Group 2

Depressive text classification.

26th March 2021

TEAM MEMBERS

- Davith Lon
- Bryan Khoo
- Ami Khalsa
- Ashish Pant
- Saurav Pawar

Project Description

Introduction

With a global pandemic in place, everyone is advised to be isolated to prevent the further spread of COVID. Along with schools, work places, and restaurants closing down, people have no choice but to stay sheltered with minimal interaction. This builds up frustration and depressive thoughts. Fortunately, with the advancements of technology, we are able to express our emotions on the internet. Twitter is one of the social media platforms that allows users to express their thoughts[1].

However, because of the popularity of Twitter, it may be easy for “tweets” that are seeking help to get lost in the sea of data. In this project, we will be building a text classifier that will take in a text and give a rating on how depressive that particular text is. We believe that with this classifier, it will be easier to identify those “tweets” who are depressive and provide appropriate help to the user.

[1] https://esrc.ukri.org/research/impact-toolkit/social-media/twitter/what-is-twitter/

Background

Twitter is widely used for research. One of the research done recently was on migraine tweets[2]. The methods used in this research, such as natural language processing, will be observed and examined to verify whether they are appropriate for our project.

This increment is an extension of the previous increment. It will include data extraction, data cleaning, analysis and visualization of the analysis.

GOALS AND OBJECTIVES

Motivation

Depression is “a common and serious medical illness that negatively affects how you feel, the way you think and how you act.”[3] The symptoms of depression ranges from changes in appetite, loss of interest or pleasure in activities once enjoyed, thoughts of death or suicide, and etc.[4] Leaving these symptoms unidentified and untreated has unfortunately claimed the lives of 800,000 people every year.[5] It also has lasting impacts on education. High school students with recent symptoms of depression are more than twice as likely as their peers to drop out.[5] This is troubling, because during the COVID-19 Pandemic there has been an elevated amount of adverse mental health conditions with depression being the lead condition. This is why our project wants to investigate the significance of variables that cause depression to create a better understanding of their impacts.

Significance

The following are the reasonings for the significance of this project.

1. The project can help health organizations to determine whether a particular text is depressed and take appropriate actions based on the result.
2. A text classification model that can score how depressed a piece of text could be used for monitoring service.
3. The project can help provide infographics that medical organizations can use to educate the public.

[2] https://journals.sagepub.com/doi/full/10.1177/2515816319898867
[3] https://www.psychiatry.org/patients-families/depression/what-is-depression
[4] https://www.nimh.nih.gov/health/topics/depression/index.shtml
[5] https://www.who.int/news-room/fact-sheets/detail/depression
[5]https://pubmed.ncbi.nlm.nih.gov/29195763/

4. “Globally, more than 264 million people of all ages suffer from depression.”[3] So, with researching this topic we’d help understand an issue that impacts a significant amount of people globally.

Objectives

In this project, we are determined to identify a depressed text and give a rating on how depressive that particular text is. By using the different attributes provided in the datasets, we will be able to determine specific patterns and identifying words that may strongly suggest that a given text is depressive, Also, we’ll identify how depressive a text is based on the specific words and patterns in the text.

Features

This project will be able to determine whether there is a relationship between the stages of depression with the “tweets” written by a person. With this finding, we can potentially identify different stages of depression a user might be in depending on their recent “tweets”. However, this project is not to be taken as a professional advice nor should it be used as a main source to identify symptoms of depression.

Dataset

The dataset that we will be using in this project will be a dataset from Kaggle[6]. The dataset contains posts and threads they’re from, r/SuicideWatch or r/depression. This dataset was selected because the posts made on these subreddits were most likely from individuals who are depressed and may be suffering from depression. Therefore, there will be valuable information for us to mine.

This data set only contains the post and what thread it comes from. We will be interested in looking at the distribution of posts from the two reddit threads. We will also be interested in mining information and patterns from the posts themselves.

Additionally, we will also be pulling live data from Twitter using their Twitter API. We are scraping tweets based off of the hashtags included with the posts. We are currently only interested in three hashtags. Those being #depression, #suicidal, and #anxiety. We chose these specific hashtags as they would most likely yield the type of text we would want to analyze.

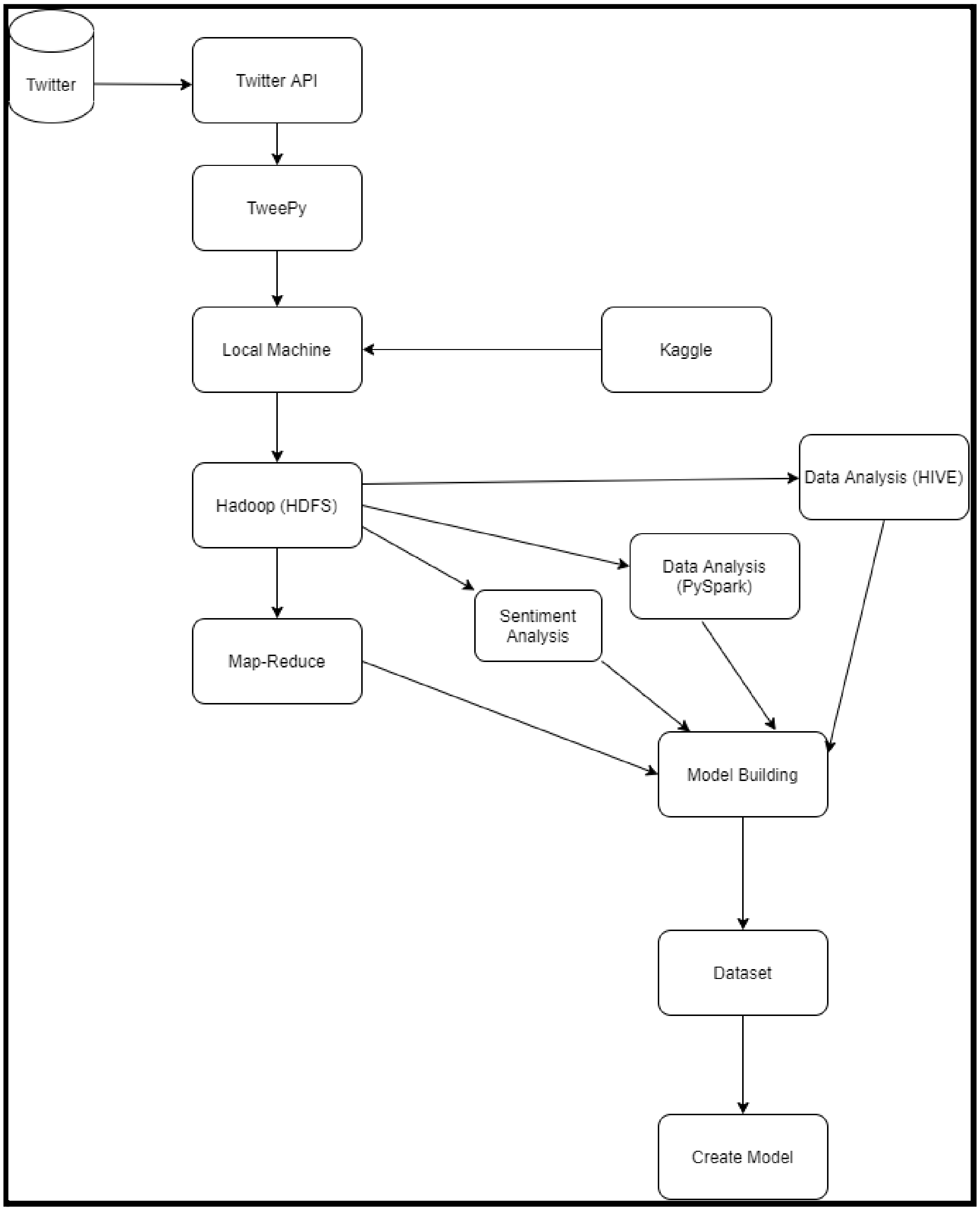
From the Twitter API, we will only be concerned with the full_text, created_at, followers_count, possibly_sensitive, verified, and location.

[6] https://www.kaggle.com/nikhileswarkomati/suicide-watch

Detailed Design of Features

Workflow

Our data starts with scraping Twitter using TweePy and bringing it to our local machines. From there, we push them onto Hadoop for file storage and management. Additionally, we pulled static data from Kaggle to our local machines and then pushed it to Hadoop as well. Then, from Hadoop we pull and use the data for processing, analyzing, and model building.

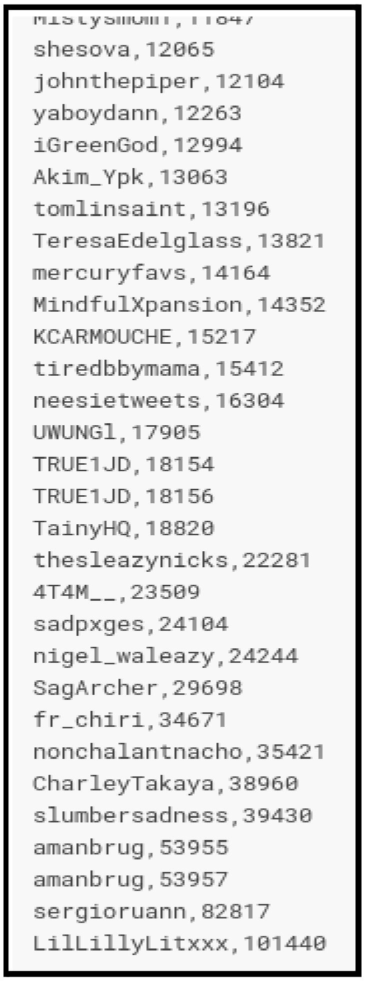


Analysis

1. Most Followed Tweeters (HIVE):

This query look at which users present in the dataset had the most followers. This analytic is important as follower count can provide a means of identifying influential tweeters. Additionally, this could show that an account may be botting followers. The following is the query used and the results.

- INSERT OVERWRITE DIRECTORY '/user/cloudera/Project/Query Results/' ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' SELECT screen_name, followers FROM tweets GROUP BY followers, screen_name;

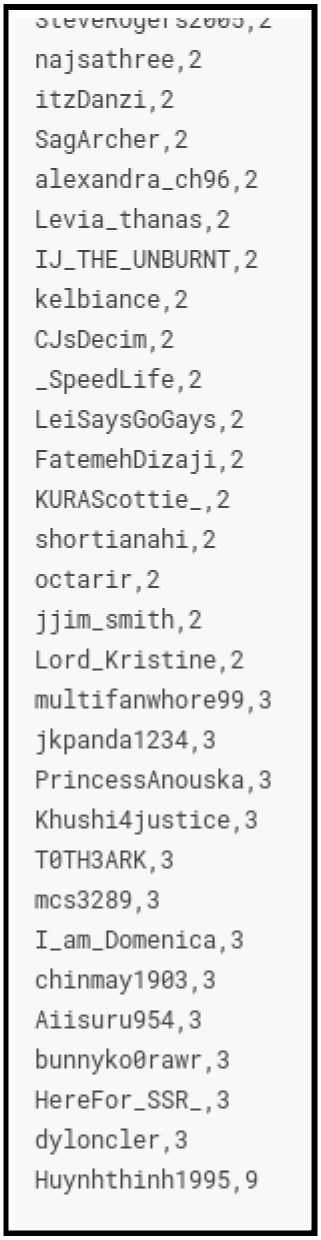


```
MistysMom1,11847
shesova,12065
johnthepiper,12104
yaboydann,12263
iGreenGod,12994
Akim_Ypk,13063
tomlinsaint,13196
TeresaEdelglass,13821
mercuryfavs,14164
MindfulXpansion,14352
KCARMOUCHE,15217
tiredbbymama,15412
neesietweets,16304
UWUNGl,17905
TRUE1JD,18154
TRUE1JD,18156
TainyHQ,18820
thesleazynicks,22281
4T4M__,23509
sadpxges,24104
nigel_waleazy,24244
SagArcher,29698
fr_chiri,34671
nonchalantnacho,35421
CharleyTakaya,38960
slumbersadness,39430
amanbrug,53955
amanbrug,53957
sergioruann,82817
LilLillyLitxxx,101440
```

2. Tweets Per User (HIVE):

This query looked at how many posts from the same users appeared within our data set. This statistic is important as it may be used to indicate high users of Twitter and as well as spam bots. The following are the query used and the results.

- INSERT OVERWRITE DIRECTORY '/user/cloudera/Project/Tweets Per User Results/' ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' SELECT screen_name, count(*) as total_tweets from tweets GROUP BY screen_name HAVING screen_name IS NOT NULL ORDER BY total_tweets;

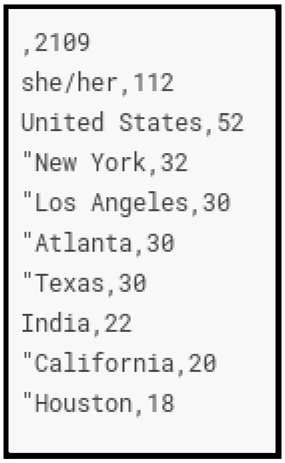


```
SteveRogers2005,2
najsathree,2
itzDanzi,2
SagArcher,2
alexandra_ch96,2
Levia_thanas,2
IJ_THE_UNBURNT,2
kelbiance,2
CJsDecim,2
_SpeedLife,2
LeiSaysGoGays,2
FatemehDizaji,2
KURAScottie_,2
shortianahi,2
octarir,2
jjim_smith,2
Lord_Kristine,2
multifanwhore99,3
jkpanda1234,3
PrincessAnouska,3
Khushi4justice,3
T0TH3ARK,3
mcs3289,3
I_am_Domenica,3
chinmay1903,3
Aiisuru954,3
bunnyko0rawr,3
HereFor_SSR_,3
dyloncler,3
Huynhthinh1995,9
```

3. Top 10 Tweeted Locations (HIVE):

This query allows us to identify the top 10 locations where the majority of our data is from. This information may play a vital role as we will be able to identify the major geological area that has a higher concentration of depressive tweets, which may link to the total number of depression in that location. This is the following query that outputs the desired output:

- SELECT location, count(*) as count FROM tweets GROUP BY location HAVING location is not NULL ORDER BY count DESC LIMIT 10;

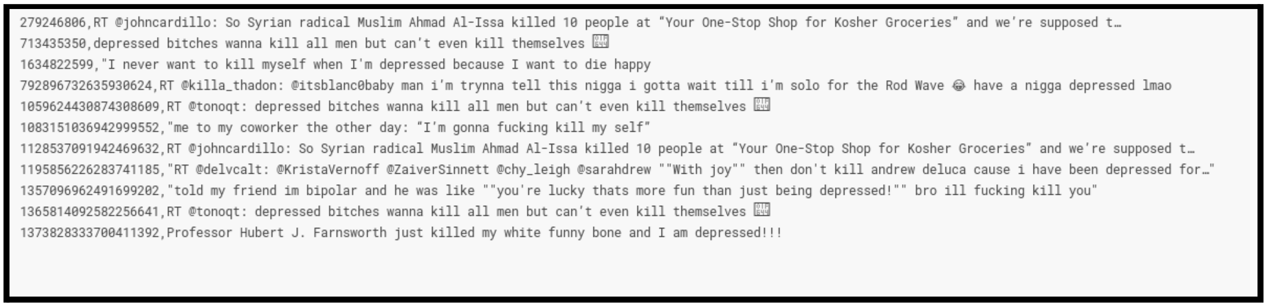


```
,2109
she/her,112
United States,52
"New York,32
"Los Angeles,30
"Atlanta,30
"Texas,30
India,22
"California,20
"Houston,18
```

4. Tweets that Include the Word “kill” (HIVE):

This query shows us the tweets that include the word “kill” and the user id of the tweet. This information collected is vital for us to identify users who may have suicidal thoughts. As the dataset includes a lot of different tweets, this query allows us to pinpoint users who specifically tweeted with a more likely intention of harming others/themselves. Below is the following query:

- SELECT user_id, text FROM tweets WHERE text RLIKE '.*(kill).*' AND text NOT RLIKE '.*(https).*' ORDER BY user_id;

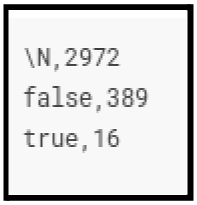


```
279246806,RT @johncardillo: So Syrian radical Muslim Ahmad Al-Issa killed 10 people at “Your One-Stop Shop for Kosher Groceries” and we’re supposed t…
713435350,depressed bitches wanna kill all men but can’t even kill themselves □
1634822599,"I never want to kill myself when I'm depressed because I want to die happy
792896732635930624,RT @killa_thadon: @itsblanc0baby man i’m trynna tell this nigga i gotta wait till i’m solo for the Rod Wave 😂 have a nigga depressed lmao
1059624430874308609,RT @tonoqt: depressed bitches wanna kill all men but can’t even kill themselves □
1083151036942999552,"me to my coworker the other day: “I’m gonna fucking kill my self”
1128537091942469632,RT @johncardillo: So Syrian radical Muslim Ahmad Al-Issa killed 10 people at “Your One-Stop Shop for Kosher Groceries” and we’re supposed t…
1195856226283741185,"RT @delvcalt: @KristaVernoff @ZaiverSinnett @chy_leigh @sarahdrew ""With joy"" then don't kill andrew deluca cause i have been depressed for…"
1357096962491699202,"told my friend im bipolar and he was like ""you're lucky thats more fun than just being depressed!"" bro ill fucking kill you"
1365814092582256641,RT @tonoqt: depressed bitches wanna kill all men but can’t even kill themselves □
1373828333700411392,Professor Hubert J. Farnsworth just killed my white funny bone and I am depressed!!!
```

5. Sensitive Flagged Posts (HIVE):

This query shows us the tweets that were flagged as containing sensitive material. This information can be used to filter out NSFW posts that may not be relevant to our project.

