Project title: Spam Detection System
Your name: Sarath Mukundan Adavakkat
Student ID number: 23267908
Email address: sarath.mukundanadavakkat2@mail.dcu.ie
Program of study: MCM1
Module code: CA675
Date of submission: 16/11/2023

I understand that the University regards breaches of academic integrity and plagiarism as grave and serious. I have read and understood the DCU Academic Integrity and Plagiarism Policy. I accept the penalties that may be imposed should I engage in practice or practices that breach this policy. I have identified and included the source of all facts, ideas, opinions, viewpoints of others in the assignment references. Direct quotations, paraphrasing, discussion of ideas from books, journal articles, internet sources, module text, or any other source whatsoever are acknowledged and the sources cited are identified in the assignment references. Any use of generative AI or search will be described in a one-page appendix including prompt queries.

I declare that this material, which I now submit for assessment, is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work. By signing this form or by submitting this material online I confirm that this assignment, or any part of it, has not been previously submitted by me or any other person for assessment on this or any other course of study. By signing this form or by submitting material for assessment online I confirm that I have read and understood DCU Academic Integrity and Plagiarism Policy.

Signed Name: Sarath Mukundan Adavakkat

Date: 16/11/2023

Table of Contents

Tables of Figures:

Link to the Git repository

https://gitlab.computing.dcu.ie/mukunds2/cloud_assignment_1

Acquired Dataset

Spam and ham-labelled video comments from the YouTube dataset from a public repository Kaggle are used. Spam and Ham are the class labels; ham is denoted by a "0" and spam reviews by a "1".
Link to the Dataset: https://www.kaggle.com/datasets/lakshmi25npathi/images/data

Task 1: Cloud Infrastructure Setup in GCP

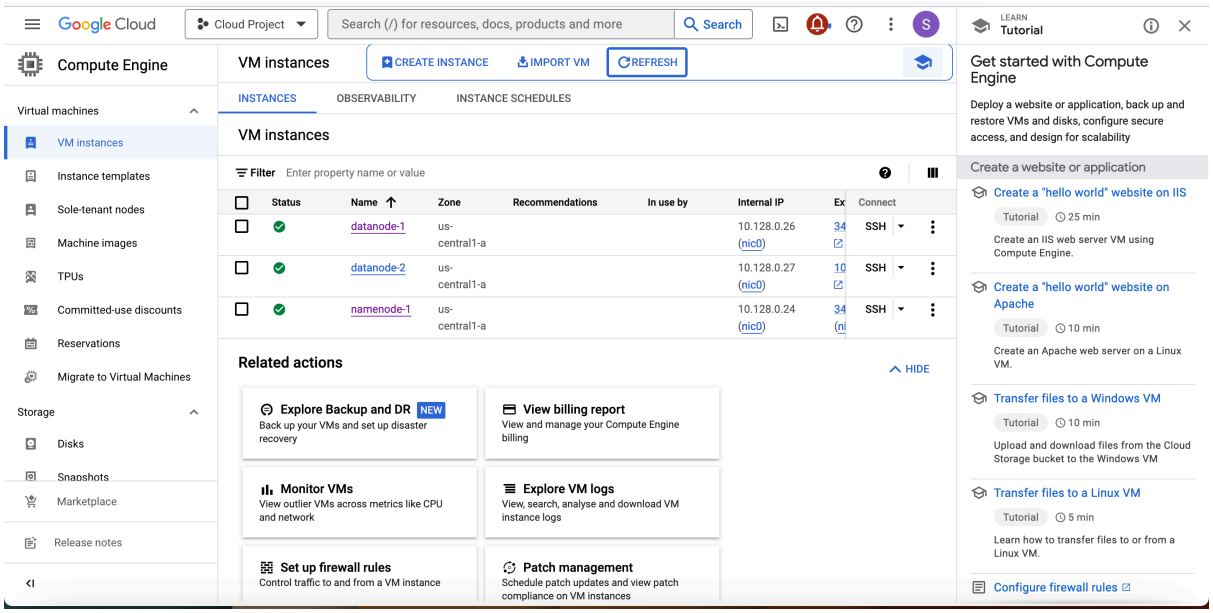

Figure 1: Multinode cluster

Here, a Multi-node cluster is created using VM instances, with namenode-1 as master and datanode-1, datanode-2 as slave nodes respectively.

