Mini Project Report

on

"Twitter Data Sentiment Analysis on Fungus Live Dataset"

Submitted By:

Ahbaz Memon	S1032180046	-	PB 04
Snehalraj Chugh	S1032181182	-	PB 27
Janhavi Chavan	S1032181707	-	PB 55
Harshit Srivastava	S1032181703	-	PB 54
Saahil Malge	S1032191699	-	PB 62

Under the guidance of

Prof. Shakti Kinger

MIT-World Peace University (MIT-WPU)

Faculty of Engineering & Technology
School of Computer Engineering & Technology
*** 2020-2021 ***

ABSTRACT

In today ‘s highly developed world, every minute, people around the globe express themselves via various platforms on the Web. And in each minute, a huge amount of unstructured data is generated. Such data is termed as big data. Twitter, one of the largest social media site receives millions of tweets every day on variety of important issues. This huge amount of raw data can be used for industrial, social, economic, government policies or business purpose by organizing according to our requirement and processing. Hadoop is one of the best tool options for twitter data analysis as it works for distributed big data, streaming data, time stamped data, text data etc. Hence, Flume is used to extract real time twitter data into HDFS. Hive and Pig which is SQL like query language is used for some extraction and analysis. People’s psychic and emotional wellbeing has been strongly proportional to this pandemic and they are suffering from panic, terror, and anxiety as the number of cases are increasing at an alarming rate around the world, we thus have retrieved data from related to the Yellow Fungus, Black & White Fungus.

These tweets have proven to be a valuable source of information in the recent years, playing key roles in success of brands, businesses and politicians. We have tackled Sentiment Analysis with a lexicon-based approach for extracting positive, negative, and neutral tweets by using part-of-speech tagging from natural language processing. We begin by collecting datasets for analysis, and then, depending on the method, we apply cleaning to it. After cleaning the dataset, we convert the rows into tokenized words and reflect the important words with the help of stemming & lemmatizations. The approach manifests in the design of a software toolkit that facilitates the sentiment analysis.

KEYWORDS: Healthcare · Hadoop · MapReduce · HDFS · Tableau · Big data · COVID-19 · Corona virus

TABLE OF CONTENTS

INTRODUCTION 1

1.1. Introduction 1

1.2. Motivation 3

1.3. Problem definition.. 4

1.4. Objectives 4

LITERATURE SURVEY 5

2.1. Existing and proposed system 5

2.2. Background study.. 6

***2.2.1.** Big Data...6*

***2.2.2.** Hadoop 8*

***2.2.3.** Technologies Used...10*

2.3. Dataset Description 11

2.4. System architecture 12

METHODOLOGY 14

3.1. Data Analysis (Any algorithm, queries or tools used) 14

***3.1.1.** Creating Twitter Application 15*

***3.1.2.** Getting Data using Python 17*

3.2. Hive queries 19

3.3. Visualization screenshots 29

FUTURE SCOPE 33

4.1. Conclusion 33

4.2. Future Work.. 35

REFERENCES...36

LIST OF FIGURES

Figure 1: Big Data 6
Figure 2: Types of Big Data 7
Figure 3: Architecture of Hadoop 9
Figure 4: Daemons of Hadoop 13
Figure 5: CSV format of our Dataset 14
Figure 6: Develop.Twitter 15
Figure 7: Creating new app 15
Figure 8: Agreement........ 16
Figure 9: Manage Keys & Tokens........ 16
Figure 10: Create your own access tokens........ 17
Figure 11: Getting Data using Python - 1 17
Figure 12: Getting Data using Python – 2 18
Figure 13: Getting Data using Python - 3........ 18
Figure 14: Loading data into tweets table........ 19
Figure 15: Displaying the table........ 20
Figure 16: Description of table dictionary........ 22
Figure 17: Viewing the dictionary table........ 22
Figure 18: View temp 1 23
Figure 19: View temp 2 24
Figure 20: View temp 3-1........ 25
Figure 21: View temp 3 -2........ 25
Figure 22: Creating sentiments table -1........ 26
Figure 23: Creating sentiments table -2........ 27
Figure 24: Creating sentiments table - 3........ 27
Figure 25: Output with tweets & ID - 1........ 28
Figure 26: Output with tweets & ID - 2........ 28
Figure 27: User Followers vs User Location........ 29
Figure 28: Count of User by Sentiment........ 29
Figure 29: User verified by Location 29
Figure 30: Count of Sentiment by User and User Location 30
Figure 31: Favorite Count by sentiment and user location........ 30
Figure 32: Retweet Count by Sentiment 30
Figure 33: User Followers by User........ 31
Figure 34: Retweet count and Favorite count by User........ 31
Figure 35: User verified and count of tweets by Sentiment........ 31
Figure 36: Retweet Count by User 32
Figure 37: Count of User by Sentiment........ 32

CHAPTER 1

INTRODUCTION

1.1. Introduction

Blogging and networking platforms like Facebook, Reddit, Twitter and LinkedIn are social media channels where users can share their thoughts and opinions. Since online chatter is a vital and exhaustive source of information, these thoughts and opinions hold the key to the success of any endeavour. Tweets which are posted by millions all over the world can be used to analyse consumers ‟opinions about individual products, services and campaigns.

Lately, people's psychic and emotional wellbeing has been strongly proportional to this pandemic and they are suffering from panic, terror, and anxiety as the number of cases are increasing at an alarming rate around the world. When the second wave of covid emerged this year, it was notable not just because of the disruption it caused, but rather because it occurred suddenly at a time when people were returning to normal. Information of this kind assists public health authorities in identifying popular issues and get a glimpse of all the global health messages sensed in the globe across Twitter. All of these will aid in improved policy design while keeping social perspectives in mind, as well as the government's awareness of issues such as food poverty, vaccination shortages, and so on.

While the country is still dealing with the coronavirus pandemic, cases of black fungus or mucormycosis are on the rise in the country. The cases of white fungus and yellow fungus have also been reported from some parts of the country. On one hand, due to the suddenness and novelty of these fungus, there was a serious lack of knowledge of how deadly this can bw, and it is difficult to quickly develop a cure for it. On the other hand, governments have not taken timely preventive measures to suppress the spread of this, resulting in severe damage to human health and societal stability. In addition, before and after the outbreak, because of the lack of an effective early warning, rapid response mechanisms, implementation of effective prevention and control decisions, the best prevention time have been missed. In order to respond efficiently and propose a preventative control plan, not only do we need a complete emergency management system, but also scientific data analysis for decision support.

However, with the spread of these diseases, the analysis results based on the gradually generated large amounts of data have played an important role in the tracking of people's movements, early warning of high-risk areas, screening of asymptomatic potential infections, drug development, information release, and policy support. They have also become an important basis for the implementation of preventive control programs and have played an important role in enhancing the modernization level of national governance, promoting protection and improving people's livelihood.

Black Fungus is a term given to the disease called Mucormycosis. It is a fungal infection caused by the Mucorales order in which the species most commonly implicated are Rhizopus, Mucor and Absidia. Mycosis is a term for fungal infection and hence Mucormycosis-Mycosis caused by Mucor.

White fungus is a genus of Yeast called Candida. It grows in the lab as white/ creamy white spots on plates of agar. In humans, they again appear as white, creamy spots on the mucosa of the oral cavity most commonly. They are commensals in the oral cavity and gastrointestinal tract and are present on skin as well. Candida albicans is the most commonly isolated species. Albicans means white. Hence the term white fungus.

Yellow fungus is a term coined by someone and unfortunately does not have any significance. Yellow is the colour of pus, which is formed whenever there is any bacterial infection. Superadded on that, there may be fungal growth of a wide variety of species. It does not mean that the fungal infection is causing the colour. It also most definitely does not mean that there is a "new yellow fungus which is more dangerous than black and white fungus"

These infections have been prevalent before, and we have been dealing with fewer cases, and in select groups having low immunity. These are not new diseases. The sheer numbers though are significant due to the number of patients with lowered immunity and additional risk factors. There may soon be a “Green fungus” or “Multicoloured fungus” variant reports coming in from different parts of the country.

While the country is still dealing with the coronavirus pandemic, cases of black fungus or mucormycosis are on the rise in the country. Meanwhile, cases of white fungus and yellow fungus have also been reported from some parts of the country.

Moreover, there is still a lack of a theoretical framework for big data analytics in the prevention and control of Major Public Health Incidents (including these fungus). Thus, it is necessary to propose such a framework to focus on the prevention and control of it using big data, which is also applicative in instances of other epidemic diseases. The proposed definitions, characteristics, data sources, applications, and framework can also enlighten and support the decisions of governments, enterprise, medical institutions, users, and researchers.

1.2. Motivation

Little work has been done to actually expand on the topic of the correlation between Twitter sentiment and these new diseases. Today we are living in the world which is surrounded by 99% of data. There are different microblogging sites where users express their views about different products these views are nothing but opinions of people and it will go waste if it is not used in proper way so there is a need to use opinions of people in improving productivity, usefulness, functionality of particular product or application or technique or any entertainment resource. Thus, it is a good idea to use Big Data technologies to perform sentiment analysis.

1.3. Problem definition

The project focuses on using Twitter, the most popular micro blogging platform, for the task of sentiment analysis. The tweets are important for analysis because data arrive at a high frequency and algorithms that process them must do so under very strict constraints of storage and time. It will be shown how to automatically collect a corpus for sentiment analysis and opinion mining purposes and then perform linguistic analysis of the collected corpus. All public tweets posted on twitter are freely available through a set of APIs provided by Twitter. Using the corpus, a sentiment classifier, is constructed that is able to determine positive, negative and neutral sentiments.

1.4. Objectives

Twitter has over a billion users and everyday people generate billions of tweets over 100 hours per minute and this number is ever increasing. To analyse and understand the activity occurring on such a massive scale, a relational SQL database is not enough. Such kind of data is well suited to a massively parallel and distributed system like Hadoop. The main objective of this project is to focus on how data generated from Twitter can be mined and utilized by different companies to make targeted, real time and informed decisions about their product that can decrease the disease or to find out the views of people on a specific topic of interest. This can be done by using Hadoop concepts. The given project will focus on how data generated from Twitter can be mined and utilized. There are multiple applications of this project. Companies, Doctors, Politicians, Decision Makers, etc can use this project to understand how effective and penetrative their plans & programs are through sentiment analysis.

CHAPTER 2

LITERATURE SURVEY

2.1. Existing and proposed system

The major issues involved in big data are the following:

- The first challenge faced is storing and accessing the information from the large huge amount of data sets from the clusters. We need a standard computing platform to manage large data since the data is growing, and data stores in different data storage locations in a centralized system, which will scale down the huge data into sizable data for computing.

- The second challenge is retrieving the data from the large social media data sets. In the scenarios where the data is growing daily, it’s somewhat difficult to accessing the data from the large networks if we want to do specific action to be performed.

- The third challenge concentrates on the algorithm design for handling the problems raised by the huge data volume and the dynamic data characteristics.

- The main scope of the project is to fetching and analysing the tweets on Types of fungal disease and to perform sentiment analysis to find the most popular tweets which are trending and finding the sentiment rating of each tweet based on that topic. Sentiment Analysis is the process of detecting the contextual polarity of text. A common use case for this technology is to discover how people feel about a particular topic.

2.2. Background study

These days internet is being widely used than it was used a few years back. Billions of people are using social media and social networking every day all across the globe. Such a huge number of people generate a flood of data which have become quite complex to manage. Considering this enormous data, a term has been coined to represent it. This term is called Big Data. Big Data is the term coined to refer this huge amount of data. The concept of big data is fast spreading its arms all over the world.

2.2.1. Big Data

Data which is very large in size and yet growing exponentially with time is called as big data. It refers to the large volume of data which may be structured or unstructured and which make use of certain new technologies and techniques to handle it

Hadoop is a programming framework used to support the processing of large data sets in a distributed computing environment. It provides storage for a large volume of data along with advanced processing power. It also gives the ability to handle multiple tasks and jobs. Hadoop was developed by Google's MapReduce that is a software framework where an application break down into various parts. The Apache Hadoop ecosystem consists of the Hadoop Kernel, MapReduce, HDFS and numbers of various components like Apache Flume, Apache Hive and Apache Pig which are being used in this project.

Figure 1: Big Data

2.2.1.1. Categories of Big Data

Structured Data: The data which can be stored and processed in table (rows and column) format is called as a structured data. Structured data is relatively simple to enter, store and analyze. Example - Relational database management system.

