In this document per-user TF-IDF of the top 10 terms for each of the top 10 users is being caluculated.

Steps Prior to calculate TF-IDF:

1. Dataset Final_PostsData.csv is cleaned using basic sed command to achieve a clean csv file. Using the sed command new lines with spaces have been replaced.

   Command: sed ':a;N;$!ba;s/\n//g' Final_PostData.csv > Clean_Final.csv

```
shradha_shivani2@cluster-hadoop-m:~$ sudo sed ':a;N;$!ba;s/\n//g' Final_PostData.csv > Clean_Final.csv
shradha_shivani2@cluster-hadoop-m:~$ ls -lrt
total 756088
-rw-rw-r--  1 shradha_shivani2 shradha_shivani2   2727761 Jul  3  2020 apache-ivy-2.5.0-bin.tar.gz.1
-rw-rw-r--  1 shradha_shivani2 shradha_shivani2 278813748 Jul  3  2020 apache-hive-3.1.2-bin.tar.gz
drwxrwxr-x 10 shradha_shivani2 shradha_shivani2      4096 Oct 25 18:36 apache-hive-3.1.2-bin
-rw-rw-r--  1 shradha_shivani2 shradha_shivani2     28469 Oct 25 21:33 apache-ivy-2.5.0-bin.tar.gz
drwxrwxr-x  3 shradha_shivani2 shradha_shivani2      4096 Oct 25 23:06 Pig
-rw-r--r--  1 root             root                  2615 Oct 25 23:15 pig_1635203662479.log
drwxr-xr-x  2 root             root                  4096 Oct 25 23:24 combined
-rw-r--r--  1 root             root                  2617 Oct 25 23:24 pig_1635203850422.log
-rwxr-xr-x  1 shradha_shivani2 shradha_shivani2    223258 Oct 26 00:00 hiveResults.csv
drwxrwxr-x  3 shradha_shivani2 shradha_shivani2      4096 Oct 26 15:14 MapReduce
-rw-rw-r--  1 shradha_shivani2 shradha_shivani2 248559656 Oct 26 15:36 Final_PostData.csv
-rw-rw-r--  1 shradha_shivani2 shradha_shivani2 243835097 Oct 26 15:37 Clean_Final.csv
shradha_shivani2@cluster-hadoop-m:~$
```

2. Basic cleaned file 'Clean_Final.csv' is moved to '/user/CA1' in HDFS
   Command: hadoop fs -put Clean_Final.csv /user/CA1

```
shradha_shivani2@cluster-hadoop-m:~$ hadoop fs -put Clean_Final.csv /user/CA1/
shradha_shivani2@cluster-hadoop-m:~$ hadoop fs -ls
ls: `.': No such file or directory
shradha_shivani2@cluster-hadoop-m:~$ hadoop fs -ls /user/CA1
Found 4 items
-rw-r--r--   2 shradha_shivani2 hadoop  243835097 2021-10-26 15:48 /user/CA1/Clean_Final.csv
-rw-r--r--   2 shradha_shivani2 hadoop  248559656 2021-10-25 23:16 /user/CA1/Final_PostData.csv
drwxr-xr-x   - shradha_shivani2 hadoop          0 2021-10-26 15:43 /user/CA1/TFIDF
drwxr-xr-x   - root             hadoop          0 2021-10-25 23:28 /user/CA1/combined
shradha_shivani2@cluster-hadoop-m:~$
```