Task 1.1: Create a Hadoop Cluster and install Java and Hadoop

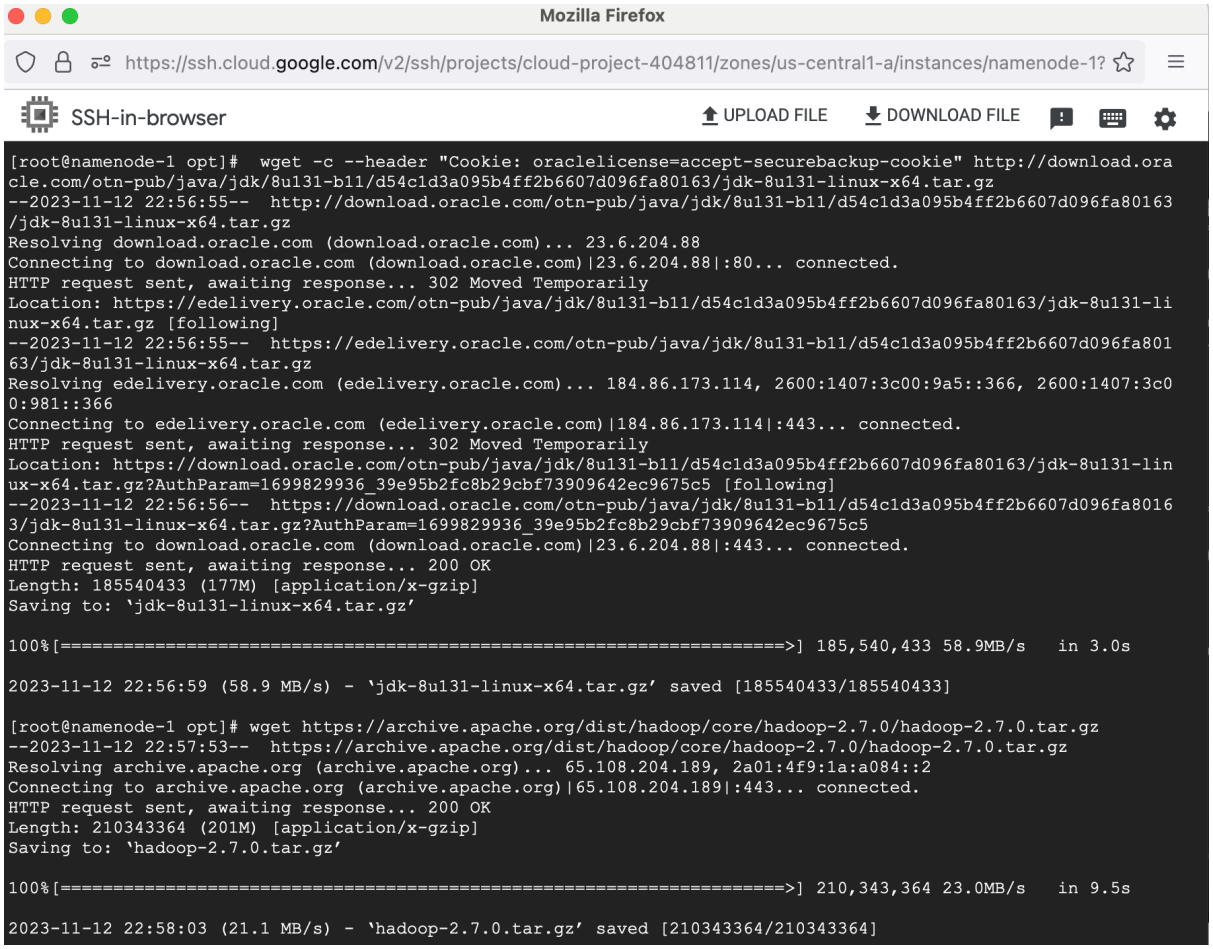

Figure 2: Installation of Java

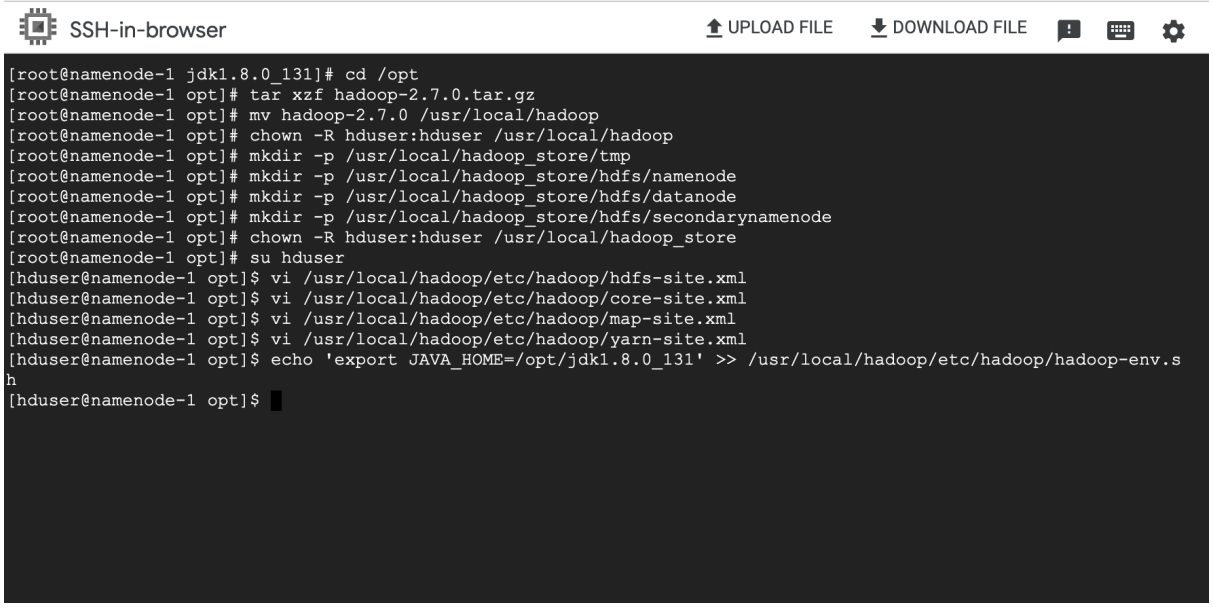

Figure 3: Installation of Hadoop

Task 1.2: Install Mapreduce.Hive, Pig in the cluster

snippets of code:

hduser@namenode-1 $ wget -b http://www-eu.apache.org/dist/hive/hive-1.2.2/apache-hive-1.2.2-bin.tar.gz

hduser@namenode-1 $ tar -xvf apache-hive-1.2.2-bin.tar.gz

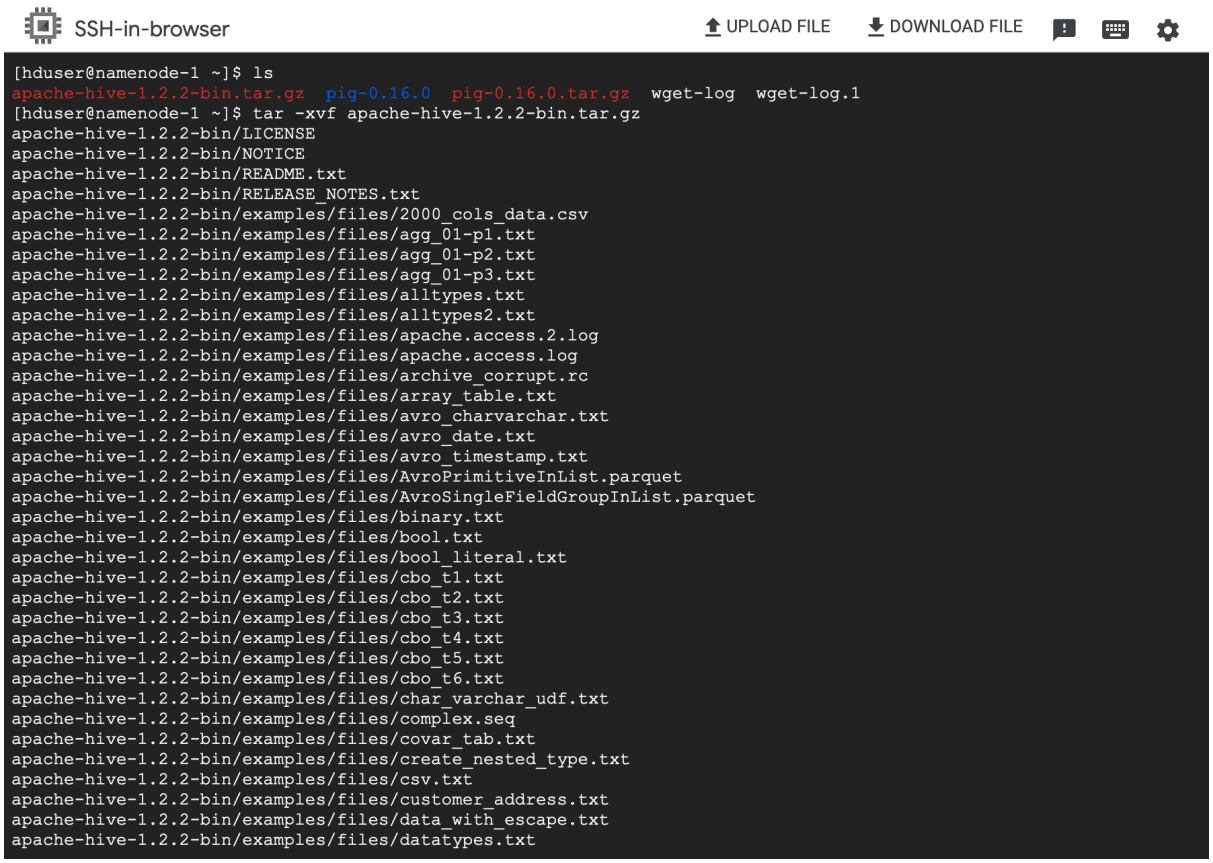

Figure 4: Installation of hive

```
apache-hive-1.2.2-bin/hcatalog/share/webhcat/svr/lib/stax-api-1.0-2.jar
apache-hive-1.2.2-bin/hcatalog/share/webhcat/svr/lib/jackson-core-asl-1.9.2.jar
apache-hive-1.2.2-bin/hcatalog/share/webhcat/svr/lib/jackson-jaxrs-1.9.2.jar
apache-hive-1.2.2-bin/hcatalog/share/webhcat/svr/lib/jackson-xc-1.9.2.jar
apache-hive-1.2.2-bin/hcatalog/share/webhcat/svr/lib/jersey-core-1.14.jar
apache-hive-1.2.2-bin/hcatalog/share/webhcat/svr/lib/jersey-server-1.14.jar
apache-hive-1.2.2-bin/hcatalog/share/webhcat/svr/lib/asm-3.1.jar
apache-hive-1.2.2-bin/hcatalog/share/webhcat/svr/lib/hive-webhcat-1.2.2.jar
apache-hive-1.2.2-bin/hcatalog/share/webhcat/svr/lib/jersey-servlet-1.14.jar
apache-hive-1.2.2-bin/hcatalog/share/webhcat/svr/lib/wadl-resourcedoc-doclet-1.4.jar
apache-hive-1.2.2-bin/hcatalog/share/webhcat/svr/lib/xercesImpl-2.9.1.jar
apache-hive-1.2.2-bin/hcatalog/share/webhcat/svr/lib/xml-apis-1.3.04.jar
apache-hive-1.2.2-bin/hcatalog/share/webhcat/svr/lib/commons-exec-1.1.jar
apache-hive-1.2.2-bin/hcatalog/share/webhcat/svr/lib/jul-to-slf4j-1.7.5.jar
apache-hive-1.2.2-bin/hcatalog/share/webhcat/java-client/hive-webhcat-java-client-1.2.2.jar
apache-hive-1.2.2-bin/conf/hive-log4j.properties.template
apache-hive-1.2.2-bin/conf/hive-exec-log4j.properties.template
apache-hive-1.2.2-bin/conf/beeline-log4j.properties.template
apache-hive-1.2.2-bin/hcatalog/share/doc/hcatalog/README.txt
[hduser@namenode-1 ~]$ vi .bash_profile
[hduser@namenode-1 ~]$ source .bash_profile
[hduser@namenode-1 ~]$ ls
apache-hive-1.2.2-bin  apache-hive-1.2.2-bin.tar.gz  pig-0.16.0  pig-0.16.0.tar.gz  wget-log  wget-log.1
[hduser@namenode-1 ~]$ sudo apache-hive-1.2.2-bin /opt/hive
[sudo] password for hduser:
hduser is not in the sudoers file.  This incident will be reported.
[hduser@namenode-1 ~]$ sudo mv apache-hive-1.2.2-bin /opt/hive
[sudo] password for hduser:
hduser is not in the sudoers file.  This incident will be reported.
[hduser@namenode-1 ~]$ vi .bashrc
[hduser@namenode-1 ~]$ vi .bashrc
[hduser@namenode-1 ~]$ source .bashrc
[hduser@namenode-1 ~]$ hive --version
Hive 1.2.2
Subversion git://vgumashta.local/Users/vgumashta/Documents/workspace/hive-git -r 395368fc6478c7e2a1e84a5a2a8aac
45e4399a9e
Compiled by vgumashta on Sun Apr 2 13:12:26 PDT 2017
From source with checksum bd47834e727562aab36c8282f8161030
[hduser@namenode-1 ~]$
```

