ASSIGNMENT-11.3

Write a Pig UDF to find the tweets posted by @realDonaldTrump from the given dataset.

DATA SET DESCRIPTION -

Column 1 : id

Column 2 : handle

Column 3 : text

Column 4 : is\_retweet

Column 5 : original\_author

Column 6 : time

Column 7 : in\_reply\_to\_screen\_name

Column 8 : in\_reply\_to\_status\_id

Column 9 : in\_reply\_to\_user\_id

Column 10 : is\_quote\_status

Column 11 : lang

Column 12 : retweet\_count

Column 13 : favorite\_count

Column 14 : Longitude

Column 15 : Latitude

Column 16 : Place\_id

Column 17 : Place\_full\_name

Column 18 : Place\_name

Column 19 : Place\_type

Column 20 : Place\_country\_code

Column 21 : Place\_country

Column 22 : Place\_contained\_within

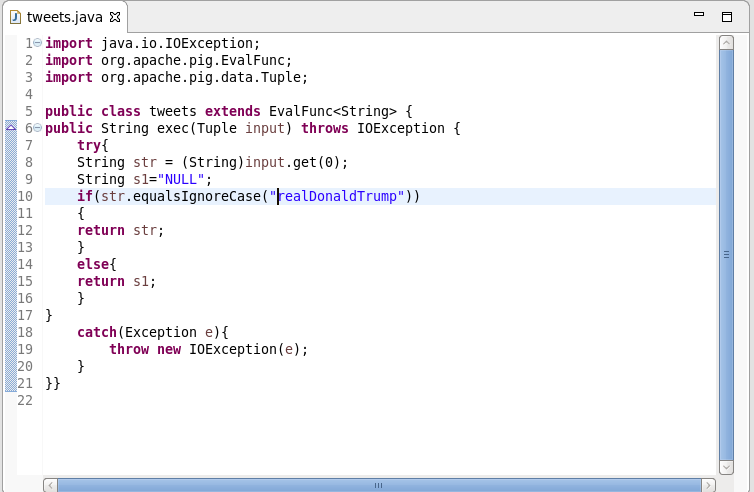
Column 23 : Place\_attributes

Column 24 : Place\_bounding\_box

Column 25 : Source\_url

Column 26 : Truncated

CODE :



PIG CODE :

REGISTER ‘/home/acadgild/Desktop/udf.jar’;

DEFINE tweets tweets();

A1 = LOAD ‘tweets.csv’ USING PigStorage (‘,’) AS (id:chararray,handle:chararray,text:chararray,is\_retweet:chararray,original\_author:chararray,time:chararray, in\_reply\_to\_screen\_name:chararray,in\_reply\_to\_status\_id:chararray, in\_reply\_to\_user\_id:chararray , is\_quote\_status:chararray,lang:chararray, retweet\_count:chararray,favorite\_count:chararray,Longitude:chararray,Latitude:chararray, Place\_id:chararray, Place\_full\_name:chararray, Place\_name:chararray, Place\_type:chararray, Place\_country\_code:chararray, Place\_country:chararray,Place\_contained\_within:chararray, Place\_attributes:chararray, Place\_bounding\_box:chararray, Source\_url:chararray, Truncated:chararray;

A2 = FOREACH A1 GENERATE handle;

A3 = FILTER A2 BY handle IS NOT NULL;

A4 = FOREACH A3 GENERATE tweets(handle) as handle1;

A5 = FILTER A4 by handle1 IS NOT NULL;

A6 = GROUP A5 by handle1;

A7 = FOREACH A6 GENERATE COUNT(E.handle1);

DUMP A7;

OUTPUT :