3. Logged in to pig terminal using HCatalog. The Command used is pig -useHCatalog.

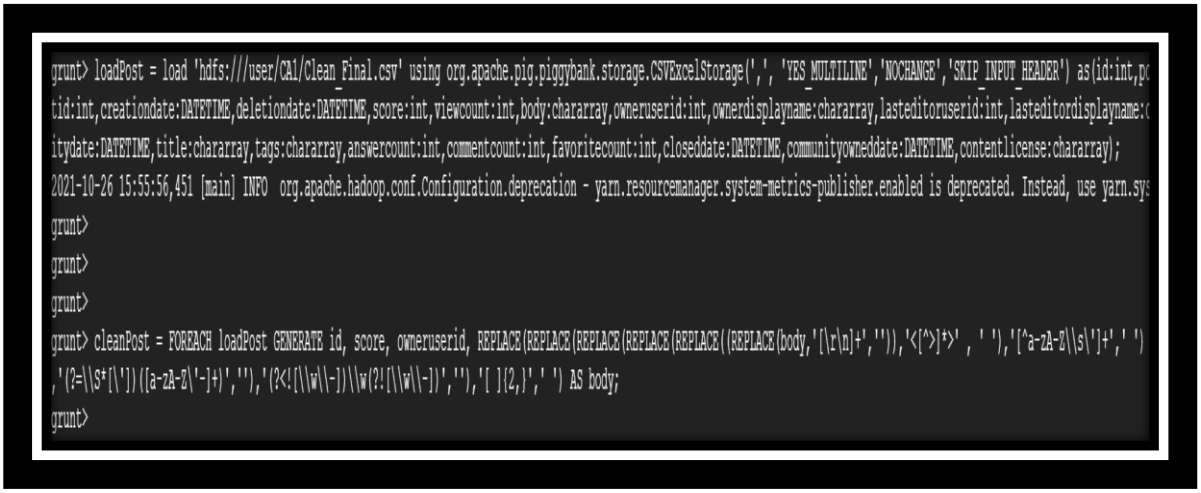
```
shradha_shivani2@cluster-hadoop-m:~$ pig -useHCatalog
ls: cannot access '/usr/lib/hive/lib/slf4j-api-*.jar': No such file or directory
ls: cannot access '/usr/lib/hive/lib/hive-hbase-handler-*.jar': No such file or directory
ls: cannot access '/usr/lib/hive-hcatalog/lib/*hbase-storage-handler-*.jar': No such file or directory
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
2021-10-26 15:53:42,219 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2021-10-26 15:53:42,220 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2021-10-26 15:53:42,221 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2021-10-26 15:53:42,268 [main] INFO  org.apache.pig.Main - Apache Pig version 0.18.0-SNAPSHOT (r: unknown) compiled Dec 21 1969, 14:26:39
2021-10-26 15:53:42,268 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/shradha_shivani2/pig_1635263622266.log
2021-10-26 15:53:42,297 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/shradha_shivani2/.pigbootup not found
2021-10-26 15:53:42,658 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2021-10-26 15:53:42,658 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://cluster-hadoop-m
2021-10-26 15:53:43,882 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-575f959b-ef5b-4523-9b71-e5dc25127fa3
2021-10-26 15:53:44,079 [main] INFO  org.apache.hadoop.yarn.client.api.impl.TimelineClientImpl - Timeline service address: cluster-hadoop-m:8188
2021-10-26 15:53:44,400 [main] INFO  org.apache.pig.backend.hadoop.PigATSClient - Created ATS Hook
2021-10-26 15:53:44,427 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
grunt>
```

4. Using Pig, the data present in Clean_Final.csv is further cleaned. New lines, tab or carriage return characters, single characters, special symbols etc are replaced with space in 'Body' column. Additionally, the dataset is filtered for NOT NULL OWNERUSERID tuples.

