

## CA675 - Assignment 1

---

**Name :** Bhargav Anant Athavale

**Student ID :** 20210278

**Email :** [bhargav.athavale2@mail.dcu.ie](mailto:bhargav.athavale2@mail.dcu.ie)

---

**Git Repository Link -** <https://github.com/bhargavdcu/CA675Assignment1>

---

### Task 1 – Data Acquisition

Fetch a total of 200,000 records. The website restricts us to obtain a maximum of 50,000 records in each query. Therefore, run a total of 4 queries and combine the CSV files to obtain a single CSV file. Order the records by viewcount Obtain the dataset from the below link.

<https://data.stackexchange.com/stackoverflow/query/new>

#### Query 1

```
select top 50000 pos.*,usr.DisplayName from posts AS pos join users AS usr on  
pos.OwnerUserId=usr.Id ORDER BY pos.ViewCount DESC
```

#### Query 2

```
select top 50000 pos.*,usr.DisplayName from posts AS pos join users AS usr on  
pos.OwnerUserId=usr.Id  
and pos.ViewCount<124974 ORDER BY pos.ViewCount DESC
```

#### Query 3

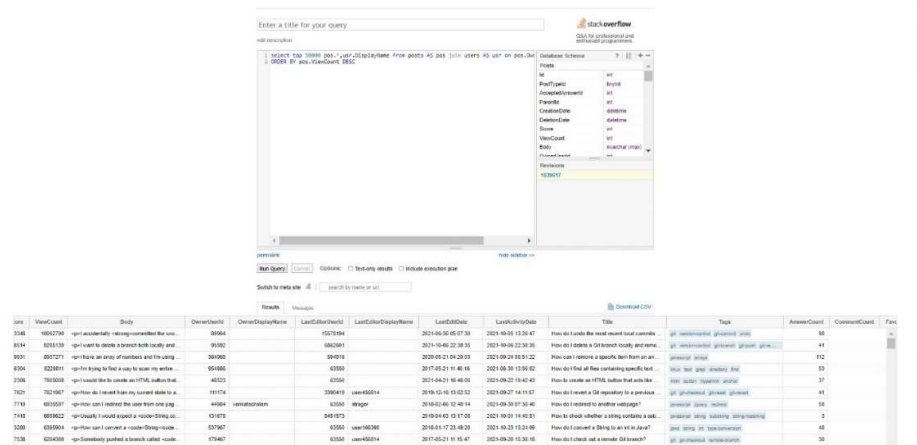
```
select top 50000 pos.*,usr.DisplayName from posts AS pos join users AS usr on  
pos.OwnerUserId=usr.Id  
and pos.ViewCount<73139 ORDER BY pos.ViewCount DESC
```

#### Query 4

```
select top 50000 pos.*,usr.DisplayName from posts AS pos join users AS usr on  
pos.OwnerUserId=usr.Id  
and pos.ViewCount<52110 ORDER BY pos.ViewCount DESC
```

The files were combined using the below command

```
cat FQ1.csv FQ2.csv FQ3.csv FQ4.csv > FQ.csv
```