Unstructured Data: The data with unknown form or structure is called as unstructured data. They are difficult for nontechnical users and data analysts to understand and process. Example - Text files, images, videos, email, webpages, PDF files, PPT, social media data etc.

Semi-structured Data: Semi-structured data is data that is neither raw data nor organized in a rational model like a table. XML and JSON documents are semi structured documents.

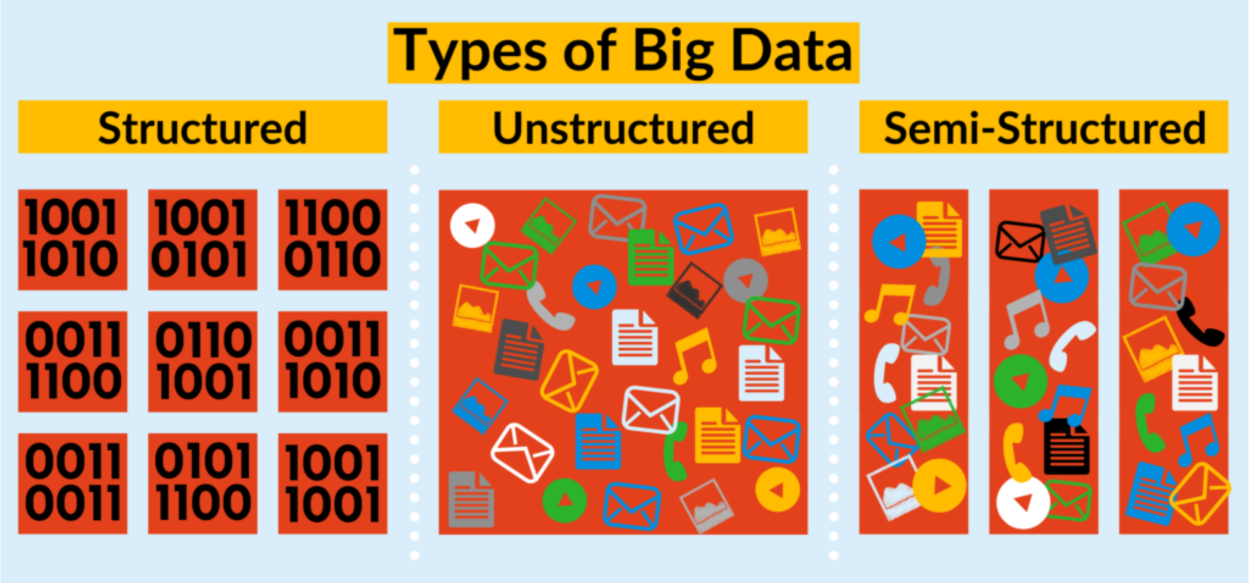


Figure 2: Types of Big Data

2.2.1.2. *Characteristics of Big Data*

The characteristics of Big Data are defined by three V's:

- **Volume –** It refers to the amount of data that is generated. The data can be low density, high volume, structured/unstructured or data with unknown value. The data can range from terabytes to petabytes.
- **Velocity –** It refers to the rate at which the data is generated. The data is received at an unprecedented speed and is acted upon in a timely manner.
- **Variety –** Variety refers to different formats of data. It may be structured, unstructured or semi- structured. The data can be audio, video, text or email.

2.2.2. Hadoop

As organizations are getting flooded with massive amount of raw data, the challenge here is that traditional tools are poorly equipped to deal with the scale and complexity of such kind of data. That's where Hadoop comes in. Hadoop is well suited to meet many Big Data challenges, especially with high volumes of data and data with a variety of structures.

Hadoop is a framework for storing data on large clusters of commodity hardware, everyday computer hardware that is affordable and easily available and running applications against that data. A cluster is a group of interconnected computers (known as nodes) that can work together on the same problem. The Current Apache Hadoop ecosystem consists of the Hadoop Kernel, Map-Reduce, HDFS and numbers of various components like Apache Hive, Pig, Flume etc.

Hadoop consists of two main components:

- HDFS (Data Storage)
- Map-Reduce (Analysing and Processing)

2.2.2.1. Architecture of Hadoop

HDFS is the main component of Hadoop architecture. It stands for Hadoop Distributed File Systems. It is used to store a large amount of data and multiple machines are used for this storage. MapReduce is another component of big data architecture. The data is processed here in a distributed manner across multiple machines. So, HDFS works as a storage part and MapReduce works as a processing part. Hive and Pig are the components of Hadoop ecosystem. These are high level data flow languages. MapReduce is the inner most layer of Hadoop ecosystem.

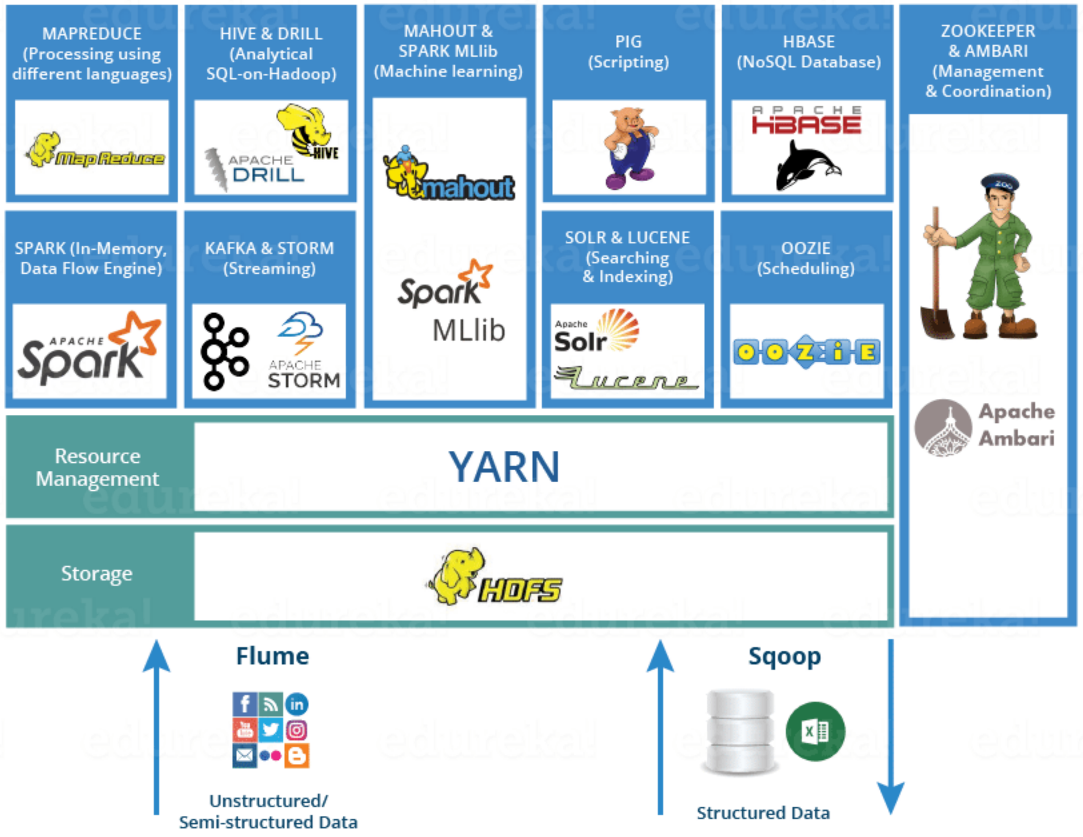

MAPREDUCE (Processing using different languages)
HIVE & DRILL (Analytical SQL-on-Hadoop)
MAHOUT & SPARK MLlib (Machine learning)
PIG (Scripting)
HBASE (NoSQL Database)
ZOOKEEPER & AMBARI (Management & Coordination)
SPARK (In-Memory, Data Flow Engine)
KAFKA & STORM (Streaming)
SOLR & LUCENE (Searching & Indexing)
OOZIE (Scheduling)
Spark MLlib
Apache Ambari
Resource Management
YARN
Storage
HDFS
Flume
Unstructured/ Semi-structured Data
Sqoop
Structured Data

Figure 3: Architecture of Hadoop

2.2.3. Technologies Used

Apache Flume: Apache Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of streaming data into the Hadoop Distributed File System (HDFS). It can be used for dumping twitter data in Hadoop HDFS. It has a simple and flexible architecture based on streaming data flows; and is robust and fault tolerant with tuneable reliability mechanisms for failover and recovery. Flume lets Hadoop users ingest high-volume streaming data into HDFS for storage.

Apache Hive: Hive is a data warehouse infrastructure tool to process structured data in Hadoop. It resides on top of Hadoop to summarize Big Data, and makes querying and analysing easy. Hive provides the ability to store large amounts of data in HDFS. Hive was designed to appeal to a community comfortable with SQL. Hive uses an SQL like language known as HIVEQL. Its philosophy is that we don't need yet another scripting language. Hive supports maps and reduced transform scripts in the language of the user's choice which can be embedded with SQL. Supporting SQL syntax also means that it is possible to integrate with existing tools like. Hive has an ODBC (Open Database Connectivity JDBC (Java Database Connectivity) driver that allows and facilitates easy queries. It also adds support for indexes which allows support for queries common in such environment. Hive is a framework for performing analytical queries. Big Data enterprises require fast analysis of data collected over a period of time. Hive is an excellent tool for analytical querying of historical data. It is to be noted that the data needs to be well organized, which would allow Hive to fully unleash its processing and analytical powers.

When you are looking to process clusters of unorganized, unstructured, decentralized data and don't want to deviate too much from your solid SQL foundation, Pig is the option to go with. You no longer need to get into writing core MapReduce jobs. If you already have SQL background, the learning curve will be smooth and development time will be faster.

2.3. Dataset Description

In this proposed method, Live Dataset is retrieved from twitter using flume related to the keywords: ***black fungus, white fungus, yellow fungus, fungus, fungai, Mucormycosis, Mucor, zygomycosis, mucoromycetes, Candidiasis, Candida, aspergillosis.*** The dataset comprises of factual information, results of examinations and information given by the users. The coloumns which are present in the dataset are ***Tweets, User, user statuses count, user followers, user location, user verified, fav count, retweet count, tweet date, sentiment.*** Some of the subjective features like fav count and retweet count binary outputs, user status count which is an examination feature showcases three different scenarios wherein helping in understanding user's activity. Tweets reflects the chances of understanding what are people's sentiment and ideation over fungal infection.

The dataset is divided into smaller blocks which are primarily processed by "Map Phase" in parallel and then by "Reduce Phase". Hadoop framework has sorted out the output of the Map phase which are then given as an input to Reduce Phase to initiate parallel reduce tasks. These input and output files are stored in file system.

Input dataset is adapted by MapReduce framework from HDFS. Input dataset is taken as key-value pair and broken down for effective analysis. Then, the number of positive cases is mapped to corresponding diseases and shuffled accordingly. The number of positive cases is the reduced output. The final result is sent to the authorities and hospitals for preventive measures.

2.4. System architecture

In this work five MapReduce key functions are used to get the desired output/outcome. The functionalities of each functions are as follows:

Hadoop offers five daemons with each daemon possessing a Java Virtual Machine-

- Data Node
- Name Node
- Job Tracker
- Secondary Name Node
- Task- Tracker

Demons which store data and metadata, i.e.; DataNode and NameNode, come under a part of Hadoop Distributed File System (HDFS). The TaskTracker and JobTracker, which keep track and actually execute the job, come under MapReduce layer. The HDFS is used in this proposed research work because of the following reasons.

Large Dataset: Current population of India is more than 1.3 billion, if we want to analyze data for that amount of people then the dataset will be huge. That much huge amount of data can’t be processed by normal file system. That’s why to get a smooth workflow the HDFS is used for analysis.

Data Replication: For working with a large dataset, occurrences of unfortunate situations like hardware failure, crashing of a node are pretty common. In such situations data loss is occurred. To overcome this kind of problem HDFS is providing a feature called data replication. The data is copied across numerous nodes in the cluster by the creation of duplicates. This methodology is maintained across stipulated time intervals by HDFS and the duplication process is taken care of by the same. The moment as machine in the cluster crashes, the data should be retrieved from other machines. Loss of data is far sighted threat and almost negligible.