Command:

```
grunt> loadPost = load 'hdfs:///user/CA1/Clean_Final.csv' using
org.apache.pig.piggybank.storage.CSVExcelStorage(',',
'YES_MULTILINE','NOCHANGE','SKIP_INPUT_HEADER')
as(id:int,posttypeid:int,acceptedanswerid:int,
parentid:int,creationdate:DATETIME,deletiondate:DATETIME,score:int,viewcount:int,body
:chararray,owneruserid:int,ownerdisplayname:chararray,lasteditoruserid:int,lasteditordisplay
name:chararray,lasteditdate:DATETIME,lastactivitydate:DATETIME,title:chararray,tags:cha
rarray,answercount:int,commentcount:int,favoritecount:int,closeddate:DATETIME,commun
ityowneddate:DATETIME,contentlicense:chararray);
2021-10-26 15:55:56,451 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation -
yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use
yarn.system-metrics-publisher.enabled
grunt>

grunt> cleanPost = FOREACH loadPost GENERATE id, score, owneruserid,
REPLACE(REPLACE(REPLACE(REPLACE(REPLACE((REPLACE(body,'[\r\n]+',")),'<[
^>]*>' , ' '),'[^a-zA-Z\\s\']+',' '),'(?=\\S*[\'])([a-zA-
Z\'-]+)',''),'(?<![\\w\\-])\\w(?![\\w\\-])',''),'[ ]{2,}',' ') AS body;
```



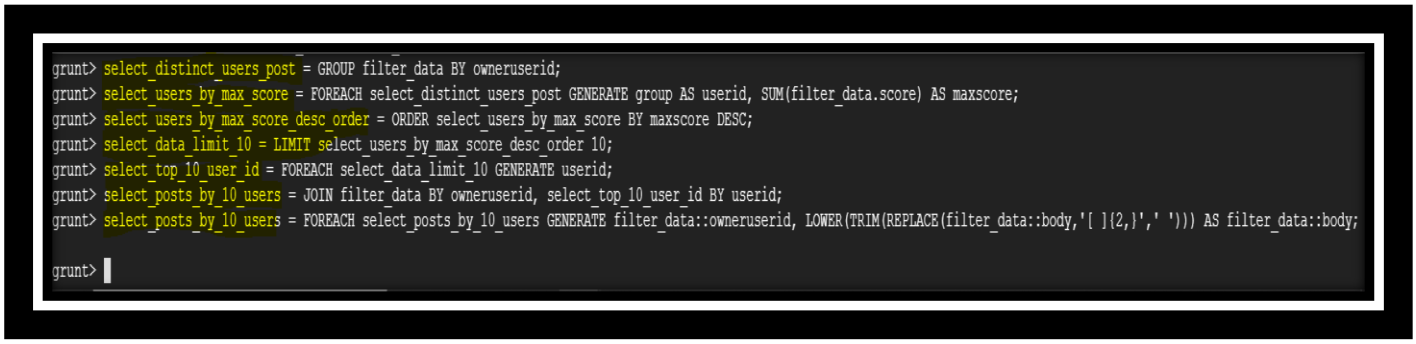
grunt> filter_data = FILTER post_data BY (owneruserid is not null);

Computation of TF-IDF:

1. In this section, the users are grouped together, their total score of their various posts is calculated and arranged in descending order . The result is just limited to top 10.So, in other terms top 10 posts of top 10 users are selected as per their total score.

   Command:

   grunt> select_distinct_users_post = GROUP filter_data BY owneruserid;
   grunt> select_users_by_max_score = FOREACH select_distinct_users_post GENERATE group AS userid, SUM(filter_data.score) AS maxscore;
   grunt> select_users_by_max_score_desc_order = ORDER select_users_by_max_score BY maxscore DESC;
   grunt> select_data_limit_10 = LIMIT select_users_by_max_score_desc_order 10;
   grunt> select_top_10_user_id = FOREACH select_data_limit_10 GENERATE userid;
   grunt> select_posts_by_10_users = JOIN filter_data BY owneruserid, select_top_10_user_id BY userid;
   grunt> select_posts_by_10_users = FOREACH select_posts_by_10_users GENERATE filter_data::owneruserid, LOWER(TRIM(REPLACE(filter_data::body,'[ ]{2,}',' '))) AS filter_data::body

