

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

Task 1: Cloud Infrastructure Setup in GCP	4
Task 1.1: Create a Hadoop Cluster and install Java and Hadoop	5
Task 1.2: Install Mapreduce.Hive, Pig in the cluster	6
Task 2: Dataset	8
Task 2.1: Choose a relevant data set.....	8
Task 2.2: Get data from the public repository	8
Task 2.3: Load the data into Google Cloud (GCP).....	8
Task 3: Data Cleaning & Processing	10
Task 4: Ham & Spam using PIG	11
Task 4.1: Query proceed data to differentiate the ham and spam part of the data set.....	11
Task 4.2: Top 10 spam accounts	13
Task 4.3: Top 10 ham accounts.....	14
Task 5: TF-IDF using MapReduce	15
Task 5.1: MapReduce to calculate the TF-IDF of the top 10 spam keywords for each top 10 spam accounts.....	16
Task 5.2: MapReduce to calculate the TF-IDF of the top 10 ham keywords for each top 10 ham accounts.....	18

Tables of Figures:

Figure 1: Multinode cluster	4
Figure 2: Installation of Java	5
Figure 3: Installation of Hadoop.....	5
Figure 4: Installation of hive	6
Figure 5: installation of Hive.....	7
Figure 6: Installation of pig	8
Figure 7: Load the data into Google Cloud	9
Figure 8: Load the dataset from Local path.....	10
Figure 9: Dataset stored in hdfs location	10
Figure 10: Data cleaning using Pig.....	11
Figure 11: separate datasets from the existing dataset.....	12
Figure 12: Query to differentiate into spam dataset.....	13
Figure 13: Top 10 spam accounts	14
Figure 14:Query to differentiate into ham dataset	14
Figure 15: Top 10 ham accounts.....	15

Link to the Git repository

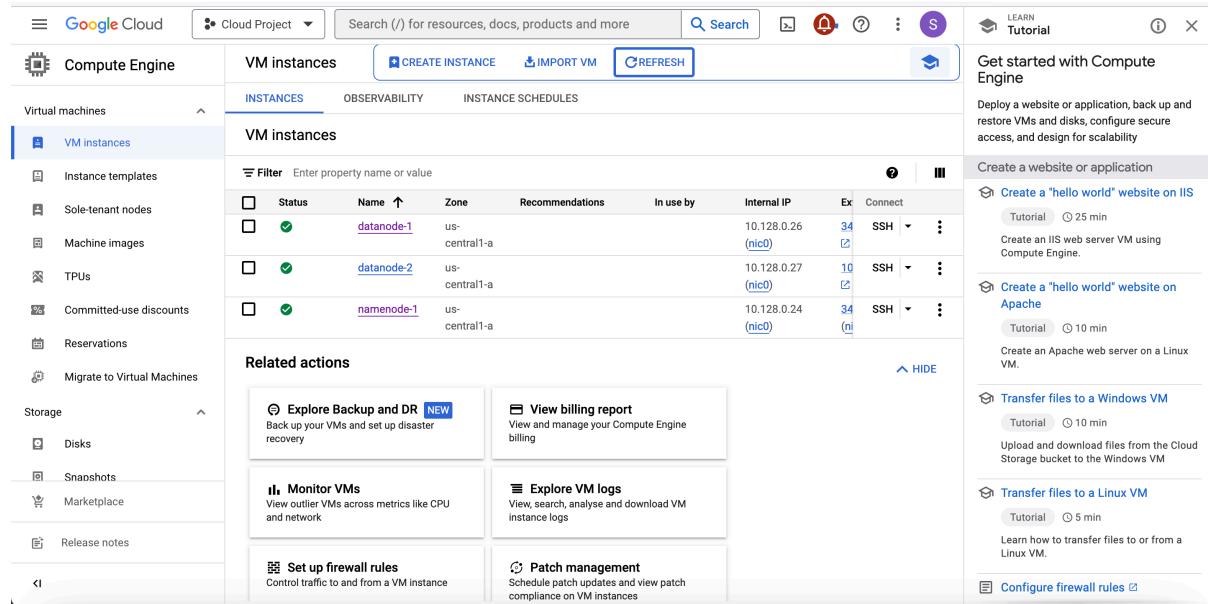
https://gitlab.com/mukunds2/cloud_assignment_1

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



The screenshot shows the Google Cloud Compute Engine interface. On the left, there's a sidebar with sections for Compute Engine, Virtual machines, Storage, and Marketplace. The main area is titled 'VM instances' and shows a table of three instances:

Status	Name	Zone	Recommendations	In use by	Internal IP	Ex	SSH	⋮
✓	datanode-1	us-central1-a			10.128.0.26 (nic0)	34	SSH	⋮
✓	datanode-2	us-central1-a			10.128.0.27 (nic0)	10	SSH	⋮
✓	namenode-1	us-central1-a			10.128.0.24 (nic0)	34	SSH	⋮

Below the table, there are 'Related actions' cards: 'Explore Backup and DR', 'View billing report', 'Monitor VMs', 'Explore VM logs', 'Set up firewall rules', and 'Patch management'. To the right, there's a sidebar titled 'Get started with Compute Engine' containing links to tutorials for creating websites, IIS, Apache, and Windows/Linux VMs.

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

```
[root@namenode-1 opt]# wget -c --header "Cookie: oraclelicense=accept-securebackup-cookie" http://download.oracle.com/otn-pub/java/jdk/8u131-b11/d54c1d3a095b4ff2b6607d096fa80163/jdk-8u131-linux-x64.tar.gz
--2023-11-12 22:56:55-- http://download.oracle.com/otn-pub/java/jdk/8u131-b11/d54c1d3a095b4ff2b6607d096fa80163/jdk-8u131-linux-x64.tar.gz
Resolving download.oracle.com (download.oracle.com)... 23.6.204.88
Connecting to download.oracle.com (download.oracle.com)|23.6.204.88|:80... connected.
HTTP request sent, awaiting response... 302 Moved Temporarily
Location: https://edelivery.oracle.com/otn-pub/java/jdk/8u131-b11/d54c1d3a095b4ff2b6607d096fa80163/jdk-8u131-linux-x64.tar.gz [following]
--2023-11-12 22:56:55-- https://edelivery.oracle.com/otn-pub/java/jdk/8u131-b11/d54c1d3a095b4ff2b6607d096fa80163/jdk-8u131-linux-x64.tar.gz
Resolving edelivery.oracle.com (edelivery.oracle.com)... 184.86.173.114, 2600:1407:3c00:9a5::366, 2600:1407:3c00:981::366
Connecting to edelivery.oracle.com (edelivery.oracle.com)|184.86.173.114|:443... connected.
HTTP request sent, awaiting response... 302 Moved Temporarily
Location: https://download.oracle.com/otn-pub/java/jdk/8u131-b11/d54c1d3a095b4ff2b6607d096fa80163/jdk-8u131-linux-x64.tar.gz?AuthParam=1699829936_39e95b2fc8b29cbf73909642ec9675c5 [following]
--2023-11-12 22:56:56-- https://download.oracle.com/otn-pub/java/jdk/8u131-b11/d54c1d3a095b4ff2b6607d096fa80163/jdk-8u131-linux-x64.tar.gz?AuthParam=1699829936_39e95b2fc8b29cbf73909642ec9675c5
Connecting to download.oracle.com (download.oracle.com)|23.6.204.88|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 185540433 (177M) [application/x-gzip]
Saving to: 'jdk-8u131-linux-x64.tar.gz'

100%[=====] 185,540,433 58.9MB/s in 3.0s

2023-11-12 22:56:59 (58.9 MB/s) - 'jdk-8u131-linux-x64.tar.gz' saved [185540433/185540433]

[root@namenode-1 opt]# wget https://archive.apache.org/dist/hadoop/core/hadoop-2.7.0/hadoop-2.7.0.tar.gz
--2023-11-12 22:57:53-- https://archive.apache.org/dist/hadoop/core/hadoop-2.7.0/hadoop-2.7.0.tar.gz
Resolving archive.apache.org (archive.apache.org)... 65.108.204.189, 2a01:4f9:1:a084::2
Connecting to archive.apache.org (archive.apache.org)|65.108.204.189|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 210343364 (201M) [application/x-gzip]
Saving to: 'hadoop-2.7.0.tar.gz'

100%[=====] 210,343,364 23.0MB/s in 9.5s

2023-11-12 22:58:03 (21.1 MB/s) - 'hadoop-2.7.0.tar.gz' saved [210343364/210343364]
```