- INSERT OVERWRITE DIRECTORY '/user/cloudera/Project/Sensitive Query/' ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' SELECT possibly_sensitive, count(*) AS total FROM tweets GROUP BY possibly_sensitive;

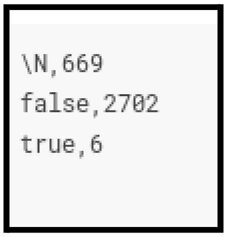


```
\N,2972
false,389
true,16
```

6. Verified Users (HIVE):

This query shows the total users that are either verified or not. Similar to the most followed twitter user query, this information helps us identify influential users.

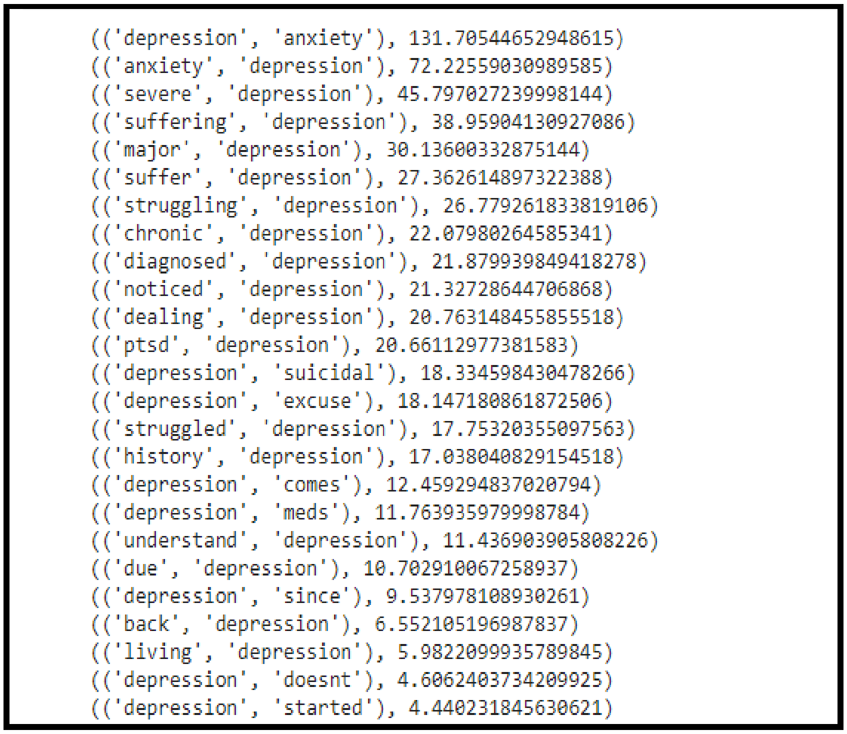
- INSERT OVERWRITE DIRECTORY '/user/cloudera/Project/Verified Query/' ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' SELECT verified, count(*) AS total FROM tweets GROUP BY verified;



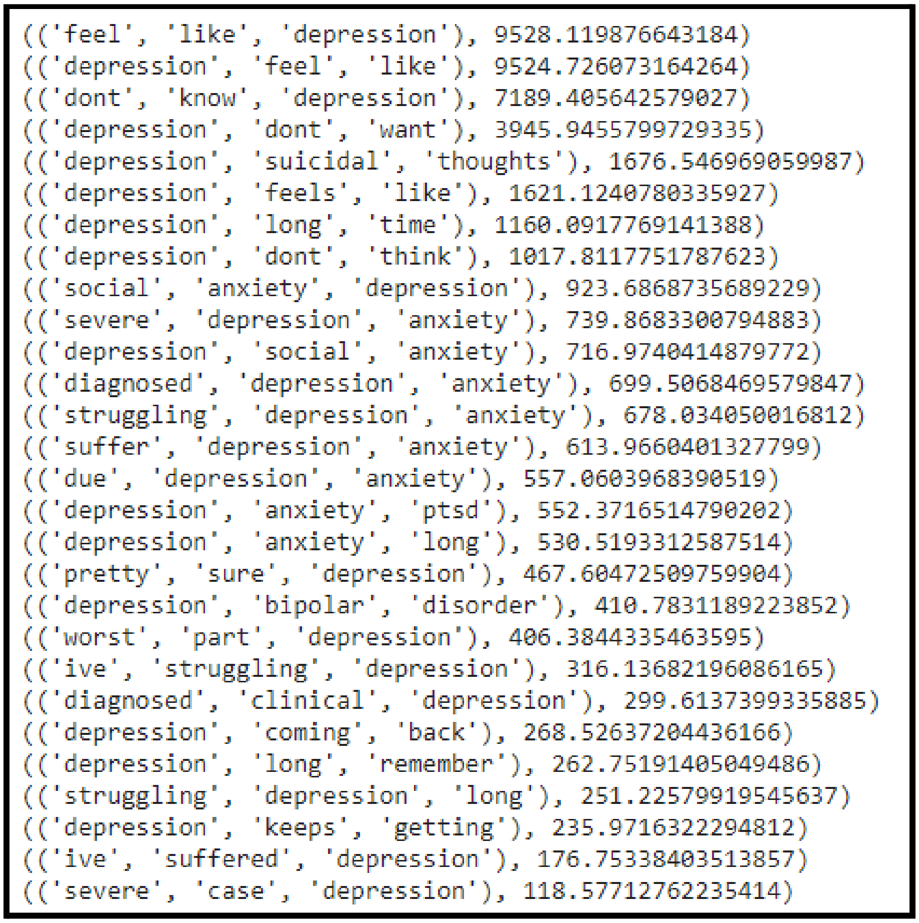
```
\N,669
false,2702
true,6
```

7. Bigrams and Trigrams on “depression” (PySpark)

This analysis looked at pairs of words that were likely to come up together. We then calculated the likelihood ratio for the key word “depression”. With this type of analysis we can see interesting patterns like how the likelihood ratio of “depression” coming before “anxiety” is higher than the vice versa. The following are the top results of performing a Bigram and Trigram on the key word “depression”.



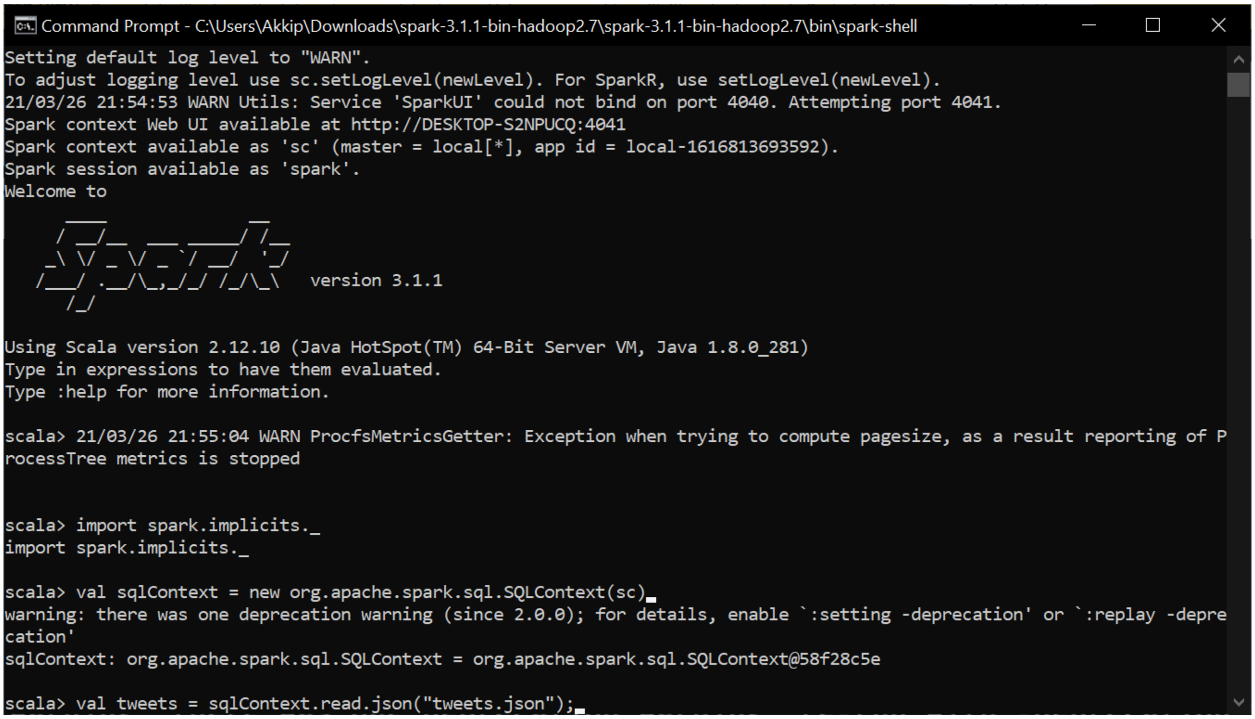
```
(('depression', 'anxiety'), 131.70544652948615)
(('anxiety', 'depression'), 72.22559030989585)
(('severe', 'depression'), 45.797027239998144)
(('suffering', 'depression'), 38.95904130927086)
(('major', 'depression'), 30.13600332875144)
(('suffer', 'depression'), 27.362614897322388)
(('struggling', 'depression'), 26.779261833819106)
(('chronic', 'depression'), 22.07980264585341)
(('diagnosed', 'depression'), 21.879939849418278)
(('noticed', 'depression'), 21.32728644706868)
(('dealing', 'depression'), 20.763148455855518)
(('ptsd', 'depression'), 20.66112977381583)
(('depression', 'suicidal'), 18.334598430478266)
(('depression', 'excuse'), 18.147180861872506)
(('struggled', 'depression'), 17.75320355097563)
(('history', 'depression'), 17.038040829154518)
(('depression', 'comes'), 12.459294837020794)
(('depression', 'meds'), 11.763935979998784)
(('understand', 'depression'), 11.436903905808226)
(('due', 'depression'), 10.702910067258937)
(('depression', 'since'), 9.537978108930261)
(('back', 'depression'), 6.552105196987837)
(('living', 'depression'), 5.9822099935789845)
(('depression', 'doesnt'), 4.6062403734209925)
(('depression', 'started'), 4.440231845630621)
```



```
(('feel', 'like', 'depression'), 9528.119876643184)
(('depression', 'feel', 'like'), 9524.726073164264)
(('dont', 'know', 'depression'), 7189.405642579027)
(('depression', 'dont', 'want'), 3945.9455799729335)
(('depression', 'suicidal', 'thoughts'), 1676.546969059987)
(('depression', 'feels', 'like'), 1621.1240780335927)
(('depression', 'long', 'time'), 1160.0917769141388)
(('depression', 'dont', 'think'), 1017.8117751787623)
(('social', 'anxiety', 'depression'), 923.6868735689229)
(('severe', 'depression', 'anxiety'), 739.8683300794883)
(('depression', 'social', 'anxiety'), 716.9740414879772)
(('diagnosed', 'depression', 'anxiety'), 699.5068469579847)
(('struggling', 'depression', 'anxiety'), 678.034050016812)
(('suffer', 'depression', 'anxiety'), 613.9660401327799)
(('due', 'depression', 'anxiety'), 557.0603968390519)
(('depression', 'anxiety', 'ptsd'), 552.3716514790202)
(('depression', 'anxiety', 'long'), 530.5193312587514)
(('pretty', 'sure', 'depression'), 467.60472509759904)
(('depression', 'bipolar', 'disorder'), 410.7831189223852)
(('worst', 'part', 'depression'), 406.3844335463595)
(('ive', 'struggling', 'depression'), 316.13682196086165)
(('diagnosed', 'clinical', 'depression'), 299.6137399335885)
(('depression', 'coming', 'back'), 268.52637204436166)
(('depression', 'long', 'remember'), 262.75191405049486)
(('struggling', 'depression', 'long'), 251.22579919545637)
(('depression', 'keeps', 'getting'), 235.9716322294812)
(('ive', 'suffered', 'depression'), 176.75338403513857)
(('severe', 'case', 'depression'), 118.57712762235414)
```

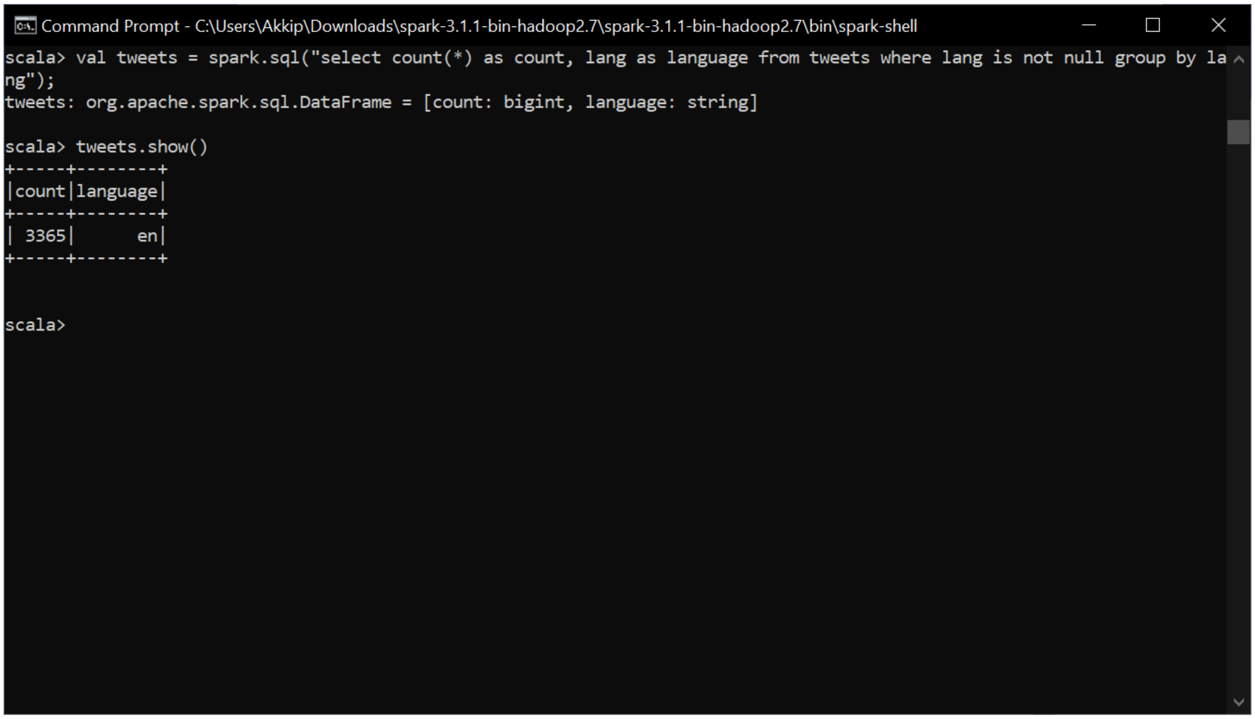
8. Load JSON and Analysis (Spark)

Start the Spark-Shell.