Figure 5: installation of Hive

Installation of pig:

snippets of code:

hduser@namenode-1 $ wget http://www-eu.apache.org/dist/pig/pig-0.16.0/pig-0.16.0.tar.gz

hduser@namenode-1 $ tar -xvf pig-0.16.0.tar.gz

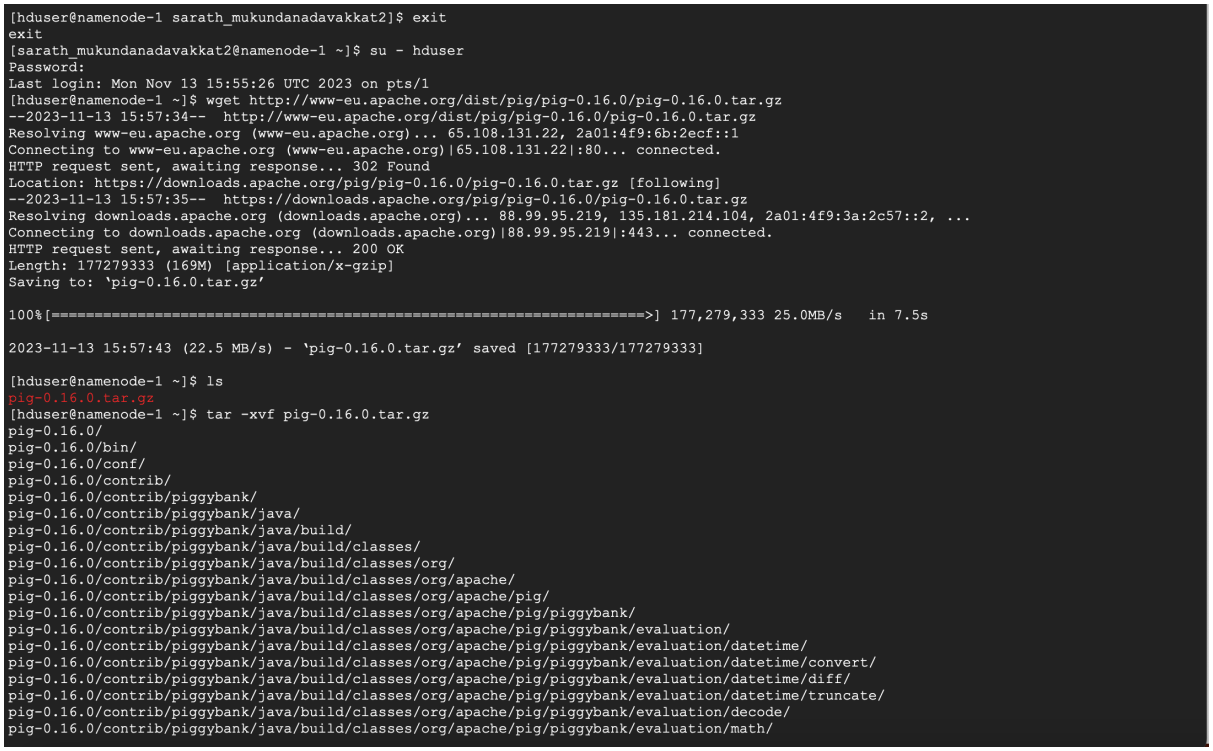

Figure 6: Installation of pig

Task 2: Dataset

Task 2.1: Choose a relevant data set

The dataset contains spam and ham comments from YouTube. It is a public dataset and consists of Comment ID, Author, Date, Content, and Tag attributes. It's in CSV format.

Task 2.2: Get data from the public repository

The dataset is from Kaggle and the link is provided below:

Link: https://www.kaggle.com/datasets/lakshmi25npathi/images/data

Task 2.3: Load the data into Google Cloud (GCP)

The dataset from the local path is transferred into GCP using scp command to securely transfer files by establishing ssh connection between the local machine and GCP Here, the RSA key is generated and the keys are used for login in remote.

snippets of code:

hduser@namenode-1 $ cat ~/.ssh/id_rsa.pub

hduser@namenode-1 $ vim ~/.ssh/authorized_keys

hduser@namenode-1 $ ifconfig

hduser@namenode-1 $ curl whatismyip.akamai.com

hduser@namenode-1 $ pwd

sarath@Saraths-MacBook-Pro ~ % ssh-keygen

sarath@Saraths-MacBook-Pro ~ %cat ~/.ssh/id_rsa.pub

sarath@Saraths-MacBook-Pro ~ % ssh hduser@34.69.135.86 -v

sarath@Saraths-MacBook-Pro ~ % scp Youtube01-Psy.csv hduser@34.69.135.86:/home/hduser

