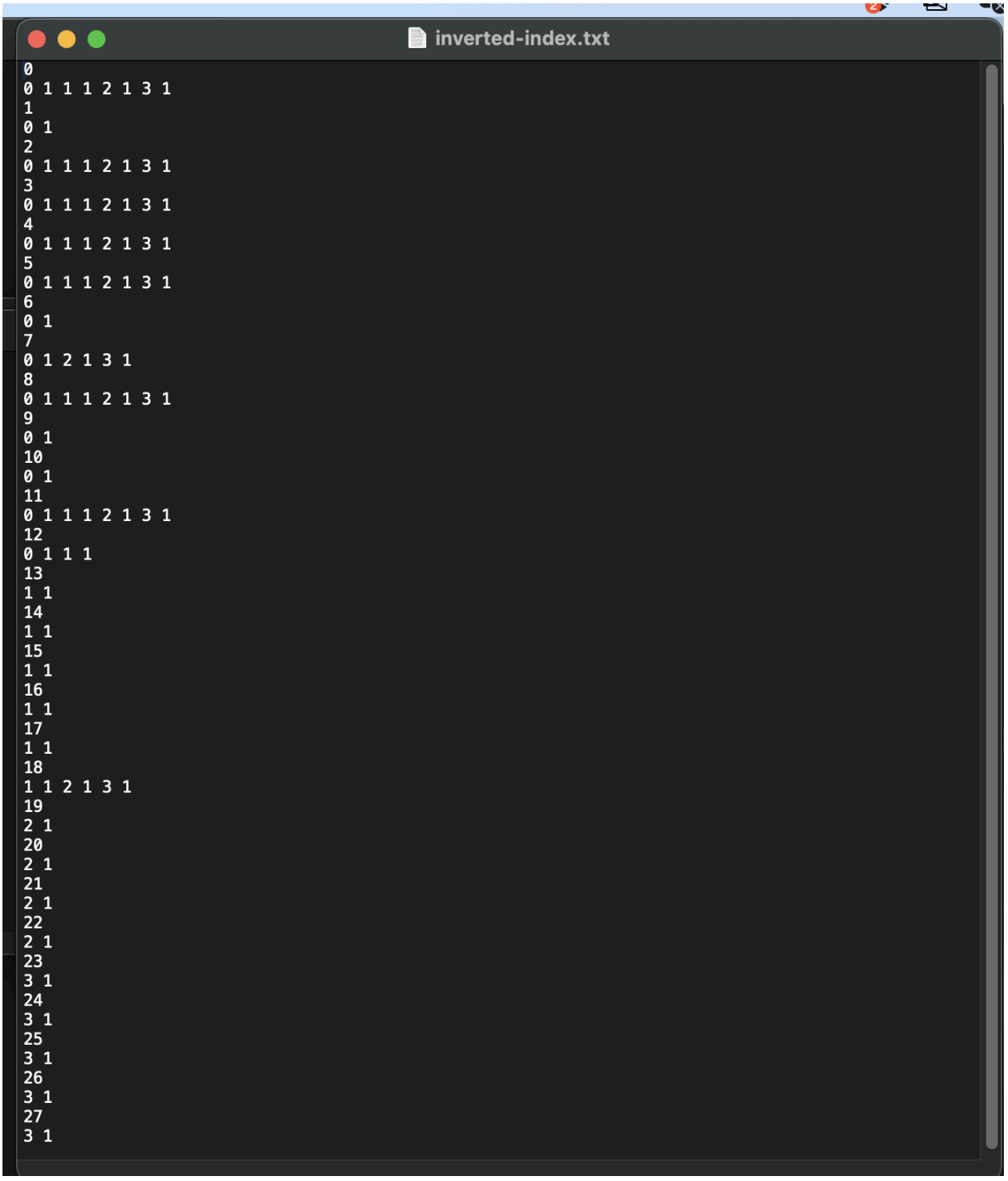
The doc lengths are saved in doc-lengths.txt, one length per line, with the line corresponding to the docID + 1:



The lexicon, stored under lexicon.txt, operates under the same structure, with the line in the doc corresponding to the termID + 1. These files contain 28 unique tokens:



The inverted index is stored in the file inverted-index.txt, with the first line corresponding to the first term id, and the next line representing the integer posting list (docID, count, docID, count etc.) with each term delimited by a space. The following line is the next term ID, then posting list, and so on:



The queries.txt file is then put into the latimes-index directory, and now BooleanAND is ready to be run.