Scalability: Our main goal of the work is to analyze healthcare dataset using Hadoop and facilitate a smoother conduct of the fight against COVID-19. So, the proposed work is scalable in order make it a dynamic project. This is achieved by using HDFS. In HDFS the infrastructure is scaled up by adding more racks or clusters to this system.

Data Locality: In older systems, the data is brought at the application layer and then worked upon. In this proposed research work, as a consequence of the huge bulk of data, bringing data to the application layer has lowered down the overall performance.

In HDFS, the computation part is brought to the Data Nodes where data resides. Hence, with Hadoop HDFS, computation logic is not moved to the data, rather than data is moved to the computation logic.

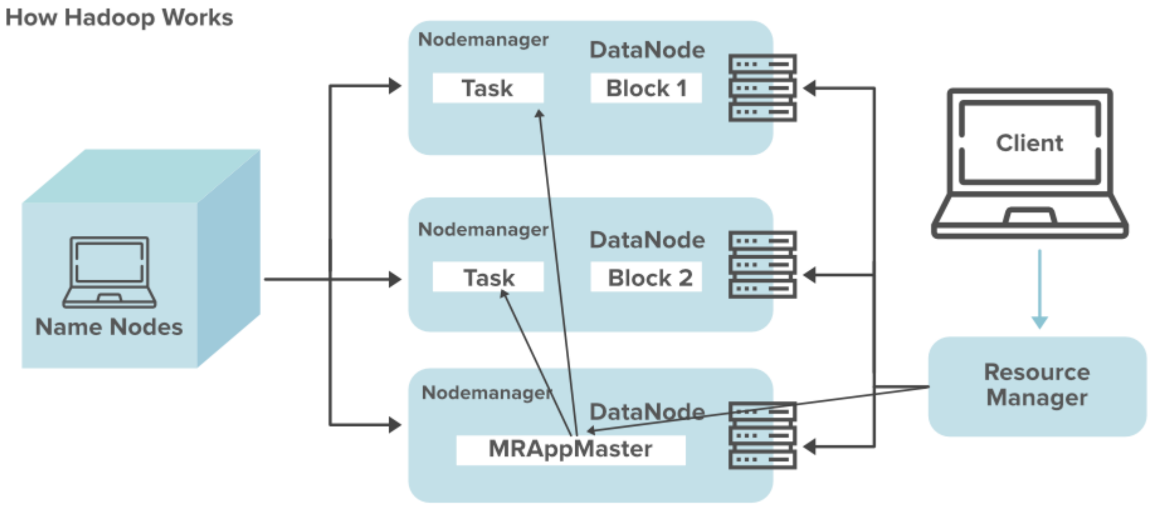


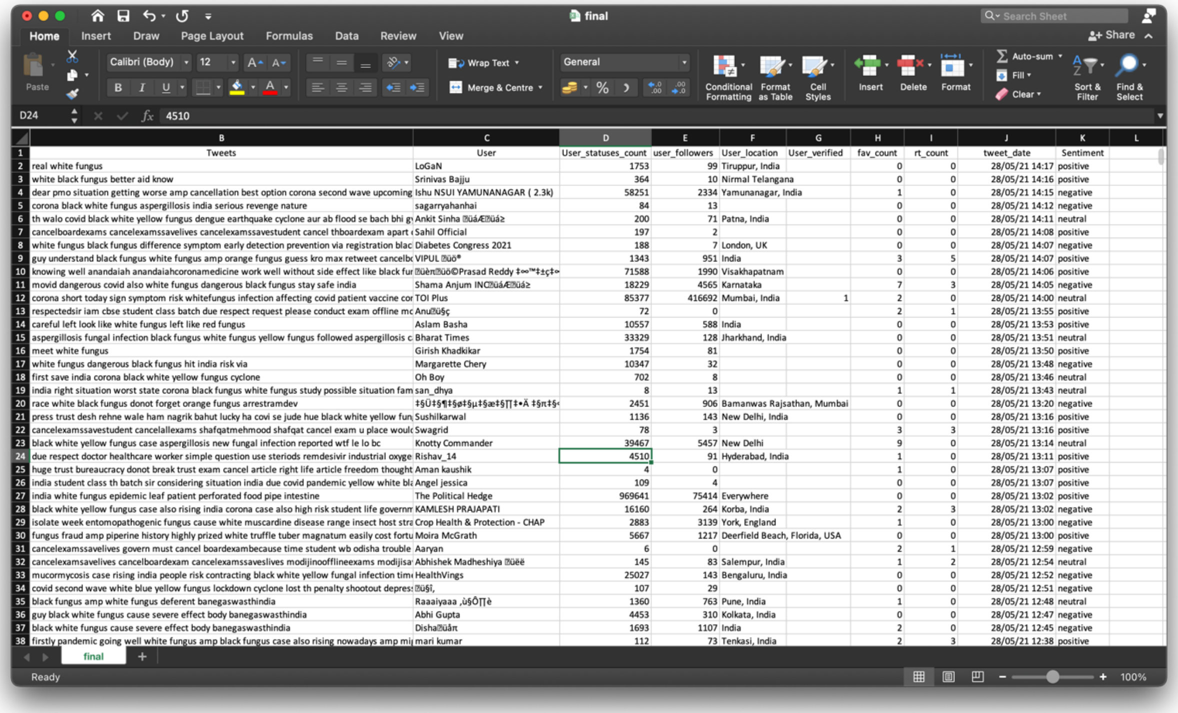
Figure 4: Daemons of Hadoop

CHAPTER 3

METHODOLOGY

3.1. Data Analysis (Any algorithm, queries or tools used)

Given below is a screenshot of our live retrieved dataset:

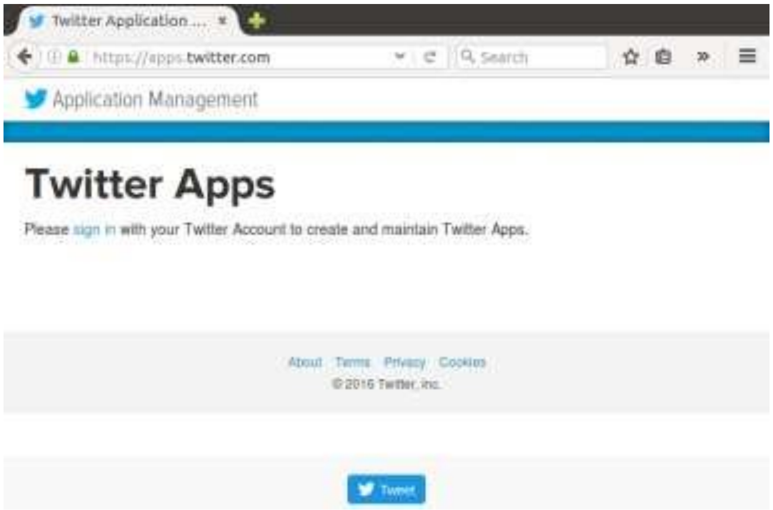


	B	C	D	E	F	G	H	I	J	K	L
1	Tweets	User	User_statuses_count	user_followers	User_location	User_verified	fav_count	rt_count	tweet_date	Sentiment	
2	real white fungus	LoGaN	1753	99	Tiruppur, India		0	0	28/05/21 14:17	positive	
3	white black fungus better aid know	Srinivas Bajju	364	10	Nirmal Telangana		0	0	28/05/21 14:16	positive	
4	dear pmo situation getting worse amp cancellation best option corona second wave upcoming	Ishu NSUI YAMUNANAGAR (2.3k)	58251	2334	Yamunanagar, India		1	0	28/05/21 14:15	negative	
5	corona black white fungus aspergillosis india serious revenge nature	sagarryahanhai	84	13			0	0	28/05/21 14:12	negative	
6	th walo covid black white yellow fungus dengue earthquake cyclone aur ab flood se bach bhi gy	Ankit Sinha ῦüáÆῦüá≥	200	71	Patna, India		0	0	28/05/21 14:11	neutral	
7	cancelboardexams cancelexamssavelives cancelexamssavestudent cancel thboardexam apart	Sahil Official	197	2			0	0	28/05/21 14:08	positive	
8	white fungus black fungus difference symptom early detection prevention via registration blac	Diabetes Congress 2021	188	7	London, UK		0	0	28/05/21 14:07	negative	
9	guy understand black fungus white fungus amp orange fungus guess kro max retweet cancelbo	VIPUL ῦüö®	1343	951	India		3	5	28/05/21 14:07	positive	
10	knowing well anandaiah anandaiahcoronamedicine work well without side effect like black fur	ῦüèπῦüö©Prasad Reddy ‡∞™‡±ç‡∞	71588	1990	Visakhapatnam		0	0	28/05/21 14:06	positive	
11	movid dangerous covid also white fungus dangerous black fungus stay safe india	Shama Anjum INCῦüáÆῦüá≥	18229	4565	Karnataka		7	3	28/05/21 14:05	negative	
12	corona short today sign symptom risk whitefungus infection affecting covid patient vaccine cor	TOI Plus	85377	416692	Mumbai, India	1	2	0	28/05/21 14:00	neutral	
13	respectedsir iam cbse student class batch due respect request please conduct exam offline mc	Anuῦü§ç	72	0			2	1	28/05/21 13:55	positive	
14	careful left look like white fungus left like red fungus	Aslam Basha	10557	588	India		0	0	28/05/21 13:53	positive	
15	aspergillosis fungal infection black fungus white fungus yellow fungus followed aspergillosis c	Bharat Times	33329	128	Jharkhand, India		0	0	28/05/21 13:51	neutral	
16	meet white fungus	Girish Khadkikar	1754	81			0	0	28/05/21 13:50	positive	
17	white fungus dangerous black fungus hit india risk via	Margarette Chery	10347	32			0	0	28/05/21 13:48	negative	
18	first save india corona black white yellow fungus cyclone	Oh Boy	702	8			0	0	28/05/21 13:46	neutral	
19	india right situation worst state corona black fungus white fungus study possible situation fam	san_dhya	8	13			1	1	28/05/21 13:43	neutral	
20	race white black fungus donot forget orange fungus arrestramdev	‡§Ü‡§¶‡§ø‡§µ‡§æ‡§∏‡•Ä ‡§π‡§	2451	906	Bamanwas Rajsathan, Mumbai		0	0	28/05/21 13:20	negative	
21	press trust desh rehne wale ham nagrik bahut lucky ha covi se jude hue black white yellow fun	Sushilkarwal	1136	143	New Delhi, India		0	0	28/05/21 13:16	positive	
22	cancelexamssavestudent cancelallexams shafqatmehmood shafqat cancel exam u place woulc	Swagrid	78	3			3	3	28/05/21 13:16	positive	
23	black white yellow fungus case aspergillosis new fungal infection reported wtf le lo bc	Knotty Commander	39467	5457	New Delhi		9	0	28/05/21 13:14	neutral	
24	due respect doctor healthcare worker simple question use steriods remdesivir industrial oxyge	Rishav_14	4510	91	Hyderabad, India		1	0	28/05/21 13:11	positive	
25	huge trust bureaucracy donot break trust exam cancel article right life article freedom thought	Aman kaushik	4	0			1	0	28/05/21 13:07	positive	
26	india student class th batch sir considering situation india due covid pandemic yellow white bla	Angel jessica	109	4			0	0	28/05/21 13:07	positive	
27	india white fungus epidemic leaf patient perforated food pipe intestine	The Political Hedge	969641	75414	Everywhere		0	0	28/05/21 13:02	positive	
28	black white yellow fungus case also rising india corona case also high risk student life governm	KAMLESH PRAJAPATI	16160	264	Korba, India		2	3	28/05/21 13:02	negative	
29	isolate week entomopathogenic fungus cause white muscardine disease range insect host stra	Crop Health & Protection - CHAP	2883	3139	York, England		1	0	28/05/21 13:00	negative	
30	fungus fraud amp piperine history highly prized white truffle tuber magnatum easily cost fortu	Moira McGrath	5667	1217	Deerfield Beach, Florida, USA		0	0	28/05/21 13:00	positive	
31	cancelexamssavelives govern must cancel boardexambecause time student wb odisha trouble	Aaryan	6	0			2	1	28/05/21 12:59	negative	
32	cancelexamsavelives cancelboardexam cancelexamssaveslives modijinoofflineexams modijisa	Abhishek Madheshiya ῦüëë	145	83	Salempur, India		1	2	28/05/21 12:54	neutral	
33	mucormycosis case rising india people risk contracting black white yellow fungal infection tim	HealthVings	25027	143	Bengaluru, India		0	0	28/05/21 12:52	negative	
34	covid second wave white blue yellow fungus lockdown cyclone lost th penalty shootout depres	ῦü§î,	107	29			0	0	28/05/21 12:51	negative	
35	black fungus amp white fungus deferent banegaswasthindia	Raaaiyaaa ,ù§Ô∏è	1360	763	Pune, India		1	0	28/05/21 12:48	neutral	
36	guy black white fungus cause severe effect body banegaswasthindia	Abhi Gupta	4453	310	Kolkata, India		0	0	28/05/21 12:47	negative	
37	black white fungus cause severe effect body banegaswasthindia	Dishaῦüåπ	1693	1107	India		2	0	28/05/21 12:45	negative	
38	firstly pandemic going well white fungus amp black fungus case also rising nowadays amp mi	mari kumar	112	73	Tenkasi, India		2	3	28/05/21 12:38	positive	