Command Prompt - C:\Users\Akkip\Downloads\spark-3.1.1-bin-hadoop2.7\spark-3.1.1-bin-hadoop2.7\bin\spark-shell

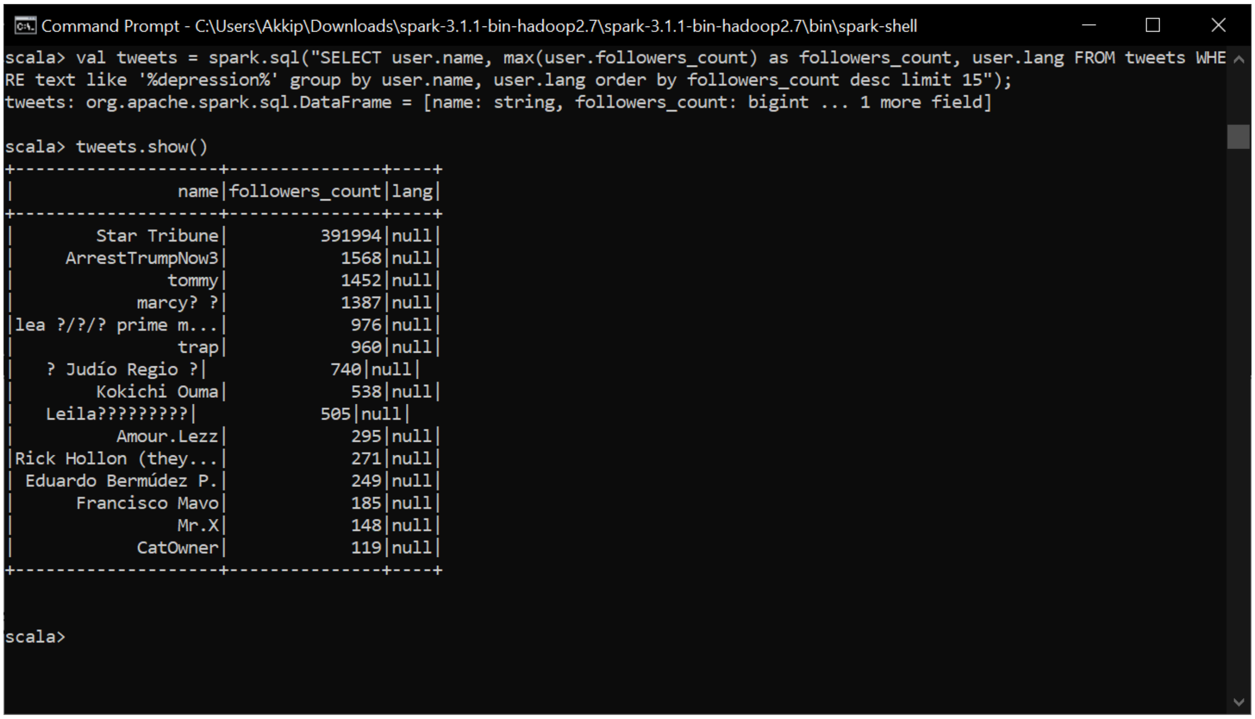
```
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
21/03/26 21:54:53 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.
Spark context Web UI available at http://DESKTOP-S2NPUCQ:4041
Spark context available as 'sc' (master = local[*], app id = local-1616813693592).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 3.1.1
      /_/

Using Scala version 2.12.10 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_281)
Type in expressions to have them evaluated.
Type :help for more information.

scala> 21/03/26 21:55:04 WARN ProcfsMetricsGetter: Exception when trying to compute pagesize, as a result reporting of P
rocessTree metrics is stopped


scala> import spark.implicits._
import spark.implicits._

scala> val sqlContext = new org.apache.spark.sql.SQLContext(sc)
warning: there was one deprecation warning (since 2.0.0); for details, enable `:setting -deprecation' or `:replay -depre
cation'
sqlContext: org.apache.spark.sql.SQLContext = org.apache.spark.sql.SQLContext@58f28c5e

scala> val tweets = sqlContext.read.json("tweets.json");
```

Query1: Show the different language the tweets are in and their count.


Command Prompt - C:\Users\Akkip\Downloads\spark-3.1.1-bin-hadoop2.7\spark-3.1.1-bin-hadoop2.7\bin\spark-shell

```
scala> val tweets = spark.sql("select count(*) as count, lang as language from tweets where lang is not null group by la
ng");
tweets: org.apache.spark.sql.DataFrame = [count: bigint, language: string]

scala> tweets.show()
+-----+--------+
|count|language|
+-----+--------+
| 3365|      en|
+-----+--------+


scala>
```

Query2: Show the users who have most followers and who are talking about depression.

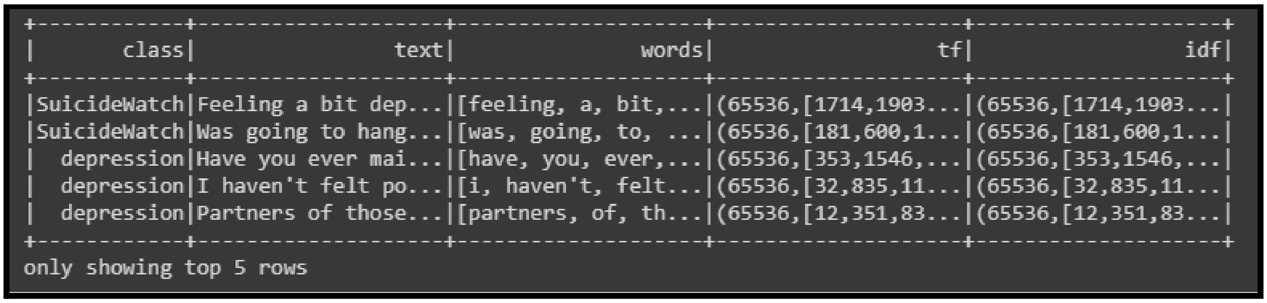

Command Prompt - C:\Users\Akkip\Downloads\spark-3.1.1-bin-hadoop2.7\spark-3.1.1-bin-hadoop2.7\bin\spark-shell

```
scala> val tweets = spark.sql("SELECT user.name, max(user.followers_count) as followers_count, user.lang FROM tweets WHE
RE text like '%depression%' group by user.name, user.lang order by followers_count desc limit 15");
tweets: org.apache.spark.sql.DataFrame = [name: string, followers_count: bigint ... 1 more field]

scala> tweets.show()
+-------------------+--------------+----+
|               name|followers_count|lang|
+-------------------+--------------+----+
|       Star Tribune|        391994|null|
|     ArrestTrumpNow3|          1568|null|
|              tommy|          1452|null|
|          marcy? ?|          1387|null|
|lea ?/?/? prime m...|           976|null|
|               trap|           960|null|
|    ? Judío Regio ?|         740|null|
|       Kokichi Ouma|           538|null|
|    Leila?????????|         505|null|
|          Amour.Lezz|           295|null|
|Rick Hollon (they...|           271|null|
| Eduardo Bermúdez P.|           249|null|
|     Francisco Mavo|           185|null|
|               Mr.X|           148|null|
|           CatOwner|           119|null|
+-------------------+--------------+----+


scala>
```

Some of the problems that I faced working with spark is that I got a lot of null values when loading the data from JSON, so I need to figure that out.

9. TF-IDF (PySpark)

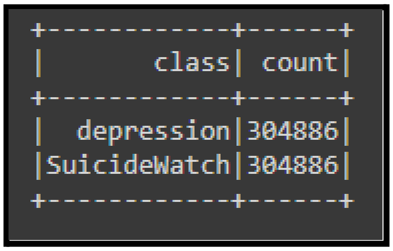
This analysis uses the statistical measure TF-IDF (Term Frequency - Inverse Document Frequency). The TF represents the number of occurrences for a specific term and IDF tells us the log of the total number of documents divided by the number of documents with the term. The product of TF and IDF provides us a numeric value which indicates the level of importance of the particular term. This plays a critical role in further analysis as this method allows us to determine whether we should consider a specific word in our examination or disregard it depending on its significance.



```
+------------+--------------------+--------------------+--------------------+--------------------+
|       class|                text|               words|                  tf|                 idf|
+------------+--------------------+--------------------+--------------------+--------------------+
|SuicideWatch|Feeling a bit dep...|[feeling, a, bit,...|(65536,[1714,1903...|(65536,[1714,1903...|
|SuicideWatch|Was going to hang...|[was, going, to, ...|(65536,[181,600,1...|(65536,[181,600,1...|
|  depression|Have you ever mai...|[have, you, ever,...|(65536,[353,1546,...|(65536,[353,1546,...|
|  depression|I haven't felt po...|[i, haven't, felt...|(65536,[32,835,11...|(65536,[32,835,11...|
|  depression|Partners of those...|[partners, of, th...|(65536,[12,351,83...|(65536,[12,351,83...|
+------------+--------------------+--------------------+--------------------+--------------------+
only showing top 5 rows
```

10. Subreddit Summations (PySpark)

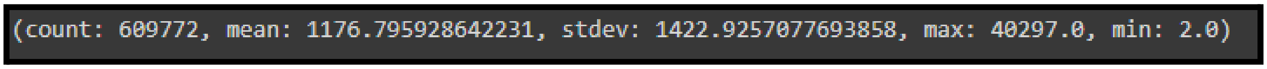
This analysis provides us the total number of texts the dataset contains for each subreddit. This is a crucial detail on which particular subreddit will be having a greater influence on the results than the other. This information will be vital to keep in mind as we progress with deeper analysis on the dataset.



```
+------------+------+
|       class| count|
+------------+------+
|  depression|304886|
|SuicideWatch|304886|
+------------+------+
```

11. Word Count (PySpark)

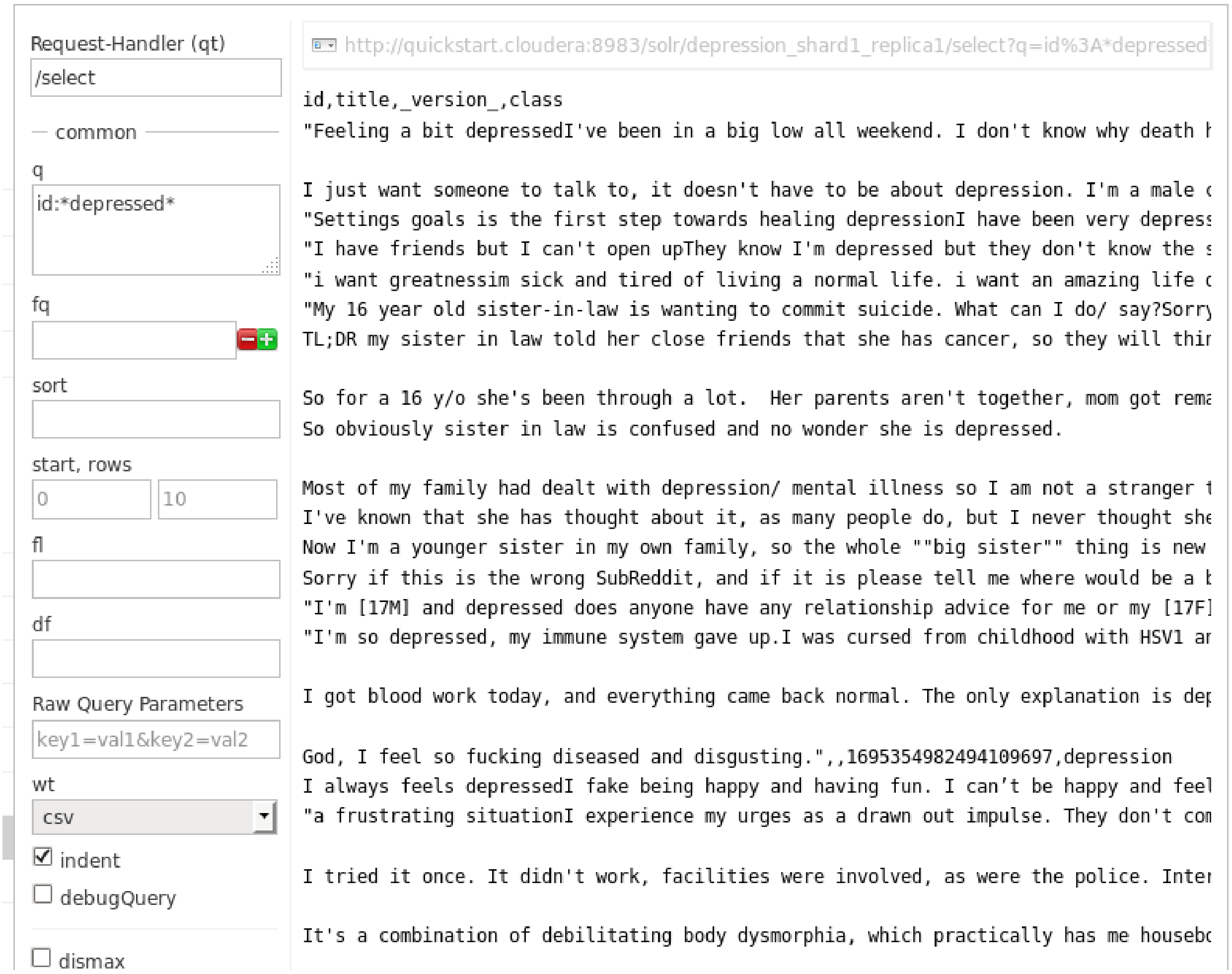
This analysis provides us with basic statistics on the maximum and minimum length for a text in the dataset, the mean length for all texts, total number of texts.



```
(count: 609772, mean: 1176.795928642231, stdev: 1422.9257077693858, max: 40297.0, min: 2.0)
```

12. Solr Queries (Solr)

This helped us run a wildcard search on various words associated with depression and suicide risk. We ran queries on the words depressed and hurt. These returned posts with the words depressed and hurt respectively included in the post.