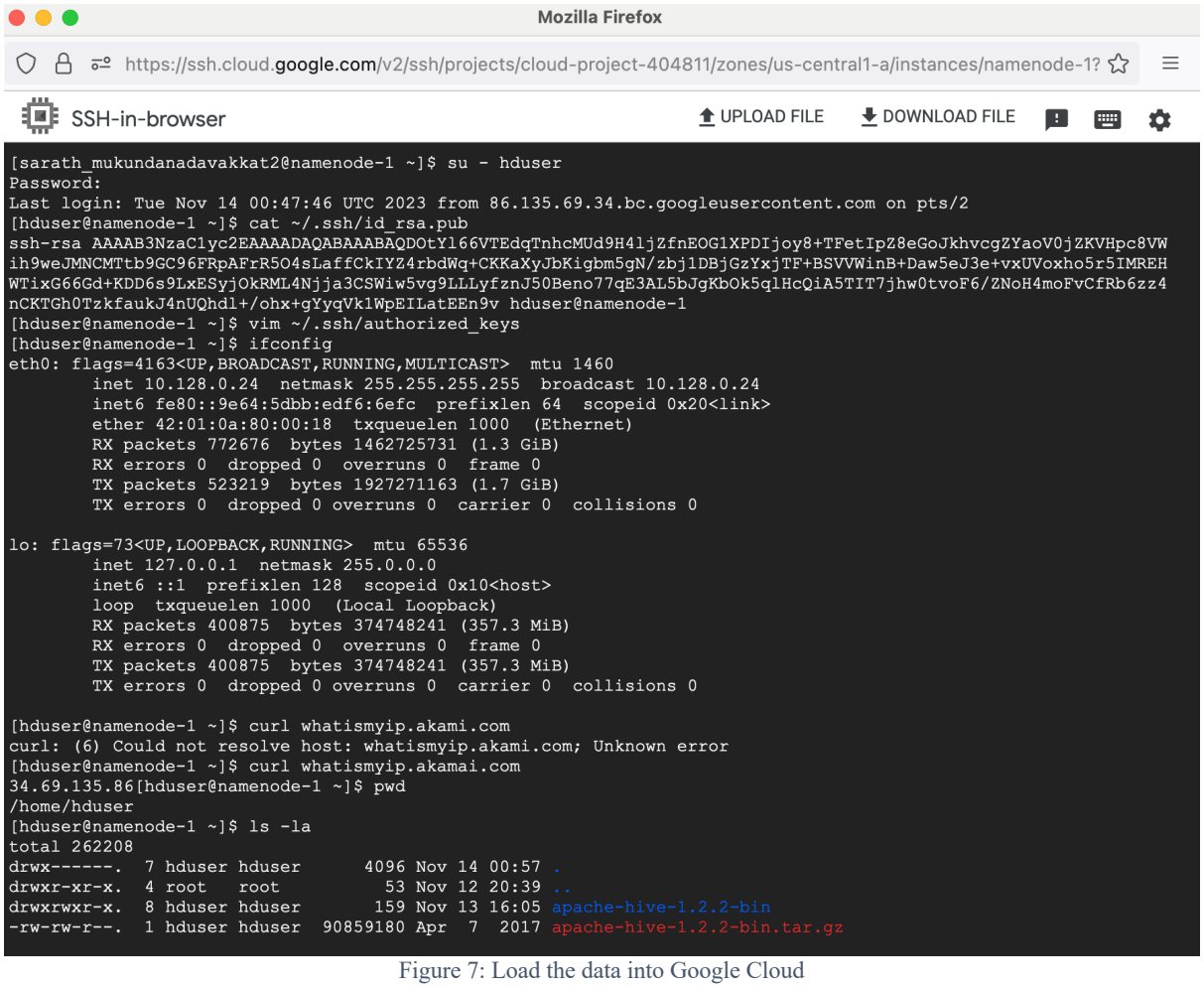

Figure 7: Load the data into Google Cloud

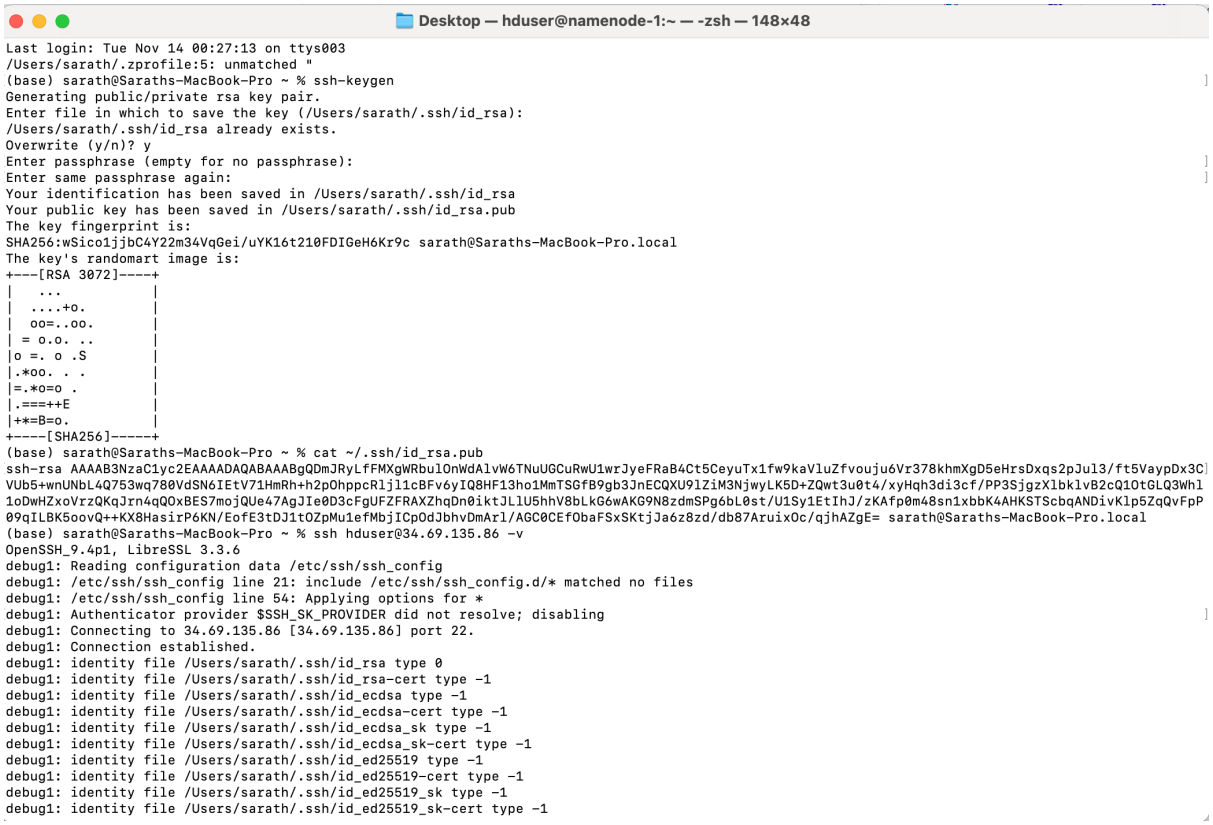

Figure 8: Load the dataset from Local path

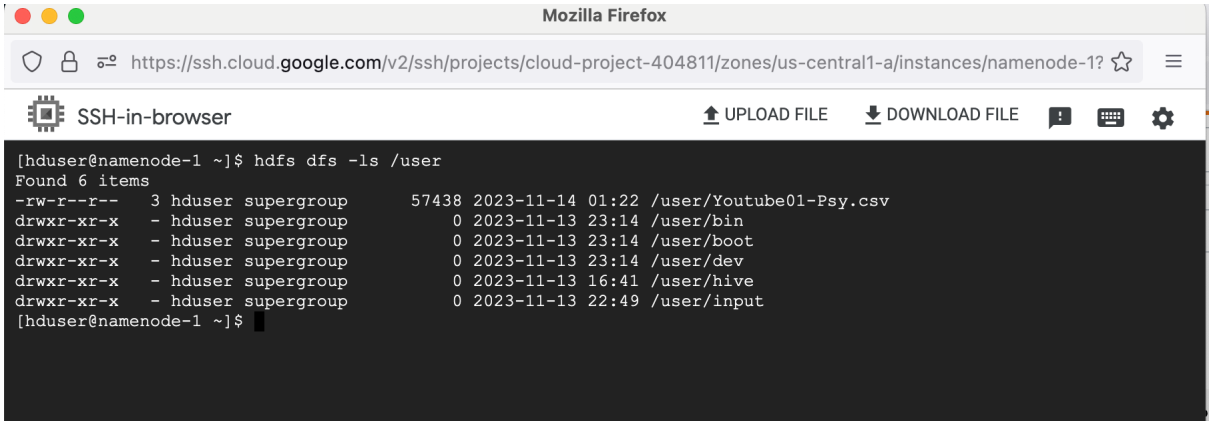

Figure 9: Dataset stored in hdfs location

Task 3: Data Cleaning & Processing

The dataset required some processing as it consists of symbols in the author's name and unwanted smiley, symbols in their comment field. To achieve this task, the dataset was cleaned using PIG.

The cleaned dataset is load stored into the hdfs dfs location as ‘data_cleaned_full1.

snippets of code:

grunt>DEFINE CSVLoader org.apache.pig.piggybank.storage.CSVLoader();

grunt>data = LOAD '/user/Youtube01-Psy.csv ' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',') as (comment_id:chararray, author:chararray, d:chararray, content:chararray, class:int);

grunt>cleaned_data = FOREACH data GENERATE REPLACE($1, '[^a-zA-Z]', '')as author, REPLACE($3, '[^a-zA-Z0-9]', ' ') as content, $4;

grunt>STORE cleaned_data INTO ‘data_cleaned_full1’;