Figure 5: CSV format of our Dataset

3.1.1. Creating Twitter Application

1. ***Open the website dev.twitter.com/apps in the Browser.***

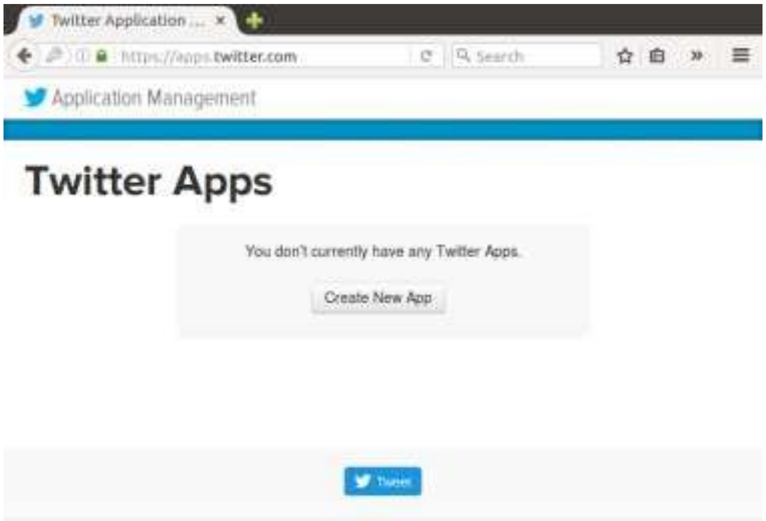


Please sign in with your Twitter Account to create and maintain Twitter Apps.

Figure 6: Develop.Twitter

2. ***We will now see the website suggesting us to sign in. So, we sign into our twitter account.***

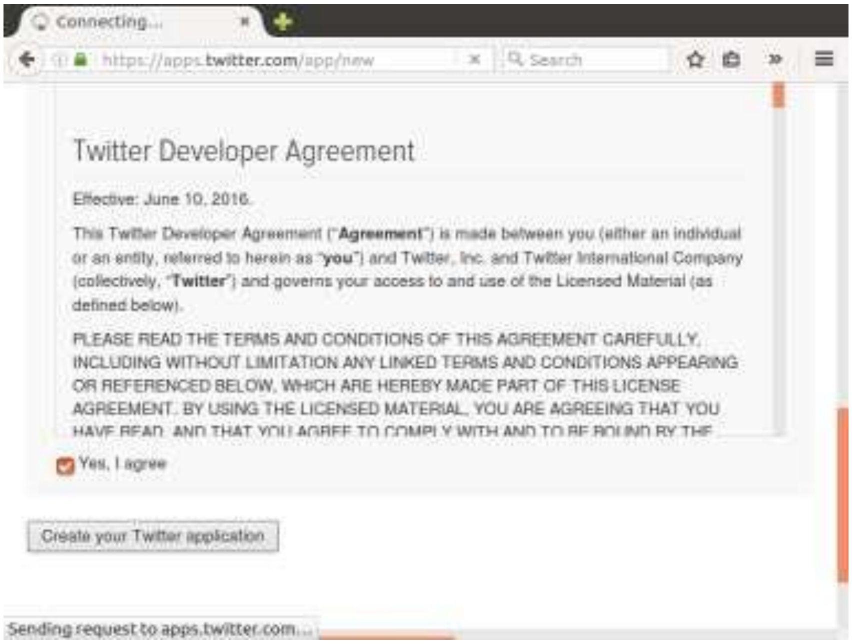
   Click on Create New App.



You don't currently have any Twitter Apps.

Figure 7: Creating new app

3. ***Now, scroll down and tick the option Yes, I agree and then click Create your Twitter application.***



Twitter Developer Agreement

Effective: June 10, 2016.

This Twitter Developer Agreement ("**Agreement**") is made between you (either an individual or an entity, referred to herein as "**you**") and Twitter, Inc. and Twitter International Company (collectively, "**Twitter**") and governs your access to and use of the Licensed Material (as defined below).

PLEASE READ THE TERMS AND CONDITIONS OF THIS AGREEMENT CAREFULLY, INCLUDING WITHOUT LIMITATION ANY LINKED TERMS AND CONDITIONS APPEARING OR REFERENCED BELOW, WHICH ARE HEREBY MADE PART OF THIS LICENSE AGREEMENT. BY USING THE LICENSED MATERIAL, YOU ARE AGREEING THAT YOU HAVE READ, AND THAT YOU AGREE TO COMPLY WITH AND TO BE BOUND BY THE

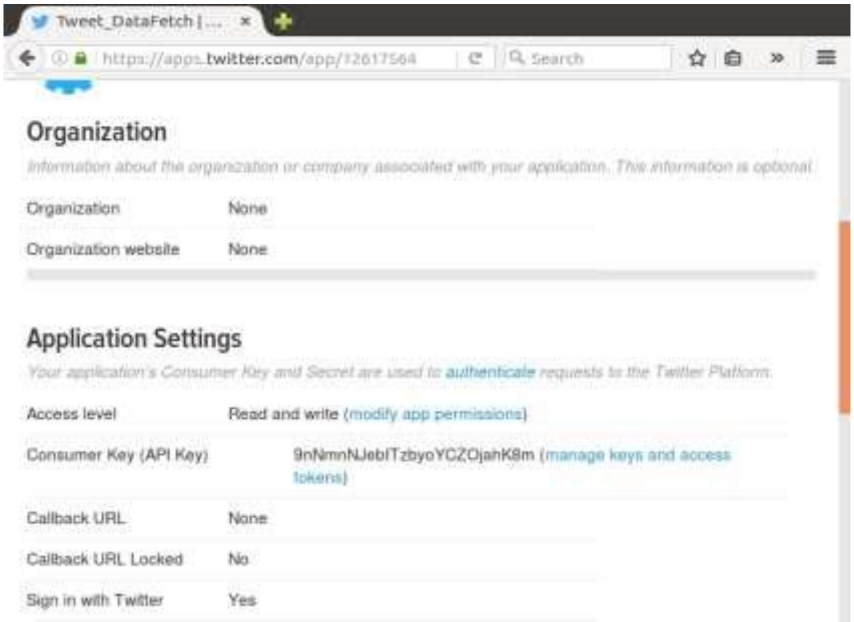
☑ Yes, I agree

Create your Twitter application

Sending request to apps.twitter.com...

Figure 8: Agreement

4. ***Click on manage keys and access tokens.***



Organization

Information about the organization or company associated with your application. This information is optional.

Organization	None
Organization website	None

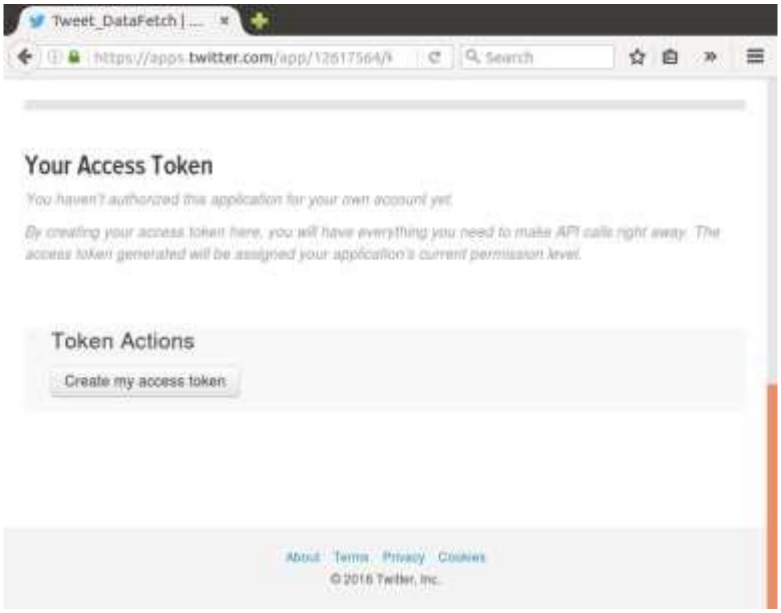
Application Settings

Your application's Consumer Key and Secret are used to authenticate requests to the Twitter Platform.

Access level	Read and write (modify app permissions)
Consumer Key (API Key)	9nNmnNJebITzbyoYCZOjahK8m (manage keys and access tokens)
Callback URL	None
Callback URL Locked	No
Sign in with Twitter	Yes

Figure 9: Manage Keys & Tokens

5. ***Now click on Create my access token.***



Your Access Token

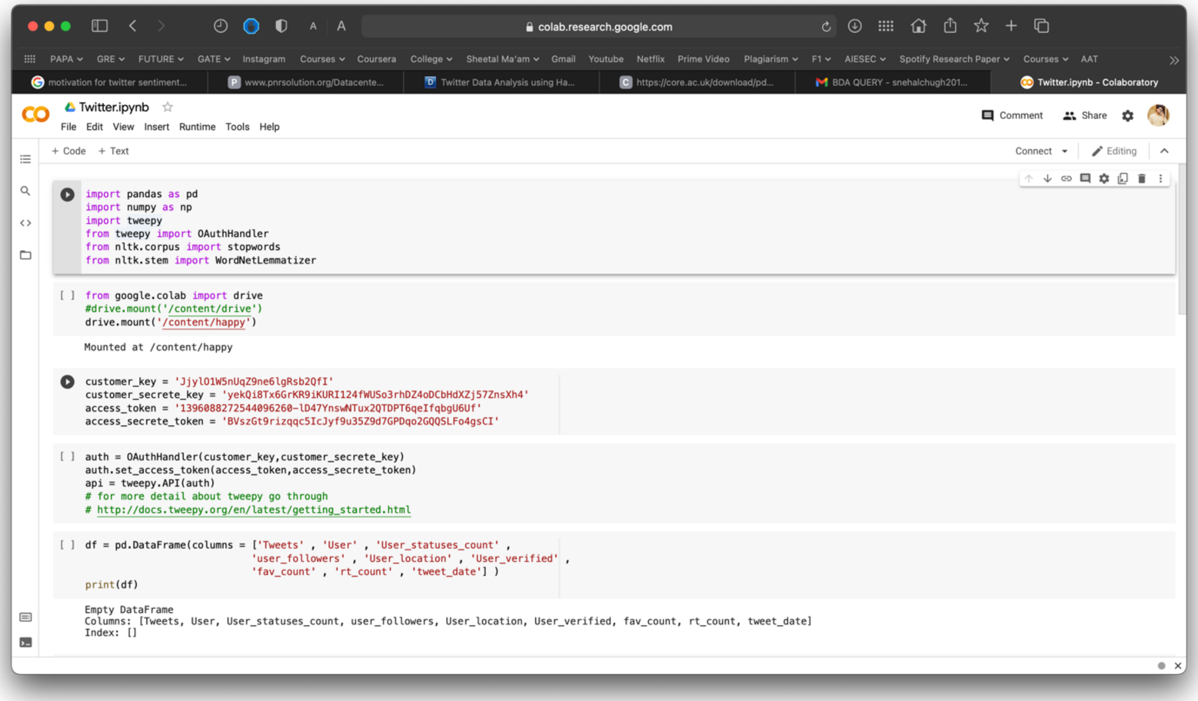
You haven't authorized this application for your own account yet.

By creating your access token here, you will have everything you need to make API calls right away. The access token generated will be assigned your application's current permission level.

