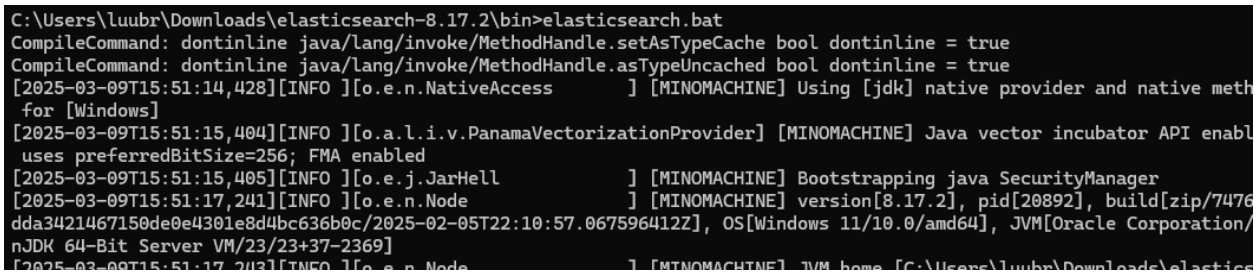
Brandon Luu - Assignment 2

Part 1: Elasticsearch

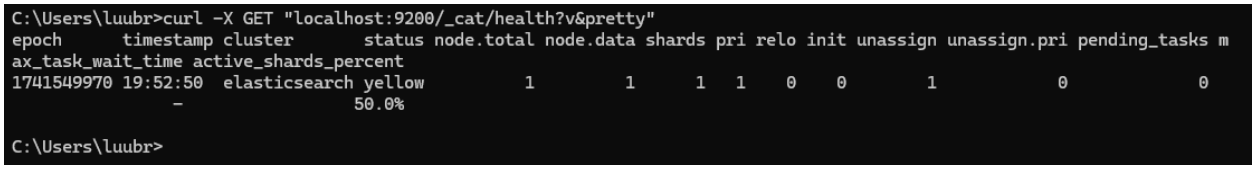
This part was done on my Windows machine. Apparently we need to have both part 1 and part 2 in this so I've included both.

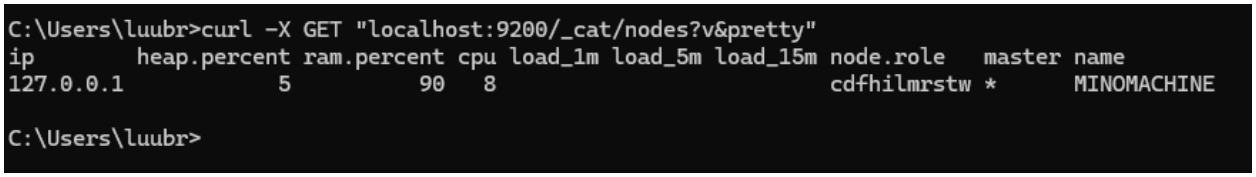
I am starting it up in this image.

