Thomas Enns - 20823674
MSCI 541
2022-10-06

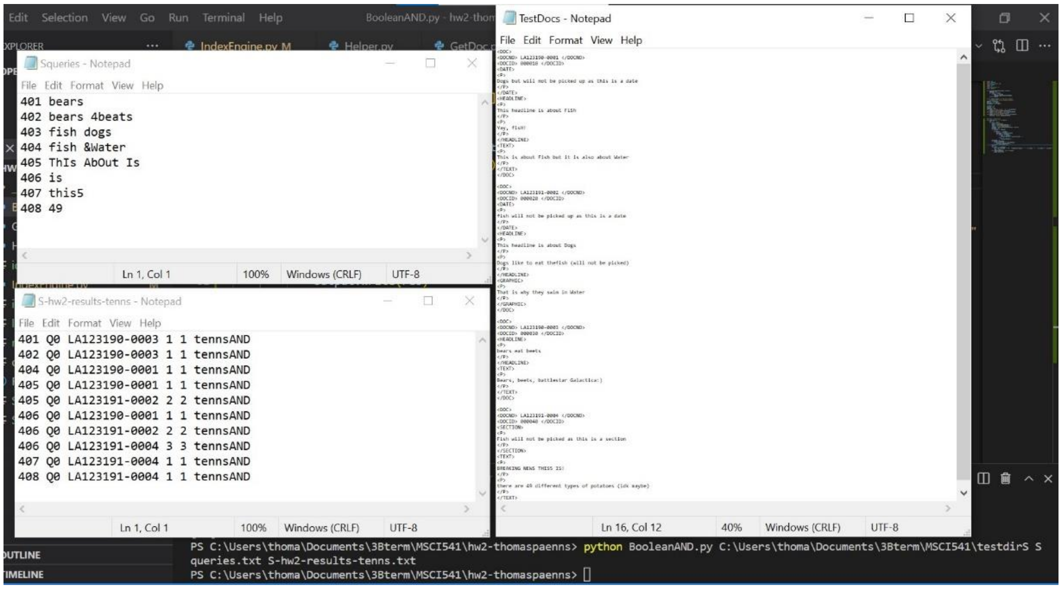
HW2 – Report

2A: A set of 4 documents were created as a test set for this program. These were created in one txt file which was the gzipped and read with the IndexEngine in the same way as the LA Times data. The content of the test set document is below:

```
<DOC>
<DOCNO> LA123190-0001 </DOCNO>
<DOCID> 000010 </DOCID>
<DATE>
<P>
Dogs but will not be picked up as this is a date
</P>
</DATE>
<HEADLINE>
<P>
This headline is about FiSh
</P>
<P>
Yay, fisH!
</P>
</HEADLINE>
<TEXT>
<P>
This is about Fish but it is also about Water
</P>
</TEXT>
</DOC>

<DOC>
<DOCNO> LA123191-0002 </DOCNO>
<DOCID> 000020 </DOCID>
<DATE>
<P>
fish will not be picked up as this is a date
</P>
</DATE>
<HEADLINE>
<P>
This headline is about Dogs
</P>
<P>
Dogs like to eat thefish (will not be picked)
</P>
</HEADLINE>
<GRAPHIC>
<P>
That is why they swim in Water
</P>
</GRAPHIC>
</DOC>

<DOC>
<DOCNO> LA123190-0003 </DOCNO>
<DOCID> 000030 </DOCID>
<HEADLINE>
```

```
<P>
bears eat beets
</P>
</HEADLINE>
<TEXT>
<P>
Bears, beets, battlestar Galactica:)
</P>
</TEXT>
</DOC>

<DOC>
<DOCNO> LA123191-0004 </DOCNO>
<DOCID> 000040 </DOCID>
<SECTION>
<P>
Fish will not be picked as this is a section
</P>
</SECTION>
<TEXT>
<P>
BREAKING NEWS THIS5 IS!
</P>
<P>
there are 49 different types of potatoes (idk maybe)
</P>
</TEXT>
</DOC>
```