## Task 2 – Extract, Transform and Load

The body, title columns of the data need to be cleaned before loading it into the hive database. **Python, Pig** was used for this purpose. **Pig was not able to clean the 'title' column efficiently.** Hence we used python for the purpose. After cleaning, we export the CSV. The python code for the same has been attached in the repository. Find below snippet

```
# In[13]:

#To Clean Body Column

stack_posts['Body'] = stack_posts['Body'].str.replace(r'\n', '', regex=True)

stack_posts['Body'] = stack_posts['Body'].str.replace('[^A-Za-z0-9 ]+', '', regex=True)

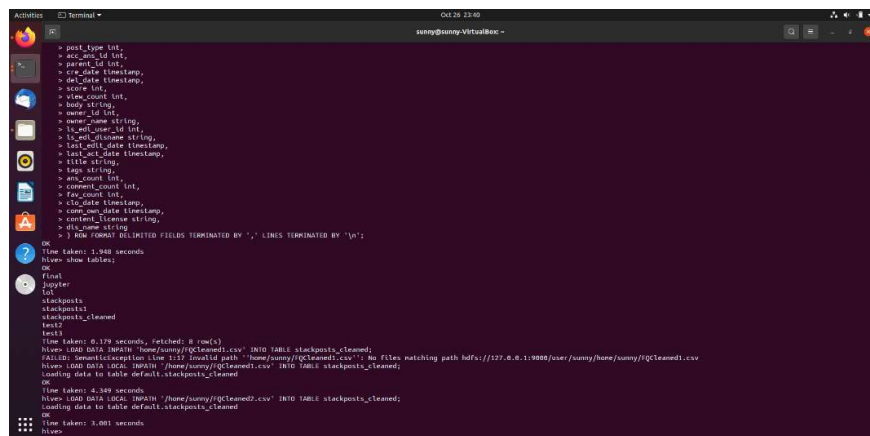
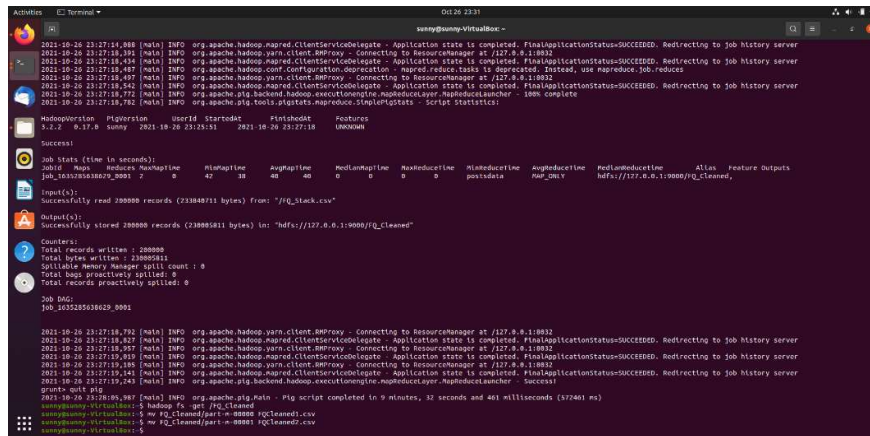
#To Clean Title Column

stack_posts['Title'] = stack_posts['Title'].str.replace(r'\n', '', regex=True)

stack_posts['Title'] = stack_posts['Title'].str.replace('[^A-Za-z0-9 ]+', '', regex=True)

Exporting to CSV

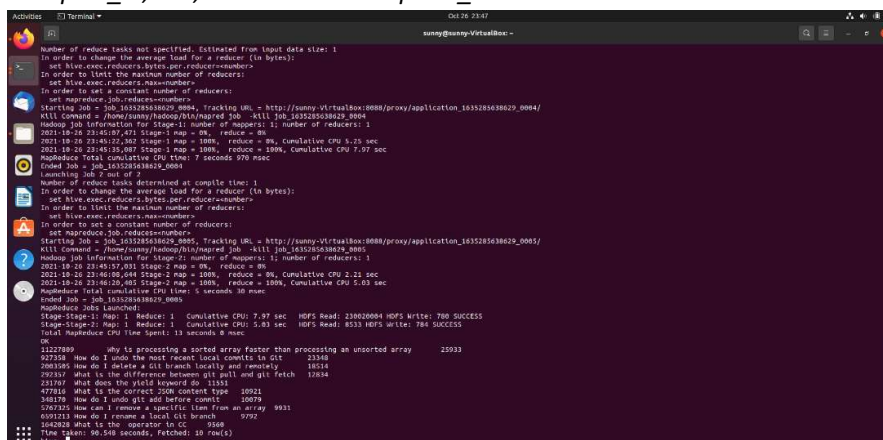
stack_posts.to_csv('FQ_Stack.csv', index=False)
```



### Task 3 – Queries

### 2.2.1. The top 10 posts by score

```
SELECT post_id,title,score FROM stackposts_cleaned SORT BY score DESC LIMIT 10
```