Figure 2: Installation of Java

```
[root@namenode-1 jdk1.8.0_131]# cd /opt
[root@namenode-1 opt]# tar xzf hadoop-2.7.0.tar.gz
[root@namenode-1 opt]# mv hadoop-2.7.0 /usr/local/hadoop
[root@namenode-1 opt]# chown -R hduser:hduser /usr/local/hadoop
[root@namenode-1 opt]# mkdir -p /usr/local/hadoop_store/tmp
[root@namenode-1 opt]# mkdir -p /usr/local/hadoop_store/hdfs/namenode
[root@namenode-1 opt]# mkdir -p /usr/local/hadoop_store/hdfs/data-node
[root@namenode-1 opt]# mkdir -p /usr/local/hadoop_store/hdfs/secondarynamenode
[root@namenode-1 opt]# chown -R hduser:hduser /usr/local/hadoop_store
[root@namenode-1 opt]# su hduser
[hduser@namenode-1 opt]$ vi /usr/local/hadoop/etc/hadoop/hdfs-site.xml
[hduser@namenode-1 opt]$ vi /usr/local/hadoop/etc/hadoop/core-site.xml
[hduser@namenode-1 opt]$ vi /usr/local/hadoop/etc/hadoop/map-site.xml
[hduser@namenode-1 opt]$ vi /usr/local/hadoop/etc/hadoop/yarn-site.xml
[hduser@namenode-1 opt]$ echo 'export JAVA_HOME=/opt/jdk1.8.0_131' >> /usr/local/hadoop/etc/hadoop/hadoop-env.sh
[hduser@namenode-1 opt]$
```