Request-Handler (qt)

/select

common

q

id:*depressed*

fq

sort

start, rows

0 10

fl

df

Raw Query Parameters

key1=val1&key2=val2

wt

csv

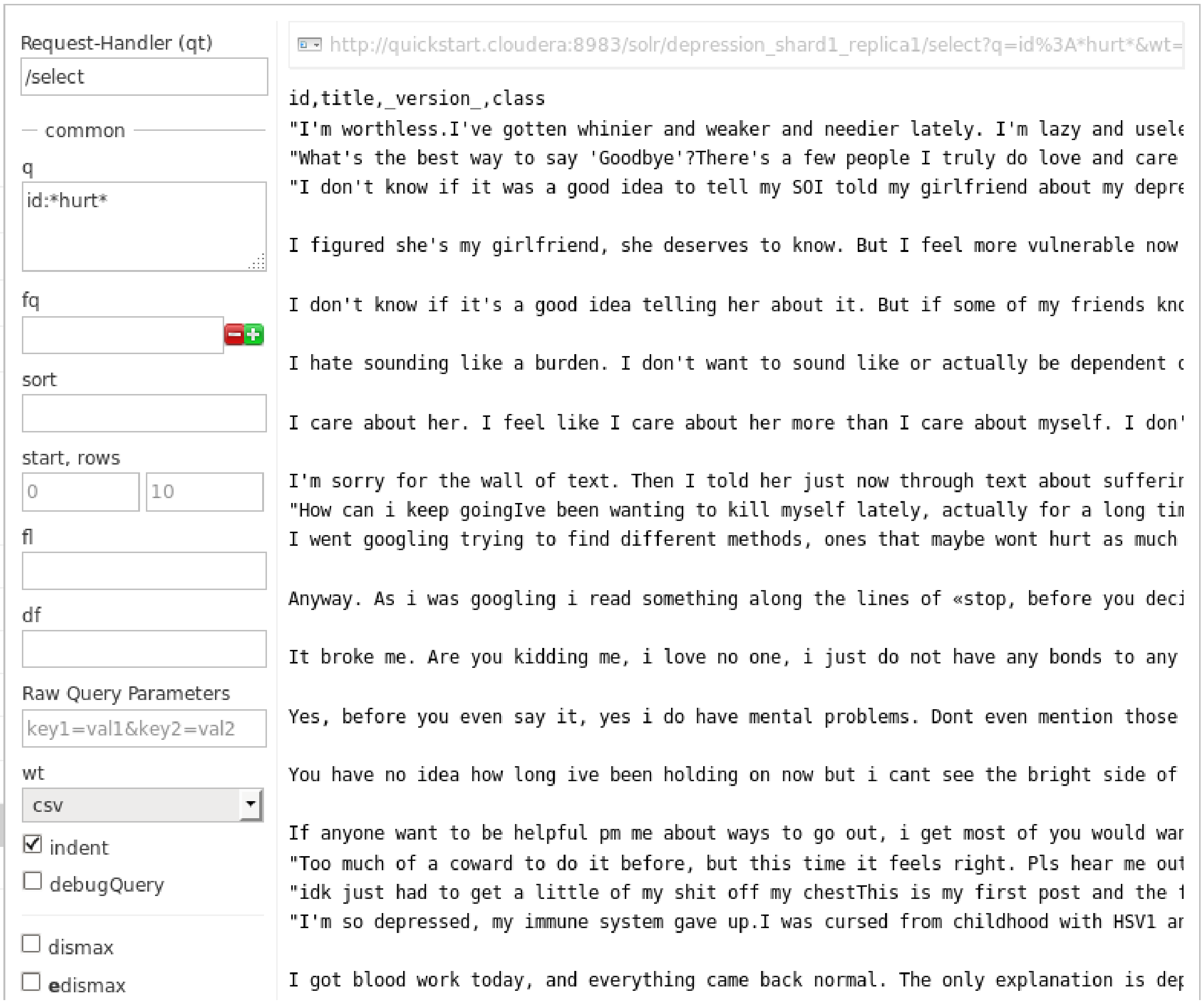
☑ indent

□ debugQuery

□ dismax

http://quickstart.cloudera:8983/solr/depression_shard1_replica1/select?q=id%3A*depressed

```
id,title,_version_,class
"Feeling a bit depressedI've been in a big low all weekend. I don't know why death h

I just want someone to talk to, it doesn't have to be about depression. I'm a male c
"Settings goals is the first step towards healing depressionI have been very depress
"I have friends but I can't open upThey know I'm depressed but they don't know the s
"i want greatnessim sick and tired of living a normal life. i want an amazing life c
"My 16 year old sister-in-law is wanting to commit suicide. What can I do/ say?Sorry
TL;DR my sister in law told her close friends that she has cancer, so they will thir

So for a 16 y/o she's been through a lot.  Her parents aren't together, mom got rema
So obviously sister in law is confused and no wonder she is depressed.

Most of my family had dealt with depression/ mental illness so I am not a stranger t
I've known that she has thought about it, as many people do, but I never thought she
Now I'm a younger sister in my own family, so the whole ""big sister"" thing is new
Sorry if this is the wrong SubReddit, and if it is please tell me where would be a b
"I'm [17M] and depressed does anyone have any relationship advice for me or my [17F]
"I'm so depressed, my immune system gave up.I was cursed from childhood with HSV1 ar

I got blood work today, and everything came back normal. The only explanation is dep

God, I feel so fucking diseased and disgusting.",,1695354982494109697,depression
I always feels depressedI fake being happy and having fun. I can’t be happy and feel
"a frustrating situationI experience my urges as a drawn out impulse. They don't com

I tried it once. It didn't work, facilities were involved, as were the police. Inter

It's a combination of debilitating body dysmorphia, which practically has me housebc
```



Request-Handler (qt)
/select

common

q
id:*hurt*

fq

sort

start, rows
0 10

fl

df

Raw Query Parameters
key1=val1&key2=val2

wt
csv

☑ indent

☐ debugQuery

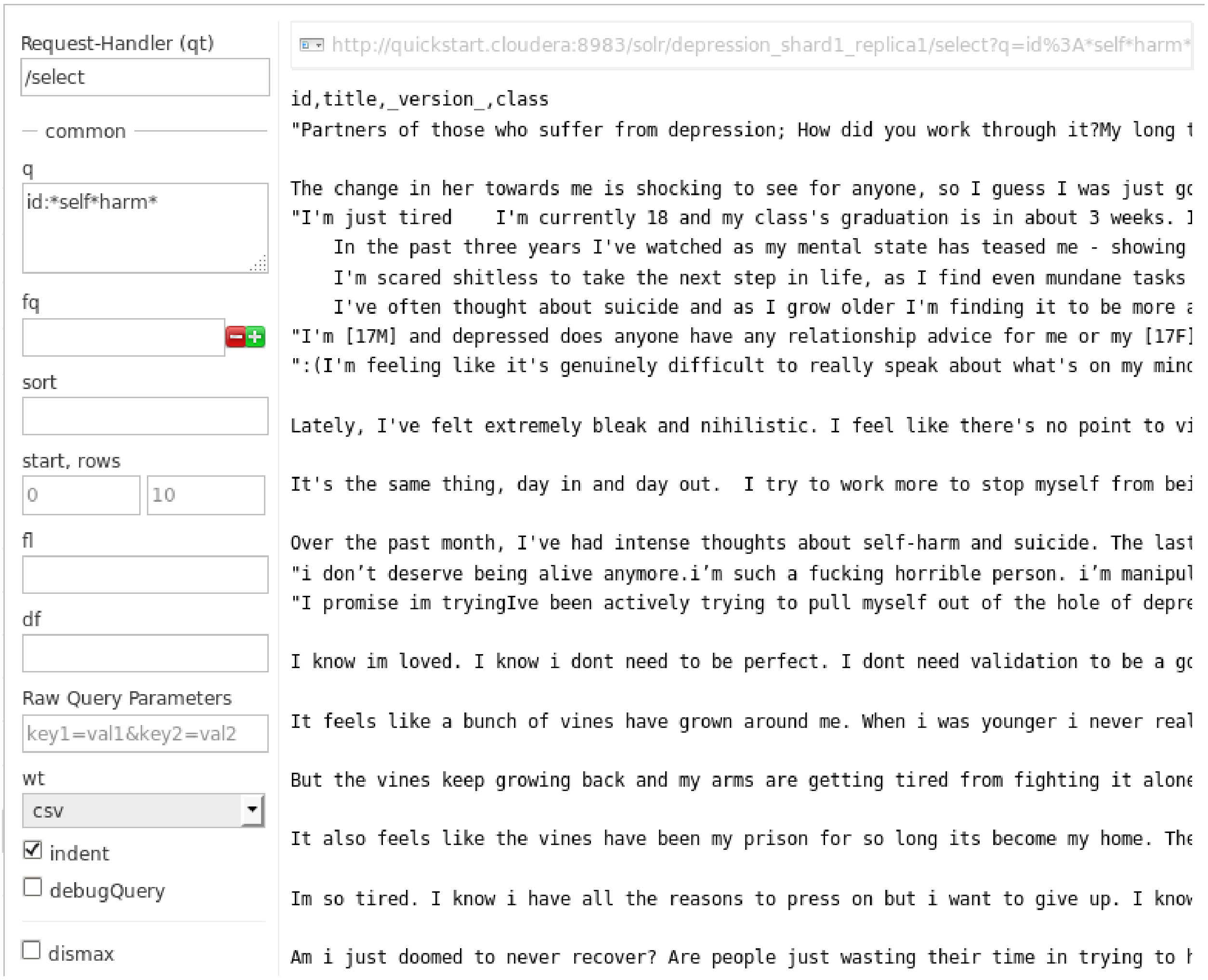
☐ dismax

☐ edismax

http://quickstart.cloudera:8983/solr/depression_shard1_replica1/select?q=id%3A*hurt*&wt=

```
id,title,_version_,class
"I'm worthless.I've gotten whinier and weaker and needier lately. I'm lazy and usele
"What's the best way to say 'Goodbye'?There's a few people I truly do love and care
"I don't know if it was a good idea to tell my SOI told my girlfriend about my depre

I figured she's my girlfriend, she deserves to know. But I feel more vulnerable now

I don't know if it's a good idea telling her about it. But if some of my friends kno

I hate sounding like a burden. I don't want to sound like or actually be dependent o

I care about her. I feel like I care about her more than I care about myself. I don'

I'm sorry for the wall of text. Then I told her just now through text about sufferin
"How can i keep goingIve been wanting to kill myself lately, actually for a long tim
I went googling trying to find different methods, ones that maybe wont hurt as much

Anyway. As i was googling i read something along the lines of «stop, before you deci

It broke me. Are you kidding me, i love no one, i just do not have any bonds to any

Yes, before you even say it, yes i do have mental problems. Dont even mention those

You have no idea how long ive been holding on now but i cant see the bright side of

If anyone want to be helpful pm me about ways to go out, i get most of you would wan
"Too much of a coward to do it before, but this time it feels right. Pls hear me out
"idk just had to get a little of my shit off my chestThis is my first post and the f
"I'm so depressed, my immune system gave up.I was cursed from childhood with HSV1 an

I got blood work today, and everything came back normal. The only explanation is dep
```

We did an additional wildcard search on “self harm” since it is common that people suffering from depression also struggle with self harm.



Request-Handler (qt): /select

common

q: id:*self*harm*

fq:

sort:

start, rows: 0, 10

fl:

df:

Raw Query Parameters: key1=val1&key2=val2

wt: csv

☑ indent

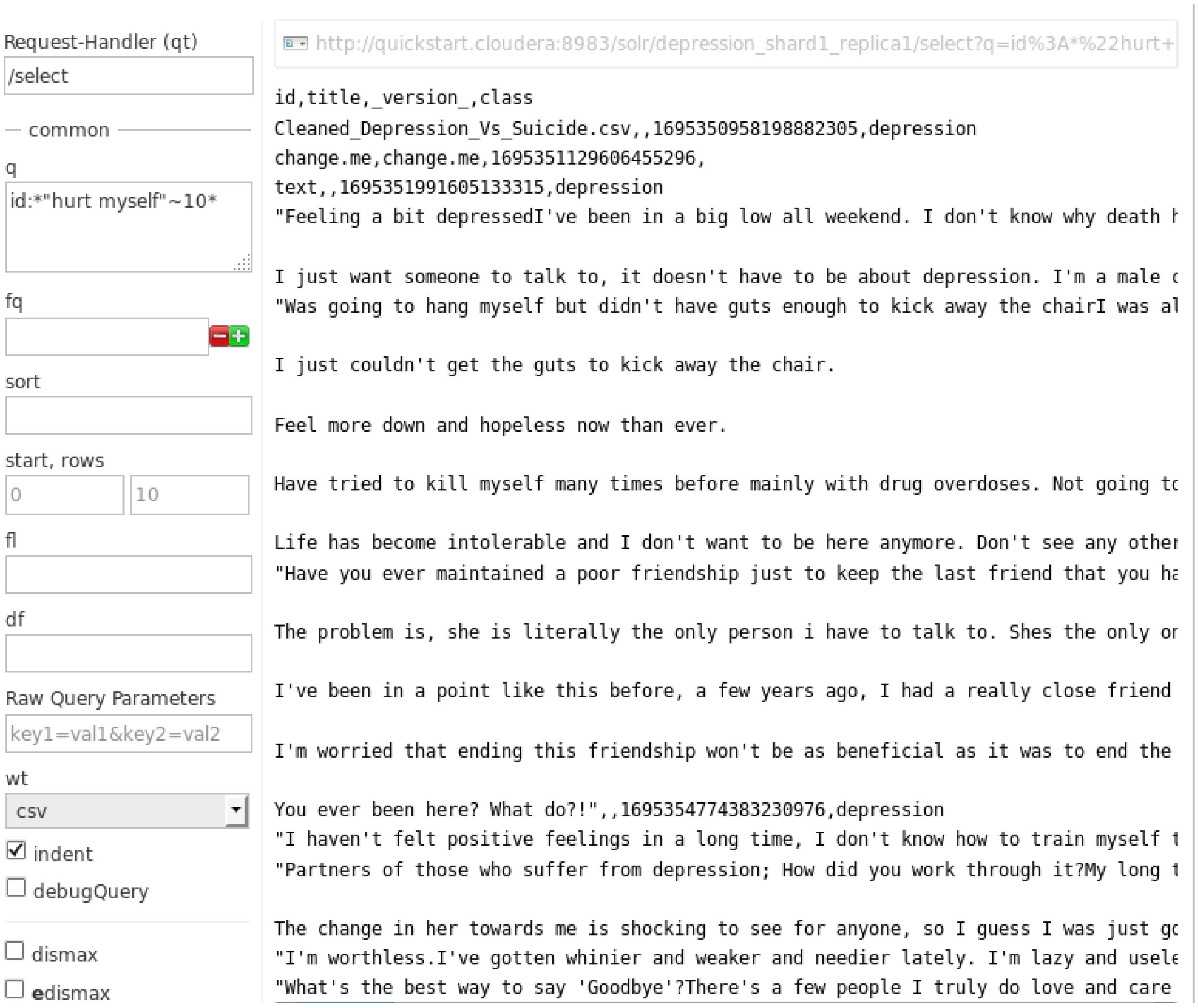
□ debugQuery

□ dismax

http://quickstart.cloudera:8983/solr/depression_shard1_replica1/select?q=id%3A*self*harm*

```
id,title,_version_,class
"Partners of those who suffer from depression; How did you work through it?My long t

The change in her towards me is shocking to see for anyone, so I guess I was just gc
"I'm just tired    I'm currently 18 and my class's graduation is in about 3 weeks. I
    In the past three years I've watched as my mental state has teased me - showing
    I'm scared shitless to take the next step in life, as I find even mundane tasks
    I've often thought about suicide and as I grow older I'm finding it to be more a
"I'm [17M] and depressed does anyone have any relationship advice for me or my [17F]
":(I'm feeling like it's genuinely difficult to really speak about what's on my minc

Lately, I've felt extremely bleak and nihilistic. I feel like there's no point to vi

It's the same thing, day in and day out.  I try to work more to stop myself from bei

Over the past month, I've had intense thoughts about self-harm and suicide. The last
"i don’t deserve being alive anymore.i’m such a fucking horrible person. i’m manipul
"I promise im tryingIve been actively trying to pull myself out of the hole of depre

I know im loved. I know i dont need to be perfect. I dont need validation to be a gc

It feels like a bunch of vines have grown around me. When i was younger i never real

But the vines keep growing back and my arms are getting tired from fighting it alone

It also feels like the vines have been my prison for so long its become my home. The

Im so tired. I know i have all the reasons to press on but i want to give up. I know

Am i just doomed to never recover? Are people just wasting their time in trying to h
```