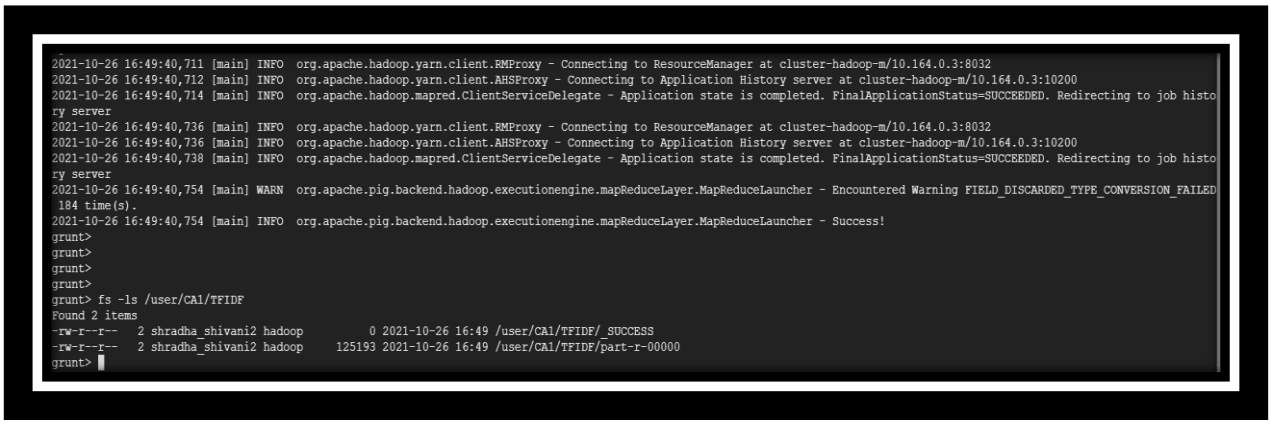


2. Next, the resultant set is stored in a folder named 'TFIDF' under '/user/CA1/'

   Command:

   Grunt>STORE select_posts_by_10_users INTO '/user/CA1/TFIDF' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','NOCHANGE','SKIP_OUTPUT_HEADER');

3. Successful logs are obtained for data storing in TFIDF location and the files part-r-00000 and _SUCCESS log can be seen under this location.



```
2021-10-26 16:49:40,711 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster-hadoop-m/10.164.0.3:8032
2021-10-26 16:49:40,712 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster-hadoop-m/10.164.0.3:10200
2021-10-26 16:49:40,714 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job histo
ry server
2021-10-26 16:49:40,736 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster-hadoop-m/10.164.0.3:8032
2021-10-26 16:49:40,736 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster-hadoop-m/10.164.0.3:10200
2021-10-26 16:49:40,738 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job histo
ry server
2021-10-26 16:49:40,754 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered Warning FIELD_DISCARDED_TYPE_CONVERSION_FAILED
 184 time(s).
2021-10-26 16:49:40,754 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt>
grunt>
grunt>
grunt>
grunt> fs -ls /user/CA1/TFIDF
Found 2 items
-rw-r--r--   2 shradha_shivani2 hadoop          0 2021-10-26 16:49 /user/CA1/TFIDF/_SUCCESS
-rw-r--r--   2 shradha_shivani2 hadoop     125193 2021-10-26 16:49 /user/CA1/TFIDF/part-r-00000
grunt>
```

Implementing TFIDF:

In this section, TFIDF is implemented in Hadoop using Python scripts. Altogether there are 4 mapper and 3 reducer python program files. The implementation takes place in four phases. The first phase uses three mappers and three reducers. The last phase uses the fourth mapper to generate a single file with 10 users word list and its TF-IDF value. The output of one phase is fed as an input for the next phase.