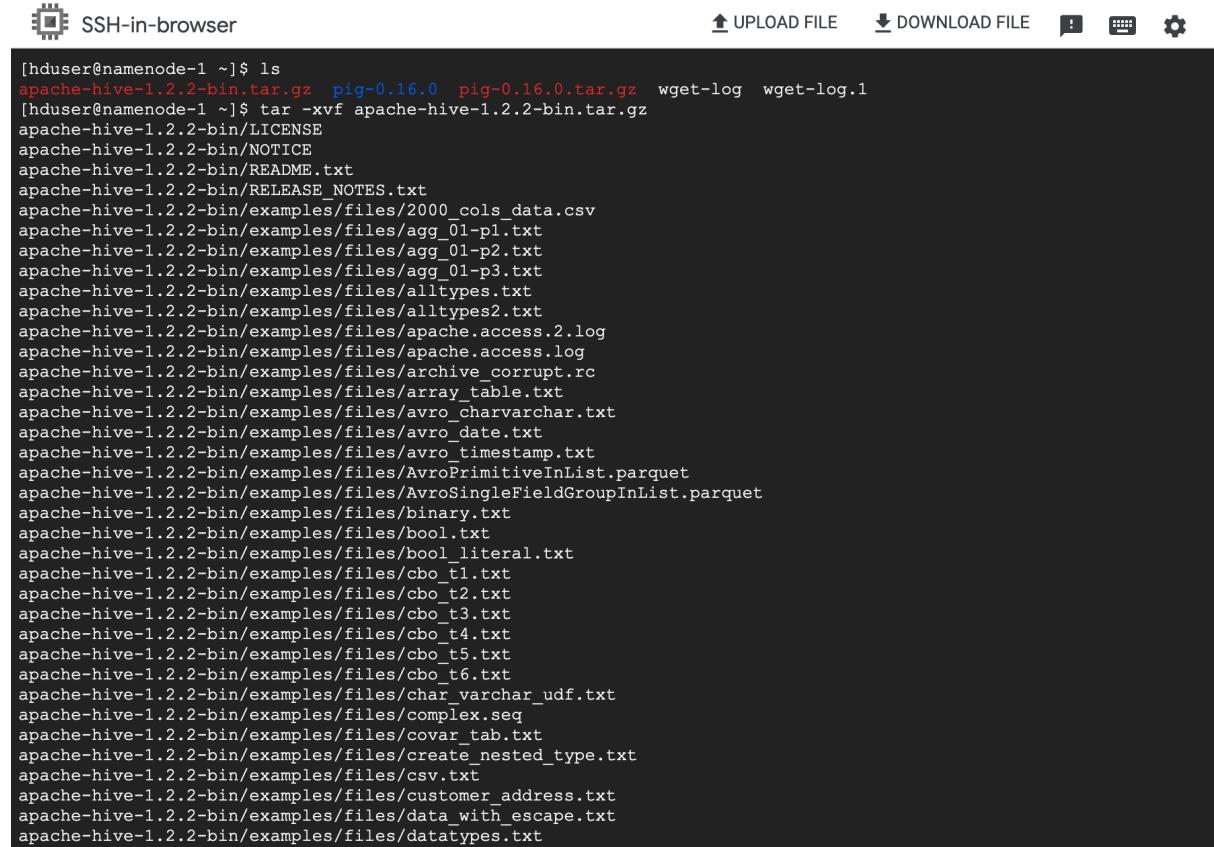
Figure 3: Installation of Hadoop

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
```



```
[hduser@namenode-1 ~]$ ls
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 ~]$ tar -xvf apache-hive-1.2.2-bin.tar.gz
apache-hive-1.2.2-bin/LICENSE
apache-hive-1.2.2-bin/NOTICE
apache-hive-1.2.2-bin/README.txt
apache-hive-1.2.2-bin/RELEASE_NOTES.txt
apache-hive-1.2.2-bin/examples/files/2000_cols_data.csv
apache-hive-1.2.2-bin/examples/files/agg_01-p1.txt
apache-hive-1.2.2-bin/examples/files/agg_01-p2.txt
apache-hive-1.2.2-bin/examples/files/agg_01-p3.txt
apache-hive-1.2.2-bin/examples/files/alltypes.txt
apache-hive-1.2.2-bin/examples/files/alltypes2.txt
apache-hive-1.2.2-bin/examples/files/apache.access.2.log
apache-hive-1.2.2-bin/examples/files/apache.access.log
apache-hive-1.2.2-bin/examples/files/archive_corrupt.rc
apache-hive-1.2.2-bin/examples/files/array_table.txt
apache-hive-1.2.2-bin/examples/files/avro_charvarchar.txt
apache-hive-1.2.2-bin/examples/files/avro_date.txt
apache-hive-1.2.2-bin/examples/files/avro_timestamp.txt
apache-hive-1.2.2-bin/examples/files/AvroPrimitiveInList.parquet
apache-hive-1.2.2-bin/examples/files/AvroSingleFieldGroupInList.parquet
apache-hive-1.2.2-bin/examples/files/binary.txt
apache-hive-1.2.2-bin/examples/files/bool.txt
apache-hive-1.2.2-bin/examples/files/bool_literal.txt
apache-hive-1.2.2-bin/examples/files/cbo_t1.txt
apache-hive-1.2.2-bin/examples/files/cbo_t2.txt
apache-hive-1.2.2-bin/examples/files/cbo_t3.txt
apache-hive-1.2.2-bin/examples/files/cbo_t4.txt
apache-hive-1.2.2-bin/examples/files/cbo_t5.txt
apache-hive-1.2.2-bin/examples/files/cbo_t6.txt
apache-hive-1.2.2-bin/examples/files/char_varchar_udf.txt
apache-hive-1.2.2-bin/examples/files/complex.seq
apache-hive-1.2.2-bin/examples/files/covar_tab.txt
apache-hive-1.2.2-bin/examples/files/create_nested_type.txt
apache-hive-1.2.2-bin/examples/files/csv.txt
apache-hive-1.2.2-bin/examples/files/customer_address.txt
apache-hive-1.2.2-bin/examples/files/data_with_escape.txt
apache-hive-1.2.2-bin/examples/files/datatypes.txt
```

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
```

```
[hduser@namenode-1 sarath_mukundanadavakkat2]$ exit
exit
[sarath_mukundanadavakkat2@namenode-1 ~]$ su - hduser
Password:
Last login: Mon Nov 13 15:55:26 UTC 2023 on pts/1
[hduser@namenode-1 ~]$ wget http://www-eu.apache.org/dist/pig/pig-0.16.0/pig-0.16.0.tar.gz
--2023-11-13 15:57:34-- http://www-eu.apache.org/dist/pig/pig-0.16.0/pig-0.16.0.tar.gz
Resolving www-eu.apache.org (www-eu.apache.org)... 65.108.131.22, 2a01:4f9:6b:2ecf::1
Connecting to www-eu.apache.org (www-eu.apache.org)|65.108.131.22|:80... connected.
HTTP request sent, awaiting response... 302 Found
Location: https://downloads.apache.org/pig/pig-0.16.0/pig-0.16.0.tar.gz [following]
--2023-11-13 15:57:35-- https://downloads.apache.org/pig/pig-0.16.0/pig-0.16.0.tar.gz
Resolving downloads.apache.org (downloads.apache.org)... 88.99.95.219, 135.181.214.104, 2a01:4f9:3a:2c57::2, ...
Connecting to downloads.apache.org (downloads.apache.org)|88.99.95.219|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 177279333 (169M) [application/x-gzip]
Saving to: 'pig-0.16.0.tar.gz'

100%[=====] 177,279,333 25.0MB/s   in 7.5s

2023-11-13 15:57:43 (22.5 MB/s) - 'pig-0.16.0.tar.gz' saved [177279333/177279333]

[hduser@namenode-1 ~]$ ls
pig-0.16.0.tar.gz
[hduser@namenode-1 ~]$ tar -xvf pig-0.16.0.tar.gz
pig-0.16.0/
pig-0.16.0/bin/
pig-0.16.0/conf/
pig-0.16.0/contrib/
pig-0.16.0/contrib/piggybank/
pig-0.16.0/contrib/piggybank/java/
pig-0.16.0/contrib/piggybank/java/build/
pig-0.16.0/contrib/piggybank/java/build/classes/
pig-0.16.0/contrib/piggybank/java/build/classes/org/
pig-0.16.0/contrib/piggybank/java/build/classes/org/apache/
pig-0.16.0/contrib/piggybank/java/build/classes/org/apache/pig/
pig-0.16.0/contrib/piggybank/java/build/classes/org/apache/pig/piggybank/
pig-0.16.0/contrib/piggybank/java/build/classes/org/apache/pig/piggybank/evaluation/
pig-0.16.0/contrib/piggybank/java/build/classes/org/apache/pig/piggybank/evaluation/datetime/
pig-0.16.0/contrib/piggybank/java/build/classes/org/apache/pig/piggybank/evaluation/datetime/convert/
pig-0.16.0/contrib/piggybank/java/build/classes/org/apache/pig/piggybank/evaluation/datetime/diff/
pig-0.16.0/contrib/piggybank/java/build/classes/org/apache/pig/piggybank/evaluation/datetime/truncate/
pig-0.16.0/contrib/piggybank/java/build/classes/org/apache/pig/piggybank/evaluation/decode/
pig-0.16.0/contrib/piggybank/java/build/classes/org/apache/pig/piggybank/evaluation/math/
```