Token Actions

Figure 10: Create your own access tokens

3.1.2. Getting Data using Python


colab.research.google.com
Twitter.ipynb
File Edit View Insert Runtime Tools Help
+ Code + Text
Comment Share
Connect Editing

```
import pandas as pd
import numpy as np
import tweepy
from tweepy import OAuthHandler
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
```

```
from google.colab import drive
#drive.mount('/content/drive')
drive.mount('/content/happy')
```

```
Mounted at /content/happy
```

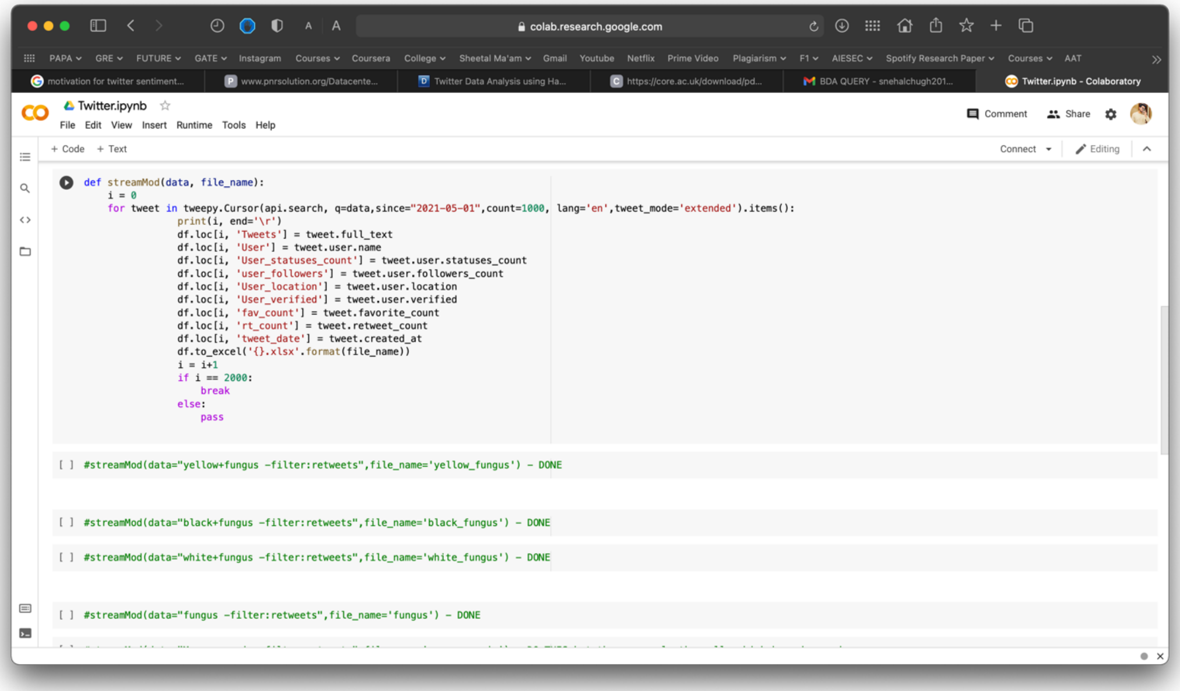
```
customer_key = 'Jjyl01W5nUqZ9ne6lgRsb2QfI'
customer_secrete_key = 'yekQi8Tx6GrKR9iKURI124fWUSo3rhDZ4oDCbHdXZj57ZnsXh4'
access_token = '1396088272544096260-lD47YnswNTux2QTDPT6qeIfqbgU6Uf'
access_secrete_token = 'BVszGt9rizqqc5IcJyf9u35Z9d7GPDqo2GQQSLFo4gsCI'
```

```
auth = OAuthHandler(customer_key,customer_secrete_key)
auth.set_access_token(access_token,access_secrete_token)
api = tweepy.API(auth)
# for more detail about tweepy go through
# http://docs.tweepy.org/en/latest/getting_started.html
```

```
df = pd.DataFrame(columns = ['Tweets' , 'User' , 'User_statuses_count' ,
                             'user_followers' , 'User_location' , 'User_verified' ,
                             'fav_count' , 'rt_count' , 'tweet_date'] )
print(df)
```

```
Empty DataFrame
Columns: [Tweets, User, User_statuses_count, user_followers, User_location, User_verified, fav_count, rt_count, tweet_date]
Index: []
```

Figure 11: Getting Data using Python - 1


colab.research.google.com
Twitter.ipynb
File Edit View Insert Runtime Tools Help
+ Code + Text
Comment Share
Connect Editing

```
def streamMod(data, file_name):
    i = 0
    for tweet in tweepy.Cursor(api.search, q=data,since="2021-05-01",count=1000, lang='en',tweet_mode='extended').items():
                print(i, end='\r')
                df.loc[i, 'Tweets'] = tweet.full_text
                df.loc[i, 'User'] = tweet.user.name
                df.loc[i, 'User_statuses_count'] = tweet.user.statuses_count
                df.loc[i, 'user_followers'] = tweet.user.followers_count
                df.loc[i, 'User_location'] = tweet.user.location
                df.loc[i, 'User_verified'] = tweet.user.verified
                df.loc[i, 'fav_count'] = tweet.favorite_count
                df.loc[i, 'rt_count'] = tweet.retweet_count
                df.loc[i, 'tweet_date'] = tweet.created_at
                df.to_excel('{}.xlsx'.format(file_name))
                i = i+1
                if i == 2000:
                    break
                else:
                    pass
```

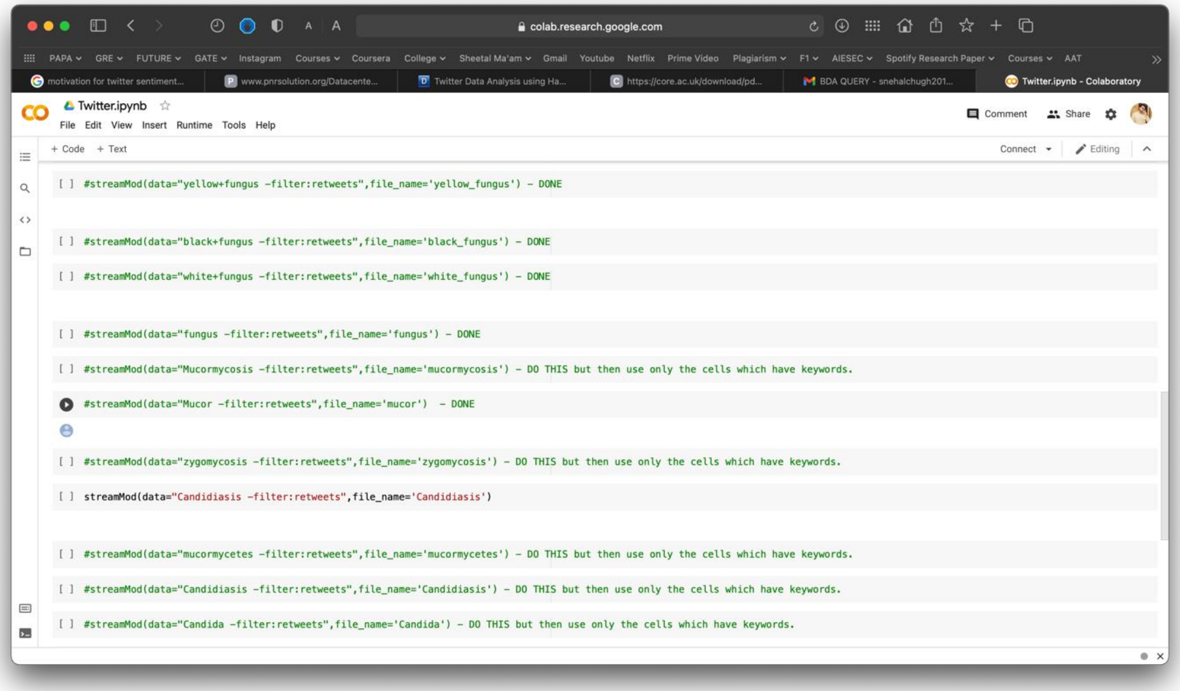
```
[ ] #streamMod(data="yellow+fungus -filter:retweets",file_name='yellow_fungus') - DONE

[ ] #streamMod(data="black+fungus -filter:retweets",file_name='black_fungus') - DONE

[ ] #streamMod(data="white+fungus -filter:retweets",file_name='white_fungus') - DONE

[ ] #streamMod(data="fungus -filter:retweets",file_name='fungus') - DONE
```

Figure 12: Getting Data using Python – 2


colab.research.google.com
Twitter.ipynb
File Edit View Insert Runtime Tools Help
+ Code + Text
Comment Share
Connect Editing

```
[ ] #streamMod(data="yellow+fungus -filter:retweets",file_name='yellow_fungus') - DONE

[ ] #streamMod(data="black+fungus -filter:retweets",file_name='black_fungus') - DONE

[ ] #streamMod(data="white+fungus -filter:retweets",file_name='white_fungus') - DONE

[ ] #streamMod(data="fungus -filter:retweets",file_name='fungus') - DONE

[ ] #streamMod(data="Mucormycosis -filter:retweets",file_name='mucormycosis') - DO THIS but then use only the cells which have keywords.

    #streamMod(data="Mucor -filter:retweets",file_name='mucor')  - DONE

[ ] #streamMod(data="zygomycosis -filter:retweets",file_name='zygomycosis') - DO THIS but then use only the cells which have keywords.

[ ] streamMod(data="Candidiasis -filter:retweets",file_name='Candidiasis')

[ ] #streamMod(data="mucormycetes -filter:retweets",file_name='mucormycetes') - DO THIS but then use only the cells which have keywords.

[ ] #streamMod(data="Candidiasis -filter:retweets",file_name='Candidiasis') - DO THIS but then use only the cells which have keywords.

[ ] #streamMod(data="Candida -filter:retweets",file_name='Candida') - DO THIS but then use only the cells which have keywords.
```

Figure 13: Getting Data using Python - 3

3.2. Hive queries

1. Create table tweets_exp.

CREATE EXTERNAL TABLE IF NOT EXISTS tweets_exp

(

id BIGINT,

tweet STRING,

user_name STRING,

user_statuses_count INT,

user_followers_count INT,

user_location STRING,

user_verified BOOLEAN,

tweet_fav_count INT,

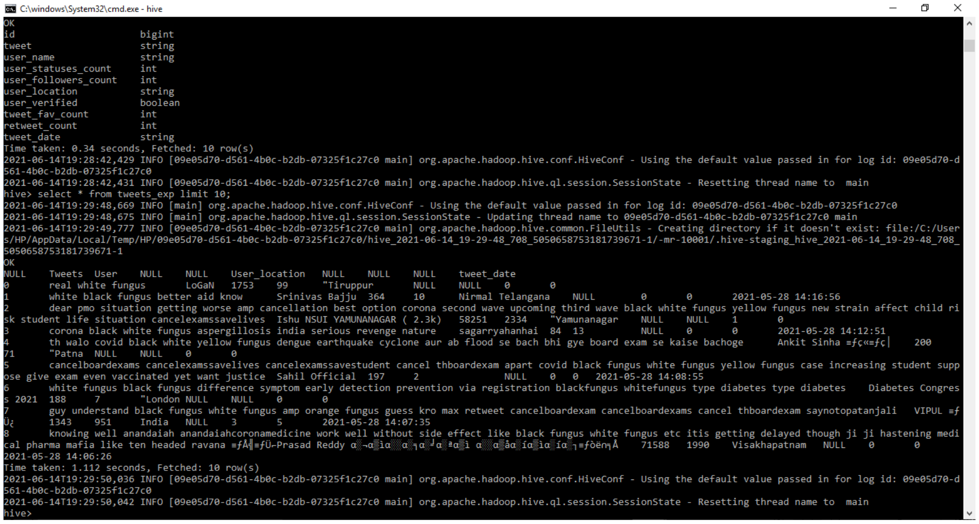
retweet_count INT,

tweet_date STRING

)