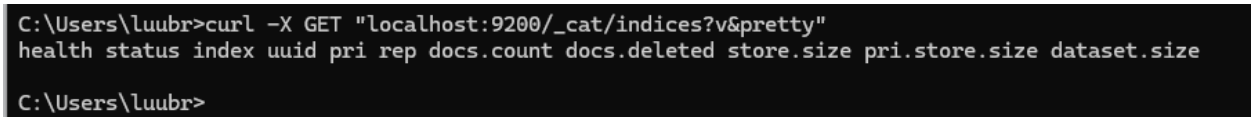


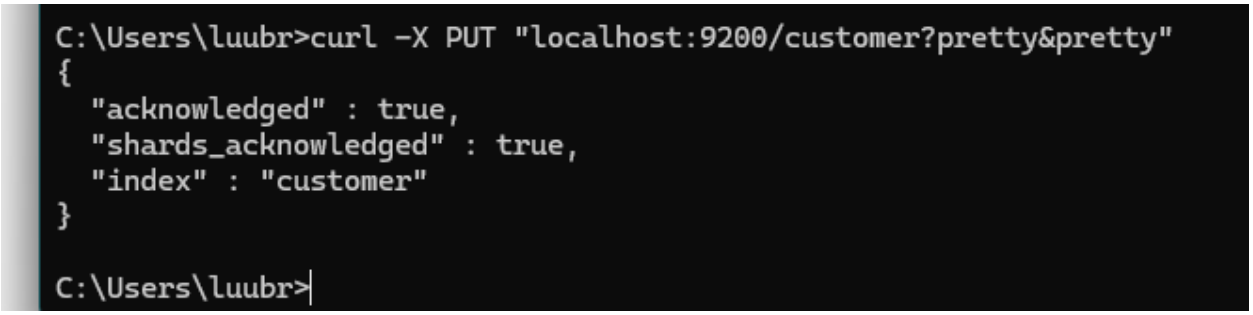
1.1 Basic Indexing

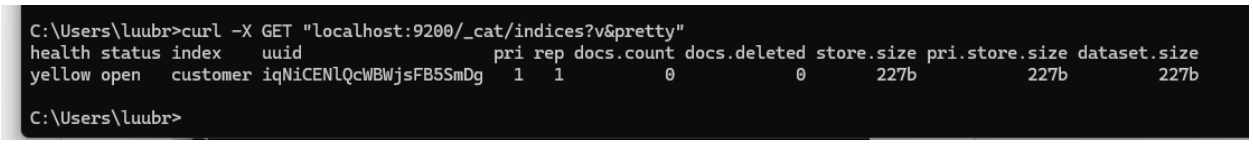
These commands I put in the terminal from the starting guide.



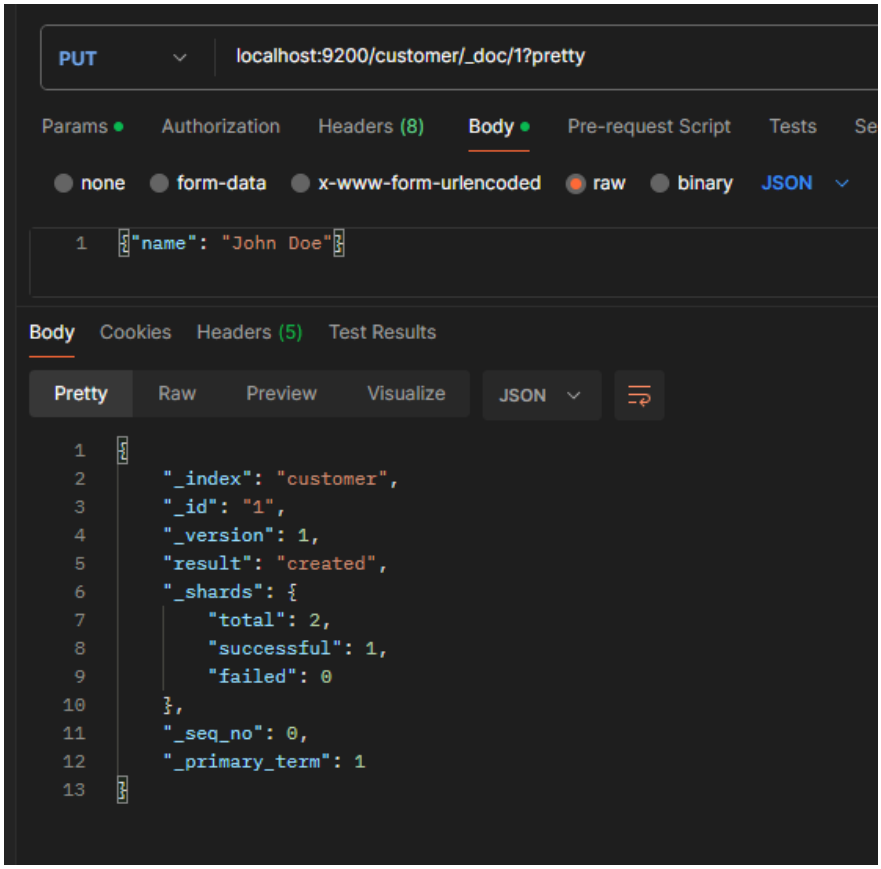








This command didn’t work in Command Prompt so I did it through Postman instead:




```
C:\Users\luubr>curl -X GET "localhost:9200/customer/_doc/1?pretty&pretty"
{
  "_index" : "customer",
  "_id" : "1",
  "_version" : 1,
  "_seq_no" : 0,
  "_primary_term" : 1,
  "found" : true,
  "_source" : {
    "name" : "John Doe"
  }
}

C:\Users\luubr>
```

```
C:\Users\luubr>curl -X DELETE "localhost:9200/customer?pretty&pretty"
{
  "acknowledged" : true
}
```

```
C:\Users\luubr>
C:\Users\luubr>curl -X GET "localhost:9200/_cat/indices?v&pretty"
health status index uuid pri rep docs.count docs.deleted store.size pri.store.size dataset.size
```

1.2 Indexing Reddit

Adding test.json into the cluster

```
C:\Users\luubr\Downloads\Assignment2\Part1>curl -H "Content-Type: application/json" -XPOST localhost:9200/_bulk --data-bi
nary @test.json > NUL
```

