Task 4: Spam & Ham Detection

First all load the dataset of spam words named as mySpamBag and myHamBag through manually uploading option on terminal SSH and then log onto hive for pre-processing and then finding the top 10 spam and ham accounts.

#Since at task 3 I used Pig, so for spam and ham detection I'll be using Hive. For this we will first create a database by the following command.

1 >>> create database dbElectronics_task4;

Since we will be using the same database, so another command to enable would be required.

2 >>> use database dbElectronics_task4

#Then we will create a new table, which will be based on the data we cleaned at the task 3 level. The data file can be accessed from the local master VM disk. To create table and link that table with the database location following command shall be used:

```
3>>> CREATE external TABLE IF not exists ElectronicsReviews (ID
string, reviewerID string, asin string, reviewerName string,helpful
string, reviewText string,overall string, summary string) ROW FORMAT
SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde' WITH
SERDEPROPERTIES ("separatorChar"= ",", "quoteChar" = "\"") LOCATION
'/home/hamza_qadeer2/MergeCleanElectronics.csv'
tblproperties("skip.header.line.count"="1");
```

#After this we will load the dataset by

```
4>>> Load data local inpath '/home/hamza_qadeer2/
MergeCleanElectronics.csv' overwrite into table ElectronicsReviews;
```

4.1: 10 top Spam account

The spam word dataset we uploaded in the previous step will be used to create a new table in Hive by the following query

```
5>>>CREATE external TABLE IF not exists mySpamBag_tb (words string)
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde' WITH
SERDEPROPERTIES ("separatorChar"= ",", "quoteChar" = "\"") LOCATION
'/home/hamza_qadeer2/mySpamBag.csv'
tblproperties("skip.header.line.count"="1");
```

#After this we will load the dataset by

```
6>>>Load data local inpath '/home/hamza_qadeer2/mySpamBag.csv'
overwrite into table mySpamBag_tb;
```

Now we will create two more tables, first one would store words matched by following command while second will store count.

```
7>>>create table matchWord as select word from (select explode
(split(LCASE(REGEXP_REPLACE(reviewText,'[\\p{Punct},\
\p{Cntrl}]','')),' )) as word from
```

```
dbElectronics_task4.electronicsreviews) words;

8>>>create table tb_count as select word, COUNT(*) AS count FROM
(SELECT * FROM matchWord LEFT OUTER JOIN mySpamBag_tb on
(matchWord.word = mySpamBag_tb.words) WHERE words IS NULL)
removestopwords GROUP BY word ORDER BY count DESC, word ASC;
```

To find the top 20 spam words use this query as follows:

```
9>>> select * from tb_count where length(word)>5 order by count desc
limit 20;
```

Then create two new tables linked with your database for spam and ham.

```
10>>> CREATE TABLE spamSelected AS select
ID,reviewerID,reviewerName,helpful,overall,summary,reviewText from
dbElectronics_task4.ElectronicsReviews where (LOWER(reviewText) LIKE
'%product%' OR LOWER(reviewText) LIKE '%quality %' OR
LOWER(reviewText) LIKE '%purchased%' OR LOWER(reviewText) LIKE
'%pictures%' OR LOWER(reviewText) LIKE '%battery%' OR
LOWER(reviewText) LIKE '%computer%' OR UPPER(reviewText) LIKE
'%PRODUCT%' OR UPPER(reviewText) LIKE '%QUALITY%' OR
UPPER(reviewText) LIKE '%PURCHASED%' OR UPPER(reviewText) LIKE
'%PICTURES%' OR UPPER(reviewText) LIKE '%BATTERY%' OR
UPPER(reviewText) LIKE '%COMPUTER%');

11>>>CREATE TABLE hamSelected AS select
ID,reviewerID,reviewerName,helpful,overall,summary,reviewText from
dbElectronics_task4.ElectronicsReviews where (LOWER(reviewText) LIKE
'%give%' OR LOWER(reviewText) LIKE '%download%' OR LOWER(reviewText)
LIKE '%link%' OR LOWER(reviewText) LIKE '%good%' OR
UPPER(reviewText) LIKE '%GIVE%' OR UPPER(reviewText) LIKE
'%DOWNLOAD%' OR UPPER(reviewText) LIKE '%LINK%' OR UPPER(reviewText)
LIKE '%GOOD%');
```

**Query to find top 10 spam account:**

```
12>>> select id,overall,helpful,reviewername,helpful from
dbElectronics_task4.spamSelected order by summary desc limit 10;
```

**Query to find top 10 ham account:**

```
13>>> select id,overall,helpful,reviewername,helpful from
dbElectronics_task4.hamSelected order by summary desc limit 10;
```