2. Load the data in tweets table

load data inpath '/BDAFINAL/final_white_plus_yellow_final.csv' into TABLE tweets_exp;



```
OK
id                      bigint
tweet                   string
user_name               string
user_statuses_count     int
user_followers_count    int
user_location           string
user_verified           boolean
tweet_fav_count         int
retweet_count           int
tweet_date              string
Time taken: 0.34 seconds, Fetched: 10 row(s)
2021-06-14T19:28:42,429 INFO [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.hive.conf.HiveConf - Using the default value passed in for log id: 09e05d70-d
561-4b0c-b2db-07325f1c27c0
2021-06-14T19:28:42,431 INFO [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.hive.ql.session.SessionState - Resetting thread name to  main
hive> select * from tweets_exp limit 10;
2021-06-14T19:29:48,669 INFO [main] org.apache.hadoop.hive.conf.HiveConf - Using the default value passed in for log id: 09e05d70-d561-4b0c-b2db-07325f1c27c0
2021-06-14T19:29:48,675 INFO [main] org.apache.hadoop.hive.ql.session.SessionState - Updating thread name to 09e05d70-d561-4b0c-b2db-07325f1c27c0 main
2021-06-14T19:29:49,777 INFO [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.hive.common.FileUtils - Creating directory if it doesn't exist: file:/C:/User
s/HP/AppData/Local/Temp/HP/09e05d70-d561-4b0c-b2db-07325f1c27c0/hive_2021-06-14_19-29-48_708_5050658753181739671-1/-mr-10001/.hive-staging_hive_2021-06-14_19-29-48_708_
5050658753181739671-1
OK
NULL    Tweets  User    NULL    NULL    User_location   NULL    NULL    NULL    tweet_date
0       real white fungus       LoGaN   1753    99      "Tiruppur       NULL    NULL    0       0
1       white black fungus better aid know      Srinivas Bajju  364     10      Nirmal Telangana    NULL        0       0       2021-05-28 14:16:56
2       dear pmo situation getting worse amp cancellation best option corona second wave upcoming third wave black white fungus yellow fungus new strain affect child ri
sk student life situation cancelexamssavelives  Ishu NSUI YAMUNANAGAR ( 2.3k)   58251   2334    "Yamunanagar    NULL    NULL    1       0
3       corona black white fungus aspergillosis india serious revenge nature    sagarryahanhai  84  13          NULL    0       0       2021-05-28 14:12:51
4       th walo covid black white yellow fungus dengue earthquake cyclone aur ab flood se bach bhi gye board exam se kaise bachoge      Ankit Sinha ≡ƒç«≡ƒç|    200
71      "Patna  NULL    NULL    0       0
5       cancelboardexams cancelexamssavelives cancelexamssavestudent cancel thboardexam apart covid black fungus white fungus yellow fungus case increasing student supp
ose give exam even vaccinated yet want justice  Sahil Official  197     2               NULL    0   0   2021-05-28 14:08:55
6       white fungus black fungus difference symptom early detection prevention via registration blackfungus whitefungus type diabetes type diabetes    Diabetes Congres
s 2021  188     7       "London NULL    NULL    0       0
7       guy understand black fungus white fungus amp orange fungus guess kro max retweet cancelboardexam cancelboardexams cancel thboardexam saynotopatanjali   VIPUL ≡ƒ
Ü¿      1343    951     India   NULL    3       5       2021-05-28 14:07:35
8       knowing well anandaiah anandaiahcoronamedicine work well without side effect like black fungus white fungus etc itis getting delayed though ji ji hastening medi
cal pharma mafia like ten headed ravana ≡ƒÅ╢≡ƒÜ−Prasad Reddy α░¬α▒ìα░░α░╕α░╛α░ªα▒ì α░░α▒åα░íα▒ìα░íα░┐≡ƒòëπÅ    71588   1990    Visakhapatnam   NULL    0       0
2021-05-28 14:06:26
Time taken: 1.112 seconds, Fetched: 10 row(s)
2021-06-14T19:29:50,036 INFO [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.hive.conf.HiveConf - Using the default value passed in for log id: 09e05d70-d
561-4b0c-b2db-07325f1c27c0
2021-06-14T19:29:50,042 INFO [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.hive.ql.session.SessionState - Resetting thread name to  main
hive>
```

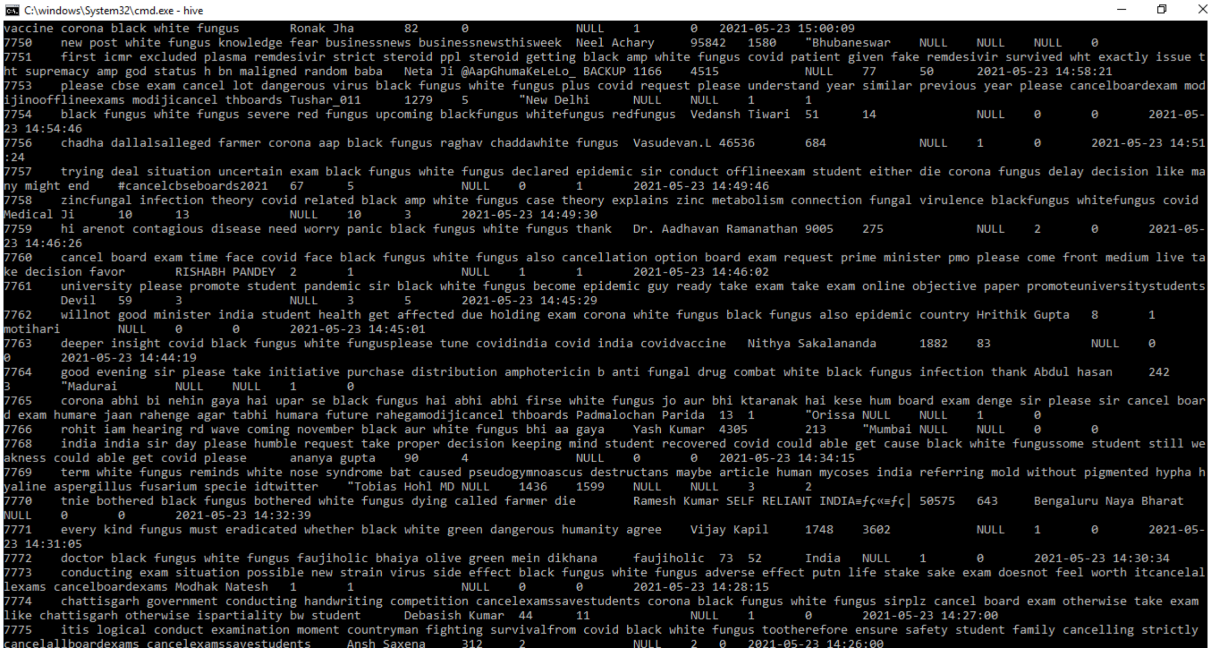
Figure 14: Loading data into tweets table

3. *Display the table*

Select * from table tweets_exp limit 10;

OUTPUT:

```
7750    new post white fungus knowledge fear businessnews businessnewsthisweek  Neel Achary    95842  1580    "Bhubaneswar    NULL    NULL    NULL    0
7751    first icmr excluded plasma remdesivir strict steroid ppl steroid getting black amp white fungus covid patient given fake remdesivir survived wht exactly issue tht supremacy amp god status h bn maligned random baba   Neta Ji
@AapGhumaKeLeLo_ BACKUP 1166    4515            NULL    77      50      2021-05-23 14:58:21
7753    please cbse exam cancel lot dangerous virus black fungus white fungus plus covid request please understand year similar previous year please cancelboardexam modijinoofflineexams modijicancel thboards Tushar_011
1279    5       "New Delhi      NULL    NULL    1       1
7754    black fungus white fungus severe red fungus upcoming blackfungus whitefungus redfungus  Vedansh Tiwari  51      14              NULL    0       0       2021-05-23 14:54:46
7756    chadha dallalsalleged farmer corona aap black fungus raghav chaddawhite fungus  Vasudevan.L 46536       684             NULL    1       0       2021-05-23 14:51:24
7757    trying deal situation uncertain exam black fungus white fungus declared epidemic sir conduct offlineexam student either die corona fungus delay decision like many might end    #cancelcbseboards2021   67      5
NULL    0       1       2021-05-23 14:49:46
7758    zincfungal infection theory covid related black amp white fungus case theory explains zinc metabolism connection fungal virulence blackfungus whitefungus covid
Medical Ji      10      13              NULL    10      3       2021-05-23 14:49:30
7759    hi arenot contagious disease need worry panic black fungus white fungus thank   Dr. Aadhavan Ramanathan 9005    275             NULL    2       0       2021-05-23 14:46:26
7760    cancel board exam time face covid face black fungus white fungus also cancellation option board exam request prime minister pmo please come front medium live take decision favor       RISHABH PANDEY  2       1
NULL    1       1       2021-05-23 14:46:02
7761    university please promote student pandemic sir black white fungus become epidemic guy ready take exam take exam online objective paper promoteuniversitystudents       Devil   59      3               NULL    3       5       2021-05-
23 14:45:29
7762    willnot good minister india student health get affected due holding exam corona white fungus black fungus also epidemic country Hrithik Gupta   8       1
motihari        NULL    0       0       2021-05-23 14:45:01
7763    deeper insight covid black fungus white fungusplease tune covidindia covid india covidvaccine   Nithya Sakalananda      1882    83              NULL    0
        0               2021-05-23 14:44:19
```



```
vaccine corona black white fungus       Ronak Jha       82      0               NULL    1       0   2021-05-23 15:00:09
7750    new post white fungus knowledge fear businessnews businessnewsthisweek  Neel Achary     95842   1580    "Bhubaneswar    NULL    NULL    NULL    0
7751    first icmr excluded plasma remdesivir strict steroid ppl steroid getting black amp white fungus covid patient given fake remdesivir survived wht exactly issue t
ht supremacy amp god status h bn maligned random baba   Neta Ji @AapGhumaKeLeLo_ BACKUP 1166    4515            NULL    77      50      2021-05-23 14:58:21
7753    please cbse exam cancel lot dangerous virus black fungus white fungus plus covid request please understand year similar previous year please cancelboardexam mod
ijinoofflineexams modijicancel thboards Tushar_011      1279    5       "New Delhi      NULL    NULL    1       1
7754    black fungus white fungus severe red fungus upcoming blackfungus whitefungus redfungus  Vedansh Tiwari  51      14              NULL    0       0       2021-05-
23 14:54:46
7756    chadha dallalsalleged farmer corona aap black fungus raghav chaddawhite fungus  Vasudevan.L 46536       684             NULL    1       0       2021-05-23 14:51
:24
7757    trying deal situation uncertain exam black fungus white fungus declared epidemic sir conduct offlineexam student either die corona fungus delay decision like ma
ny might end    #cancelcbseboards2021   67      5               NULL    0       1       2021-05-23 14:49:46
7758    zincfungal infection theory covid related black amp white fungus case theory explains zinc metabolism connection fungal virulence blackfungus whitefungus covid
Medical Ji      10      13              NULL    10      3       2021-05-23 14:49:30
7759    hi arenot contagious disease need worry panic black fungus white fungus thank   Dr. Aadhavan Ramanathan 9005    275             NULL    2       0       2021-05-
23 14:46:26
7760    cancel board exam time face covid face black fungus white fungus also cancellation option board exam request prime minister pmo please come front medium live ta
ke decision favor       RISHABH PANDEY  2       1               NULL    1       1       2021-05-23 14:46:02
7761    university please promote student pandemic sir black white fungus become epidemic guy ready take exam take exam online objective paper promoteuniversitystudents
        Devil   59      3               NULL    3       5       2021-05-23 14:45:29
7762    willnot good minister india student health get affected due holding exam corona white fungus black fungus also epidemic country Hrithik Gupta   8       1
motihari        NULL    0       0       2021-05-23 14:45:01
7763    deeper insight covid black fungus white fungusplease tune covidindia covid india covidvaccine   Nithya Sakalananda      1882    83              NULL    0
0       2021-05-23 14:44:19
7764    good evening sir please take initiative purchase distribution amphotericin b anti fungal drug combat white black fungus infection thank Abdul hasan     242
3       "Madurai        NULL    NULL    1       0
7765    corona abhi bi nehin gaya hai upar se black fungus hai abhi abhi firse white fungus jo aur bhi ktaranak hai kese hum board exam denge sir please sir cancel boar
d exam humare jaan rahenge agar tabhi humara future rahegamodijicancel thboards Padmalochan Parida  13  1       "Orissa NULL    NULL    1       0
7766    rohit iam hearing rd wave coming november black aur white fungus bhi aa gaya    Yash Kumar  4305        213     "Mumbai NULL    NULL    0       0
7768    india india sir day please humble request take proper decision keeping mind student recovered covid could able get cause black white fungussome student still we
akness could able get covid please      ananya gupta    90      4               NULL    0       0   2021-05-23 14:34:15
7769    term white fungus reminds white nose syndrome bat caused pseudogymnoascus destructans maybe article human mycoses india referring mold without pigmented hypha h
yaline aspergillus fusarium specie idtwitter    "Tobias Hohl MD NULL    1436    1599    NULL    NULL    3       2
7770    tnie bothered black fungus bothered white fungus dying called farmer die        Ramesh Kumar SELF RELIANT INDIA≡ƒç«≡ƒç│ 50575   643     Bengaluru Naya Bharat
NULL    0       0       2021-05-23 14:32:39
7771    every kind fungus must eradicated whether black white green dangerous humanity agree    Vijay Kapil     1748    3602            NULL    1       0       2021-05-
23 14:31:05
7772    doctor black fungus white fungus faujiholic bhaiya olive green mein dikhana     faujiholic  73  52      India   NULL    1       0       2021-05-23 14:30:34
7773    conducting exam situation possible new strain virus side effect black fungus white fungus adverse effect putn life stake sake exam doesnot feel worth itcancelal
lexams cancelboardexams Modhak Natesh   1       1               NULL    0       0       2021-05-23 14:28:15
7774    chattisgarh government conducting handwriting competition cancelexamssavestudents corona black fungus white fungus sirplz cancel board exam otherwise take exam
like chattisgarh otherwise ispartiality bw student      Debasish Kumar  44      11              NULL    1       0       2021-05-23 14:27:00
7775    itis logical conduct examination moment countryman fighting survivalfrom covid black white fungus tootherefore ensure safety student family cancelling strictly
cancelallboardexams cancelexamssavestudents     Ansh Saxena     312     2               NULL    2   0   2021-05-23 14:26:00
```