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

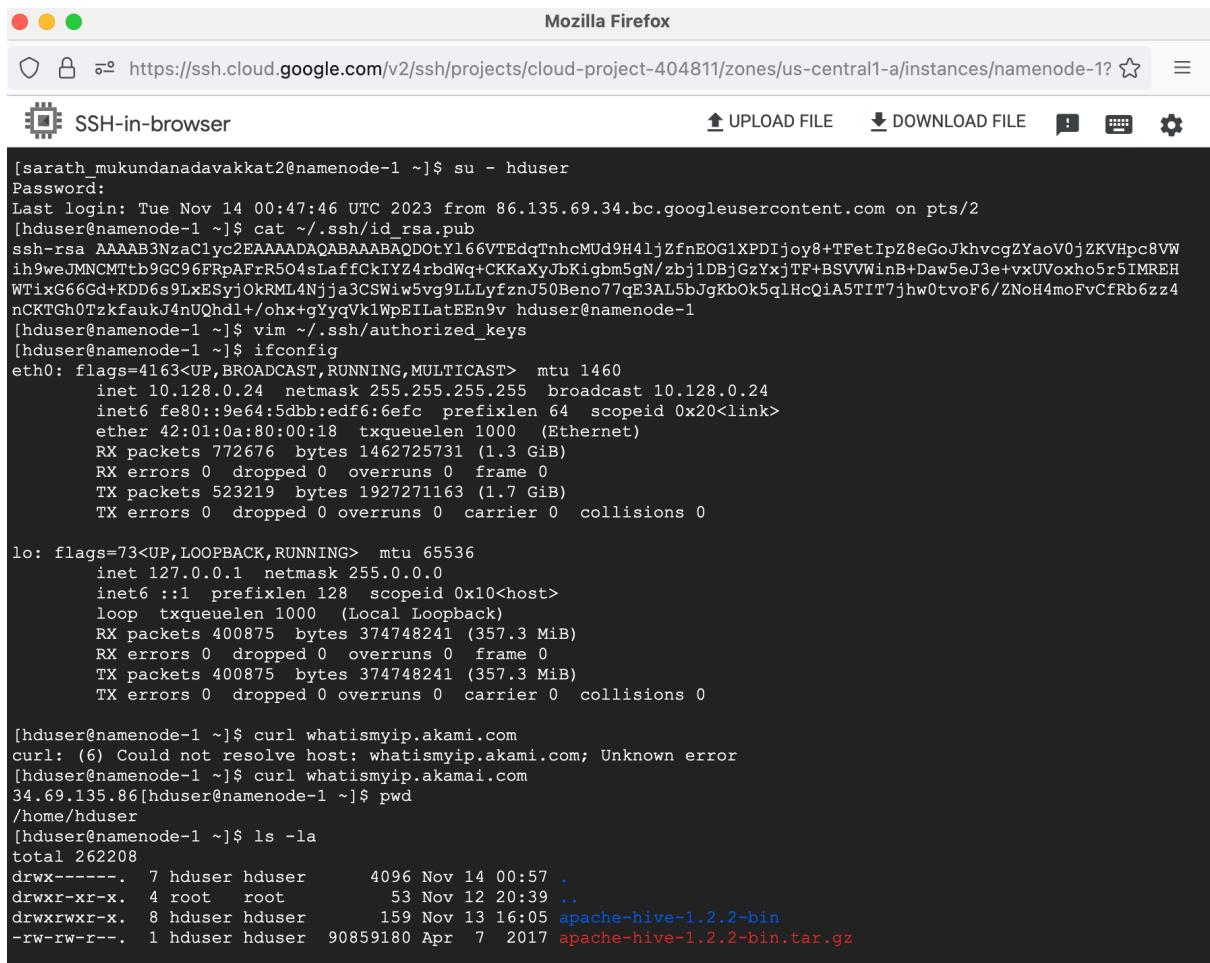
sarath@Saraths-MacBook-Pro ~ % ssh-keygen

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

```



The screenshot shows a Mozilla Firefox window with the title bar "Mozilla Firefox". Below the title bar, the address bar displays the URL "https://ssh.cloud.google.com/v2/ssh/projects/cloud-project-404811/zones/us-central1-a/instances/namenode-1?". The main content area is an iframe with the title "SSH-in-browser". Inside the iframe, a terminal session is running on a Linux system. The terminal output includes:

```

[sarath_mukundanadavakkat2@namenode-1 ~]$ su - hduser
Password:
Last login: Tue Nov 14 00:47:46 UTC 2023 from 86.135.69.34.bc.googleusercontent.com on pts/2
[hduser@namenode-1 ~]$ cat ~/.ssh/id_rsa.pub
ssh-rsa AAAAB3NzaC1yc2EAAAQABAAQD0tY166VTEdqTnhcMUD9H4ljZfnEOG1XPDIjoy8+TFetIpZ8eGoJkhvcgZYaoV0jZKVHpc8VW
ih9weJMNCMTtb9GC96FRpAFrR5O4sLaffCKtYZ4rbdWq+CKKaXyJbKigbm5gN/zbj1DBjGzYxjTF+BSVWWinB+Daw5eJ3e+vxUVoxho5r5IMRH
WTIxG66Gd+KDD6s9LxEsyjOKRML4Njjja3CSwiw5vg9LLLyfznJ50Beno77qE3AL5bJgKbOk5qlHcQiA5TIT7jhw0tvoF6/ZNoH4moFvCrRb6zz4
nCkTGh0TzkfaukJ4nUQhdl+/ohx+gYqVklWpEILatEEEn9v hduser@namenode-1
[hduser@namenode-1 ~]$ vim ~/.ssh/authorized_keys
[hduser@namenode-1 ~]$ ifconfig
eth0: flags=4163<UP,BROADCAST,RUNNING,MULTICAST> mtu 1460
    inet 10.128.0.24 netmask 255.255.255.255 broadcast 10.128.0.24
    inet6 fe80::9e64:5dbb:edf6:6efc prefixlen 64 scopeid 0x20<link>
        ether 42:01:0a:80:00:18 txqueuelen 1000 (Ethernet)
            RX packets 772676 bytes 1462725731 (1.3 GiB)
            RX errors 0 dropped 0 overruns 0 frame 0
            TX packets 523219 bytes 1927271163 (1.7 GiB)
            TX errors 0 dropped 0 overruns 0 carrier 0 collisions 0

lo: flags=73<UP,LOOPBACK,RUNNING> mtu 65536
    inet 127.0.0.1 netmask 255.0.0.0
    inet6 ::1 prefixlen 128 scopeid 0x10<host>
        loop txqueuelen 1000 (Local Loopback)
            RX packets 400875 bytes 374748241 (357.3 MiB)
            RX errors 0 dropped 0 overruns 0 frame 0
            TX packets 400875 bytes 374748241 (357.3 MiB)
            TX errors 0 dropped 0 overruns 0 carrier 0 collisions 0

[hduser@namenode-1 ~]$ curl whatismyip.akami.com
curl: (6) Could not resolve host: whatismyip.akami.com; Unknown error
[hduser@namenode-1 ~]$ curl whatismyip.akamai.com
34.69.135.86[hduser@namenode-1 ~]$ pwd
/home/hduser
[hduser@namenode-1 ~]$ ls -la
total 262208
drwx----- 7 hduser hduser 4096 Nov 14 00:57 .
drwxr-xr-x 4 root root 53 Nov 12 20:39 ..
drwxrwxr-x 8 hduser hduser 159 Nov 13 16:05 apache-hive-1.2.2-bin
-rw-rw-r-- 1 hduser hduser 90859180 Apr 7 2017 apache-hive-1.2.2-bin.tar.gz

```

Figure 7: Load the data into Google Cloud

