Exploring Vector Data Models with Lucene

By the end of this activity, you will be able to:

1. Query text documents with Lucene
2. Perform weighted queries to see how rankings change
3. View the Term Frequency-Inverse Document Frequency (TF-IDF)

Step 1. Open a terminal shell. Open your local terminal shell and go to your *big-data-2/vector* directory

```
PS C:\Users\                \Desktop\coursera\big-data-2\vector>
```

Run *ls* to see the scripts:

```
    Directory: C:\Users\                \Desktop\coursera\big-data-2\vector


Mode                 LastWriteTime         Length Name
----                 -------------         ------ ----
-a----         2/29/2024   2:00 PM             35 runLuceneQuery.sh
-a----         2/29/2024   1:54 PM             35 runLuceneTFIDF.sh
```

Step 2. Start Docker. Make sure to start Docker by opening Docker Desktop.

Once you have started Docker, go back to your terminal and run *docker pull pramonettivega/lucene-coursera:latest* to pull a Docker image for this activity.

```
docker pull pramonettivega/lucene-coursera:latest
```

Step 3. Start your container. Run *docker run -it --rm pramonettivega/lucene-coursera ./runLuceneQuery.sh data* to start a container.

```
docker run -it --rm pramonettivega/lucene-coursera ./runLuceneQuery.sh data
```

You should see the Lucene's CLI after starting the container:

```
Index Location:data/index
Skipping (not csv/htm/html/xml/txt) : write.lock
Indexed : data/news3.csv
Indexed : data/news1.csv
Indexed : data/news2.csv

**********************************************************************
3 new documents added.
**********************************************************************
Enter query for Lucene (q=quit):
```

Notice that 3 documents have been added. These are textual data from the news.

Enter *voters* to query for that term:

```
Enter query for Lucene (q=quit):
voters
****************************************
Displaying 3 results.
****************************************
1) data/news1.csv score :0.043995064
2) data/news2.csv score :0.024887364
3) data/news3.csv score :0.011129968
```

The output shows the rankings and score for each of the three CSV files for the term *voters*. This shows that *news1.csv* is ranked first, *news2.csv* is second, and *news3.csv* is third.

Next, enter *delegates* to query for that term:

```
Enter query for Lucene (q=quit):
delegates
****************************************
Displaying 2 results.
****************************************
1) data/news2.csv score :0.041339863
2) data/news1.csv score :0.01953125
```

The output shows that *news2.csv* is ranked first, *news1.csv* is ranked second, and *news3.csv* is not shown since the term *delegates* does not appear in this document.

We can query for multiple terms by entering them together; enter *voters delegates* to query for both terms:

```
Enter query for Lucene (q=quit):
voters delegates
****************************************
Displaying 3 results.
****************************************
1) data/news2.csv score :0.04811
2) data/news1.csv score :0.041432917
3) data/news3.csv score :0.0032286723
```

The output shows that *news2.csv* is ranked first, *news1.csv* ranked second, and *news3.csv* ranked third.

Step 4. Perform weighted queries. We can perform a weighted query (or "boosting") to give one term more importance than the others. Enter *voters^5 delegates* to give the term *voters* a boost factor of 5:

```
Enter query for Lucene (q=quit):
voters^5 delegates
****************************************
Displaying 3 results.
****************************************
1) data/news1.csv score :0.047636837
2) data/news2.csv score :0.035135828
3) data/news3.csv score :0.005357802
```

The output shows that *news1.csv* is ranked first and *news2.csv* is ranked second. Note that these two rankings are reversed from when we performed the same query without boosting.

Enter *q* to quit this script and exit the container (the container will be automatically deleted).

Step 5. View the TF-IDF. Run *docker run -it --rm pramonettivega/lucene-coursera ./runLuceneTFIDF.sh data* to start a new container to see the TF-IDF for terms in the documents:

```
docker run -it --rm pramonettivega/lucene-coursera ./runLuceneTFIDF.sh data
```

```
[cloudera@quickstart vector]$ ./runLuceneTFIDF.sh  data
Index Location:data/index
Skipping (not csv,htm,html,xml,txt : write.lock
Indexed : data/news1.csv
Indexed : data/news2.csv
Indexed : data/news3.csv
**************************************************
3 new documents added.
**************************************************
```

Enter *voters* to see the TF-IDF for that term:

```
Enter a term to calculate TF-IDF (q=quit):
voters
Doc # 0: data/news1.csv   TF-IDF = 2.252547264099121
Doc # 1: data/news2.csv   TF-IDF = 1.5927913188934326
Doc # 2: data/news3.csv   TF-IDF = 0.712317943572998
```

Enter *delegates* to see the TF-IDF for that term:

```
Enter a term to calculate TF-IDF (q=quit):
delegates
Doc # 0: data/news1.csv   TF-IDF = 1.0
Doc # 1: data/news2.csv   TF-IDF = 2.6457512378692627
```

Enter *q* to quit this script and exit the container (the container will be automatically deleted).

Go to next item ✓ Completed