We also did proximity searching for posts that talked about hurting themselves by searching for the words “hurt myself” within 10 words of each other:



Request-Handler (qt)
/select

common

q
id:*"hurt myself"~10*

fq

sort

start, rows
0 10

fl

df

Raw Query Parameters
key1=val1&key2=val2

wt
csv

☑ indent

☐ debugQuery

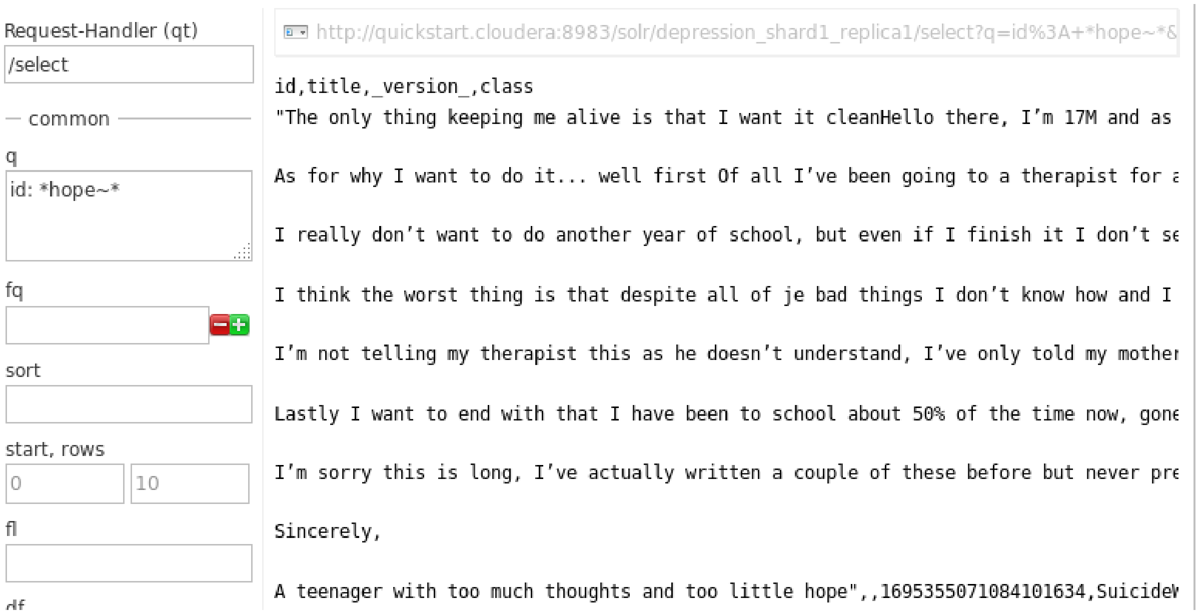
☐ dismax

☐ edismax

http://quickstart.cloudera:8983/solr/depression_shard1_replica1/select?q=id%3A*%22hurt+

```
id,title,_version_,class
Cleaned_Depression_Vs_Suicide.csv,,1695350958198882305,depression
change.me,change.me,1695351129606455296,
text,,1695351991605133315,depression
"Feeling a bit depressedI've been in a big low all weekend. I don't know why death h

I just want someone to talk to, it doesn't have to be about depression. I'm a male c
"Was going to hang myself but didn't have guts enough to kick away the chairI was al

I just couldn't get the guts to kick away the chair.

Feel more down and hopeless now than ever.

Have tried to kill myself many times before mainly with drug overdoses. Not going to

Life has become intolerable and I don't want to be here anymore. Don't see any other
"Have you ever maintained a poor friendship just to keep the last friend that you ha

The problem is, she is literally the only person i have to talk to. Shes the only on

I've been in a point like this before, a few years ago, I had a really close friend

I'm worried that ending this friendship won't be as beneficial as it was to end the

You ever been here? What do?!",,1695354774383230976,depression
"I haven't felt positive feelings in a long time, I don't know how to train myself t
"Partners of those who suffer from depression; How did you work through it?My long t

The change in her towards me is shocking to see for anyone, so I guess I was just go
"I'm worthless.I've gotten whinier and weaker and needier lately. I'm lazy and usele
"What's the best way to say 'Goodbye'?There's a few people I truly do love and care
```

Fuzzy search on the word hope:



Request-Handler (qt)

/select

q

id: *hope~*

fq

sort

start, rows

0 10

fl

df

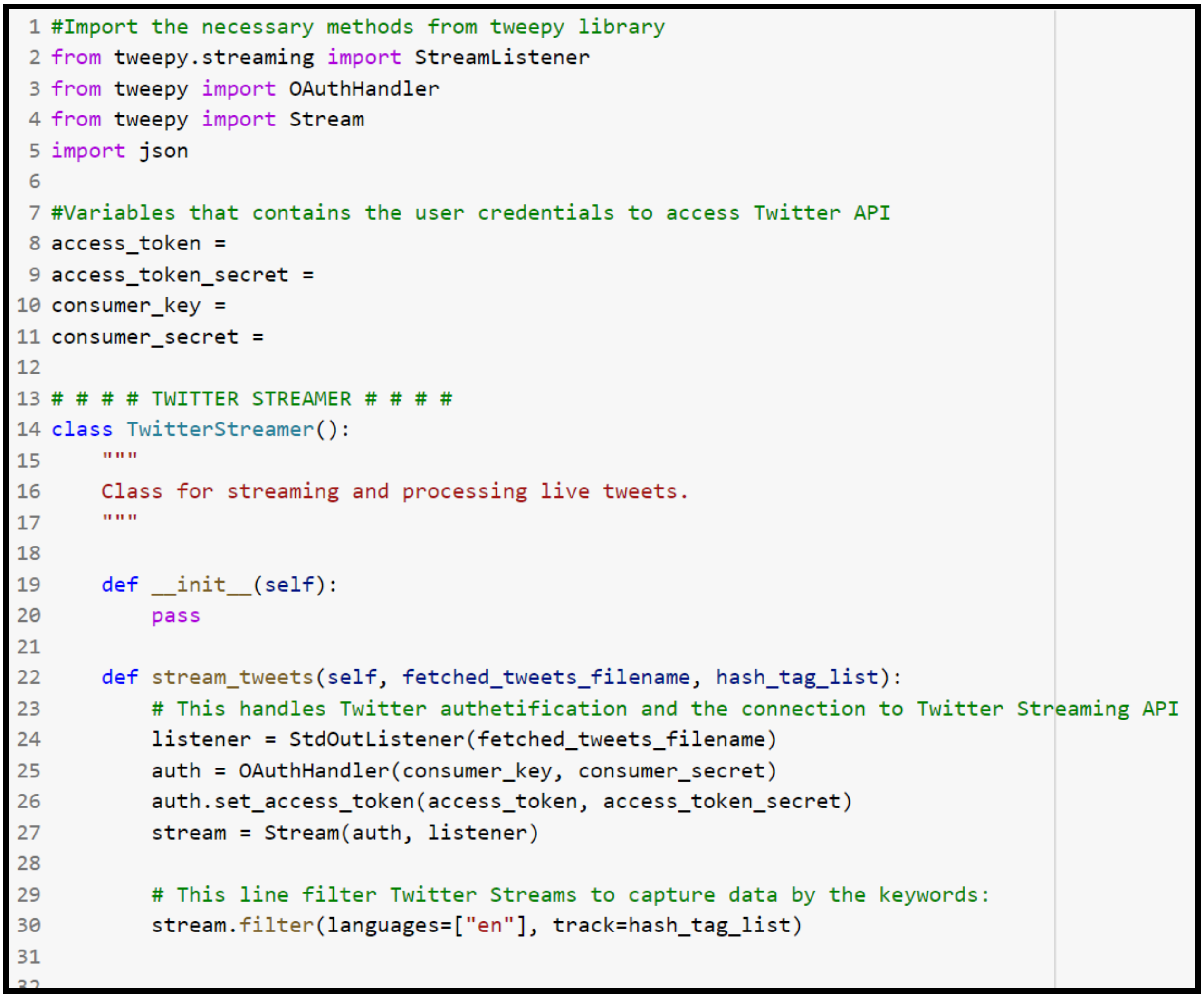
```
http://quickstart.cloudera:8983/solr/depression_shard1_replica1/select?q=id%3A+*hope~*&
```

```
id,title,_version_,class
"The only thing keeping me alive is that I want it cleanHello there, I'm 17M and as

As for why I want to do it... well first Of all I've been going to a therapist for a

I really don't want to do another year of school, but even if I finish it I don't se

I think the worst thing is that despite all of je bad things I don't know how and I

I'm not telling my therapist this as he doesn't understand, I've only told my mother

Lastly I want to end with that I have been to school about 50% of the time now, gone

I'm sorry this is long, I've actually written a couple of these before but never pre

Sincerely,

A teenager with too much thoughts and too little hope",,1695355071084101634,SuicideW
```

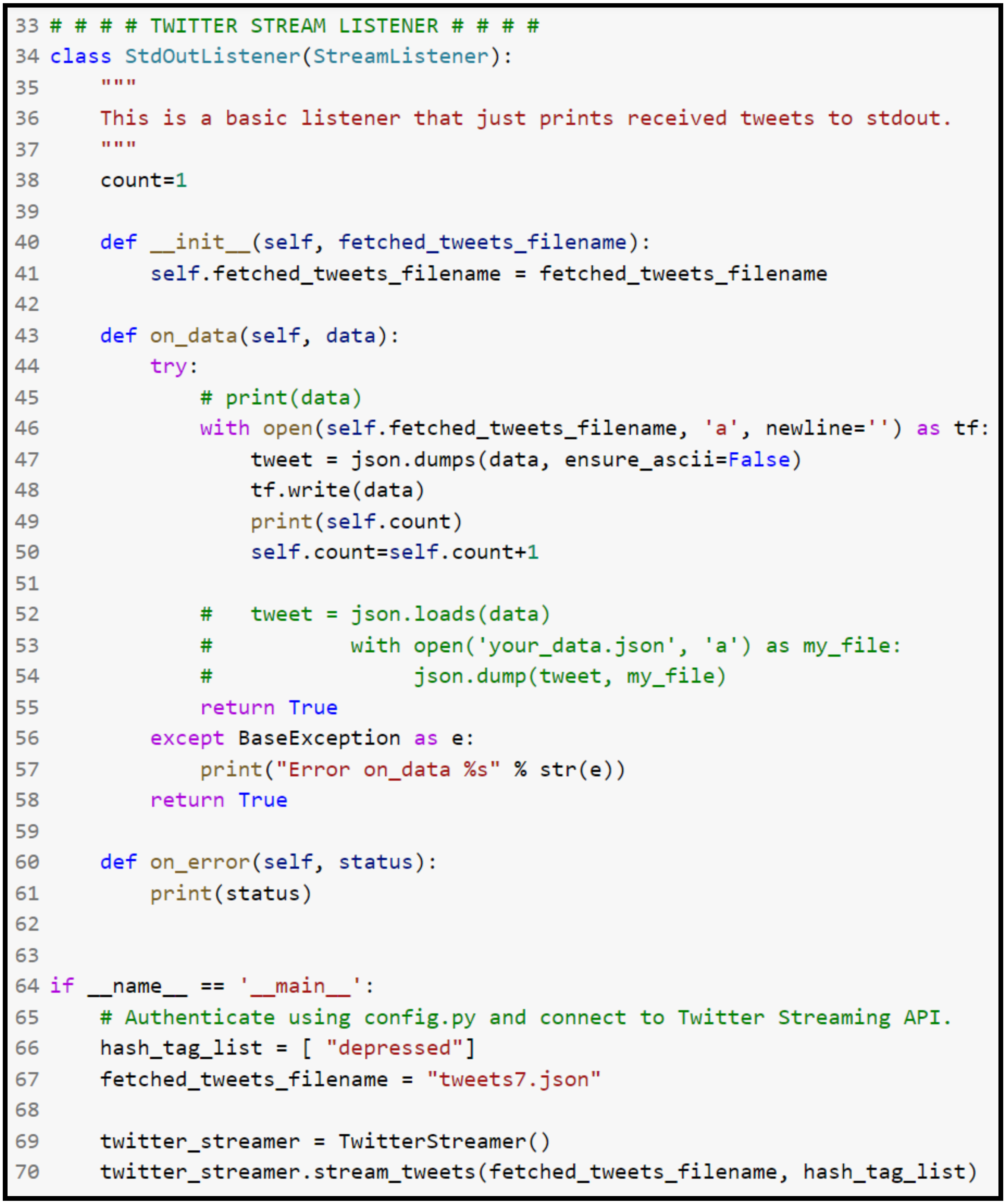
Implementation

Data Collection (Tweets using TwitterAPI):

First I created a twitter developer account and then created an application and got the tokens from there.
Collecting tweets using python tweepy library. First I authenticated the twitter API and then I used a listener to get the tweets as JSON object and then store it into a JSON file.