### 2.2.2. The top 10 users by post score

```
SELECT owner_id,sum(score) AS sum_score FROM stackposts_cleaned WHERE owner_id IS NOT NULL GROUP BY owner_id ORDER BY sum_score DESC LIMIT 10;
```

```

Time taken: 120.327 seconds, fetched: 10 row(s)
hive> SELECT owner_id, SUM(score) AS sum_score FROM stackposts_cleaned WHERE owner_id IS NOT NULL GROUP BY owner_id ORDER BY sum_score DESC LIMIT 10;
Query ID = sunny_2021102700051_35009478-dca2-4993-aea-27abb10b0cf
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reducers=<number>
Starting Job = job_1635285638629_0009, Tracking URL = http://sunny-VirtualBox:8080/proxy/application_1635285638629_0009/
Kill Command = /home/sunny/hadoop/bin/mapred job -kill job_1635285638629_0009
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-10-27 00:09:19.810 Stage-1 map = 0%, reduce = 0%
2021-10-27 00:09:31.815 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 8.4 sec
2021-10-27 00:09:36.270 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 14.61 sec
MapReduce Total cumulative CPU time: 14 seconds 610 msec
Ended Job = job_1635285638629_0009
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reducers=<number>
Starting Job = job_1635285638629_0009, Tracking URL = http://sunny-VirtualBox:8080/proxy/application_1635285638629_0009/
Kill Command = /home/sunny/hadoop/bin/mapred job -kill job_1635285638629_0009
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2021-10-27 00:10:15.766 Stage-2 map = 0%, reduce = 0%
2021-10-27 00:10:28.721 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 4.06 sec
2021-10-27 00:10:47.633 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 7.31 sec
MapReduce Total cumulative CPU time: 7 seconds 310 msec
Ended Job = job_1635285638629_0009
Hadoop Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 14.61 sec HDFS Read: 238027205 HDFS Write: 2715724 SUCCESS
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 7.31 sec HDFS Read: 2773400 HDFS Write: 325 SUCCESS
Total MapReduce CPU Time Spent: 21 seconds 920 msec
OK
87234 37672
4883 28812
9951 26878
6060 25944
89904 24024
51818 23783
49153 20203
179736 15603
95202 19479
63851 15362
Time taken: 112.447 seconds, fetched: 10 row(s)
hive> SELECT COUNT(DISTINCT(owner_id)) FROM stackposts_cleaned WHERE lower(body) like '%cloud%' OR lower(title) like '%cloud%' OR lower(tags) like '%cloud%';
Query ID = sunny_20211027002020_ee0f923-ee0d-4632-b098-118f4f0e222
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reducers=<number>
Starting Job = job_1635285638629_0010, Tracking URL = http://sunny-VirtualBox:8080/proxy/application_1635285638629_0010/
Kill Command = /home/sunny/hadoop/bin/mapred job -kill job_1635285638629_0010
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-10-27 00:18:44.482 Stage-1 map = 0%, reduce = 0%
2021-10-27 00:19:04.431 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 9.11 sec
2021-10-27 00:19:18.352 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 12.17 sec
MapReduce Total cumulative CPU time: 12 seconds 170 msec
Ended Job = job_1635285638629_0010
Hadoop Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 12.17 sec HDFS Read: 238019814 HDFS Write: 103 SUCCESS
Total MapReduce CPU Time Spent: 12 seconds 170 msec
OK
988
Time taken: 53.835 seconds, fetched: 1 row(s)
hive>

```

Figure 5: Top 10 Users By Post Score

### 2.2.3. The number of distinct users, who used the word “cloud” in one of their posts

`SELECT COUNT(DISTINCT(owner_id)) FROM stackposts_cleaned WHERE lower(body) like '%cloud%' OR lower(title) like '%cloud%' OR lower(tags) like '%cloud%';`

```

Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2021-10-27 00:10:15.766 Stage-2 map = 0%, reduce = 0%
2021-10-27 00:10:28.721 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 4.06 sec
2021-10-27 00:10:47.633 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 7.31 sec
MapReduce Total cumulative CPU time: 7 seconds 310 msec
Ended Job = job_1635285638629_0009
Hadoop Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 14.61 sec HDFS Read: 238027205 HDFS Write: 2715724 SUCCESS
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 7.31 sec HDFS Read: 2773400 HDFS Write: 325 SUCCESS
Total MapReduce CPU Time Spent: 21 seconds 920 msec
OK
87234 37672
4883 28812
9951 26878
6060 25944
89904 24024
51818 23783
49153 20203
179736 15603
95202 19479
63851 15362
Time taken: 112.447 seconds, fetched: 10 row(s)
hive> SELECT COUNT(DISTINCT(owner_id)) FROM stackposts_cleaned WHERE lower(body) like '%cloud%' OR lower(title) like '%cloud%' OR lower(tags) like '%cloud%';
Query ID = sunny_20211027002020_ee0f923-ee0d-4632-b098-118f4f0e222
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reducers=<number>
Starting Job = job_1635285638629_0010, Tracking URL = http://sunny-VirtualBox:8080/proxy/application_1635285638629_0010/
Kill Command = /home/sunny/hadoop/bin/mapred job -kill job_1635285638629_0010
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-10-27 00:18:44.482 Stage-1 map = 0%, reduce = 0%
2021-10-27 00:19:04.431 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 9.11 sec
2021-10-27 00:19:18.352 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 12.17 sec
MapReduce Total cumulative CPU time: 12 seconds 170 msec
Ended Job = job_1635285638629_0010
Hadoop Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 12.17 sec HDFS Read: 238019814 HDFS Write: 103 SUCCESS
Total MapReduce CPU Time Spent: 12 seconds 170 msec
OK
988
Time taken: 53.835 seconds, fetched: 1 row(s)
hive>