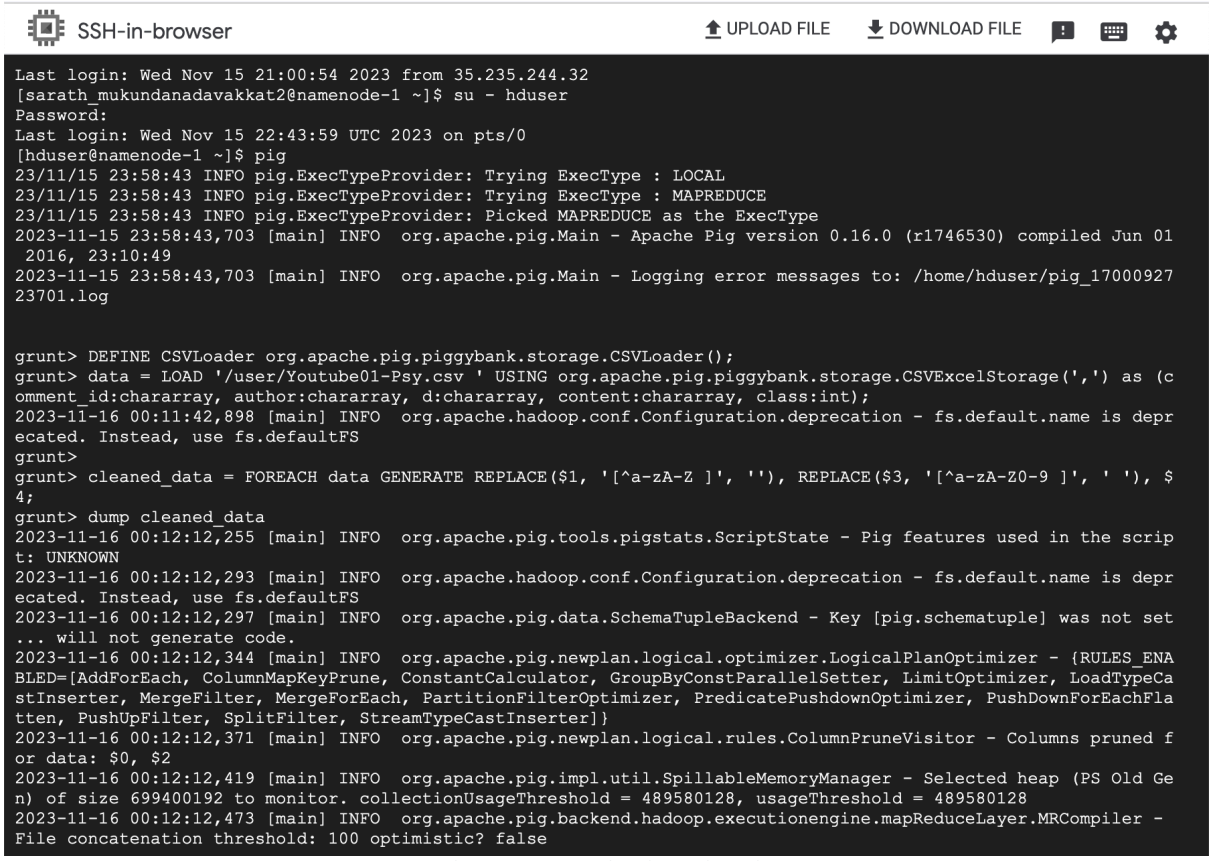

Figure 10: Data cleaning using Pig

Task 4: Ham & Spam using PIG

Task 4.1: Query proceed data to differentiate the ham and spam part of the data set

The cleaned data from the hdfs location is retrieved into relation data. The spam dataset is obtained by filtered from the cleaned dataset by using Bag of words such as ‘subscribe’, ‘my channel’, ‘facebook’, ‘http’ and is stored as ‘Data1’ and the file is stored as Spam dataset in

the form CSV file (spam.csv). Similarly in the case of ham dataset, the bag of words such as ‘views’, ‘OPPA’, ‘good’ is taken and the ham dataset is stored as a ‘ Data2 ’ file and stored as ham dataset in the form of a CSV file (ham.csv).

snippets of code:

grunt>Data1 = filter cleaned_data by (content matches '.*(subscribe | my channel| facebook| http).*');

grunt>store Data1 into '/user/hadoop/spam_dataset' using PigStorage('\t','-schema');

grunt>hadoop fs -getmerge /user/hadoop/spams_dataset ./spams.csv

grunt>Data2 = filter cleaned_data by (content matches '.*(views | OPPA | good).*');

grunt>store Data2 into '/user/hadoop/ham_dataset' using PigStorage('\t','-schema');

grunt>hadoop fs -getmerge /user/hadoop/ham_dataset ./ham.csv

```
grunt> cleaned_data = FOREACH data GENERATE REPLACE($1, '[^a-zA-Z ]', '')as author, REPLACE($3, '[^a-zA-Z0-9 ]'
, ' ') as content, $4;
grunt>
grunt>
grunt> Data1 = filter cleaned_data by (content matches '.*( subscribe | my channel| facebook|http).*');
grunt> store Data1 into '/user/hadoop/spam_dataset' using PigStorage('\t','-schema');
2023-11-16 00:54:59,419 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is depr
ecated. Instead, use fs.defaultFS
2023-11-16 00:54:59,459 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the scrip
t: FILTER
2023-11-16 00:54:59,483 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is depr
ecated. Instead, use fs.defaultFS
2023-11-16 00:54:59,484 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set
... will not generate code.
2023-11-16 00:54:59,484 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENA
BLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCa
stInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFla
tten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2023-11-16 00:54:59,485 [main] INFO  org.apache.pig.newplan.logical.rules.ColumnPruneVisitor - Columns pruned f
or data: $0, $2
2023-11-16 00:54:59,488 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler -
File concatenation threshold: 100 optimistic? false
2023-11-16 00:54:59,491 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOpt
imizer - MR plan size before optimization: 1
2023-11-16 00:54:59,491 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOpt
imizer - MR plan size after optimization: 1
2023-11-16 00:54:59,505 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is depr
ecated. Instead, use fs.defaultFS
2023-11-16 00:54:59,506 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics w
ith processName=JobTracker, sessionId= - already initialized
2023-11-16 00:54:59,508 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script setting
s are added to the job
2023-11-16 00:54:59,508 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCom
piler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2023-11-16 00:54:59,508 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCom
piler - This job cannot be converted run in-process
2023-11-16 00:54:59,615 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCom
piler - Added jar file:/home/hduser/pig-0.16.0/pig-0.16.0-core-h2.jar to DistributedCache through /tmp/temp3126
63614/tmp198189607/pig-0.16.0-core-h2.jar
```

Figure 11: separate datasets from the existing dataset

Task 4.2: Top 10 spam accounts

The following code was executed to find the top 10 spam accounts from the dataset.

snippets of code:

grunt>spam = GROUP Data1 by author;

grunt>spam_count = FOREACH spam GENERATE group as author, COUNT(Data1) as total_comments;

grunt>top_ten_spam1 = ORDER spam_count BY total_comments DESC;

grunt>top_ten_spam = LIMIT top_ten_spam1 10 ;

```
[hduser@namenode-1 ~]$ pig
23/11/16 00:41:07 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
23/11/16 00:41:07 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
23/11/16 00:41:07 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2023-11-16 00:41:07,192 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01
 2016, 23:10:49
2023-11-16 00:41:07,193 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/hduser/pig_17000952
67190.log
2023-11-16 00:41:07,218 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/hduser/.pigboot
up not found
2023-11-16 00:41:08,038 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is d
eprecated. Instead, use mapreduce.jobtracker.address
2023-11-16 00:41:08,038 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is depr
ecated. Instead, use fs.defaultFS
2023-11-16 00:41:08,038 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connectin
g to hadoop file system at: hdfs://namenode-1:54310
2023-11-16 00:41:08,637 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-ad0f
3419-d9c4-455b-9cf2-fd74a3ba8768
2023-11-16 00:41:08,637 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.ena
bled set to false
grunt> data = LOAD '/user/Youtube01-Psy.csv ' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',') as (c
omment_id:chararray, author:chararray, d:chararray, content:chararray, class:int);
2023-11-16 00:41:21,379 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is depr
ecated. Instead, use fs.defaultFS
grunt>
grunt> cleaned_data = FOREACH data GENERATE REPLACE($1, '[^a-zA-Z ]', '')as author, REPLACE($3, '[^a-zA-Z0-9 ]'
, ' ') as content, $4;
grunt>
grunt> Data1 = filter cleaned_data by (content matches '.*( subscribe | my channel| facebook|http).*');
grunt>
grunt> spam = GROUP Data1 by author;
grunt> spam_count = FOREACH spam GENERATE group as author, COUNT(Data1) as total_comments;
grunt>
grunt> top_ten_spam1 = ORDER spam_count BY total_comments DESC;
grunt> top_ten_spam = LIMIT top_ten_spam1 10 ;
grunt> dump top_ten_spam
```