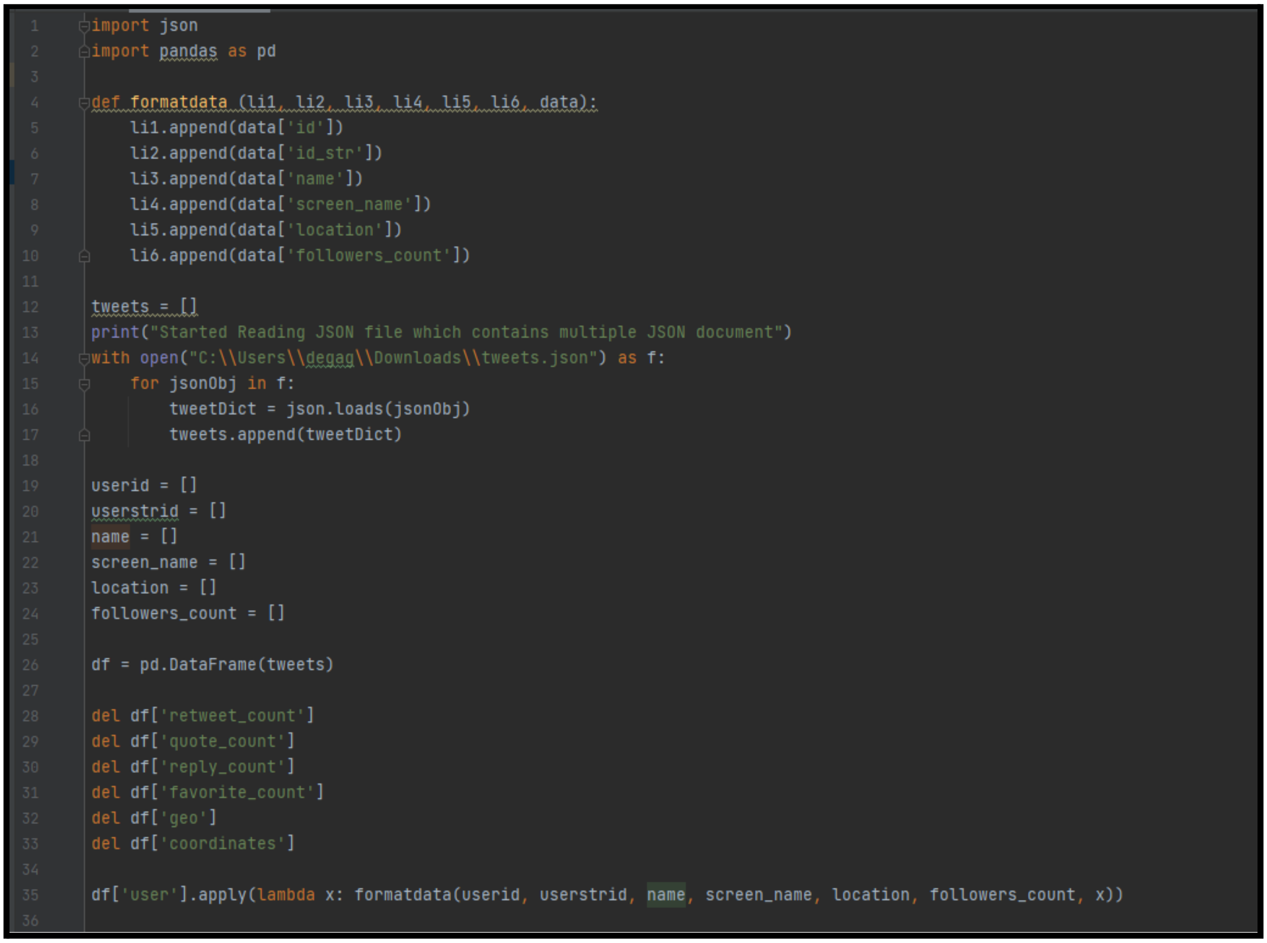
```
#Import the necessary methods from tweepy library
from tweepy.streaming import StreamListener
from tweepy import OAuthHandler
from tweepy import Stream
import json

#Variables that contains the user credentials to access Twitter API
access_token =
access_token_secret =
consumer_key =
consumer_secret =

# # # # TWITTER STREAMER # # # #
class TwitterStreamer():
    """
    Class for streaming and processing live tweets.
    """

    def __init__(self):
        pass

    def stream_tweets(self, fetched_tweets_filename, hash_tag_list):
        # This handles Twitter authetification and the connection to Twitter Streaming API
        listener = StdOutListener(fetched_tweets_filename)
        auth = OAuthHandler(consumer_key, consumer_secret)
        auth.set_access_token(access_token, access_token_secret)
        stream = Stream(auth, listener)

        # This line filter Twitter Streams to capture data by the keywords:
        stream.filter(languages=["en"], track=hash_tag_list)

```



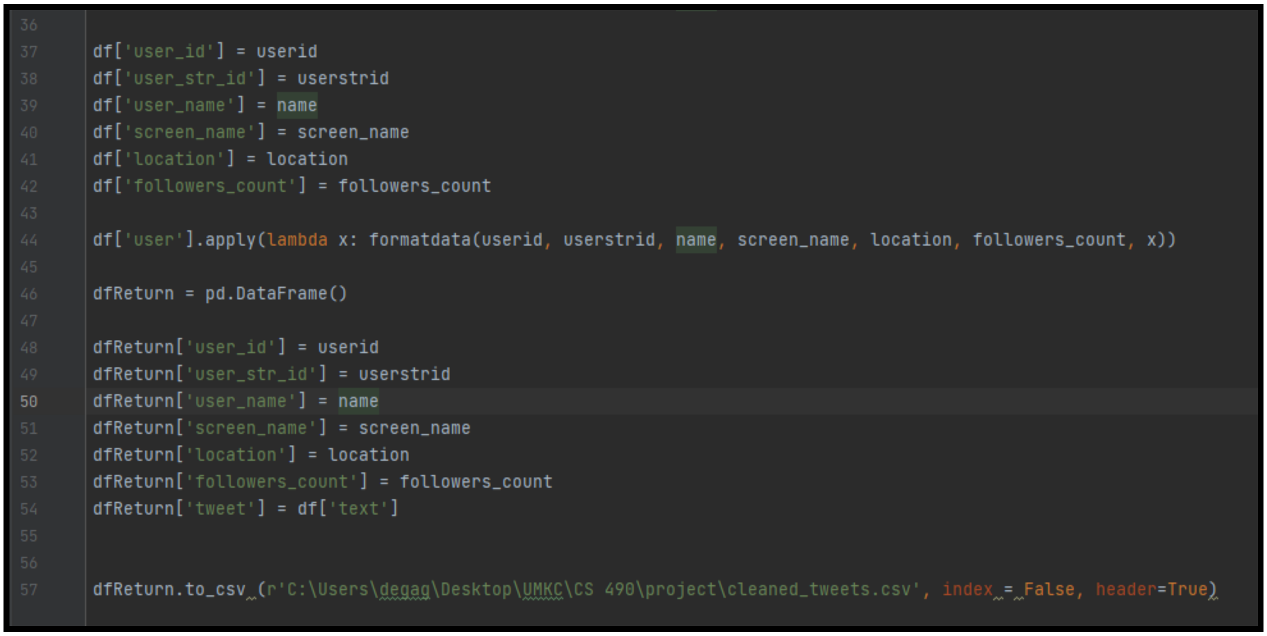
```
# # # # TWITTER STREAM LISTENER # # # #
class StdOutListener(StreamListener):
    """
    This is a basic listener that just prints received tweets to stdout.
    """
    count=1

    def __init__(self, fetched_tweets_filename):
        self.fetched_tweets_filename = fetched_tweets_filename

    def on_data(self, data):
        try:
            # print(data)
            with open(self.fetched_tweets_filename, 'a', newline='') as tf:
                tweet = json.dumps(data, ensure_ascii=False)
                tf.write(data)
                print(self.count)
                self.count=self.count+1

            #   tweet = json.loads(data)
            #           with open('your_data.json', 'a') as my_file:
            #                json.dump(tweet, my_file)
            return True
        except BaseException as e:
            print("Error on_data %s" % str(e))
        return True

    def on_error(self, status):
        print(status)


if __name__ == '__main__':
    # Authenticate using config.py and connect to Twitter Streaming API.
    hash_tag_list = [ "depressed"]
    fetched_tweets_filename = "tweets7.json"

    twitter_streamer = TwitterStreamer()
    twitter_streamer.stream_tweets(fetched_tweets_filename, hash_tag_list)
```

Data Cleaning (Python JSON to CSV):

After extracting the data from Twitter and outputting a JSON file, we want to clean the data to only contain data that are useful for our analysis. Before uploading our dataset onto Hadoop, we first filter through the data and extract specific data and fields in the JSON file and migrate the fields onto a dataframe and export the DataFrame as a CSV. Below is the source code for our data cleaning:



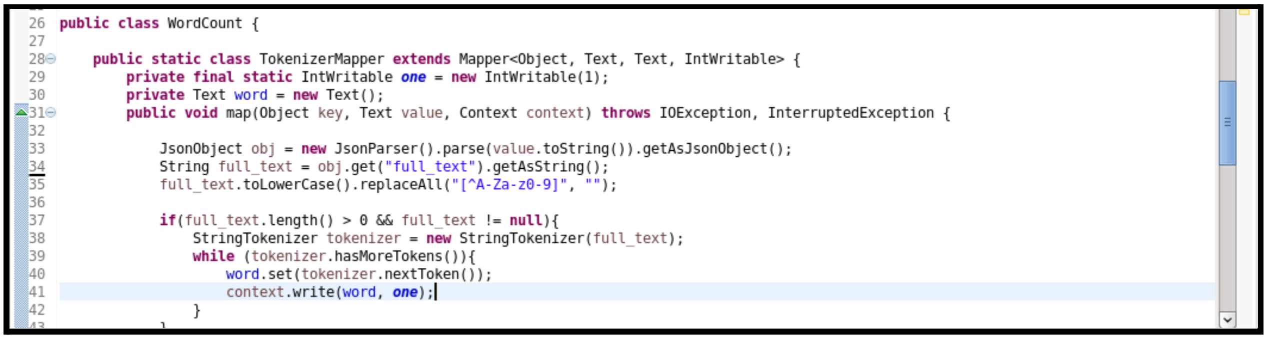
```
import json
import pandas as pd

def formatdata (li1, li2, li3, li4, li5, li6, data):
    li1.append(data['id'])
    li2.append(data['id_str'])
    li3.append(data['name'])
    li4.append(data['screen_name'])
    li5.append(data['location'])
    li6.append(data['followers_count'])

tweets = []
print("Started Reading JSON file which contains multiple JSON document")
with open("C:\\Users\\degag\\Downloads\\tweets.json") as f:
    for jsonObj in f:
        tweetDict = json.loads(jsonObj)
        tweets.append(tweetDict)

userid = []
userstrid = []
name = []
screen_name = []
location = []
followers_count = []

df = pd.DataFrame(tweets)

del df['retweet_count']
del df['quote_count']
del df['reply_count']
del df['favorite_count']
del df['geo']
del df['coordinates']

df['user'].apply(lambda x: formatdata(userid, userstrid, name, screen_name, location, followers_count, x))
```



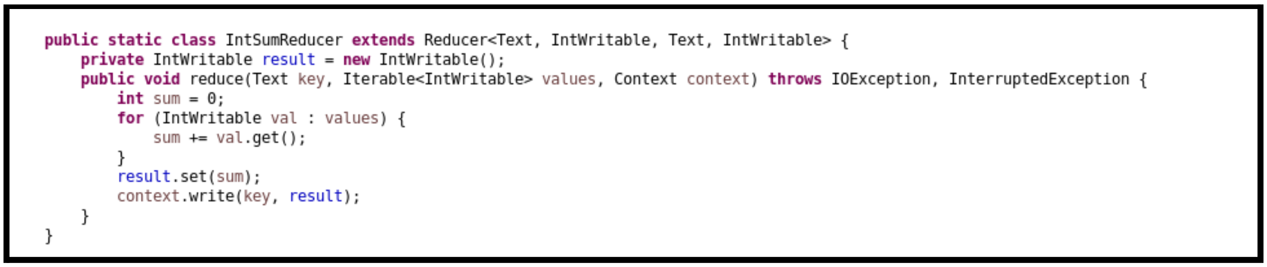
```python

df['user_id'] = userid
df['user_str_id'] = userstrid
df['user_name'] = name
df['screen_name'] = screen_name
df['location'] = location
df['followers_count'] = followers_count

df['user'].apply(lambda x: formatdata(userid, userstrid, name, screen_name, location, followers_count, x))

dfReturn = pd.DataFrame()

dfReturn['user_id'] = userid
dfReturn['user_str_id'] = userstrid
dfReturn['user_name'] = name
dfReturn['screen_name'] = screen_name
dfReturn['location'] = location
dfReturn['followers_count'] = followers_count
dfReturn['tweet'] = df['text']


dfReturn.to_csv (r'C:\Users\degag\Desktop\UMKC\CS 490\project\cleaned_tweets.csv', index = False, header=True)
```

Pre Processing (Hadoop Map-Reduce):

Inorder to help us develop a model we need to generate key words that attribute to if a piece of text is depressive. To create this list, we utilised a Map-Reduce method to generate a word count and order it by the count of each word so the most common words appear at the top. The WordCount Map-Reduce stayed fairly traditional, with only having to handle a JSON file from our twitter scraping. During the Mapping we remove non character characters and bring everything to lowercase for processing.

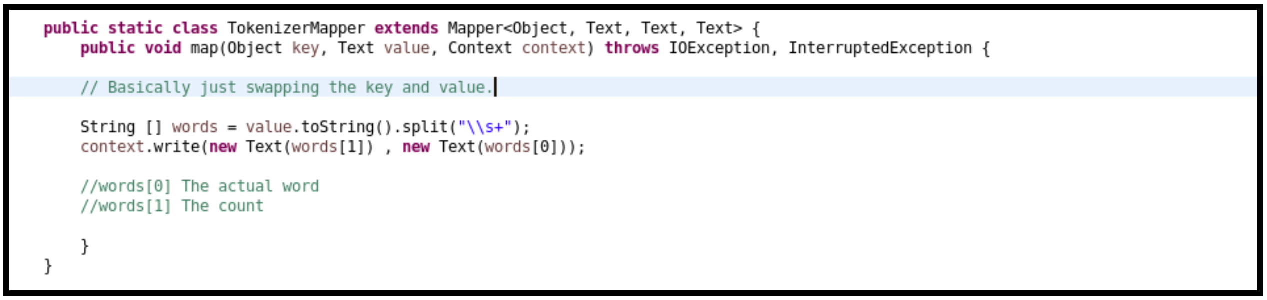


```java
public class WordCount {

    public static class TokenizerMapper extends Mapper<Object, Text, Text, IntWritable> {
        private final static IntWritable one = new IntWritable(1);
        private Text word = new Text();
        public void map(Object key, Text value, Context context) throws IOException, InterruptedException {

            JsonObject obj = new JsonParser().parse(value.toString()).getAsJsonObject();
            String full_text = obj.get("full_text").getAsString();
            full_text.toLowerCase().replaceAll("[^A-Za-z0-9]", "");

            if(full_text.length() > 0 && full_text != null){
                StringTokenizer tokenizer = new StringTokenizer(full_text);
                while (tokenizer.hasMoreTokens()){
                    word.set(tokenizer.nextToken());
                    context.write(word, one);
                }
```



```java
public static class IntSumReducer extends Reducer<Text, IntWritable, Text, IntWritable> {
    private IntWritable result = new IntWritable();
    public void reduce(Text key, Iterable<IntWritable> values, Context context) throws IOException, InterruptedException {
        int sum = 0;
        for (IntWritable val : values) {
            sum += val.get();
        }
        result.set(sum);
        context.write(key, result);
    }
}
```

After we get the initial WordCount, we run another Map-Reduce job on the WordCount results to sort it by the count amounts. This was done by swapping the key and value pairs and having it auto sort it as a byproduct in end.