```

Figure 6: Number of distinct users, who used the word “cloud”

### Task 4 - Calculate the per-user TF-IDF of the top 10 terms for each of the top 10 users

TF – IDF is Term Frequency/Inverse Document Frequency. To calculate TF-IDF , first get the whole data for top ten users. Using 2<sup>nd</sup> query, get 10 users and store into table. Using mappers and reducer files (reference given below), get the output and store into table. Iterate the table to fetch TF/IDF of top 10 terms for each of the top 10 users. We have changed Hadoop commands from source according to our file path and jar version. Mapper and Reducer programs remain unchanged due to logic being consistent The mapreduce programs are from

<https://github.com/SatishUC15/TFIDF-HadoopMapReduce>

```
top_ten_users
Time taken: 2.287 seconds, Fetched: 11 row(s)
hive> insert overwrite local directory '/home/sunny/top_ten_data' row format delimited fields terminated by ',' select owner_id,body,title from stackposts_cleaned where owner_id in (select owner_id from top_ten_users);
Query ID = sunny_20211028025413_9c4e5eb-2b7c-4278-a754-304157dd5277
Total Jobs = 1
2021-10-28 02:54:39 Dump the side-table for tag: 1 with group count: 10 into file: file:/tmp/sunny/c3945aca-dad5-425f-bd95-4b0041030569/hive_2021-10-28_02-54-13_955_6807310979988769518-1/-local-10000/HashTable-Stage-3/MapJoin-mapfile01--hashtable
Execution completed successfully
MapReduce task succeeded
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = Job_1635358540846_0016, Tracking URL = http://sunny-VirtualBox:8088/proxy/application_1635358540846_0016/
Kill Command = /home/sunny/hadoop/bin/mapred job -kill job_1635358540846_0016
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 0
2021-10-28 02:55:04,084 Stage-3 map = 0%, reduce = 0%
2021-10-28 02:55:22,217 Stage-3 map = 100%, reduce = 0%, Cumulative CPU 6.24 sec
MapReduce Total cumulative CPU time: 6 seconds 240 msec
Ended Job = Job_1635358540846_0016
Moving data to local directory /home/sunny/top_ten_data
MapReduce Jobs Launched:
Stage:Stage-3: Map: 1 Cumulative CPU: 6.24 sec HDFS Read: 230017462 HDFS Write: 194503 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 240 msec
OK
Time taken: 7.789 seconds
```