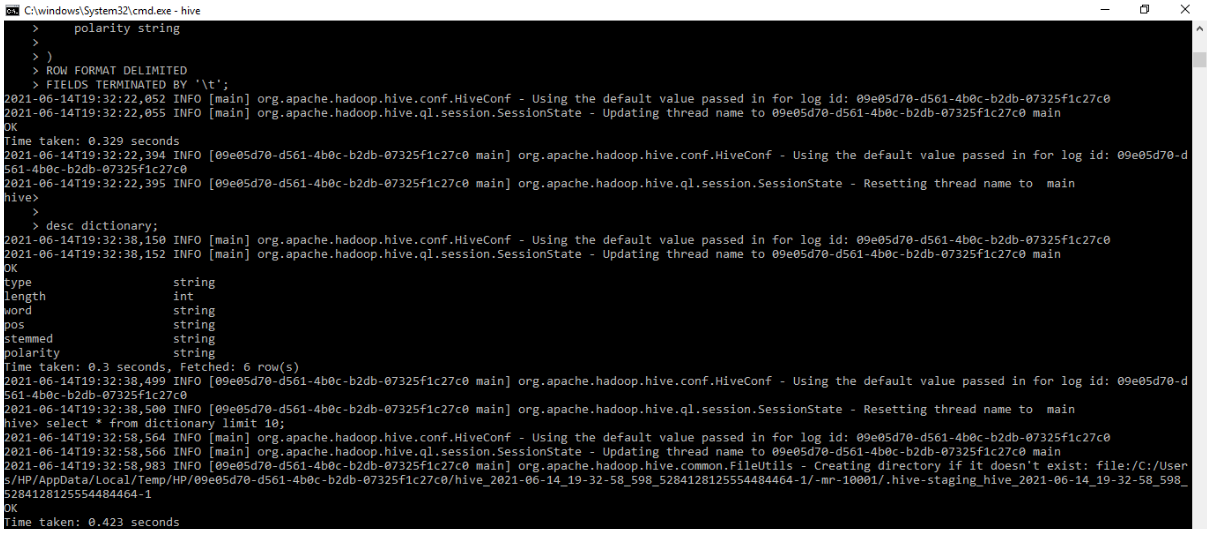
Figure 15: Displaying the table

4. Create table dictionary

```
CREATE EXTERNAL TABLE dictionary
  (
     type string,
     length int,
     word string,
     pos string,
     stemmed string,
     polarity string
   )
  ROW FORMAT DELIMITED
  FIELDS TERMINATED BY '\t';
```

5. DESC table

```
desc dictionary;
OK
type                string
length               int
word                 string
pos                 string
stemmed                string
polarity              string
Time taken: 0.315 seconds, Fetched: 6 row(s)
```

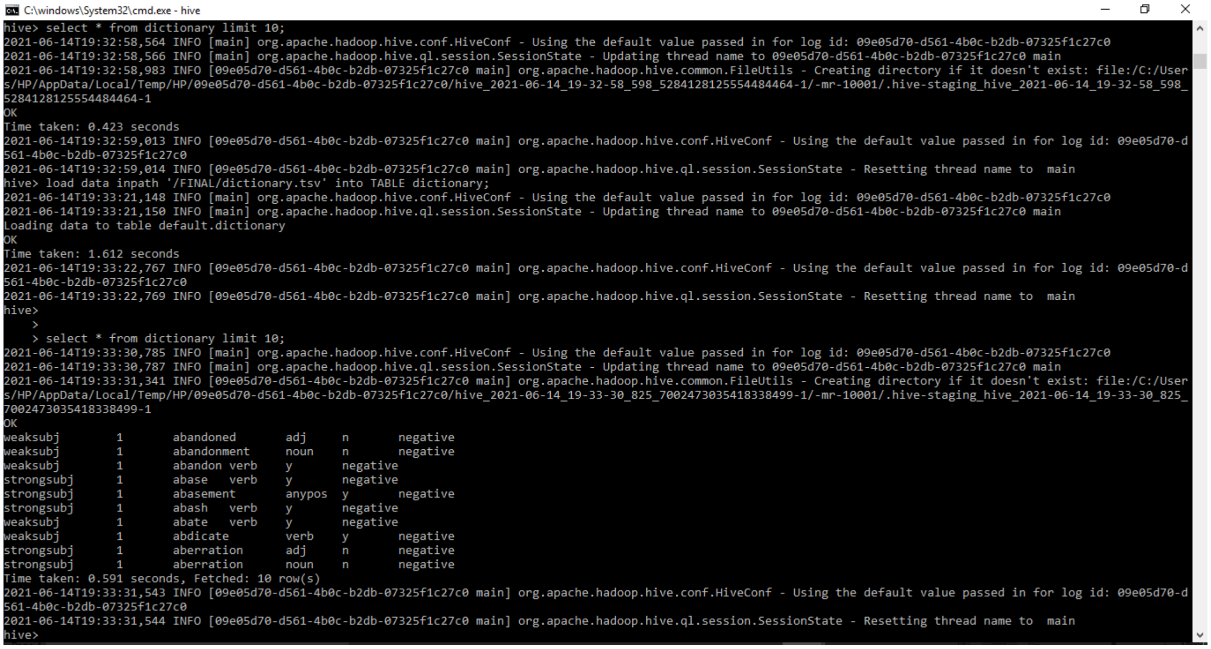


```
    >     polarity string
    >
    > )
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY '\t';
2021-06-14T19:32:22,052 INFO [main] org.apache.hadoop.hive.conf.HiveConf - Using the default value passed in for log id: 09e05d70-d561-4b0c-b2db-07325f1c27c0
2021-06-14T19:32:22,055 INFO [main] org.apache.hadoop.hive.ql.session.SessionState - Updating thread name to 09e05d70-d561-4b0c-b2db-07325f1c27c0 main
OK
Time taken: 0.329 seconds
2021-06-14T19:32:22,394 INFO [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.hive.conf.HiveConf - Using the default value passed in for log id: 09e05d70-d
561-4b0c-b2db-07325f1c27c0
2021-06-14T19:32:22,395 INFO [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.hive.ql.session.SessionState - Resetting thread name to  main
hive>
    >
    > desc dictionary;
2021-06-14T19:32:38,150 INFO [main] org.apache.hadoop.hive.conf.HiveConf - Using the default value passed in for log id: 09e05d70-d561-4b0c-b2db-07325f1c27c0
2021-06-14T19:32:38,152 INFO [main] org.apache.hadoop.hive.ql.session.SessionState - Updating thread name to 09e05d70-d561-4b0c-b2db-07325f1c27c0 main
OK
type                    string
length                  int
word                    string
pos                     string
stemmed                 string
polarity                string
Time taken: 0.3 seconds, Fetched: 6 row(s)
2021-06-14T19:32:38,499 INFO [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.hive.conf.HiveConf - Using the default value passed in for log id: 09e05d70-d
561-4b0c-b2db-07325f1c27c0
2021-06-14T19:32:38,500 INFO [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.hive.ql.session.SessionState - Resetting thread name to  main
hive> select * from dictionary limit 10;
2021-06-14T19:32:58,564 INFO [main] org.apache.hadoop.hive.conf.HiveConf - Using the default value passed in for log id: 09e05d70-d561-4b0c-b2db-07325f1c27c0
2021-06-14T19:32:58,566 INFO [main] org.apache.hadoop.hive.ql.session.SessionState - Updating thread name to 09e05d70-d561-4b0c-b2db-07325f1c27c0 main
2021-06-14T19:32:58,983 INFO [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.hive.common.FileUtils - Creating directory if it doesn't exist: file:/C:/User
s/HP/AppData/Local/Temp/HP/09e05d70-d561-4b0c-b2db-07325f1c27c0/hive_2021-06-14_19-32-58_598_5284128125554484464-1/-mr-10001/.hive-staging_hive_2021-06-14_19-32-58_598_
5284128125554484464-1
OK
Time taken: 0.423 seconds
```

Figure 16: Description of table dictionary

6. Display the table dictionary

Select * from dictionary limit 10



```
hive> select * from dictionary limit 10;
2021-06-14T19:32:58,564 INFO [main] org.apache.hadoop.hive.conf.HiveConf - Using the default value passed in for log id: 09e05d70-d561-4b0c-b2db-07325f1c27c0
2021-06-14T19:32:58,566 INFO [main] org.apache.hadoop.hive.ql.session.SessionState - Updating thread name to 09e05d70-d561-4b0c-b2db-07325f1c27c0 main
2021-06-14T19:32:58,983 INFO [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.hive.common.FileUtils - Creating directory if it doesn't exist: file:/C:/User
s/HP/AppData/Local/Temp/HP/09e05d70-d561-4b0c-b2db-07325f1c27c0/hive_2021-06-14_19-32-58_598_5284128125554484464-1/-mr-10001/.hive-staging_hive_2021-06-14_19-32-58_598_
5284128125554484464-1
OK
Time taken: 0.423 seconds
2021-06-14T19:32:59,013 INFO [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.hive.conf.HiveConf - Using the default value passed in for log id: 09e05d70-d
561-4b0c-b2db-07325f1c27c0
2021-06-14T19:32:59,014 INFO [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.hive.ql.session.SessionState - Resetting thread name to  main
hive> load data inpath '/FINAL/dictionary.tsv' into TABLE dictionary;
2021-06-14T19:33:21,148 INFO [main] org.apache.hadoop.hive.conf.HiveConf - Using the default value passed in for log id: 09e05d70-d561-4b0c-b2db-07325f1c27c0
2021-06-14T19:33:21,150 INFO [main] org.apache.hadoop.hive.ql.session.SessionState - Updating thread name to 09e05d70-d561-4b0c-b2db-07325f1c27c0 main
Loading data to table default.dictionary
OK
Time taken: 1.612 seconds
2021-06-14T19:33:22,767 INFO [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.hive.conf.HiveConf - Using the default value passed in for log id: 09e05d70-d
561-4b0c-b2db-07325f1c27c0
2021-06-14T19:33:22,769 INFO [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.hive.ql.session.SessionState - Resetting thread name to  main
hive>
    >
    > select * from dictionary limit 10;
2021-06-14T19:33:30,785 INFO [main] org.apache.hadoop.hive.conf.HiveConf - Using the default value passed in for log id: 09e05d70-d561-4b0c-b2db-07325f1c27c0
2021-06-14T19:33:30,787 INFO [main] org.apache.hadoop.hive.ql.session.SessionState - Updating thread name to 09e05d70-d561-4b0c-b2db-07325f1c27c0 main
2021-06-14T19:33:31,341 INFO [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.hive.common.FileUtils - Creating directory if it doesn't exist: file:/C:/User
s/HP/AppData/Local/Temp/HP/09e05d70-d561-4b0c-b2db-07325f1c27c0/hive_2021-06-14_19-33-30_825_7002473035418338499-1/-mr-10001/.hive-staging_hive_2021-06-14_19-33-30_825_
7002473035418338499-1
OK
weaksubj        1       abandoned       adj     n       negative
weaksubj        1       abandonment     noun    n       negative
weaksubj        1       abandon verb    y       negative
strongsubj      1       abase   verb    y       negative
strongsubj      1       abasement       anypos  y       negative
strongsubj      1       abash   verb    y       negative
weaksubj        1       abate   verb    y       negative
weaksubj        1       abdicate        verb    y       negative
strongsubj      1       aberration      adj     n       negative
strongsubj      1       aberration      noun    n       negative
Time taken: 0.591 seconds, Fetched: 10 row(s)
2021-06-14T19:33:31,543 INFO [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.hive.conf.HiveConf - Using the default value passed in for log id: 09e05d70-d
561-4b0c-b2db-07325f1c27c0
2021-06-14T19:33:31,544 INFO [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.hive.ql.session.SessionState - Resetting thread name to  main
hive>
```