This test set of documents includes words repeated across many documents as well as words stored in tags that should not be treated as text. Many letters are haphazardly capitalized to ensure that tokens will be matched regardless of capitalization. Other words have non-alphanumeric adjoined to ensure that these will not be parsed and that the word will still become a token without the symbol. There are also areas where two words have no space between them to ensure that the program will tokenize them together. The below command, queries, and their results indicate that all the above input was handled as expected:



```
401 bears
402 bears 4beats
403 fish dogs
404 fish &Water
405 ThIs AbOut Is
406 is
407 this5
408 49
```

```
401 Q0 LA123190-0003 1 1 tennsAND
402 Q0 LA123190-0003 1 1 tennsAND
404 Q0 LA123190-0001 1 1 tennsAND
405 Q0 LA123190-0001 1 1 tennsAND
405 Q0 LA123191-0002 2 2 tennsAND
406 Q0 LA123190-0001 1 1 tennsAND
406 Q0 LA123191-0002 2 2 tennsAND
406 Q0 LA123191-0004 3 3 tennsAND
407 Q0 LA123191-0004 1 1 tennsAND
408 Q0 LA123191-0004 1 1 tennsAND
```

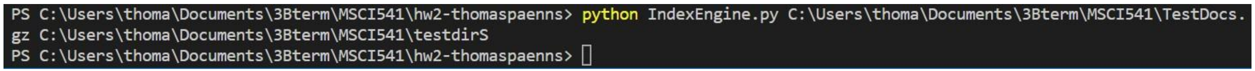
```
<DOC>
<DOCNO> LA123190-0001 </DOCNO>
<DOCID> 000010 </DOCID>
<DATE>
<P>
Dogs but will not be picked up as this is a date
</P>
</DATE>
<HEADLINE>
<P>
This headline is about Fish
</P>
<P>
Yay, fish!
</P>
</HEADLINE>
<TEXT>
<P>
This is about Fish but it is also about Water
</P>
</TEXT>
</DOC>

<DOC>
<DOCNO> LA123191-0002 </DOCNO>
<DOCID> 000020 </DOCID>
<DATE>
<P>
fish will not be picked up as this is a date
</P>
</DATE>
<HEADLINE>
<P>
This headline is about Dogs
</P>
<P>
Dogs like to eat thefish (will not be picked)
</P>
</HEADLINE>
<GRAPHIC>
<P>
That is why they swim in Water
</P>
</GRAPHIC>
</DOC>

<DOC>
<DOCNO> LA123190-0003 </DOCNO>
<DOCID> 000030 </DOCID>
<HEADLINE>
<P>
bears eat beets
</P>
</HEADLINE>
<TEXT>
<P>
Bears, beets, battlestar Galactica:)
</P>
</TEXT>
</DOC>

<DOC>
<DOCNO> LA123191-0004 </DOCNO>
<DOCID> 000040 </DOCID>
<SECTION>
<P>
Fish will not be picked as this is a section
</P>
</SECTION>
<TEXT>
<P>
BREAKING NEWS THIS5 IS!
</P>
<P>
there are 49 different types of potatoes (idk maybe)
</P>
</TEXT>
```

```
PS C:\Users\thoma\Documents\3Bterm\MSCI541\hw2-thomaspaenns> python BooleanAND.py C:\Users\thoma\Documents\3Bterm\MSCI541\testdirS S
queries.txt S-hw2-results-tenns.txt
PS C:\Users\thoma\Documents\3Bterm\MSCI541\hw2-thomaspaenns> 
```

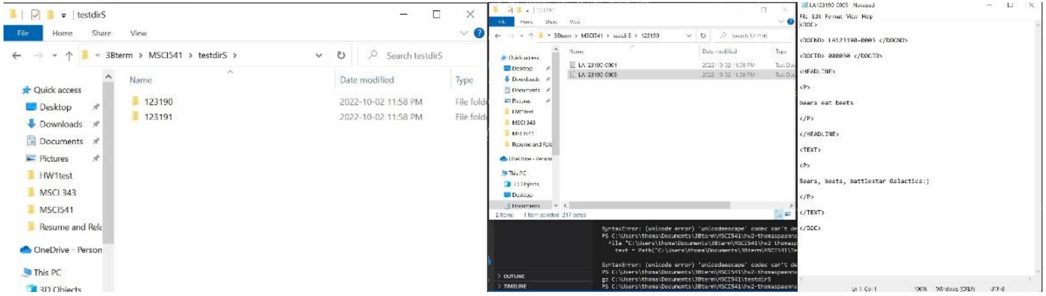
To properly test the new modifications to the code. Various data was printed out while modifying the IndexEngine file using the test set of documents. The integrity of this data directly effects the integrity of the BooleanAND program. As such, the following testing was done to ensure that the data used by the BooleanAND program was clean.