Queries:

queries.txt

```
1
Brampton
2
Engineering, data
3
Management Engineering
4
Management engineering, Ontario
5
Management engineering vacation Halifax
6
New York
7
Los Angeles
```

1. One search term, only exists for Abhinav, (termid = 20), one result (docid = 2): LA010189-0009
2. Two search terms, engineering is in all four docs, however data is only in the first two: LA010189-0007, LA010189-0008
3. Two search terms, they are in every single doc, so all four docs are returned: LA010189-0007, LA010189-0008, LA010189-0009, LA010189-0010
4. Three search terms, management and engineering are in all four docs, but ontario is only in the first, third, and fourth: LA010189-0007, LA010189-0009, LA010189-0010
5. Four search terms, management engineering and vacation are in all docs, but halifax is only in the second: LA010189-0008
6. Two search terms, both are only present in the last doc: LA010189-0010
7. Two search terms, neither are in any docs, will not return any results

Results:

hw2-results-tkleinkn.txt

```
1 Q0 LA010189-0009 1 0 tkleinknAND
2 Q0 LA010189-0007 1 1 tkleinknAND
2 Q0 LA010189-0008 2 0 tkleinknAND
3 Q0 LA010189-0007 1 3 tkleinknAND
3 Q0 LA010189-0008 2 2 tkleinknAND
3 Q0 LA010189-0009 3 1 tkleinknAND
3 Q0 LA010189-0010 4 0 tkleinknAND
4 Q0 LA010189-0007 1 2 tkleinknAND
4 Q0 LA010189-0009 2 1 tkleinknAND
4 Q0 LA010189-0010 3 0 tkleinknAND
5 Q0 LA010189-0008 1 0 tkleinknAND
6 Q0 LA010189-0010 1 0 tkleinknAND
```

This demonstrates how the program will only return the docs in which every term in present, and how it can iterate through queries such as 3 and 4, and only return those specific doc(s), even though the other terms appear in all of the documents. It is also able to downcase everything, as seen by the variation in capitalization in the query, and remove non-alphanumerics, such as the colons in the doc, or the comma in query 4.

Resilience/Input Testing:

1. Three arguments aren't provided

```
topic id: 7, query: Los Angeles
Thomas-MacBook:msci-541-f23-hw2-thomask902 thomaskleinknecht$ java BooleanAND "/Users/thomaskleinknecht/Desktop/MSCI 541/latimes-index" queries.txt
Please provide a path to your latimes-index directory as well as your queries file name and output file name.
Thomas-MacBook:msci-541-f23-hw2-thomask902 thomaskleinknecht$
```

2. Path to latimes-index is invalid

```
Please provide a path to your latimes-index directory as well as your queries file name and output file name.
Thomas-MacBook:msci-541-f23-hw2-thomask902 thomaskleinknecht$ java BooleanAND "/Users/thomaskleinknecht/Desktop/MSCI/laimes-index" queries.txt hw2-results-tkleinkn.txt
Please provide the proper path to the latimes-index file. This directory does not exist.
Thomas-MacBook:msci-541-f23-hw2-thomask902 thomaskleinknecht$
```

3. Query file name is invalid

```
Please provide the proper path to the latimes-index file. This directory does not exist.
Thomas-MacBook:msci-541-f23-hw2-thomask902 thomaskleinknecht$ java BooleanAND "/Users/thomaskleinknecht/Desktop/MSCI 541/latimes-index" queriesnot.txt hw2-results-tkleinkn.txt
Please provide the proper path to the queries file. This file does not exist.
Thomas-MacBook:msci-541-f23-hw2-thomask902 thomaskleinknecht$
```

4. Output file name is invalid

```
Please provide the proper path to the queries file. This file does not exist.
Thomas-MacBook:msci-541-f23-hw2-thomask902 thomaskleinknecht$ java BooleanAND "/Users/thomaskleinknecht/Desktop/MSCI 541/latimes-index" queries.txt hw2-results-tkleink
Please enter a valid output file name.
Thomas-MacBook:msci-541-f23-hw2-thomask902 thomaskleinknecht$
```

Problem 3 - Judging Relevance

Topic 401:
Search:
Title: foreign minorities, Germany

Description:
What language and cultural differences impede the integration of foreign minorities in Germany?