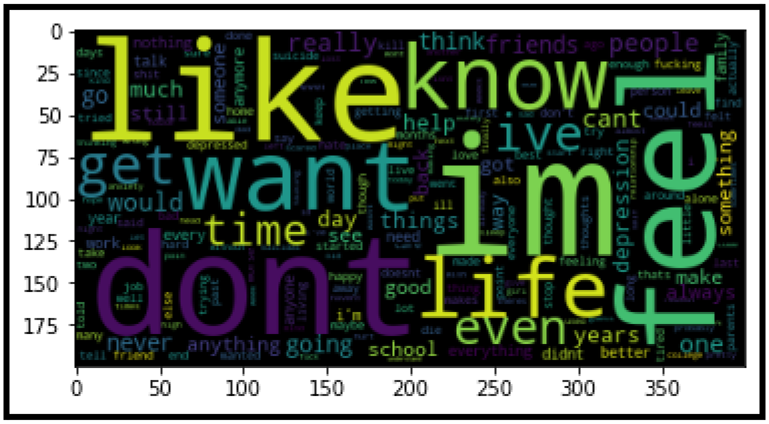


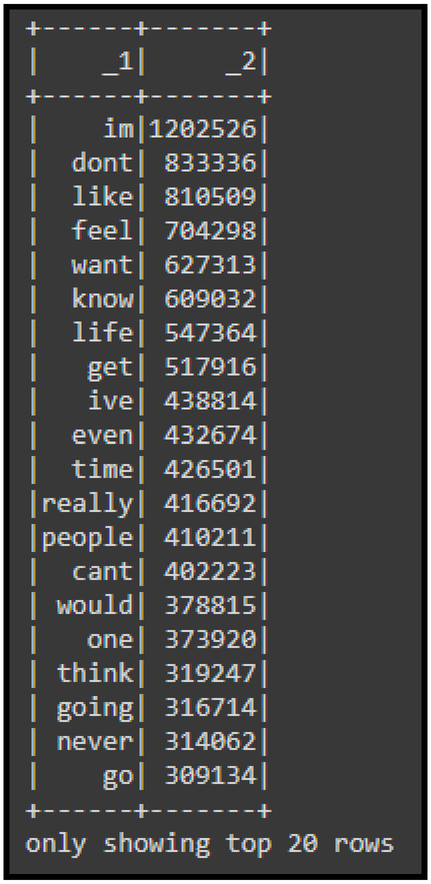
```
public static class TokenizerMapper extends Mapper<Object, Text, Text, Text> {
    public void map(Object key, Text value, Context context) throws IOException, InterruptedException {

    // Basically just swapping the key and value.

    String [] words = value.toString().split("\\s+");
    context.write(new Text(words[1]) , new Text(words[0]));

    //words[0] The actual word
    //words[1] The count

    }
}
```

Preliminary Results

Word Cloud (Pyspark):

For the Word Count of the Reddit posts we created the following Word Cloud graphic to show high frequency words.





```
+------+-------+
|    _1|     _2|
+------+-------+
|    im|1202526|
|  dont| 833336|
|  like| 810509|
|  feel| 704298|
|  want| 627313|
|  know| 609032|
|  life| 547364|
|   get| 517916|
|   ive| 438814|
|  even| 432674|
|  time| 426501|
|really| 416692|
|people| 410211|
|  cant| 402223|
| would| 378815|
|   one| 373920|
| think| 319247|
| going| 316714|
| never| 314062|
|    go| 309134|
+------+-------+
only showing top 20 rows
```

Implementation Status Report

Ami Khalsa:

- Bigrams and Trigrams.
- Solr Querying: wildcard search on the word “Depression”
- Solr Querying: wildcard search on the word “hurt”
- Solr Querying: wildcard search on the words self and harm
- Solr Querying: proximity search on the phrase “hurt myself”
- Solr Querying: fuzzy search on the word hope

Ashish Pant:

- Data Extraction from twitter using Twitter API.
- Storing extracted tweets into a json file as JSON objects.
- Loading JSON into Spark and doing analysis.
- Created WordCloud Visualization.
- PySpark: Summation Subreddits
- PySpark: Word Count

Bryan Khoo:

- Hive Querying: Top 10 Tweeted Locations.
- Hive Querying: Tweets that Include the Word “kill”.
- PySpark: TF-IDF
- PySpark: Summation Subreddits
- Python: Data Cleaning

Davith Lon:

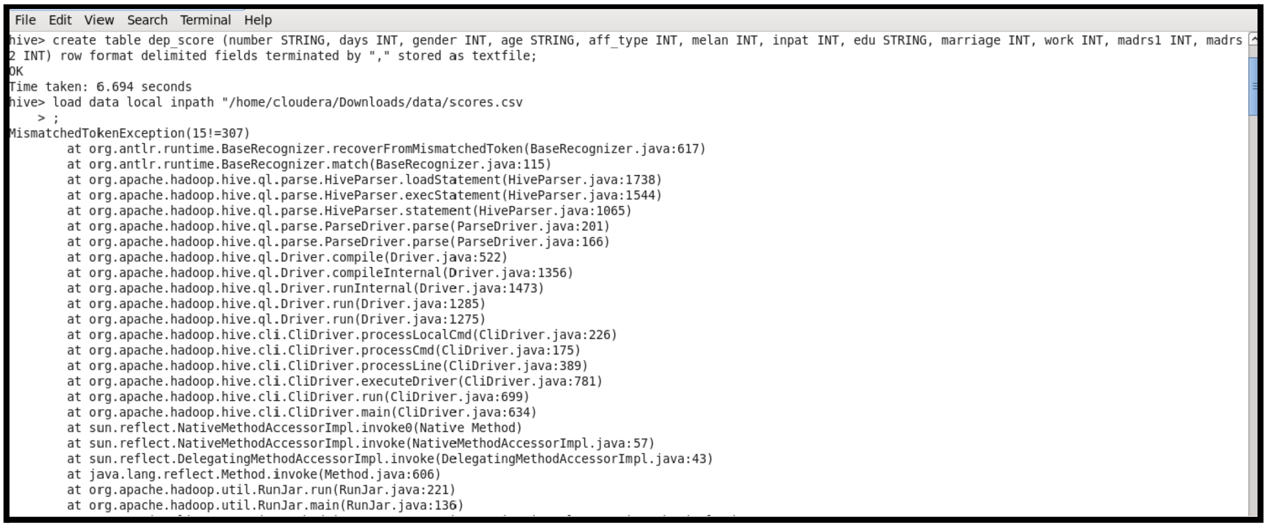
- Writing Map Reduce for Word Count.
- Writing Map Reduce for Sorting By Count.
- Hive Querying: Tweets Per User.
- Hive Querying: Most Followed Twitter User.
- Hive Query: Sensitive Flagged Posts.
- Hive Query: Verified Users.
- WorkFlow Diagram.

- PySpark: Word Count

Saurav Pawar (Hive/ PySpark):

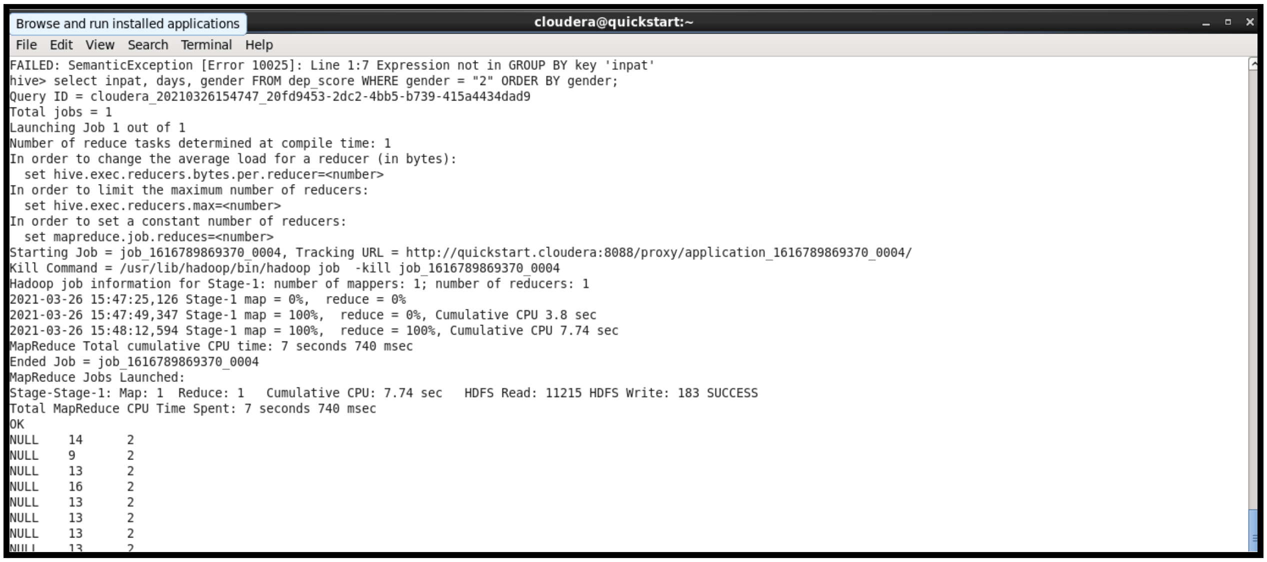
The following analysis and queries were done on our old dataset that we were basing our project on. After Saurav had done his analysis on this dataset we realized we needed to change the direction of our project and found other sources of data. We will provide his work down here to show his contributions to the project.

- Uploading dataset on Hive



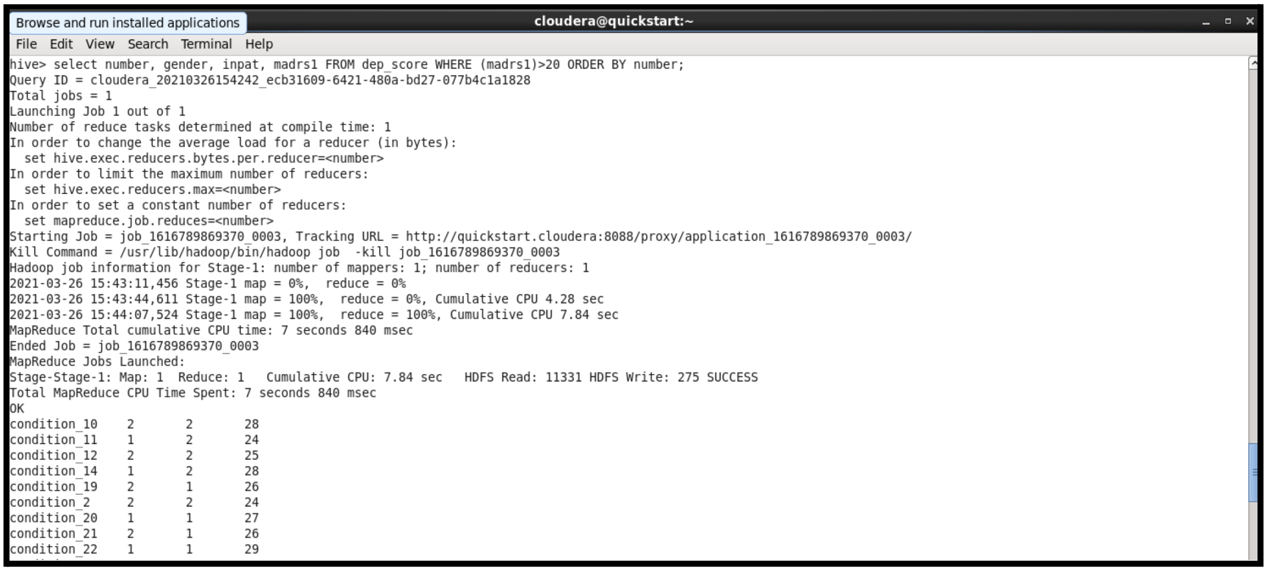
```
hive> create table dep_score (number STRING, days INT, gender INT, age STRING, aff_type INT, melan INT, inpat INT, edu STRING, marriage INT, work INT, madrs1 INT, madrs
2 INT) row format delimited fields terminated by "," stored as textfile;
OK
Time taken: 6.694 seconds
hive> load data local inpath "/home/cloudera/Downloads/data/scores.csv
    > ;
MismatchedTokenException(15!=307)
        at org.antlr.runtime.BaseRecognizer.recoverFromMismatchedToken(BaseRecognizer.java:617)
        at org.antlr.runtime.BaseRecognizer.match(BaseRecognizer.java:115)
        at org.apache.hadoop.hive.ql.parse.HiveParser.loadStatement(HiveParser.java:1738)
        at org.apache.hadoop.hive.ql.parse.HiveParser.execStatement(HiveParser.java:1544)
        at org.apache.hadoop.hive.ql.parse.HiveParser.statement(HiveParser.java:1065)
        at org.apache.hadoop.hive.ql.parse.ParseDriver.parse(ParseDriver.java:201)
        at org.apache.hadoop.hive.ql.parse.ParseDriver.parse(ParseDriver.java:166)
        at org.apache.hadoop.hive.ql.Driver.compile(Driver.java:522)
        at org.apache.hadoop.hive.ql.Driver.compileInternal(Driver.java:1356)
        at org.apache.hadoop.hive.ql.Driver.runInternal(Driver.java:1473)
        at org.apache.hadoop.hive.ql.Driver.run(Driver.java:1285)
        at org.apache.hadoop.hive.ql.Driver.run(Driver.java:1275)
        at org.apache.hadoop.hive.cli.CliDriver.processLocalCmd(CliDriver.java:226)
        at org.apache.hadoop.hive.cli.CliDriver.processCmd(CliDriver.java:175)
        at org.apache.hadoop.hive.cli.CliDriver.processLine(CliDriver.java:389)
        at org.apache.hadoop.hive.cli.CliDriver.executeDriver(CliDriver.java:781)
        at org.apache.hadoop.hive.cli.CliDriver.run(CliDriver.java:699)
        at org.apache.hadoop.hive.cli.CliDriver.main(CliDriver.java:634)
        at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
        at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:57)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:606)
        at org.apache.hadoop.util.RunJar.run(RunJar.java:221)
        at org.apache.hadoop.util.RunJar.main(RunJar.java:136)
```

- SELECT inpat, days, gender FROM dep_score WHERE gender = ‘2’ ORDER BY gender;
- Relationship of the female patients with their days of records from dataset.


Browse and run installed applications
cloudera@quickstart:~
File Edit View Search Terminal Help

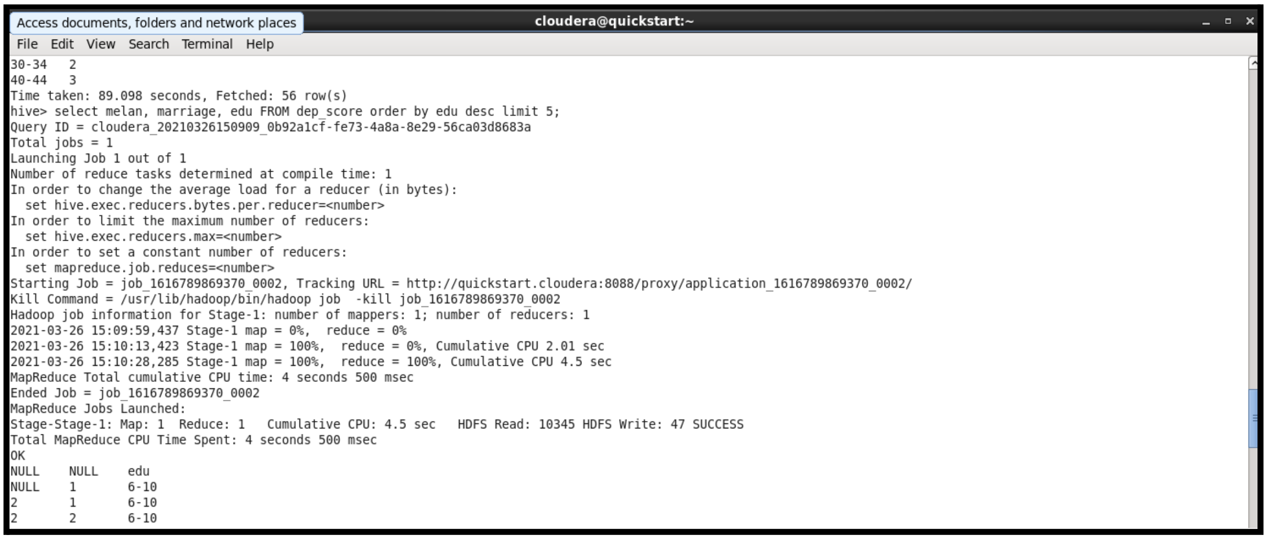
```
FAILED: SemanticException [Error 10025]: Line 1:7 Expression not in GROUP BY key 'inpat'
hive> select inpat, days, gender FROM dep_score WHERE gender = "2" ORDER BY gender;
Query ID = cloudera_20210326154747_20fd9453-2dc2-4bb5-b739-415a4434dad9
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1616789869370_0004, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1616789869370_0004/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1616789869370_0004
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-03-26 15:47:25,126 Stage-1 map = 0%,  reduce = 0%
2021-03-26 15:47:49,347 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.8 sec
2021-03-26 15:48:12,594 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 7.74 sec
MapReduce Total cumulative CPU time: 7 seconds 740 msec
Ended Job = job_1616789869370_0004
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 7.74 sec   HDFS Read: 11215 HDFS Write: 183 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 740 msec
OK
NULL    14      2
NULL    9       2
NULL    13      2
NULL    16      2
NULL    13      2
NULL    13      2
NULL    13      2
NULL    13      2
```

- SELECT number, gender, inpat, madrs1 FROM dep_score WHERE (madrs1)>20 ORDER BY number;
- Relationship between patients, their gender and melancholic type from dataset.