The mapreducer.sh script consists of below commands:

hadoop jar hadoop-streaming-3.2.2.jar -files /home/shradha_shivani2/MapReduce/mapper1.py,/home/shradha_shivani2/MapReduce/reducer1.py -mapper 'python3 mapper1.py' -reducer 'python3 reducer1.py' -input hdfs:///user/CA1/TFIDF/part-r-00000 -output hdfs:///data/output1

hadoop jar hadoop-streaming-3.2.2.jar -files /home/shradha_shivani2/MapReduce/mapper2.py,/home/shradha_shivani2/MapReduce/reducer2.py -mapper 'python3 mapper2.py' -reducer 'python3 reducer2.py' -input hdfs:///data/output1/part-0000* -output hdfs:///data/output2

hadoop jar hadoop-streaming-3.2.2.jar -files /home/shradha_shivani2/MapReduce/mapper3.py,/home/shradha_shivani2/MapReduce/reducer3.py -mapper 'python3 mapper3.py' -reducer 'python3 reducer3.py' -input hdfs:///data/output2/part-0000* -output hdfs:///data/output3

hadoop jar hadoop-streaming-3.2.2.jar -files /home/shradha_shivani2/MapReduce/mapper4.py -mapper 'python3 mapper4.py' -input hdfs:///data/output3/part-0000* -output hdfs:///data/output4

hadoop fs -getmerge hdfs:///data/output4/part-0000* /home/shradha_shivani2/MapReduce/tfidResults/result.txt

hadoop fs -rm -r /data/output*

```
shradha_shivani2@cluster-hadoop-m:~/MapReduce$ bash mapreduce.sh
packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.2.2.jar] /tmp/streamjob7763702977836067497.jar tmpDir=null
2021-10-26 17:04:25,766 INFO client.RMProxy: Connecting to ResourceManager at cluster-hadoop-m/10.164.0.3:8032
2021-10-26 17:04:26,015 INFO client.AHSProxy: Connecting to Application History server at cluster-hadoop-m/10.164.0.3:10200
2021-10-26 17:04:26,512 INFO client.RMProxy: Connecting to ResourceManager at cluster-hadoop-m/10.164.0.3:8032
2021-10-26 17:04:26,513 INFO client.AHSProxy: Connecting to Application History server at cluster-hadoop-m/10.164.0.3:10200
2021-10-26 17:04:26,692 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/shradha_shivani2/.staging/job_1635259661945_0012
2021-10-26 17:04:27,063 INFO mapred.FileInputFormat: Total input files to process : 1
2021-10-26 17:04:27,127 INFO mapreduce.JobSubmitter: number of splits:21
2021-10-26 17:04:27,284 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1635259661945_0012
2021-10-26 17:04:27,286 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-10-26 17:04:27,570 INFO conf.Configuration: resource-types.xml not found
2021-10-26 17:04:27,570 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-10-26 17:04:27,635 INFO impl.YarnClientImpl: Submitted application application_1635259661945_0012
2021-10-26 17:04:27,685 INFO mapreduce.Job: The url to track the job: http://cluster-hadoop-m:8088/proxy/application_1635259661945_0012/
2021-10-26 17:04:27,687 INFO mapreduce.Job: Running job: job_1635259661945_0012
```