Figure 12: Query to differentiate into spam dataset

```
2023-11-16 00:43:30,172 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics w
ith processName=JobTracker, sessionId= - already initialized
2023-11-16 00:43:30,173 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics w
ith processName=JobTracker, sessionId= - already initialized
2023-11-16 00:43:30,177 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics w
ith processName=JobTracker, sessionId= - already initialized
2023-11-16 00:43:30,178 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics w
ith processName=JobTracker, sessionId= - already initialized
2023-11-16 00:43:30,179 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics w
ith processName=JobTracker, sessionId= - already initialized
2023-11-16 00:43:30,183 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics w
ith processName=JobTracker, sessionId= - already initialized
2023-11-16 00:43:30,185 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics w
ith processName=JobTracker, sessionId= - already initialized
2023-11-16 00:43:30,187 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics w
ith processName=JobTracker, sessionId= - already initialized
2023-11-16 00:43:30,190 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLaun
cher - Encountered Warning FIELD_DISCARDED_TYPE_CONVERSION_FAILED 1 time(s).
2023-11-16 00:43:30,190 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLaun
cher - Success!
2023-11-16 00:43:30,193 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is depr
ecated. Instead, use fs.defaultFS
2023-11-16 00:43:30,194 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already be
en initialized
2023-11-16 00:43:30,199 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths
to process : 1
2023-11-16 00:43:30,199 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total inpu
t paths to process : 1
( ,4)
(Giang Nguyen,2)
(OutrightIgnite,2)
(KodysMan Dunt Know,1)
(Dovydas Baranauskas,1)
(Emerson Zanol Zanol,1)
(Gaming and Stuff PRO,1)
(Elieo Cardiopulmonary,1)
(Bishwaroop Bhattacharjee,1)
(The Bibliophile Flautist,1)
grunt>
```

Figure 13: Top 10 spam accounts

Task 4.3: Top 10 ham accounts

The following code was executed to find the top 10 ham accounts from the dataset.

snippets of code:

grunt>ham = GROUP Data2 by author;

grunt>ham_count = FOREACH ham GENERATE group as author, COUNT(Data2) as total_comments;

grunt>top_ten_ham1 = ORDER ham_count BY total_comments DESC;

grunt>top_ten_ham = LIMIT top_ten_ham1 10 ;

```
grunt>
grunt>
grunt> Data2 = filter cleaned_data by (content matches '.*( views | OPPA | good).*');
grunt>
grunt> ham = GROUP Data2 by author;
grunt> ham_count = FOREACH ham GENERATE group as author, COUNT(Data2) as total_comments;
grunt>
grunt> top_ten_ham1 = ORDER ham_count BY total_comments DESC;
grunt>
grunt> top_ten_ham = LIMIT top_ten_ham1 10 ;
grunt> dump top_ten_ham
```

Figure 14:Query to differentiate into ham dataset

```
2023-11-16 00:47:06,908 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics w
ith processName=JobTracker, sessionId= - already initialized
2023-11-16 00:47:06,909 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics w
ith processName=JobTracker, sessionId= - already initialized
2023-11-16 00:47:06,912 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics w
ith processName=JobTracker, sessionId= - already initialized
2023-11-16 00:47:06,913 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics w
ith processName=JobTracker, sessionId= - already initialized
2023-11-16 00:47:06,914 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics w
ith processName=JobTracker, sessionId= - already initialized
2023-11-16 00:47:06,917 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics w
ith processName=JobTracker, sessionId= - already initialized
2023-11-16 00:47:06,919 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics w
ith processName=JobTracker, sessionId= - already initialized
2023-11-16 00:47:06,919 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics w
ith processName=JobTracker, sessionId= - already initialized
2023-11-16 00:47:06,921 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLaun
cher - Encountered Warning FIELD_DISCARDED_TYPE_CONVERSION_FAILED 1 time(s).
2023-11-16 00:47:06,921 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLaun
cher - Success!
2023-11-16 00:47:06,922 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is depr
ecated. Instead, use fs.defaultFS
2023-11-16 00:47:06,923 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already be
en initialized
2023-11-16 00:47:06,929 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths
to process : 1
2023-11-16 00:47:06,929 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total inpu
t paths to process : 1
(MrTuizentfloot,1)
(The Silhouette,1)
(jayson calzado,1)
(Carmen Racasanu,1)
(Alucard Hellsing,1)
(DropItLikeItsSloth,1)
(Kincaid Liebenberg,1)
(Digital Media Butterfly,1)
(The Silent Troll Defuser HD,1)
(Oopsthenameistoolong Oh well,1)
grunt>
```

Figure 15: Top 10 ham accounts

Task 5: TF-IDF using MapReduce

```
SSH-in-browser                                                                    UPLOAD FILE  DOWNLOAD FILE

[hduser@namenode-1 ~]$ hadoop jar /usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.7.0.jar -input /user/hadoop/new_spam1 -mapper "/home/hduser/mapper-tfidf.py" -reducer
"/home/hduser/reducer-tfidf.py" -output ./tf-idf-output
23/11/16 20:37:06 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
23/11/16 20:37:06 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
23/11/16 20:37:06 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
23/11/16 20:37:06 INFO mapred.FileInputFormat: Total input paths to process : 1
23/11/16 20:37:06 INFO mapreduce.JobSubmitter: number of splits:1
23/11/16 20:37:07 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local2079211724_0001
23/11/16 20:37:07 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
23/11/16 20:37:07 INFO mapreduce.Job: Running job: job_local2079211724_0001
23/11/16 20:37:07 INFO mapred.LocalJobRunner: OutputCommitter set in config null
23/11/16 20:37:07 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
23/11/16 20:37:07 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
23/11/16 20:37:07 INFO mapred.LocalJobRunner: Waiting for map tasks
23/11/16 20:37:07 INFO mapred.LocalJobRunner: Starting task: attempt_local2079211724_0001_m_000000_0
23/11/16 20:37:07 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
23/11/16 20:37:07 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
23/11/16 20:37:07 INFO mapred.MapTask: Processing split: hdfs://namenode-1:54310/user/hadoop/new_spam1/part-m-00000:0+17403
23/11/16 20:37:07 INFO mapred.MapTask: numReduceTasks: 1
23/11/16 20:37:07 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
23/11/16 20:37:07 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
23/11/16 20:37:07 INFO mapred.MapTask: soft limit at 83886080
23/11/16 20:37:07 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
23/11/16 20:37:07 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
23/11/16 20:37:07 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
23/11/16 20:37:07 INFO streaming.PipeMapRed: PipeMapRed exec [/home/hduser/mapper-tfidf.py]
23/11/16 20:37:07 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
23/11/16 20:37:07 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce.cluster.local.dir
23/11/16 20:37:07 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.map.input.file
23/11/16 20:37:07 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords
23/11/16 20:37:07 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce.map.input.length
23/11/16 20:37:07 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use mapreduce.task.output.dir
23/11/16 20:37:07 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce.map.input.start
23/11/16 20:37:07 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
23/11/16 20:37:07 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.user.name
23/11/16 20:37:07 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.ismap
23/11/16 20:37:07 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
23/11/16 20:37:07 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.partition
23/11/16 20:37:07 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
23/11/16 20:37:07 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
23/11/16 20:37:07 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
23/11/16 20:37:07 INFO streaming.PipeMapRed: Records R/W=109/1
23/11/16 20:37:07 INFO streaming.PipeMapRed: MRErrorThread done
23/11/16 20:37:07 INFO streaming.PipeMapRed: mapRedFinished
```

Figure 16: TF-IDF using MapReduce in spam dataset

Task 5.1: MapReduce to calculate the TF-IDF of the top 10 spam keywords for each top 10 spam accounts

Snippet of code:

hadoop jar /usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.7.0.jar -input /user/hadoop/new_spam1 -mapper "/home/hduser/mapper-tfidf.py" -reducer "/home/hduser/reducer-tfidf.py" -output ./tfidf-out