For the purposes of brevity and clarity, the screenshot corresponding to each test will be embedded in the test plan below the corresponding test. In each screenshot with the command line, the test is the most recent command, unless otherwise specified

a) After running the below command, the proper creation of files and documents for the small test set was verified:

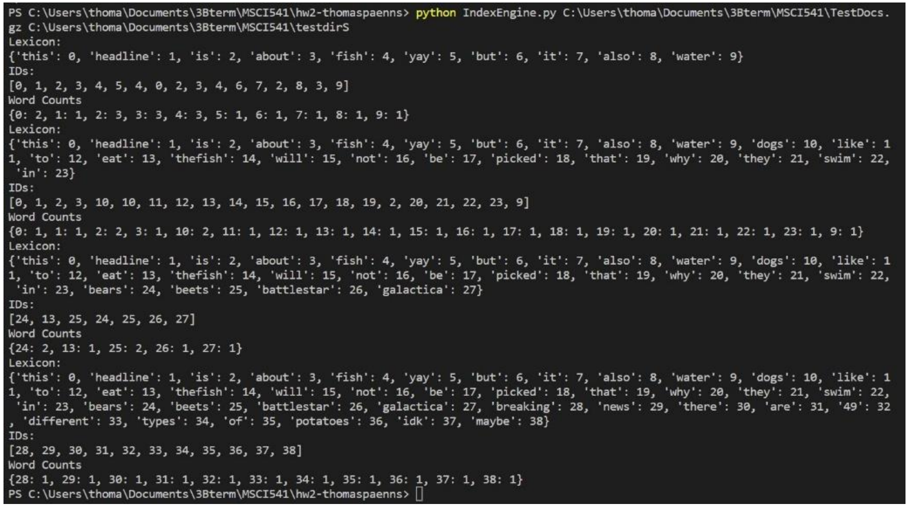


```
PS C:\Users\thoma\Documents\3Bterm\MSCI541\hw2-thomaspaenns> python IndexEngine.py C:\Users\thoma\Documents\3Bterm\MSCI541\TestDocs.
gz C:\Users\thoma\Documents\3Bterm\MSCI541\testdirS
PS C:\Users\thoma\Documents\3Bterm\MSCI541\hw2-thomaspaenns> 
```



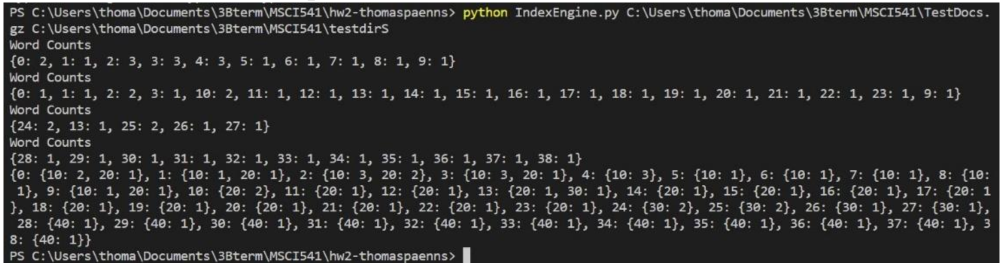
```
<DOC>

<DOCNO> LA123190-0003 </DOCNO>

<DOCID> 000030 </DOCID>

<HEADLINE>

<P>

bears eat beets

</P>

</HEADLINE>

<TEXT>

<P>

Bears, beets, battlestar Galactica:)

</P>

</TEXT>

</DOC>
```