4. Running above commands created tfidResults directory under /home/shradha_shivani2/Mapreduce and generated result.txt under tfidResults.

```
shradha_shivani2@cluster-hadoop-m:~/MapReduce$ ls -lrt
total 220
-rw-rw-r-- 1 shradha_shivani2 shradha_shivani2   1559 Oct 26 00:02 mapper1.py
-rw-rw-r-- 1 shradha_shivani2 shradha_shivani2    587 Oct 26 00:02 mapper2.py
-rw-rw-r-- 1 shradha_shivani2 shradha_shivani2    322 Oct 26 00:02 mapper3.py
-rw-rw-r-- 1 shradha_shivani2 shradha_shivani2    412 Oct 26 00:03 mapper4.py
-rw-rw-r-- 1 shradha_shivani2 shradha_shivani2   1033 Oct 26 00:03 reducer1.py
-rw-rw-r-- 1 shradha_shivani2 shradha_shivani2    769 Oct 26 00:03 reducer2.py
-rw-rw-r-- 1 shradha_shivani2 shradha_shivani2    739 Oct 26 00:03 reducer3.py
-rw-rw-r-- 1 shradha_shivani2 shradha_shivani2    345 Oct 26 00:08 splitTopUsers.py
-rw-rw-r-- 1 shradha_shivani2 shradha_shivani2    556 Oct 26 00:08 sortResults.py
-rw-r--r-- 1 shradha_shivani2 shradha_shivani2 176502 Oct 26 00:16 hadoop-streaming-3.2.2.jar
drwxr-xr-x 2 shradha_shivani2 shradha_shivani2   4096 Oct 26 00:21 tfidResults
-rw-rw-r-- 1 shradha_shivani2 shradha_shivani2   1121 Oct 26 17:01 mapreduce.sh
shradha_shivani2@cluster-hadoop-m:~/MapReduce$
shradha_shivani2@cluster-hadoop-m:~/MapReduce$
shradha_shivani2@cluster-hadoop-m:~/MapReduce$
shradha_shivani2@cluster-hadoop-m:~/MapReduce$ ls -lrt tfidResults/
total 264
-rw-r--r-- 1 shradha_shivani2 shradha_shivani2 266952 Oct 26 17:07 result.txt
shradha_shivani2@cluster-hadoop-m:~/MapReduce$
```

5. Next using the sorting script, the results of result.txt were processed and final TFIDF resultant was displayed and captured in TFIDF_Results.txt

```
shradha_shivani2@cluster-hadoop-m:~/MapReduce$ python3 sortResults.py > TFIDF_Result.txt
shradha_shivani2@cluster-hadoop-m:~/MapReduce$ ls -lrt
total 224
-rw-rw-r-- 1 shradha_shivani2 shradha_shivani2   1559 Oct 26 00:02 mapper1.py
-rw-rw-r-- 1 shradha_shivani2 shradha_shivani2    587 Oct 26 00:02 mapper2.py
-rw-rw-r-- 1 shradha_shivani2 shradha_shivani2    322 Oct 26 00:02 mapper3.py
-rw-rw-r-- 1 shradha_shivani2 shradha_shivani2    412 Oct 26 00:03 mapper4.py
-rw-rw-r-- 1 shradha_shivani2 shradha_shivani2   1033 Oct 26 00:03 reducer1.py
-rw-rw-r-- 1 shradha_shivani2 shradha_shivani2    769 Oct 26 00:03 reducer2.py
-rw-rw-r-- 1 shradha_shivani2 shradha_shivani2    739 Oct 26 00:03 reducer3.py
-rw-rw-r-- 1 shradha_shivani2 shradha_shivani2    345 Oct 26 00:08 splitTopUsers.py
-rw-rw-r-- 1 shradha_shivani2 shradha_shivani2    556 Oct 26 00:08 sortResults.py
-rw-r--r-- 1 shradha_shivani2 shradha_shivani2 176502 Oct 26 00:16 hadoop-streaming-3.2.2.jar
drwxr-xr-x 2 shradha_shivani2 shradha_shivani2   4096 Oct 26 00:21 tfidResults
-rw-rw-r-- 1 shradha_shivani2 shradha_shivani2   1121 Oct 26 17:01 mapreduce.sh
-rw-rw-r-- 1 shradha_shivani2 shradha_shivani2    344 Oct 26 17:16 TFIDF_Result.txt
shradha_shivani2@cluster-hadoop-m:~/MapReduce$ cat TFIDF_Result.txt
result Owner User ID

           word  tfidf_score
610        grab     0.019362
1175      lower     0.014061
2787       wall     0.007683
890        void     0.007453
590      figure     0.006838
1740      pyfor     0.006685
408         use     0.006300
2262      store     0.006300
1837  untracked     0.005916
3071    network     0.004687
```