```

Desktop — hduser@namenode-1:~ -- zsh -- 148x48

Last login: Tue Nov 14 00:27:13 on ttys003
/Users/sarath/.zprofile:5: unmatched "
(base) sarath@Saraths-MacBook-Pro ~ % ssh-keygen
Generating public/private rsa key pair.
Enter file in which to save the key (/Users/sarath/.ssh/id_rsa):
/Users/sarath/.ssh/id_rsa already exists.
Overwrite (y/n)? y
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been saved in /Users/sarath/.ssh/id_rsa
Your public key has been saved in /Users/sarath/.ssh/id_rsa.pub
The key fingerprint is:
SHA256:wSicoijjbC4Y22m34VqGei/uYK16t210FDIGeH6Kr9c sarath@Saraths-MacBook-Pro.local
The key's randomart image is:
+---[RSA 3072]----+
|   ...
|   ...+o.
|   o=..oo.
| = o.o. ..
|o = . o .S
| *oo. ..
|=.*o=o .
| ==+=E
| **B=o.
+---[SHA256]-----+
(base) sarath@Saraths-MacBook-Pro ~ % cat ~/.ssh/id_rsa.pub
ssh-rsa AAAAB3NzaC1yc2EAAAQABgQDmJRYlFFMXgWRbul0nWdAlvW6TNuUGCuRwU1wrJyeFRaB4Ct5CeyuTx1fw9kaVluZfvouju6Vr378khamXgD5eHrsDxqs2pJu13/ft5VaypDx3C
VUb5+wnUNbl40753wad780vdSN6IEtV71HmRh+h2p0hpcR1j1cBFv6yIQ8Hf13ho1MmTSGfb9gb3JnECQYU91ZiM3NjwyLKSd+ZQwt3u0t4/xyHqh3di3cf/PP3SjgzXlbklvB2c010tGLQ3Wh1
1oDWHzxovrzQKqJrn4QQxBES7mcjQUE47AgJle0D3cFgUFZFRAXZhqdN0iktJL1UShhV8bLK6wAKG9N8zdmSpG6bL0st/UJSy1EtIhJ/zKAfp0m48sn1xbbK4AHKSTScbqANDivKlp5zqQvFpP
09q1LBK5ooovQ++KX8HasirP6KN/EoF3tDj1t0ZMu1eMbjiCp0dJbhvDmAr1/AGC0CEF0baFSxSKtjJa6z8zd/db87AruixOc/qjhAZgE= sarath@Saraths-MacBook-Pro.local
(base) sarath@Saraths-MacBook-Pro ~ % ssh hduser@34.69.135.86 -v
OpenSSH_9.4p1, LibreSSL 3.3.6
debug1: Reading configuration data /etc/ssh/ssh_config
debug1: /etc/ssh/ssh_config line 21: include /etc/ssh/ssh_config.d/* matched no files
debug1: /etc/ssh/ssh_config line 54: Applying options for *
debug1: Authenticator provider $SSH_SK_PROVIDER did not resolve; disabling
debug1: Connecting to 34.69.135.86 [34.69.135.86] port 22.
debug1: Connection established.
debug1: identity file /Users/sarath/.ssh/id_rsa type 0
debug1: identity file /Users/sarath/.ssh/id_rsa-cert type -1
debug1: identity file /Users/sarath/.ssh/id_ecdsa type -1
debug1: identity file /Users/sarath/.ssh/id_ecdsa-cert type -1
debug1: identity file /Users/sarath/.ssh/id_ecdsa_sk type -1
debug1: identity file /Users/sarath/.ssh/id_ecdsa_sk-cert type -1
debug1: identity file /Users/sarath/.ssh/id_ed25519 type -1
debug1: identity file /Users/sarath/.ssh/id_ed25519-cert type -1
debug1: identity file /Users/sarath/.ssh/id_ed25519_sk type -1
debug1: identity file /Users/sarath/.ssh/id_ed25519_sk-cert type -1

```

Figure 8: Load the dataset from Local path

```

Mozilla Firefox

SSH-in-browser
[hduser@namenode-1 ~]$ hdfs dfs -ls /user
Found 6 items
-rw-r--r-- 3 hduser supergroup      57438 2023-11-14 01:22 /user/Youtube01-Psy.csv
drwxr-xr-x - hduser supergroup      0 2023-11-13 23:14 /user/bin
drwxr-xr-x - hduser supergroup      0 2023-11-13 23:14 /user/boot
drwxr-xr-x - hduser supergroup      0 2023-11-13 23:14 /user/dev
drwxr-xr-x - hduser supergroup      0 2023-11-13 16:41 /user/hive
drwxr-xr-x - hduser supergroup      0 2023-11-13 22:49 /user/input
[hduser@namenode-1 ~]$ 

```

Figure 9: Dataset stored in hdfs location

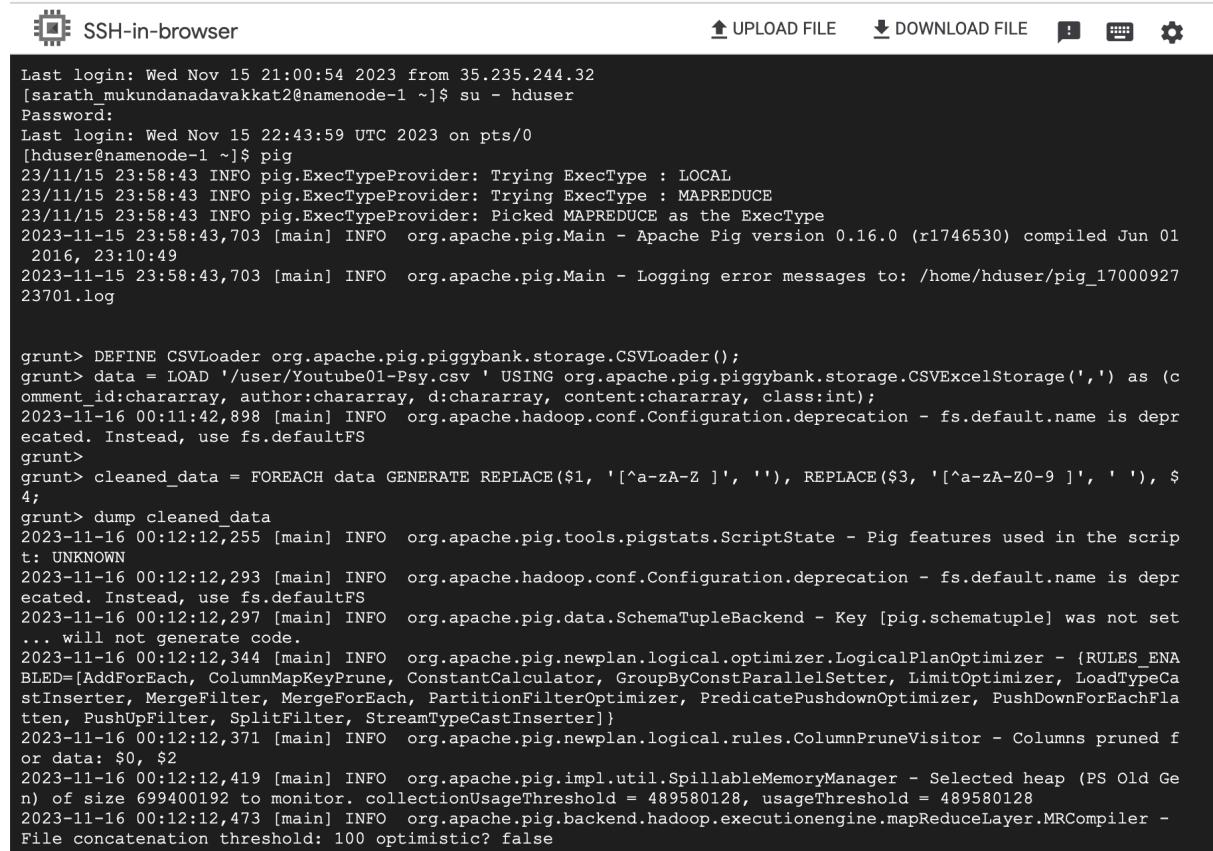
Task 3: Data Cleaning & Processing

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';
```