hadoop fs -getmerge /user/hduser/tfidf-out ./ tfidf-out

head -100 tfidf-out

```
generating tf-idf for word 48051 for user


48051,z131idupvn3yhf3mv23dwzhi4pqixvwuw 0.36824324324324326
48051,z131idupvn3yhf3mv23dwzhi4pqixvwuw 0.36824324324324326
48051,z131idupvn3yhf3mv23dwzhi4pqixvwuw 0.36824324324324326
48051,z131idupvn3yhf3mv23dwzhi4pqixvwuw 0.36824324324324326
48051,z131idupvn3yhf3mv23dwzhi4pqixvwuw 0.36824324324324326
48051,z131idupvn3yhf3mv23dwzhi4pqixvwuw 0.36824324324324326
48051,z131idupvn3yhf3mv23dwzhi4pqixvwuw 0.36824324324324326
48051,z131idupvn3yhf3mv23dwzhi4pqixvwuw 0.36824324324324326
48051,z131idupvn3yhf3mv23dwzhi4pqixvwuw 0.36824324324324326
48051,z131idupvn3yhf3mv23dwzhi4pqixvwuw 0.36824324324324326
48051,z131idupvn3yhf3mv23dwzhi4pqixvwuw 0.36824324324324326
48051,z131idupvn3yhf3mv23dwzhi4pqixvwuw 0.36824324324324326
48051,z131idupvn3yhf3mv23dwzhi4pqixvwuw 0.36824324324324326
48051,z131idupvn3yhf3mv23dwzhi4pqixvwuw 0.36824324324324326
48051,z131idupvn3yhf3mv23dwzhi4pqixvwuw 0.36824324324324326
48051,z131idupvn3yhf3mv23dwzhi4pqixvwuw 0.36824324324324326
48051,z131idupvn3yhf3mv23dwzhi4pqixvwuw 0.36824324324324326
48051,z131idupvn3yhf3mv23dwzhi4pqixvwuw 0.36824324324324326
48051,z131idupvn3yhf3mv23dwzhi4pqixvwuw 0.36824324324324326
48051,z131idupvn3yhf3mv23dwzhi4pqixvwuw 0.36824324324324326


generating tf-idf for word ru for user


ru,z131idupvn3yhf3mv23dwzhi4pqixvwuw    0.36824324324324326
ru,z131idupvn3yhf3mv23dwzhi4pqixvwuw    0.36824324324324326
ru,z131idupvn3yhf3mv23dwzhi4pqixvwuw    0.36824324324324326
ru,z131idupvn3yhf3mv23dwzhi4pqixvwuw    0.36824324324324326
ru,z131idupvn3yhf3mv23dwzhi4pqixvwuw    0.36824324324324326
ru,z131idupvn3yhf3mv23dwzhi4pqixvwuw    0.36824324324324326
ru,z131idupvn3yhf3mv23dwzhi4pqixvwuw    0.36824324324324326
ru,z131idupvn3yhf3mv23dwzhi4pqixvwuw    0.36824324324324326
ru,z131idupvn3yhf3mv23dwzhi4pqixvwuw    0.36824324324324326
ru,z131idupvn3yhf3mv23dwzhi4pqixvwuw    0.36824324324324326
ru,z131idupvn3yhf3mv23dwzhi4pqixvwuw    0.36824324324324326
ru,z131idupvn3yhf3mv23dwzhi4pqixvwuw    0.36824324324324326
ru,z131idupvn3yhf3mv23dwzhi4pqixvwuw    0.36824324324324326
ru,z131idupvn3yhf3mv23dwzhi4pqixvwuw    0.36824324324324326
ru,z131idupvn3yhf3mv23dwzhi4pqixvwuw    0.36824324324324326
ru,z131idupvn3yhf3mv23dwzhi4pqixvwuw    0.36824324324324326
ru,z131idupvn3yhf3mv23dwzhi4pqixvwuw    0.36824324324324326
```

```
generating tf-idf for word check for user Jdidk Loma


check,LZQPQhLyRh_C2cTtd9MvFRJedxydaVW-2sNg5Diuo4A       2.658536585365854
check,z121st5w5k3ui1veg22zirn4gkr5tby2v 7.785714285714286
check,z12avveb4xqiirsix04chxviiljryduwxg0       6.411764705882353
check,z12eexphzo2uslizg04cirmywyzjdnm5gqc0k     5.45
check,z12ruft4aq3thhpbl04cdxrzjvmohjyqhs00k     2.7948717948717947
check,z134d5hbckywylmj404cgnlo3kysfhsjoeg       4.1923076923076925
check,z13cyzbbqsrxyfaec23xc10rdrrqgd0ch 3.892857142857143
check,z13phrmwrkfisn5er22eyrbpbvaiwfvwf04       1.3974358974358974
check,z13phrmwrkfisn5er22eyrbpbvaiwfvwf04       1.3974358974358974
check,z13pv52hkmf4jn23g22nzx5zqr2gen1gv04       4.541666666666666
check,z13supiartrcdr4la22xc3aripu2x1z3a 4.36
check,z13ttt1jcraqexk2o234ghbgzxymz1zzi04       9.90909090909091


generating tf-idf for word game for user Jdidk Loma


game,z121st5w5k3ui1veg22zirn4gkr5tby2v  7.785714285714286


generating tf-idf for word great for user Jdidk Loma


great,z121st5w5k3ui1veg22zirn4gkr5tby2v 7.785714285714286
great,z12cehoxozfgg3nok04cjj05xznbgrlpfjo       0.42745098039215684


generating tf-idf for word cd92db3f4 for user Jdidk Loma


cd92db3f4,z121st5w5k3ui1veg22zirn4gkr5tby2v     7.785714285714286


generating tf-idf for word friend for user Jdidk Loma


friend,z121st5w5k3ui1veg22zirn4gkr5tby2v        7.785714285714286
friend,z12cehoxozfgg3nok04cjj05xznbgrlpfjo      0.42745098039215684
friend,z12kyn0qjzzzur2ai04cg5szenjxdrorp4w      5.45
```

Figure 17: top 10 spam keywords for each top 10 spam account

Task 5.2: MapReduce to calculate the TF-IDF of the top 10 ham keywords for each top 10 ham accounts

Snippet of code:

hadoop jar /usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.7.0.jar -input /user/hadoop/new_ham1 -mapper "/home/hduser/mapper-tfidf.py" -reducer "/home/hduser/reducer-tfidf.py" -output ./out11