Narrative:
A relevant document will focus on the causes of the lack of integration in a significant way; that is, the mere mention of immigration difficulties is not relevant. Documents that discuss immigration problems unrelated to Germany are also not relevant.

Results:

Rank	Docno	Relevant?	Explanation:
1	LA021890-0100	No	Discusses reunification of germany, does not mention foreign minorities or immigration at all.
2	LA040389-0047	No	Discusses NATO relations as well as soviet politics, does not mention foreign minorities or immigration to Germany at all.
3	LA040490-0003	No	Discusses soviet states, does not mention foreign minorities or immigration to Germany at all.
4	LA050590-0114	No	Discusses Latvia's independence from USSR,

			does not mention foreign minorities or immigration to Germany at all.
5	LA050789-0068	Yes	Discusses the issues of foreign minorities in western european countries, including specific mention of German difficulties, and what is causing these issues.
6	LA051390-0170	No	Discusses Baltic States' independence from USSR, does not mention foreign minorities or immigration to Germany at all.
7	LA052190-0065	No	Discusses romanian election, does not mention foreign minorities or immigration to Germany at all.
8	LA082690-0052	No	Discusses western journalists driving motorcycles across USSR, does not mention foreign minorities or immigration to Germany at all.
9	LA090490-0093	No	Discusses tensions in persian gulf as they relate to europe and the US, does not mention foreign minorities or immigration to Germany at all.
10	LA100889-0019	No	Discusses freedom of expression convention held in Canada, does not mention foreign minorities or immigration to Germany at all.

Precision (Topic 401) = 1/10 = 0.1 = 10%

Topic 403:
Search:
Title: osteoporosis

Description:
Find information on the effects of the dietary intakes of potassium, magnesium and fruits and vegetables as determinants of bone mineral density in elderly men and women thus preventing osteoporosis (bone decay).

Narrative:
A relevant document may include one or more of the dietary intakes in the prevention of osteoporosis. Any discussion of the disturbance of nutrition and mineral metabolism that results in a decrease in bone mass is also relevant.

Results:

Rank	Docno	Relevant?	Explanation:
1	LA010390-0067	No	Doc discusses osteoporosis but does not mention dietary intake, nutrition, or mineral

			metabolism as a cause or potential solution.
2	LA010490-0218	No	Discusses drug innovations of the 80s and 90s, does not mention potential causes or dietary solutions to osteoporosis.
3	LA010689-0040	No	Discusses current dresses in fashion, briefly mentions one woman's exercise routine, does not mention potential causes or dietary solutions to osteoporosis.
4	LA010790-0103	No	Calls for participants in a study about nasal medicine and osteoporosis, does not mention potential causes or dietary solutions to osteoporosis.
5	LA011289-0149	Yes	Discusses how many women's lack on dietary calcium can cause osteoporosis and encourages foods high in calcium.
6	LA011389-0029	Yes	Discusses potential treatments of osteoporosis using calcium, a mineral/nutritional issue.
7	LA012990-0041	Yes	Discusses disturbance in nutrition, specifically a lack of dietary calcium, as a cause of decrease in bone mass/osteoporosis.
8	LA020490-0136	Yes	Discusses disturbance in nutrition, specifically a lack of dietary calcium, as a cause of decrease in bone mass/osteoporosis, and a potential preventative tool.
9	LA020990-0100	No	Mentions briefly out of context, but does not discuss dietary intake or nutrition as a cause or preventative tool for osteoporosis.
10	LA021590-0062	No	Doc discusses osteoporosis and exercise, but does not mention dietary intake, nutrition, or mineral metabolism as a cause or potential solution.

Precision (Topic 403) = 4/10 = 0.4 = 40%

References

- Heavily referenced pseudocode from course content, as well as design suggestions made in campuswire.