The screenshot shows a terminal window titled "SSH-in-browser". At the top, there are icons for "UPLOAD FILE", "DOWNLOAD FILE", and other system controls. The terminal displays a series of Pig Latin commands for data cleaning. The commands define a CSVLoader, load data from a CSV file, and then use FOREACH to replace non-alphanumeric characters and non-digit characters in specific fields. Finally, it stores the cleaned data into a new relation. The log output shows the execution of these commands and the resulting error messages related to deprecated configuration settings.

```
Last login: Wed Nov 15 21:00:54 2023 from 35.235.244.32  
[sarath_mukundanadavakkat2@namenode-1 ~]$ su - hduser  
Password:  
Last login: Wed Nov 15 22:43:59 UTC 2023 on pts/0  
[hduser@namenode-1 ~]$ pig  
23/11/15 23:58:43 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL  
23/11/15 23:58:43 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE  
23/11/15 23:58:43 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType  
2023-11-15 23:58:43,703 [main] INFO org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01  
2016, 23:10:49  
2023-11-15 23:58:43,703 [main] INFO org.apache.pig.Main - Logging error messages to: /home/hduser/pig_17000927  
23701.log  
  
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);  
2023-11-16 00:11:42,898 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS  
grunt>  
grunt> cleaned_data = FOREACH data GENERATE REPLACE($1, '[^a-zA-Z ]', ''), REPLACE($3, '[^a-zA-Z0-9 ]', ' '), $4;  
grunt> dump cleaned_data  
2023-11-16 00:12:12,255 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: UNKNOWN  
2023-11-16 00:12:12,293 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS  
2023-11-16 00:12:12,297 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set ... will not generate code.  
2023-11-16 00:12:12,344 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFilter, PushUpFilter, SplitFilter, StreamTypeCastInserter]}  
2023-11-16 00:12:12,371 [main] INFO org.apache.pig.newplan.logical.rules.ColumnPruneVisitor - Columns pruned for data: $0, $2  
2023-11-16 00:12:12,419 [main] INFO org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (PS Old Gen) of size 699400192 to monitor. collectionUsageThreshold = 489580128, usageThreshold = 489580128  
2023-11-16 00:12:12,473 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
```

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 ‘Datal’ 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 deprecated. Instead, use fs.defaultFS
2023-11-16 00:54:59,459 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: FILTER
2023-11-16 00:54:59,483 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. 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_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFilter, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2023-11-16 00:54:59,485 [main] INFO org.apache.pig.newplan.logical.rules.ColumnPruneVisitor - Columns pruned for 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.MultiQueryOptimizer - MR plan size before optimization: 1
2023-11-16 00:54:59,491 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - 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 deprecated. Instead, use fs.defaultFS
2023-11-16 00:54:59,506 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2023-11-16 00:54:59,508 [main] INFO org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to the job
2023-11-16 00:54:59,508 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - 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.JobControlCompiler - This job cannot be converted run in-process
2023-11-16 00:54:59,615 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/home/hduser/pig-0.16.0/pig-0.16.0-core-h2.jar to DistributedCache through /tmp/temp312663614/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
epricated. 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.en
abled 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

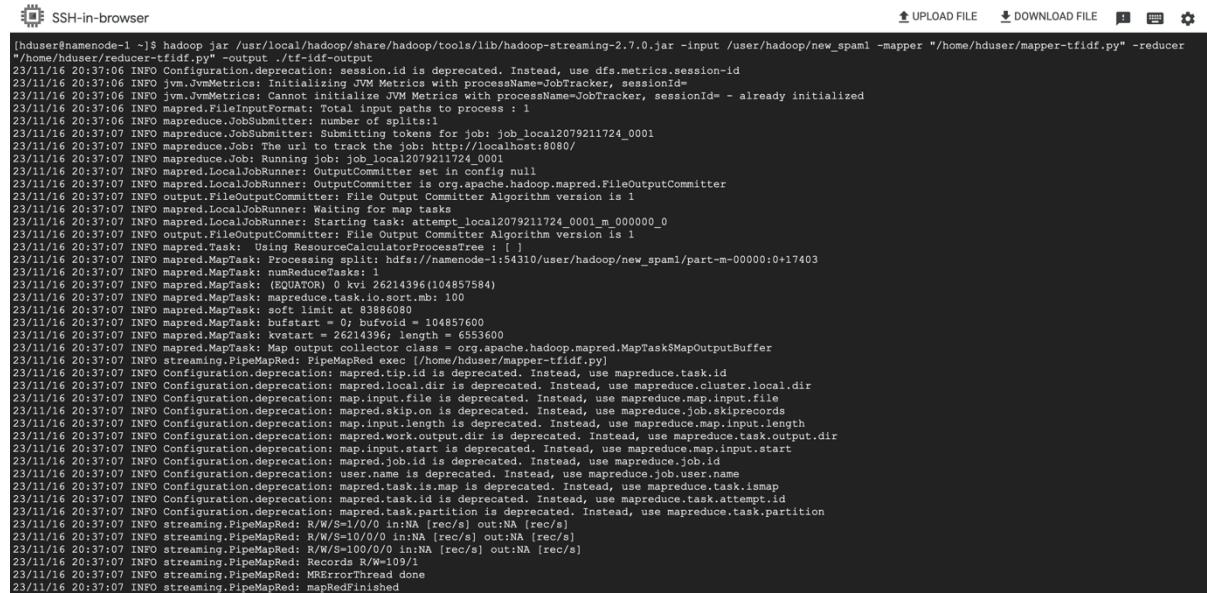
```

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 00:47:06,915 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics w
ith processName=JobTracker, sessionId= - already initialized
2023-11-16 00:47:06,916 [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 00:47:06,918 [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
(MrTuizentfloat,1)
(The Silhouette,1)
(jayson calzado,1)
(Carmen Racasanu,1)
(Alucard Hellising,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



```

[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 Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
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 mapred.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 seen 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 mapred.FileOutputCommitter: InputFormat is org.apache.hadoop.mapred.TextInputFormat
23/11/16 20:37:07 INFO mapred.FileOutputCommitter: LocalJobRunner: Waiting for 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 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: bufferToWriSize = 104857600
23/11/16 20:37:07 INFO mapred.MapTask: bufferToWriOffset = 262143961 length = 655360
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: mapred.output.dir is deprecated. Instead, use mapreduce.map.output.dir
23/11/16 20:37:07 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.jobid
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=199/1
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
```