The index is called “comments” with a size of 31.7mb

```
C:\Users\luubr\Downloads\Assignment2\Part1>curl -X GET "localhost:9200/_cat/indices?v&pretty"
health status index    uuid                   pri rep docs.count docs.deleted store.size pri.store.size dataset.size
yellow open   comments ZbaSAlvuSTSvw4LG6QBoqA   1   1      61013            0     31.7mb         31.7mb       31.7mb
```

Running the first query, the other two also just give results, so please see part1.txt for my answers.

```
C:\Users\luubr\Downloads\Assignment2\Part1>curl -XGET "localhost:9200/_search?pretty" -H "Content-Type: application/json
" -d"{ \"query\": { \"match\" : { \"body\" : { \"query\" : \"cat\"} } } }"
{
  "took" : 71,
  "timed_out" : false,
  "_shards" : {
    "total" : 1,
    "successful" : 1,
    "skipped" : 0,
    "failed" : 0
  },
  "hits" : {
    "total" : {
      "value" : 90,
      "relation" : "eq"
    },
    "max_score" : 10.503806,
    "hits" : [
      {
        "_index" : "comments",
        "_id" : "TnV_fJUBadhSyrnJ48Zy",
        "_score" : 10.503806,
        "_ignored" : [
          "body.keyword"
```

Stopword query for “a”. Hits are returned so there is no stopword removal.

```
C:\Users\luubr\Downloads\Assignment2\Part1>curl -XGET "localhost:9200/_search?pretty" -H "Content-Type: application/json
" -d"{ \"query\": { \"match\" : { \"body\" : { \"query\" : \"a\"} } } }"
{
  "took" : 12,
  "timed_out" : false,
  "_shards" : {
    "total" : 1,
    "successful" : 1,
    "skipped" : 0,
    "failed" : 0
  },
  "hits" : {
    "total" : {
      "value" : 10000,
      "relation" : "gte"
    },
    "max_score" : 1.8262576,
    "hits" : [
      {
        "_index" : "comments",
        "_id" : "mHV_fJUBadhSyrnJ49F0",
        "_score" : 1.8262576,
        "_ignored" : [
          "body.keyword"
        ],
        "_source" : {
          "retrieved_on" : 1473808874,
```

1.3 Analyzers

Remove comments index.

```
C:\Users\luubr\Downloads\Assignment2\Part1>curl -X DELETE "localhost:9200/comments?pretty&pretty"
{
  "acknowledged" : true
}
```

Running two commands.

```
C:\Users\luubr\Downloads\Assignment2\Part1>curl -XPUT "localhost:9200/comments?pretty" -H "Content-Type: application/jso
n" -d"{\"settings\":{\"analysis\":{\"analyzer\":{\"my_analyzer\":{\"tokenizer\":\"standard\",\"filter\":[\"lowercase\",\
"my_stemmer\"]}},\"filter\":{\"my_stemmer\":{\"type\":\"stemmer\",\"name\":\"english\"}}}}}"
{
  "acknowledged" : true,
  "shards_acknowledged" : true,
  "index" : "comments"
}

C:\Users\luubr\Downloads\Assignment2\Part1>curl -XPUT "localhost:9200/comments/_mapping?pretty" -H "Content-Type: applic
ation/json" -d"{\"properties\" : {\"body\" : {\"type\" : \"text\",\"analyzer\" : \"my_analyzer\"}}}"
{
  "acknowledged" : true
}
```

Test query from assignment.

```
C:\Users\luubr\Downloads\Assignment2\Part1>curl -X POST "localhost:9200/comments/_analyze?pretty" -H "Content-Type: appl
ication/json" -d"{  \"analyzer\": \"my_analyzer\", \"text\": \"I\u0027m a :) person, and you?\"}"
{
  "tokens" : [
    {
      "token" : "i'm",
      "start_offset" : 0,
      "end_offset" : 3,
      "type" : "<ALPHANUM>",
      "position" : 0
```

Finally, I re-add the comments data into the index and re-run the queries. See part1.txt for my answers.