Browse and run installed applications
cloudera@quickstart:~
File Edit View Search Terminal Help

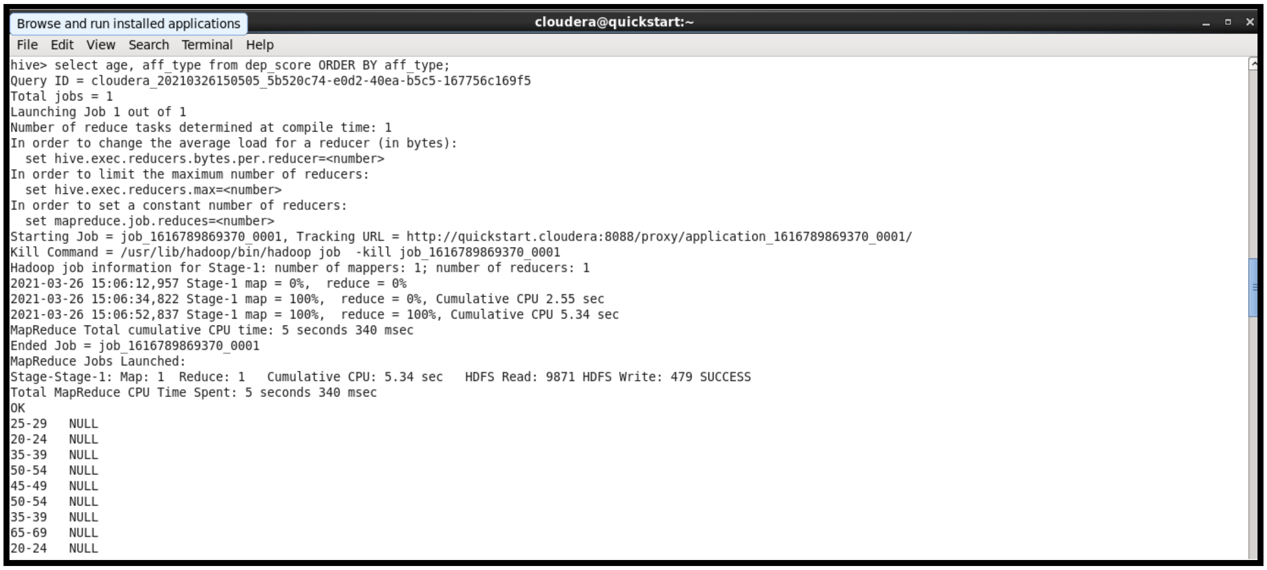
```
hive> select number, gender, inpat, madrs1 FROM dep_score WHERE (madrs1)>20 ORDER BY number;
Query ID = cloudera_20210326154242_ecb31609-6421-480a-bd27-077b4c1a1828
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1616789869370_0003, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1616789869370_0003/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1616789869370_0003
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-03-26 15:43:11,456 Stage-1 map = 0%,  reduce = 0%
2021-03-26 15:43:44,611 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.28 sec
2021-03-26 15:44:07,524 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 7.84 sec
MapReduce Total cumulative CPU time: 7 seconds 840 msec
Ended Job = job_1616789869370_0003
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 7.84 sec   HDFS Read: 11331 HDFS Write: 275 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 840 msec
OK
condition_10    2       2       28
condition_11    1       2       24
condition_12    2       2       25
condition_14    1       2       28
condition_19    2       1       26
condition_2     2       2       24
condition_20    1       1       27
condition_21    2       1       26
condition_22    1       1       29
```

- SELECT melan, mariage, edu FROM dep_score ORDER BY edu desc limit 5;
- Relationship between melancholia, marriage, and education from dataset.


Access documents, folders and network places
cloudera@quickstart:~
File Edit View Search Terminal Help

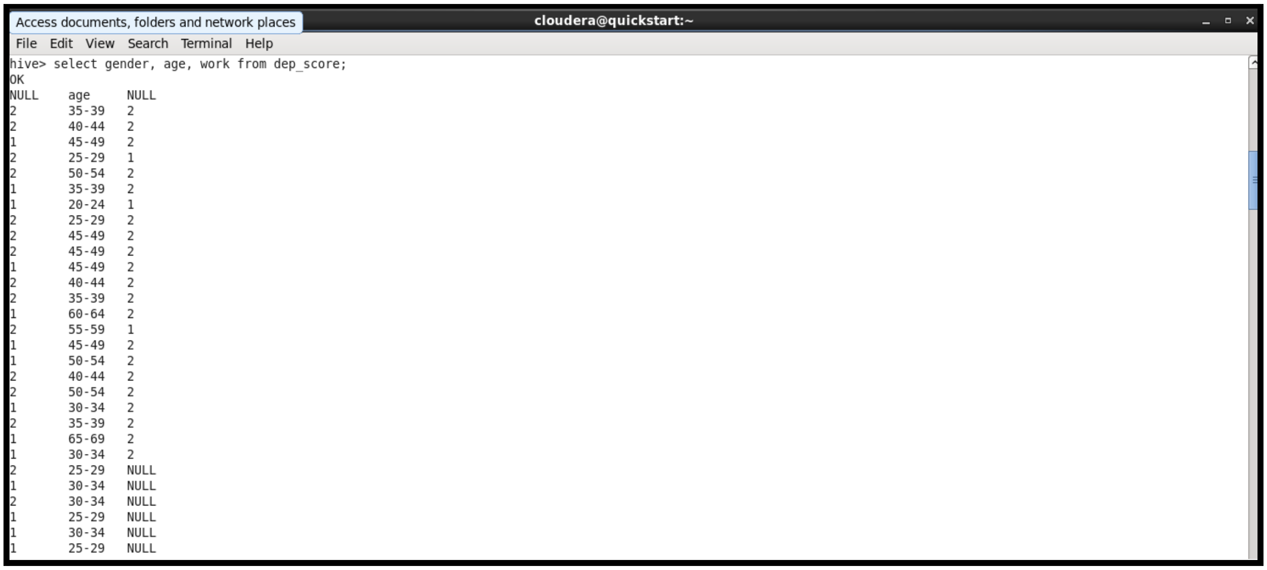
```
30-34   2
40-44   3
Time taken: 89.098 seconds, Fetched: 56 row(s)
hive> select melan, marriage, edu FROM dep_score order by edu desc limit 5;
Query ID = cloudera_20210326150909_0b92a1cf-fe73-4a8a-8e29-56ca03d8683a
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1616789869370_0002, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1616789869370_0002/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1616789869370_0002
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-03-26 15:09:59,437 Stage-1 map = 0%,  reduce = 0%
2021-03-26 15:10:13,423 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.01 sec
2021-03-26 15:10:28,285 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 4.5 sec
MapReduce Total cumulative CPU time: 4 seconds 500 msec
Ended Job = job_1616789869370_0002
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 4.5 sec   HDFS Read: 10345 HDFS Write: 47 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 500 msec
OK
NULL    NULL    edu
NULL    1       6-10
2       1       6-10
2       2       6-10
```

- SELECT age, affinity type FROM dep_score ORDER BY aff_type;
- Relationship between age and affinity type


Browse and run installed applications
cloudera@quickstart:~
File Edit View Search Terminal Help

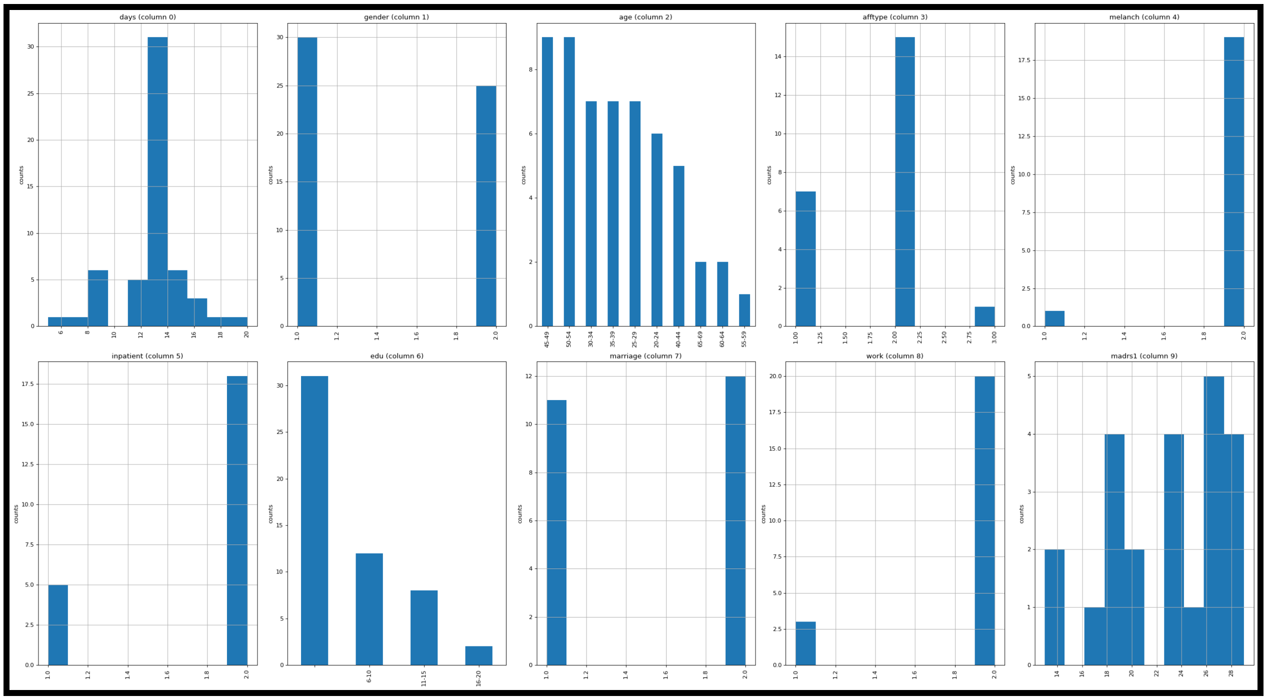
```
hive> select age, aff_type from dep_score ORDER BY aff_type;
Query ID = cloudera_20210326150505_5b520c74-e0d2-40ea-b5c5-167756c169f5
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1616789869370_0001, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1616789869370_0001/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1616789869370_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-03-26 15:06:12,957 Stage-1 map = 0%,  reduce = 0%
2021-03-26 15:06:34,822 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.55 sec
2021-03-26 15:06:52,837 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 5.34 sec
MapReduce Total cumulative CPU time: 5 seconds 340 msec
Ended Job = job_1616789869370_0001
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 5.34 sec   HDFS Read: 9871 HDFS Write: 479 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 340 msec
OK
25-29   NULL
20-24   NULL
35-39   NULL
50-54   NULL
45-49   NULL
50-54   NULL
35-39   NULL
65-69   NULL
20-24   NULL
```

- SELECT age, gender, work FROM dep_score;


Access documents, folders and network places
cloudera@quickstart:~
File Edit View Search Terminal Help

```
hive> select gender, age, work from dep_score;
OK
NULL    age     NULL
2       35-39   2
2       40-44   2
1       45-49   2
2       25-29   1
2       50-54   2
1       35-39   2
1       20-24   1
2       25-29   2
2       45-49   2
2       45-49   2
1       45-49   2
2       40-44   2
2       35-39   2
1       60-64   2
2       55-59   1
1       45-49   2
1       50-54   2
2       40-44   2
2       50-54   2
1       30-34   2
2       35-39   2
1       65-69   2
1       30-34   2
2       25-29   NULL
1       30-34   NULL
2       30-34   NULL
1       25-29   NULL
1       30-34   NULL
1       25-29   NULL
```

- (PySpark) Pandas library to visualize the relation between the patients and their contributing factors for depression that are given in the datasets. Such as Age, marital status, work, and so on.



Work to be Completed:

- More Data via Streaming Twitter Data
- Visualization of Queries and Results
- Sentiment Analysis
- Training Data Labeling
- Model Building

Remaining Work Percentage (60%)

Story Telling

Who:

This dataset includes members of the world who are struggling with depression and/or suicidal thoughts. In our dataset, r/SuicideWatch and r/depression posts from Reddit, we had posts from members of the r/SuicideWatch and r/depression subreddit communities. Both communities were represented equally, meaning 50% of the posts came from r/SuicideWatch and 50% came from r/depression. Since depression is an illness that can affect anyone, we did not filter any demographics out of our dataset.

No identifying information other than what the user provided in their own post was collected. Users are anonymous from the dataset's end. Data about the posting user (age, legal name, gender, etc.) was not collected or used.

What:

The dataset acquired from Kaggle includes posts that were made on each subreddit: r/SuicideWatch and r/depression. The dataset that was pulled from Twitter included the date when the tweet was posted, tweet ID, tweet, name of the account, user ID, followers count for the account, the location of the user, status for the tweet, and whether the twitter account is verified.

When:

In our r/SuicideWatch and r/depression posts from Reddit dataset, the dataset was last updated on Jan. 20, 2021 and covers data from 2008-12-15 to 2021-01-01. It is not real time data. The posts are from a variety of times. Many people relate to old posts just as well as new ones. For this reason, we believe it is useful to examine this data regardless of time since the content of this data is the collection of thoughts of many individuals experiencing struggles with their mental health at the time of their posting. Those thoughts are valuable for us to examine so that we can better understand the people that are currently struggling.

Where:

The dataset acquired from Kaggle was collected from an online forum called Reddit, under the subreddits of r/SuicideWatch and r/depression. Reddit is a platform that is available worldwide which allows us to have the assumption that the dataset obtained is a global dataset. However, the dataset did not include the actual location for each post. Hence, we are not able to conclude where the majority of our dataset originates from.

The dataset acquired from Twitter was collected from a social media that is available worldwide as well. Within our dataset, there is a field labeled location which helps us determine the origin for a particular tweet. A section of the data verifies that the majority of the tweets collected were from the United States while still having a few tweets from other countries as well. However, it is important to keep in mind that the location field might not be accurate as some accounts had inputs on the field that were not locatable.

Why:

Fortunately for us, the Kaggle dataset was collected for the exact same reason we're using it. The original poster of the dataset stated, "When I thought of building a text classifier to detect Suicide Ideation I couldn't find any public dataset. Hope this can be useful to anyone looking for suicide detection datasets and can save their time[7]". Additionally, the behaviour and mindset of the posters on these subreddits is exactly what we want fueling the text.

As for the Twitter scraping, the reasoning behind collecting this data is similar to the Reddit postings. The idea is that with the specific hashtags filtering our post collections, we should be able to get organic text live with the same mind set behind it. However, with the Twitter data there is a lot more to filter through. As we are collecting from the hashtags, we can't say for sure everyone is using them the way we think they would be. Also there is the problem of Twitter bots that can create junk for us.

REFERENCES

[1] https://www.psychiatry.org/patients-families/depression/what-is-depression
[2] https://www.nimh.nih.gov/health/topics/depression/index.shtml
[3] https://www.who.int/news-room/fact-sheets/detail/depression
[4]
https://www.businesstoday.in/current/world/japan-appoints-loneliness-minister-to-tackle-suicide-rates/story/432226.html
[5] https://pubmed.ncbi.nlm.nih.gov/29195763/
[6] https://pubmed.ncbi.nlm.nih.gov/27627885/
[7] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4314052/

[7] https://www.kaggle.com/nikhileswarkomati/suicide-watch