hadoop fs -getmerge /user/hduser/out11 ./out11

head -100 out11

```
[hduser@namenode-1 ~]$ hadoop jar /usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.7.0.jar -input /user/hadoop/new_ham1 -mapper "/home/hduser/mapper-tfidf.py" -reducer "
/home/hduser/reducer-tfidf.py" -output ./out11
23/11/16 20:58:48 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
23/11/16 20:58:48 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
23/11/16 20:58:48 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
23/11/16 20:58:49 INFO mapred.FileInputFormat: Total input paths to process : 1
23/11/16 20:58:49 INFO mapreduce.JobSubmitter: number of splits:1
23/11/16 20:58:49 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local716381967_0001
23/11/16 20:58:49 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
23/11/16 20:58:49 INFO mapreduce.Job: Running job: job_local716381967_0001
23/11/16 20:58:49 INFO mapred.LocalJobRunner: OutputCommitter set in config null
23/11/16 20:58:49 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
23/11/16 20:58:49 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
23/11/16 20:58:49 INFO mapred.LocalJobRunner: Waiting for map tasks
23/11/16 20:58:49 INFO mapred.LocalJobRunner: Starting task: attempt_local716381967_0001_m_000000_0
23/11/16 20:58:49 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
23/11/16 20:58:49 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
23/11/16 20:58:49 INFO mapred.MapTask: Processing split: hdfs://namenode-1:54310/user/hadoop/new_ham1/part-m-00000:0+6971
23/11/16 20:58:49 INFO mapred.MapTask: numReduceTasks: 1
23/11/16 20:58:49 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
23/11/16 20:58:49 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
23/11/16 20:58:49 INFO mapred.MapTask: soft limit at 83886080
23/11/16 20:58:49 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
23/11/16 20:58:49 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
23/11/16 20:58:49 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
23/11/16 20:58:49 INFO streaming.PipeMapRed: PipeMapRed exec [/home/hduser/mapper-tfidf.py]
23/11/16 20:58:49 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
23/11/16 20:58:49 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce.cluster.local.dir
23/11/16 20:58:49 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.map.input.file
23/11/16 20:58:49 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords
23/11/16 20:58:49 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce.map.input.length
23/11/16 20:58:49 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use mapreduce.task.output.dir
23/11/16 20:58:49 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce.map.input.start
23/11/16 20:58:49 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
23/11/16 20:58:49 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.user.name
23/11/16 20:58:49 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.ismap
23/11/16 20:58:49 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
23/11/16 20:58:49 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.partition
23/11/16 20:58:50 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
23/11/16 20:58:50 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
23/11/16 20:58:50 INFO streaming.PipeMapRed: Records R/W=51/1
23/11/16 20:58:50 INFO streaming.PipeMapRed: MRErrorThread done
23/11/16 20:58:50 INFO streaming.PipeMapRed: mapRedFinished
23/11/16 20:58:50 INFO mapred.LocalJobRunner:
23/11/16 20:58:50 INFO mapred.MapTask: Starting flush of map output
```

```
[hduser@namenode-1 ~]$ vim out11
[hduser@namenode-1 ~]$ head -100 out11
sorted user dictionary where users are soreted based on total ham count {'Vitor Belliny': 1, 'crestpee': 1, 'Warrdrew': 1, 'Bob Kanowski': 1, 'Mia Aspinall': 1, 'Romeo Sweeney': 1,
 'Luat ha hpuoc': 1, 'Agarra Mela': 1, 'Yabbamita DBH': 1, 'zhichao wang': 1, 'DropItLikeItsSloth': 1, 'Amir effect': 1, 'ArioseRose': 1, 'MrTuizentfloot': 1, 'jayson calzado': 1,
'The Silhouette': 1, 'Gio D': 1, 'Gaming ': 1, 'Kincaid Liebenberg': 1, 'Owen Lai': 1, 'Caius Ballad': 1, 'Susan Jay': 1, 'Sev': 1, 'Eanna Cusack': 1, 'Phuc Ly': 1, 'x Judda': 1, '
Brandon Wilson': 1, 'Tom Hosford': 1, 'King uzzy': 1, 'khir abqari': 1, 'Alucard Hellsing': 1, 'Adam Mudd': 1, 'richardex': 1, 'itzquiffer': 1, 'DropShotSkr': 1, 'Zielimeek': 1, 'D
igital Media Butterfly': 1, 'Diana Roque': 1, 'Praise Samuel': 1, 'Gaming Land': 1, 'Milan George': 1, 'Wert Walleet': 1, 'Oopsthenameistoolong Oh well': 1, 'Yeung Marvin': 1, 'The
 Silent Troll Defuser HD': 1, 'Carmen Racasanu': 1, 'Kochos': 1, 'Haley Harmicar': 1, 'rolle': 1, 'Emily Hamilton': 1, 'Thanh Phong': 1}
```

```
generating tf-idf for word 750 for user Vitor Belliny


750,z121yttbfpyxw1dya04cgtq4clasebvoib4 2.125


generating tf-idf for word views for user Vitor Belliny


views,z121yttbfpyxw1dya04cgtq4clasebvoib4       2.125
views,z121znjjloycwtjf2233dfo5esfggn3lj 2.5500000000000003
views,z122srjbvovwzxg0522yd1i4ty3kznxqs04       4.25
views,z122wfnzgt30fhubn04cdn3xfx2mxzngsl40k     2.217391304347826
views,z122z5pa2wyofbjj304cgfwrrmvjgn0pohc       5.666666666666666
views,z123gh5o1pibgntf222einrisrfasnha104       8.5
views,z123gnabnwbqtr1e022jejij0zzzez2os04       2.3181818181818183
views,z124elbgazqmuhvs5230ut14uta5irenh 2.3181818181818183
views,z124zbcigmbfvhpru04cepdx2vikibhp42c       1.9615384615384617
views,z12axnji5w2axxht522thb3bktvqjdlbp04       2.5500000000000003
views,z12axnji5w2axxht522thb3bktvqjdlbp04       2.5500000000000003
views,z12dixvi2qepuhb4o04chpd5exjdx3oqdic       5.1000000000000005
views,z12dzlpo0ueie1go404cfjjwsxf1g1lrtdo       3.642857142857143
views,z12gcx5hzkayhbmve23ghd3plt2czf2rp04       2.5500000000000003
views,z12ifxrbkmaechwtt22jwryqmoaefhipf04       0.9807692307692308
views,z12ifxrbkmaechwtt22jwryqmoaefhipf04       0.9807692307692308
views,z12ifxrbkmaechwtt22jwryqmoaefhipf04       0.9807692307692308
views,z12iubmp2mv1txfsy235vb2xdqirufcp304       1.8888888888888888
views,z12ivx14lye5szr4k04cddnhuujytfhz4v00k     2.217391304347826
views,z12jy1tw0unnhhdnh23yyvyhjsnljxhf4 1.2750000000000001
views,z12kttwqvzi4fd0ei23rdp4xjt2ef5hbk04       1.8214285714285714
views,z12kttwqvzi4fd0ei23rdp4xjt2ef5hbk04       1.8214285714285714
views,z12ntlcqht2bvjewi04cf1up0xjvs5lq3mc0k     10.200000000000001
views,z12oex5p1miewvugp04cgjfqiq3xsxrolhk       2.4285714285714284
views,z12oxlzh4qjicd2zu04cgfabqtipf3gq4is       1.3421052631578947
views,z12pcbng1m2shlgkp23ki1izwkreylid3 0.9272727272727272
views,z12rglllzmwnfuba4h04ch3njgy2fg3ryerg0k     5.1000000000000005
views,z12stpyqaonsercxw04cjbr4dlnrxd34soo0k     5.666666666666666
views,z12ttjopmofst1gpp04cc5ezywjrwntjrc0       3.4
views,z12uyzhazxqkzbrzz04cedqpkwzdvjyy3u40k     3.642857142857143
views,z12yivpznrezu3gyf04ce11xdvmquxzjdo00k     1.59375
views,z12yivpznrezu3gyf04ce11xdvmquxzjdo00k     1.59375
views,z130dbmz2ourjtupz04chhjrpunwcr4yjrs0k     4.25
views,z130gviqarmshdnau04cdzigvs3jepx4qw00k     2.04
views,z130xbcwfnj5vlskv23airsxqqfqvhl5504       1.7
```

```
views,z13lip25arigch3rj04cf1ko0pnczzyrtng0k     6.375
views,z13nvr2xayrwffsio04cj3zwyuf3vb1imdg       3.642857142857143
views,z13nw3lhgt2nf5wwe04cdlx5iyaydznrve0       3.9230769230769234
views,z13pfnrxuyelivdoz04cgdx4esa2dxeymck       3.4
views,z13pitkr5prbgf3ja04cjjg4qme3txjpyqc0k     1.7
views,z13th1q4yzihf1bll23qxzpjeujterydj 2.833333333333333
views,z13ucxa51rnftvu2j22if5oihq2qipjlp 3.1875
views,z13ucxdzemugi1v5n04ccjloko25drfb4js       1.9615384615384617
views,z13vx3kbgmq5fnlgj04cfdoqtpfyw5xqzuc0k     1.7
views,z13wfr2rbwytwbhf123cs1l5lz2aefdex 3.9230769230769234
views,z13xgr1w3oeycnszq04cihybwtimi3vqc40       7.285714285714286


generating tf-idf for word 2 for user Vitor Belliny


2,z121yttbfpyxw1dya04cgtq4clasebvoib4   1.0625
2,z122z5pa2wyofbjj304cgfwrrmvjgn0pohc   2.833333333333333
2,z122z5pa2wyofbjj304cgfwrrmvjgn0pohc   2.833333333333333
2,z123gnabnwbqtr1e022jejij0zzzez2os04   1.1590909090909092
2,z124elbgazqmuhvs5230ut14uta5irenh     1.1590909090909092
2,z12iubmp2mv1txfsy235vb2xdqirufcp304   0.9444444444444444
2,z12jy1tw0unnhhdnh23yyvyhjsnljxhf4     0.6375000000000001
2,z12kttwqvzi4fd0ei23rdp4xjt2ef5hbk04   0.9107142857142857
2,z12yivpznrezu3gyf04ce11xdvmquxzjdo00k 0.796875
2,z130xbcwfnj5vlskv23airsxqqfqvhl5504   0.85
2,z130xbcwfnj5vlskv23airsxqqfqvhl5504   0.85
2,z13dtz1zzkagdromt230g5cqfsejstr3p     0.6891891891891893
2,z13fgt5wtsf3znibb04cgrfgasztuv5q5wk0k 1.7
2,z13jzr151zb4cfmqs04chbrbuknchzxy40   1.1590909090909092
2,z13lip25arigch3rj04cf1ko0pnczzyrtng0k 3.1875
2,z13nw3lhgt2nf5wwe04cdlx5iyaydznrve0   1.9615384615384617
2,z13th1q4yzihf1bll23qxzpjeujterydj     1.4166666666666665
2,z13ucxdzemugi1v5n04ccjloko25drfb4js   0.9807692307692308
2,z13vx3kbgmq5fnlgj04cfdoqtpfyw5xqzuc0k 0.85
2,z13wfr2rbwytwbhf123cs1l5lz2aefdex     1.9615384615384617
2,z13xgr1w3oeycnszq04cihybwtimi3vqc40   3.642857142857143


generating tf-idf for word 126 for user Vitor Belliny


126,z121yttbfpyxw1dya04cgtq4clasebvoib4 2.125

[hduser@namenode-1 ~]$
```

Figure 18:top 10 spam keywords for each top 10 spam account