Part 2: Lucene

Done on my Mac using the terminal and IntelliJ editor. Please see part2.txt for the explanations, the output is included there too

2. Original Output

```
minokah@Naomis-MacBook-Air Part2 % java MySearchFiles < testcases.txt
Enter query:
Searching for: his fiery sword
5 total matching documents
1. ./documents/RJ3.txt
2. ./documents/RJ5.txt
3. ./documents/RJ4.txt
4. ./documents/RJ9.txt
5. ./documents/RJ6.txt
Press (q)uit or enter number to jump to a page.
Enter query:
Searching for: alas o love
5 total matching documents
1. ./documents/RJ6.txt
2. ./documents/RJ8.txt
3. ./documents/RJ7.txt
4. ./documents/RJ1.txt
5. ./documents/RJ10.txt
Press (q)uit or enter number to jump to a page.
Enter query:
Searching for: and
8 total matching documents
1. ./documents/RJ3.txt
2. ./documents/RJ10.txt
3. ./documents/RJ5.txt
4. ./documents/RJ2.txt
5. ./documents/RJ4.txt
6. ./documents/RJ8.txt
7. ./documents/RJ7.txt
8. ./documents/RJ1.txt
Press (q)uit or enter number to jump to a page.
Enter query:
Searching for: reckon
0 total matching documents
Enter query:
Searching for: q
0 total matching documents
Enter query:
Searching for: love
4 total matching documents
1. ./documents/RJ6.txt
2. ./documents/RJ7.txt
3. ./documents/RJ1.txt
4. ./documents/RJ10.txt
Press (q)uit or enter number to jump to a page.
Enter query:
Searching for: fair
3 total matching documents
1. ./documents/RJ9.txt
2. ./documents/RJ10.txt
3. ./documents/RJ1.txt
Press (q)uit or enter number to jump to a page.
Enter query:
minokah@Naomis-MacBook-Air Part2 %
```

2.1 Stopping

Adding code for stopwords in indexer and searcher

```
/*
2.1 Stopping : Adding the custom stop word list
*/

List<String> wordList = Arrays.asList("about", "dost", "from", "hath", "his", "O", "that", "the", "thou"); // CHANGE 1: Creating the word list
CharArraySet arraySet = new CharArraySet(wordList, ignoreCase: true); // CHANGE 2: Adding them to the Lucene class. Set ignoreCase to true for "O"

Analyzer analyzer = new MyStandardAnalyzer(arraySet); // CHANGE 3: Passing them into the Analyzer
```

Recompile and reindex. Results on next page

```
minokah@Naomis-MacBook-Air Part2 % javac *.java
minokah@Naomis-MacBook-Air Part2 %
minokah@Naomis-MacBook-Air Part2 %
minokah@Naomis-MacBook-Air Part2 %
minokah@Naomis-MacBook-Air Part2 % java MyIndexFiles -docs documents/
Indexing to directory 'index'...
adding documents/RJ10.txt
adding documents/RJ1.txt
adding documents/RJ2.txt
adding documents/RJ3.txt
adding documents/RJ7.txt
adding documents/RJ6.txt
adding documents/RJ4.txt
adding documents/RJ5.txt
adding documents/RJ8.txt
adding documents/RJ9.txt
185 total milliseconds
```

```
minokah@Naomis-MacBook-Air Part2 % java MySearchFiles < testcases.txt
Enter query:
Searching for: fiery sword
1 total matching documents
1. documents/RJ3.txt
Press (q)uit or enter number to jump to a page.
Enter query:
Searching for: alas love
4 total matching documents
1. documents/RJ6.txt
2. documents/RJ7.txt
3. documents/RJ1.txt
4. documents/RJ10.txt
Press (q)uit or enter number to jump to a page.
Enter query:
Searching for: and
8 total matching documents
1. documents/RJ3.txt
2. documents/RJ10.txt
3. documents/RJ5.txt
4. documents/RJ2.txt
5. documents/RJ4.txt
6. documents/RJ8.txt
7. documents/RJ7.txt
8. documents/RJ1.txt
Press (q)uit or enter number to jump to a page.
Enter query:
Searching for: reckon
0 total matching documents
Enter query:
Searching for: q
0 total matching documents
Enter query:
Searching for: love
4 total matching documents
1. documents/RJ6.txt
2. documents/RJ7.txt
3. documents/RJ1.txt
4. documents/RJ10.txt
Press (q)uit or enter number to jump to a page.
Enter query:
Searching for: fair
3 total matching documents
1. documents/RJ9.txt
2. documents/RJ10.txt
3. documents/RJ1.txt
Press (q)uit or enter number to jump to a page.
Enter query:
minokah@Naomis-MacBook-Air Part2 %
```

2.2 Stemming

Add Porter stemmer to analyzer

```
88          /*
89          2.2 Stemming: Adding the Porter Stem Filter
90           */
91
92          return new TokenStreamComponents(r -> {
93            src.setMaxTokenLength(MyStandardAnalyzer.this.maxTokenLength);
94            src.setReader(r);
95          }, new PorterStemFilter(tok)); // CHANGE 1: Add Porter stem filter
96        }
97
```