b) To ensure the integrity of token data, the word counts, token ids, and the lexicon were printed out progressively for each doc in the test set.



```
PS C:\Users\thoma\Documents\3Bterm\MSCI541\hw2-thomaspaenns> python IndexEngine.py C:\Users\thoma\Documents\3Bterm\MSCI541\TestDocs.
gz C:\Users\thoma\Documents\3Bterm\MSCI541\testdirS
Lexicon:
{'this': 0, 'headline': 1, 'is': 2, 'about': 3, 'fish': 4, 'yay': 5, 'but': 6, 'it': 7, 'also': 8, 'water': 9}
IDs:
[0, 1, 2, 3, 4, 5, 4, 0, 2, 3, 4, 6, 7, 2, 8, 3, 9]
Word Counts
{0: 2, 1: 1, 2: 3, 3: 3, 4: 3, 5: 1, 6: 1, 7: 1, 8: 1, 9: 1}
Lexicon:
{'this': 0, 'headline': 1, 'is': 2, 'about': 3, 'fish': 4, 'yay': 5, 'but': 6, 'it': 7, 'also': 8, 'water': 9, 'dogs': 10, 'like': 1
1, 'to': 12, 'eat': 13, 'thefish': 14, 'will': 15, 'not': 16, 'be': 17, 'picked': 18, 'that': 19, 'why': 20, 'they': 21, 'swim': 22,
 'in': 23}
IDs:
[0, 1, 2, 3, 10, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 2, 20, 21, 22, 23, 9]
Word Counts
{0: 1, 1: 1, 2: 2, 3: 1, 10: 2, 11: 1, 12: 1, 13: 1, 14: 1, 15: 1, 16: 1, 17: 1, 18: 1, 19: 1, 20: 1, 21: 1, 22: 1, 23: 1, 9: 1}
Lexicon:
{'this': 0, 'headline': 1, 'is': 2, 'about': 3, 'fish': 4, 'yay': 5, 'but': 6, 'it': 7, 'also': 8, 'water': 9, 'dogs': 10, 'like': 1
1, 'to': 12, 'eat': 13, 'thefish': 14, 'will': 15, 'not': 16, 'be': 17, 'picked': 18, 'that': 19, 'why': 20, 'they': 21, 'swim': 22,
 'in': 23, 'bears': 24, 'beets': 25, 'battlestar': 26, 'galactica': 27}
IDs:
[24, 13, 25, 24, 25, 26, 27]
Word Counts
{24: 2, 13: 1, 25: 2, 26: 1, 27: 1}
Lexicon:
{'this': 0, 'headline': 1, 'is': 2, 'about': 3, 'fish': 4, 'yay': 5, 'but': 6, 'it': 7, 'also': 8, 'water': 9, 'dogs': 10, 'like': 1
1, 'to': 12, 'eat': 13, 'thefish': 14, 'will': 15, 'not': 16, 'be': 17, 'picked': 18, 'that': 19, 'why': 20, 'they': 21, 'swim': 22,
 'in': 23, 'bears': 24, 'beets': 25, 'battlestar': 26, 'galactica': 27, 'breaking': 28, 'news': 29, 'there': 30, 'are': 31, '49': 32
, 'different': 33, 'types': 34, 'of': 35, 'potatoes': 36, 'idk': 37, 'maybe': 38}
IDs:
[28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38]
Word Counts
{28: 1, 29: 1, 30: 1, 31: 1, 32: 1, 33: 1, 34: 1, 35: 1, 36: 1, 37: 1, 38: 1}
PS C:\Users\thoma\Documents\3Bterm\MSCI541\hw2-thomaspaenns> 
```