```
Map-Reduce Framework
  Map input records=109
  Map output records=1947
  Map output bytes=11988
  Map spillover bytes=121888
  Input split bytes=110
  Combine input records=0
  Combine output records=0
  Reduce input groups=109
  Reduce shuffle bytes=121888
  Reduce input records=1947
  Reduce output records=1848
  Spillable Records=3994
  Shuffles=1
  Failed Shuffles=0
  Merged Map outputs=1
  GC time elapsed (ms)=0
  Total committed heap usage (bytes)=601882624
Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  NETWORK_FAILURE=0
  WRONG_MAP=0
  WRONG_REDUCE=0
File Input Format Counters
  Bytes Read=17403
File Output Format Counters
  Bytes Written=87992
23/11/16 20:46:20 INFO streaming.StreamJob: Output directory: ./tfidf-out
[hduser@namenode-1 ~]$ hadoop fs -getmerge /user/hduser/tfidf-out ./tfidf-out
[hduser@namenode-1 ~]$ cat ./tfidf-out | sort -n
sorted user dictionary where users are sorted based on total spam count {' ': 4, 'OutrightIgnite': 2, 'Giang Nguyen': 2, 'adam riyati': 1, 'Goran Theboss': 1, 'TLoUx music': 1, 'J ackal James': 1, 'Niggle Wiggly': 1, 'abdellah chafouai': 1, 'Jaidk Loma': 1, 'Squir': 1, 'Oh s': 1, 'Leonardo Baptista': 1, 'BeBe Burkey': 1, 'Elieo Cardiopulmon ary': 1, 'zakala zikd': 1, 'fab life': 1, 'AnthonyV': 1, 'Ripazha Gaming': 1, 'Ruddy Tapia': 1, 'Photo Editor': 1, 'Daniel Istrati': 1, 'Bizzle Sperg': 1, 'pro stomper': 1, 'Didie r Drogba': 1, 'Mehmet Demirel': 1, 'Dave X': 1, 'Aleksaivan Neider': 1, 'O sbio das eras': 1, 'Ariel Baptista': 1, 'Bishwaroop Bhattacharjee': 1, 'Detroit Red Wings': 1, 'Francis co Nora': 1, 'JackTheLad': 1, 'MrValentiniique': 1, 'Tony K Frazier': 1, 'DylasF5': 1, 'OverSpace': 1, 'derrick lawson': 1, 'Salim Tayara': 1, 'Markus Mairhofer': 1, 'David Boek': 1, 'JD COKE': 1, 'Kiddy Kids': 1, 'Angek': 1, 'Serkac Kaya': 1, 'Ghaz Rizvi': 1, 'CustomerService GM': 1, 'Dovydas Baranauskas': 1, 'Neely Nesley': 1, 'EleptichRage': 1, 'Rancy Gami ng': 1, 'Chinsoma Films': 1, 'MiningBip': 1, 'Adrian Skalak': 1, 'Kemal Kurtoglu': 1, 'Monwar Sarkar': 1, 'MrBrunoExtreme': 1, 'Stuart McDonald': 1, 'Huckyduck': 1, 'The Bibliophi le Plautist': 1, 'Eugene Kalinin': 1, 'DirtySouthFlorida': 1, 'Malin Linford': 1, 'Emerson Zanol Zanol': 1, 'ZodeXID': 1, 'TheHarriiii': 1, 'Wumroque lite': 1, 'KodyMan Dunt Know': 1, 'DontKnowWho': 1, 'Metz Lover': 1, 'Aiss': 1, 'Djedjelien': 1, 'noelend': 1, 'LBBProductions': 1, 'LBBProductions': 1, 'Amine moha': 1, 'Stefano Albanese': 1, 'Carlos Thegamer': 1, 'Arc hie Lewis': 1, 'Cody Tolleson': 1, 'Gaming and Stuff PRO': 1, 'LBEPproductions': 1, 'firo mota': 1, 'Ameenk Chanel': 1, 'Leonel Hernandez': 1, 'Ink Video Shorts': 1, 'proflocopter': 1, 'Lars Zaadstra': 1, 'Cony': 1, 'fad lad': 1, 'Navin Surya': 1, 'Tofik Miedzy': 1, 'Kirsty Brown': 1, 'MineCraftViasat': 1, 'Alessio Siri': 1, 'Lone Twisett': 1, 'MasterRobotTV': 1, 'Uro Slemenjak': 1, 'Patrick Peznia': 1, 'funtimekid': 1, 'ROB YSE': 1, 'Thomas sea': 1, 'JoelR Ch': 1}
```

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

check,LZQPQhLyRh_C2cTtd9MvFRJedxydaVW-2sNg5Diuo4A      2.65853658536585
check,z121st5w5k3ui1veg22zirn4gkr5tby2v    7.785714285714286
check,z12avveb4xqjirsix04chxvilyjrdwuxg0      6.411764705882353
check,z12eexphzo2uslizg04cirmywjzdm5gqc0k      5.45
check,z12r0q3t3hpnb104cdxrzjvmohjyqhs00k      2.7948717948717947
check,z134d5hbckwyblmj0404cgnl03kysfhjsjoeq      4.1923076923076925
check,z13cyzbbsqrsxyfaec23xcl0rdrqqd0ch 3.892857142857143
check,z13phrmwrkfisn5er22eyrbpbvaiwfvwf04      1.3974358974358974
check,z13phrmwrkfisn5er22eyrbpbvaiwfvwf04      1.3974358974358974
check,z13pv52hkmf4jn23g22nx5zqr2gen1gv04      4.541666666666666
check,z13supiartrcdr4la22xc3aripu2xz3a 4.36
check,z13tttljcragexz2o234ghgbgzxyzmzlzi04      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,z12cehoxozfgg3no04cj05xznbgrlpfjo 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,z12cehoxozfgg3no04cj05xznbgrlpfjo 0.42745098039215684
friend,z12kyn0qjzzur2ai04cg5szenjxdrorp4w 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.Job: Job tracking url: http://localhost:8080/
23/11/16 20:58:49 INFO mapreduce.Job: The url to track the job: http://localhost:16381967_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 mapred.LocalJobRunner: 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_local16381967_0001_m_000000_0
23/11/16 20:58:49 INFO mapred.MapTask: File Output Committer Algorithm version is 1
23/11/16 20:58:49 INFO mapred.MapTask: Processing split: hdfs://namenode-1:54310/user/hadoop/new_ham1/part-m-000000:0+6971
23/11/16 20:58:49 INFO mapred.MapTask: numReduceTasks: 1
23/11/16 20:58:49 INFO mapred.MapTask: (EQUATOR) 0 kv1 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 = 655360
23/11/16 20:58:49 INFO mapred.MapTask: collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
23/11/16 20:58:49 INFO streaming.PipeMapRed: PipeMap red exec [/home/hduser/mapper-tfidf.py]
23/11/16 20:58:49 INFO Configuration.deprecation: mapred.task.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: Record: 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 sorted based on total ham count ('Vitor Belliny': 1, 'crestpee': 1, 'Warrdrew': 1, 'Bob Kanowski': 1, 'Mia Aspinall': 1, 'Romeo Sweeney': 1, 'Lat ha hpuc': 1, 'Agarra Mela': 1, 'Yabbatma 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, 'Brando Wilson': 1, 'Tom Hosford': 1, 'King uzy': 1, 'khir abgari': 1, 'Alucard Hellsing': 1, 'Adam Mudd': 1, 'richardex': 1, 'itzquiffer': 1, 'DropShotSkr': 1, 'Zielimeek': 1, 'Digital Media Butterfly': 1, 'Diana Roque': 1, 'Praise Samuel': 1, 'Gaming Land': 1, 'Milan George': 1, 'Wert Wallheet': 1, 'Oopsthenameistoolong Oh well': 1, 'Young Marvin': 1, 'The Silent Troll Defuser HD': 1, 'Carmen Racasanu': 1, 'Kochos': 1, 'Haley Harnicar': 1, 'rolle': 1, 'Emily Hamilton': 1, 'Thanh Phong': 1)
```