Recompile and reindex. Results on next page

```
minokah@Naomis-MacBook-Air Part2 %
minokah@Naomis-MacBook-Air Part2 % javac *.java
minokah@Naomis-MacBook-Air Part2 % java MyIndexFiles -docs documents/
Indexing to directory 'index'...
adding documents/RJ10.txt
adding documents/RJ1.txt
adding documents/RJ2.txt
adding documents/RJ3.txt
adding documents/RJ7.txt
adding documents/RJ6.txt
adding documents/RJ4.txt
adding documents/RJ5.txt
adding documents/RJ8.txt
adding documents/RJ9.txt
302 total milliseconds
```

```
minokah@Naomis-MacBook-Air Part2 % java MySearchFiles < testcases.txt
Enter query:
Searching for: fieri sword
1 total matching documents
1. documents/RJ3.txt
Press (q)uit or enter number to jump to a page.
Enter query:
Searching for: ala love
5 total matching documents
1. documents/RJ6.txt
2. documents/RJ7.txt
3. documents/RJ8.txt
4. documents/RJ1.txt
5. documents/RJ10.txt
Press (q)uit or enter number to jump to a page.
Enter query:
Searching for: and
8 total matching documents
1. documents/RJ3.txt
2. documents/RJ10.txt
3. documents/RJ5.txt
4. documents/RJ2.txt
5. documents/RJ4.txt
6. documents/RJ8.txt
7. documents/RJ7.txt
8. documents/RJ1.txt
Press (q)uit or enter number to jump to a page.
Enter query:
Searching for: reckon
1 total matching documents
1. documents/RJ10.txt
Press (q)uit or enter number to jump to a page.
Enter query:
Searching for: love
5 total matching documents
1. documents/RJ6.txt
2. documents/RJ7.txt
3. documents/RJ8.txt
4. documents/RJ1.txt
5. documents/RJ10.txt
Press (q)uit or enter number to jump to a page.
Enter query:
Searching for: fair
3 total matching documents
1. documents/RJ9.txt
2. documents/RJ10.txt
3. documents/RJ1.txt
Press (q)uit or enter number to jump to a page.
Enter query:
minokah@Naomis-MacBook-Air Part2 %
```

2.3 Similarity

Add TF-IDF classic similarity to indexer and searcher

```
98      /*
99        2.3 Similarity: Add TF-IDF similarity over BM25
100     */
101     searcher.setSimilarity(new ClassicSimilarity()); // CHANGE 1 : Set it to TF-IDF
102
```

Once again, recompile and reindex. Results below

```
minokah@Naomis-MacBook-Air Part2 % javac *.java
minokah@Naomis-MacBook-Air Part2 % java MyIndexFiles -docs documents/
Indexing to directory 'index'...
adding documents/RJ10.txt
adding documents/RJ1.txt
adding documents/RJ2.txt
adding documents/RJ3.txt
adding documents/RJ7.txt
adding documents/RJ6.txt
adding documents/RJ4.txt
adding documents/RJ5.txt
adding documents/RJ8.txt
adding documents/RJ9.txt
291 total milliseconds
```

```
minokah@Naomis-MacBook-Air Part2 % java MySearchFiles < testcases.txt
Enter query:
Searching for: fieri sword
1 total matching documents
1. documents/RJ3.txt
Press (q)uit or enter number to jump to a page.
Enter query:
Searching for: ala love
5 total matching documents
1. documents/RJ6.txt
2. documents/RJ7.txt
3. documents/RJ8.txt
4. documents/RJ1.txt
5. documents/RJ10.txt
Press (q)uit or enter number to jump to a page.
Enter query:
Searching for: and
8 total matching documents
1. documents/RJ3.txt
2. documents/RJ5.txt
3. documents/RJ10.txt
4. documents/RJ2.txt
5. documents/RJ4.txt
6. documents/RJ8.txt
7. documents/RJ7.txt
8. documents/RJ1.txt
Press (q)uit or enter number to jump to a page.
Enter query:
Searching for: reckon
1 total matching documents
1. documents/RJ10.txt
Press (q)uit or enter number to jump to a page.
Enter query:
Searching for: love
5 total matching documents
1. documents/RJ6.txt
2. documents/RJ7.txt
3. documents/RJ8.txt
4. documents/RJ1.txt
5. documents/RJ10.txt
Press (q)uit or enter number to jump to a page.
Enter query:
Searching for: fair
3 total matching documents
1. documents/RJ9.txt
2. documents/RJ10.txt
3. documents/RJ1.txt
Press (q)uit or enter number to jump to a page.
Enter query:
minokah@Naomis-MacBook-Air Part2 %
```