c) To ensure that the inverted index had the proper nested data structure and was loaded with the correct postings, the word counts for each document were printed progressively with the inverted index printed at the end.



```
PS C:\Users\thoma\Documents\3Bterm\MSCI541\hw2-thomaspaenns> python IndexEngine.py C:\Users\thoma\Documents\3Bterm\MSCI541\TestDocs.
gz C:\Users\thoma\Documents\3Bterm\MSCI541\testdirS
Word Counts
{0: 2, 1: 1, 2: 3, 3: 3, 4: 3, 5: 1, 6: 1, 7: 1, 8: 1, 9: 1}
Word Counts
{0: 1, 1: 1, 2: 2, 3: 1, 10: 2, 11: 1, 12: 1, 13: 1, 14: 1, 15: 1, 16: 1, 17: 1, 18: 1, 19: 1, 20: 1, 21: 1, 22: 1, 23: 1, 9: 1}
Word Counts
{24: 2, 13: 1, 25: 2, 26: 1, 27: 1}
Word Counts
{28: 1, 29: 1, 30: 1, 31: 1, 32: 1, 33: 1, 34: 1, 35: 1, 36: 1, 37: 1, 38: 1}
{0: {10: 2, 20: 1}, 1: {10: 1, 20: 1}, 2: {10: 3, 20: 2}, 3: {10: 3, 20: 1}, 4: {10: 3}, 5: {10: 1}, 6: {10: 1}, 7: {10: 1}, 8: {10:
 1}, 9: {10: 1, 20: 1}, 10: {20: 2}, 11: {20: 1}, 12: {20: 1}, 13: {20: 1, 30: 1}, 14: {20: 1}, 15: {20: 1}, 16: {20: 1}, 17: {20: 1
}, 18: {20: 1}, 19: {20: 1}, 20: {20: 1}, 21: {20: 1}, 22: {20: 1}, 23: {20: 1}, 24: {30: 2}, 25: {30: 2}, 26: {30: 1}, 27: {30: 1},
 28: {40: 1}, 29: {40: 1}, 30: {40: 1}, 31: {40: 1}, 32: {40: 1}, 33: {40: 1}, 34: {40: 1}, 35: {40: 1}, 36: {40: 1}, 37: {40: 1}, 3
8: {40: 1}}
PS C:\Users\thoma\Documents\3Bterm\MSCI541\hw2-thomaspaenns> 
```

2B:

There is no written component for this problem, the results can be found in Github. However, note that I operated under the assumption that the program should only return the top ten results to each query.

3:

Topic 401 - foreign minorities, Germany

Rank	DOCNO	Judgement	Reasoning
1	LA021890-0100	Not Relevant	This document does not discuss minorities in Germany or their integration into society. Rather, it focuses German reunification post cold-war.
2	LA040389-0047	Not Relevant	This document does not discuss minorities in Germany. It discusses de-armament in Europe and minorities in the Soviet Union
3	LA040490-0003	Not Relevant	This document discusses breakaway soviet republics and the legality of secession given ethnic minorities wishing to remain in the Soviet Union. No mention of German minorities.
4	LA050590-0114	Not Relevant	This document speaks of Latvian Independence. The only mention of Germany is mentioning how a pact with Germany led to Lithuania under Soviet rule.
5	LA050789-0068	Relevant	This document is about European immigration policy. It mentions that “In West Berlin, three of the city's television stations now broadcast in Turkish.” However, at the time of writing, unemployment is causing animosity towards immigrants.
6	LA051390-0170	Not Relevant	This document, again only mentions minorities and Germany in the context of Baltic Soviet republics.
7	LA052190-0065	Not Relevant	This document is about the Romanian election and only makes mention of parliamentary minorities.
8	LA082690-0052	Not Relevant	This is about a motorcycle convoy on the silk road, only mentions Russian minorities and that German students were present
9	LA090490-0093	Not Relevant	This is about the gulf war. Mentions of Germany are limited to economic partnership with Iraq
10	LA100889-0019	Not Relevant	This is about an international writers conference

Topic 403 - Osteoperosis

Rank	DOCNO	Judgement	Reasoning
1	LA010390-0067	Not Relevant	This is about osteoporosis but speaks of a study that shows men are affected as well. No mention of nutrition or minerals.
2	LA010490-0218	Not Relevant	This is about new drugs, only a passing mention of osteoporosis in the conclusion.
3	LA010689-0040	Not Relevant	This only mentions osteoporosis in passing in regards to a workout regimen.
4	LA010790-0103	Not Relevant	This is an add for an osteoporosis drug study. There is no mention of nutrition or prevention.
5	LA011289-0149	Relevant	This document offers recipes for a healthy diet and mentions the importance of calcium to avoid osteoporosis.
6	LA011389-0029	Relevant	This document pertains to a sodium fluoride therapy for osteoporosis and also mentions the importance of calcium intake.
7	LA012990-0041	Not Relevant	This document is about the milk industry going in a new marketing direction. Only mentions in passing that they have had success promoting calcium as a prevention for osteoporosis.
8	LA020490-0136	Relevant	This article is on the effect of taking calcium and estrogen on osteoporosis.
9	LA020990-0100	Relevant	Mentions the link between calcium intake and a reduced risk of osteoporosis.
10	LA021590-0062	Not Relevant	Only mentions osteoporosis once in relation to exercise.