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

```
750,z121yttbfpxw1dya04cgtq4clasebvoib4 2.125
```

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

```
views,z121yttbfpxw1dya04cgtq4clasebvoib4      2.125
views,z121znj1loyctjf2233dfo5esfggn3lj 2.55000000000000003
views,z122srjbvovwzxg0522ydl14ty3kznxqs04      4.25
views,z122wfnzgt30fhuhn04cdn3xfx2mxzngs140k 2.217391304347826
views,z122z5pa2wyofbjj304cfgwrrmvjgn0pohc 5.6666666666666666
views,z123gh5olpibqntf22einrisrafashna104     8.5
views,z123gnabnwbqtrle022jeiji0zzzez2os04 2.3181818181818183
views,z124elbgazqmuhvs5230ut14uta5irenh 2.3181818181818183
views,z124zbcigmfbvhpru04cepdx2vikibhp42c 1.9615384615384617
views,z12axnj5w2axxht522thb3bktvqjdlbp04 2.55000000000000003
views,z12axnj5w2axxht522thb3bktvqjdlbp04 2.55000000000000003
views,z12dixvi2gepuhba04chpd5exjdx3oqdic 5.1000000000000005
views,z12dzlpo0ueie1go404cfjjwsxf1glrltdo 3.642857142857143
views,z12gxc5hzkayhbmve23ghd3plt2czf2rp04 2.5500000000000003
views,z12ifxrkbmaechwtt22jwrygmoaehipf04 0.9807692307692308
views,z12ifxrkbmaechwtt22jwrygmoaehipf04 0.9807692307692308
views,z12ifxrkbmaechwtt22jwrygmoaehipf04 0.9807692307692308
views,z12iubmp2mv1txfsy235vb2xdqirufcp304 1.8888888888888888
views,z12ivx14lye5szr4k04cdndhuuyjtfhz4v00k 2.217391304347826
views,z12jyltw0unnnhdh23yyvhjsnljxhf4 1.2750000000000001
views,z12kttwqz14fd0ei23rdp4xjt2ef5hbk04 1.8214285714285714
views,z12kttwqz14fd0ei23rdp4xjt2ef5hbk04 1.8214285714285714
views,z12nt1cqht2byjewi04cfllup0xjvs51q3mc0k 10.200000000000001
views,z12oe5p1mwwugp04cgjfqi3xsxr0lnk 2.4285714285714284
views,z12ox1zh4jicd2zu04cfgfabqtipf3q4is 1.3421052631578947
views,z12pcbng1m2sh1gkp23kilizwkreylid3 0.9272727272727272
views,z12rgj1zmnwfuba404ch3njgy2fg3ryerg0k 5.1000000000000005
views,z12stpyqaconsercxw04cjbr4dlrxnd34sso0k 5.6666666666666666
views,z12ttjopmfst1gpp04cc5ezywjrwntjrc0 3.4
views,z12uyhzaxzkbrzz04cedqpkzwdvjyy3u40k 3.642857142857143
views,z12yipznrreuz3gyf04ce11xdvmquxzdo00k 1.59375
views,z12yipznrreuz3gyf04ce11xdvmquxzdo00k 1.59375
views,z130dbmz2ourjtupz04chhjrpunwcr4yjrs0k 4.25
views,z130gviqarmshdnau04cdzivgs3jepx4qwh00k 2.04
views,z130xbcfwnj5vlskv23airsxqfqvh15504 1.7
```

```
views,z13lip25arigch3rj04cff1ko0pncczyrtng0k 6.375
views,z13nvr2xayrwoffsio04cj3zwuf3vblimdg 3.642857142857143
views,z13nw3lght2nf5we04cdlx5iyadzrnve0 3.9230769230769234
views,z13pfnruxyelivdzo04cgdx4esa2dxeymck 3.4
views,z13pitkr5prbgf3ja04cjjg4qme3txjpyqc0k 1.7
views,z13th1q4yzihf1b1l23qxpjieujtrydj 2.8333333333333333
views,z13ucxa1lrnftvu2j22if5oihq2qip1lp 3.1875
views,z13ucxdzemugilv5n04ccj1oko25drfb4js 1.9615384615384617
views,z13vx3kbqm5fnlgj04cfdoqtpfw5xqzuc0k 1.7
views,z13wfr2rbwytwhf123cs115l2aefdex 3.9230769230769234
views,z13xgr1w3oeycnfq04cihybtimi3vgc40 7.285714285714286
```

```
generating tf-idf for word 2 for user Vitor Belliny
```

```
2,z121yttbfpxw1dya04cgtq4clasebvoib4 1.0625
2,z122z5pa2wyofbjj304cfgwrrmvjgn0pohc 2.8333333333333333
2,z122z5pa2wyofbjj304cfgwrrmvjgn0pohc 2.8333333333333333
2,z123gnabnwbqtrle022jeiji0zzzez2os04 1.1590909090909092
2,z124elbgazqmuhvs5230ut14uta5irenh 1.1590909090909092
2,z12iubmp2mv1txfsy235vb2xdqirufcp304 0.9444444444444444
2,z12jyltw0unnnhdh23yyvhjsnljxhf4 0.6375000000000001
2,z12kttwqz14fd0ei23rdp4xjt2ef5hbk04 0.9107142857142857
2,z12yipznrreuz3gyf04ce11xdvmquxzdo00k 0.796875
2,z130xbcfwnj5vlskv23airsxqfqvh15504 0.85
2,z130xbcfwnj5vlskv23airsxqfqvh15504 0.85
2,z13dtz1zkgadromt230g5qfqsejstr3p 0.6891891891891893
2,z13fgt5wtsf3znibb04crgfgasztuv5gwk0k 1.7
2,z13jzrl51zb4cmfms04chbrbukncfhzxy40 1.1590909090909092
2,z13lip25arigch3rj04cff1ko0pncczyrtng0k 3.1875
2,z13nw3lght2nf5we04cdlx5iyadzrnve0 1.9615384615384617
2,z13th1q4yzihf1b1l23qxpjieujtrydj 1.4166666666666666
2,z13ucxdzemugilv5n04ccj1oko25drfb4js 0.9807692307692308
2,z13vx3kbqm5fnlgj04cfdoqtpfw5xqzuc0k 0.85
2,z13wfr2rbwytwhf123cs115l2aefdex 1.9615384615384617
2,z13xgr1w3oeycnfq04cihybtimi3vgc40 3.642857142857143
```

```
generating tf-idf for word 126 for user Vitor Belliny
```

```
126,z121yttbfpxw1dya04cgtq4clasebvoib4 2.125
```

```
[hduser@namenode-1 ~]$ █
```

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