Figure 7: Dump whole data of top 10 users

```
sunny@sunny-VirtualBox:~$ hadoop jar /home/sunny/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.2.2.jar -file /home/sunny/TFIDF/MapPhaseOne.py /home/sunny/TFIDF/ReducerPhaseOne.py -mapper "python MapperPhaseOne.py" -reducer "python ReducerPhaseOne.py" -input /tfidf_data/top_data -output /phase1;
2021-10-28 03:16:17,305 INFO org.apache.hadoop.mapred.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [/home/sunny/TFIDF/MapPhaseOne.py, /home/sunny/TFIDF/ReducerPhaseOne.py, /tmp/hadoop-unjar7277514311506825401/] [] /tmp/streamjob7852085881784185826.jar tmpDir=null
2021-10-28 03:16:40,485 INFO client.BMPProxy: Connecting to ResourceManager at /127.0.0.1:8032
2021-10-28 03:16:41,637 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/sunny/.staging/job_1635358540846_0018
2021-10-28 03:16:42,413 INFO mapred.FileInputFormat: Total input files to process : 1
2021-10-28 03:16:43,391 INFO mapreduce.JobSubmitter: number of splits:2
2021-10-28 03:16:43,987 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1635358540846_0018
2021-10-28 03:16:43,989 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-10-28 03:16:44,485 INFO conf.Configuration: resource-types.xml not found
2021-10-28 03:16:44,485 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-10-28 03:16:44,673 INFO impl.YarnClientImpl: Submitted application application_1635358540846_0018
2021-10-28 03:16:44,729 INFO mapreduce.Job: The url to track the job: http://sunny-VirtualBox:8088/proxy/application_1635358540846_0018/
2021-10-28 03:16:44,734 INFO mapreduce.Job: Running job: job_1635358540846_0018
2021-10-28 03:17:00,039 INFO mapreduce.Job: Job job_1635358540846_0018 running in uber mode : false
2021-10-28 03:17:00,044 INFO mapreduce.Job: map 0% reduce 0%
2021-10-28 03:17:59,714 INFO mapreduce.Job: map 100% reduce 0%
2021-10-28 03:18:12,903 INFO mapreduce.Job: map 100% reduce 100%
2021-10-28 03:18:13,923 INFO mapreduce.Job: Job job_1635358540846_0018 completed successfully
2021-10-28 03:18:14,132 INFO mapreduce.Job: Counters: 54
File System Counters
FILE: Number of bytes read=649081
FILE: Number of bytes written=2012452
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=198785
HDFS: Number of bytes written=129242
HDFS: Number of read operations=11
HDFS: Number of large read operations=0
HDFS: Number of write operations=3
HDFS: Number of bytes read erasure-coded=0
Job Counters
Launched map tasks=2
Launched reduce tasks=1
Data-Local map tasks=2
Total time spent by all maps in occupied slots (ms)=114358
Total time spent by all reduces in occupied slots (ms)=9531
Total time spent by all map tasks (ms)=114358
Total time spent by all reduce tasks (ms)=9531
Total vcore-millisecods taken by all map tasks=114358
Total vcore-millisecods taken by all reduce tasks=9531
Total megabyte-millisecods taken by all map tasks=11702592
Total megabyte-millisecods taken by all reduce tasks=9752744
Map-Reduce Framework
Map input records=425
```

Figure 8: Mapper and Reducer

```
sunny@sunny-VirtualBox:~$ hadoop jar /home/sunny/hadoop/bin/mapred job -kill job_1635389442972_0004
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-10-28 04:24:02,512 Stage-1 map = 0%, reduce = 0%
2021-10-28 04:24:16,464 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.72 sec
2021-10-28 04:24:30,772 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 8.05 sec
MapReduce Total cumulative CPU time: 8 seconds 50 msec
Ended Job = Job_1635389442972_0004
MapReduce Jobs Launched:
Stage:Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 8.05 sec HDFS Read: 272020 HDFS Write: 3744 SUCCESS
Total MapReduce CPU Time Spent: 8 seconds 50 msec
OK
1 0 179736 0.5790645
2 p 179736 0.02507928
3 I 179736 0.014817287
4 code 179736 0.01346571
5 pre 179736 0.009811447
6 a 179736 0.009268804
7 do 179736 0.008565549
8 the 179736 0.008359753
9 to 179736 0.008299578
10 how 179736 0.006618833
1 p 4883 0.059844155
2 a 4883 0.03787013
3 the 4883 0.029922077
4 showImage 4883 0.020851948
5 xargs 4883 0.023376623
6 to 4883 0.022099092
7 transaction 4883 0.018701298
8 in 4883 0.018701298
9 branch 4883 0.015384416
10 comprehensions 4883 0.014025974
1 p 49153 0.054099902
2 code 49153 0.040374264
3 the 49153 0.032483503
4 to 49153 0.02813632
5 a 49153 0.023547566
6 I 49153 0.02113243
7 validation 49153 0.015698377
8 hibernate 49153 0.014498089
9 pre 49153 0.01249296
10 in 49153 0.013102485
1 p 51816 0.059636362
2 code 51816 0.045454547
3 to 51816 0.026727272
4 I 51816 0.026
5 the 51816 0.024363637
6 a 51816 0.024181819
7 WPF 51816 0.023636363
8 in 51816 0.023636363
```

Figure 9:TFIDF

## **Appendix**

### **Completion of Tasks**

#### **Tasks Completed**

1. Data Acquisition
2. Data ETL
  - i. Top 10 posts by score
  - ii. Top 10 users by post score
  - iii. The number of distinct users, who used the word “cloud” in one of their posts
3. Calculate the per-user TF-IDF of the top 10 terms for each of the top 10 users