Figure 17: Viewing the dictionary table

7. Create view temp 1

create view temp_1 as select

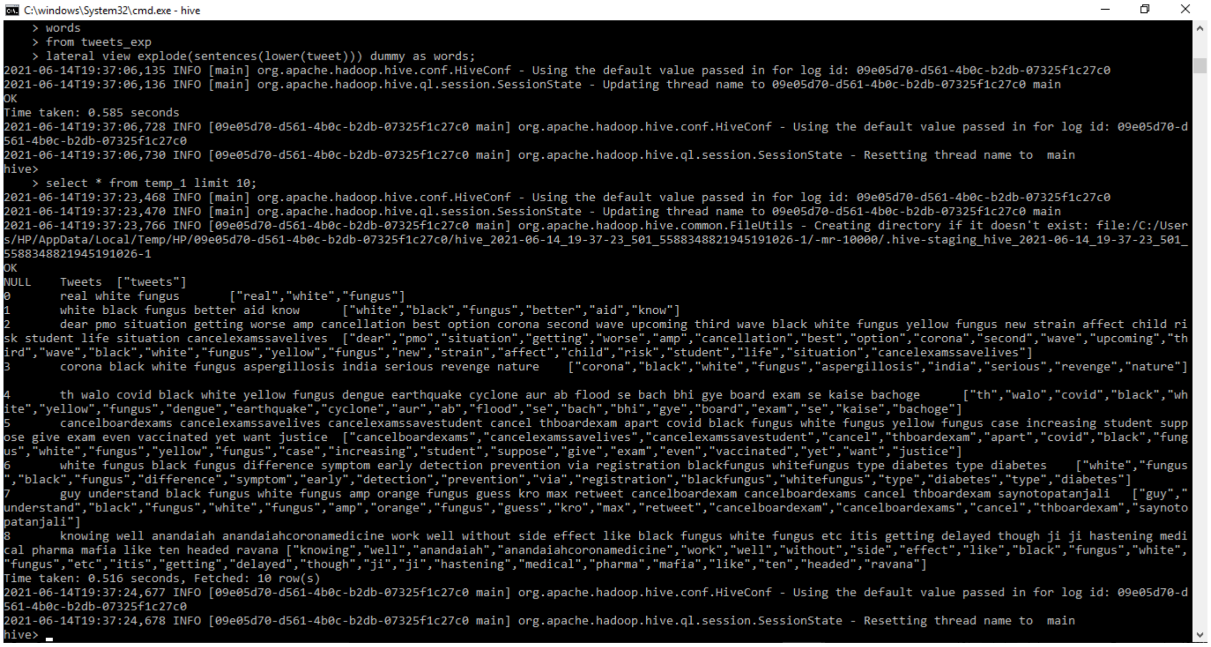
id,

tweets_exp.tweet,

words

from tweets_exp

lateral view explode(sentences(lower(tweet))) dummy as words;



```
    > words
    > from tweets_exp
    > lateral view explode(sentences(lower(tweet))) dummy as words;
2021-06-14T19:37:06,135 INFO [main] org.apache.hadoop.hive.conf.HiveConf - Using the default value passed in for log id: 09e05d70-d561-4b0c-b2db-07325f1c27c0
2021-06-14T19:37:06,136 INFO [main] org.apache.hadoop.hive.ql.session.SessionState - Updating thread name to 09e05d70-d561-4b0c-b2db-07325f1c27c0 main
OK
Time taken: 0.585 seconds
2021-06-14T19:37:06,728 INFO [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.hive.conf.HiveConf - Using the default value passed in for log id: 09e05d70-d
561-4b0c-b2db-07325f1c27c0
2021-06-14T19:37:06,730 INFO [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.hive.ql.session.SessionState - Resetting thread name to  main
hive>
    > select * from temp_1 limit 10;
2021-06-14T19:37:23,468 INFO [main] org.apache.hadoop.hive.conf.HiveConf - Using the default value passed in for log id: 09e05d70-d561-4b0c-b2db-07325f1c27c0
2021-06-14T19:37:23,470 INFO [main] org.apache.hadoop.hive.ql.session.SessionState - Updating thread name to 09e05d70-d561-4b0c-b2db-07325f1c27c0 main
2021-06-14T19:37:23,766 INFO [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.hive.common.FileUtils - Creating directory if it doesn't exist: file:/C:/User
s/HP/AppData/Local/Temp/HP/09e05d70-d561-4b0c-b2db-07325f1c27c0/hive_2021-06-14_19-37-23_501_5588348821945191026-1/-mr-10000/.hive-staging_hive_2021-06-14_19-37-23_501_
5588348821945191026-1
OK
NULL    Tweets  ["tweets"]
0       real white fungus       ["real","white","fungus"]
1       white black fungus better aid know      ["white","black","fungus","better","aid","know"]
2       dear pmo situation getting worse amp cancellation best option corona second wave upcoming third wave black white fungus yellow fungus new strain affect child ri
sk student life situation cancelexamssavelives  ["dear","pmo","situation","getting","worse","amp","cancellation","best","option","corona","second","wave","upcoming","th
ird","wave","black","white","fungus","yellow","fungus","new","strain","affect","child","risk","student","life","situation","cancelexamssavelives"]
3       corona black white fungus aspergillosis india serious revenge nature    ["corona","black","white","fungus","aspergillosis","india","serious","revenge","nature"]

4       th walo covid black white yellow fungus dengue earthquake cyclone aur ab flood se bach bhi gye board exam se kaise bachoge      ["th","walo","covid","black","wh
ite","yellow","fungus","dengue","earthquake","cyclone","aur","ab","flood","se","bach","bhi","gye","board","exam","se","kaise","bachoge"]
5       cancelboardexams cancelexamssavelives cancelexamssavestudent cancel thboardexam apart covid black fungus white fungus yellow fungus case increasing student supp
ose give exam even vaccinated yet want justice  ["cancelboardexams","cancelexamssavelives","cancelexamssavestudent","cancel","thboardexam","apart","covid","black","fung
us","white","fungus","yellow","fungus","case","increasing","student","suppose","give","exam","even","vaccinated","yet","want","justice"]
6       white fungus black fungus difference symptom early detection prevention via registration blackfungus whitefungus type diabetes type diabetes    ["white","fungus
","black","fungus","difference","symptom","early","detection","prevention","via","registration","blackfungus","whitefungus","type","diabetes","type","diabetes"]
7       guy understand black fungus white fungus amp orange fungus guess kro max retweet cancelboardexam cancelboardexams cancel thboardexam saynotopatanjali   ["guy","
understand","black","fungus","white","fungus","amp","orange","fungus","guess","kro","max","retweet","cancelboardexam","cancelboardexams","cancel","thboardexam","saynoto
patanjali"]
8       knowing well anandaiah anandaiahcoronamedicine work well without side effect like black fungus white fungus etc itis getting delayed though ji ji hastening medi
cal pharma mafia like ten headed ravana ["knowing","well","anandaiah","anandaiahcoronamedicine","work","well","without","side","effect","like","black","fungus","white",
"fungus","etc","itis","getting","delayed","though","ji","ji","hastening","medical","pharma","mafia","like","ten","headed","ravana"]
Time taken: 0.516 seconds, Fetched: 10 row(s)
2021-06-14T19:37:24,677 INFO [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.hive.conf.HiveConf - Using the default value passed in for log id: 09e05d70-d
561-4b0c-b2db-07325f1c27c0
2021-06-14T19:37:24,678 INFO [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.hive.ql.session.SessionState - Resetting thread name to  main
hive> _
```

Figure 18: View temp 1

8. Create view temp 2

create view temp_2 as select

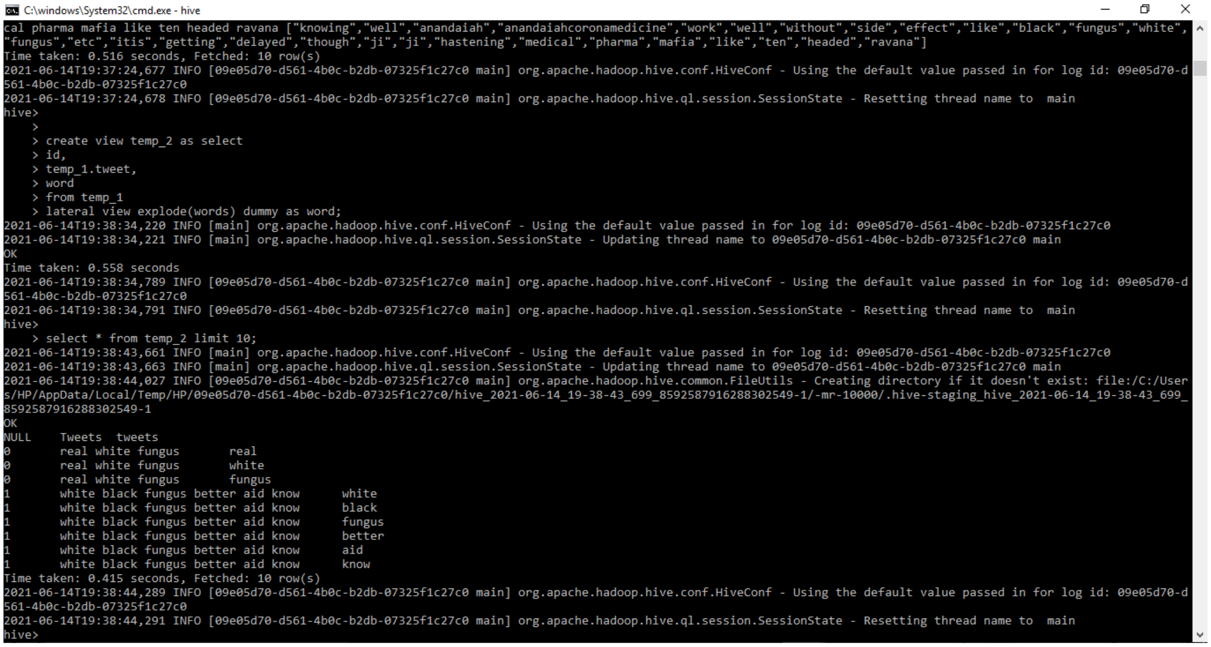
id,

temp_1.tweet,

word

from temp_1

lateral view explode(words) dummy as word;



```
cal pharma mafia like ten headed ravana ["knowing","well","anandaiah","anandaiahcoronamedicine","work","well","without","side","effect","like","black","fungus","white",
"fungus","etc","itis","getting","delayed","though","ji","ji","hastening","medical","pharma","mafia","like","ten","headed","ravana"]
Time taken: 0.516 seconds, Fetched: 10 row(s)
2021-06-14T19:37:24,677 INFO [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.hive.conf.HiveConf - Using the default value passed in for log id: 09e05d70-d
561-4b0c-b2db-07325f1c27c0
2021-06-14T19:37:24,678 INFO [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.hive.ql.session.SessionState - Resetting thread name to  main
hive>
    >
    > create view temp_2 as select
    > id,
    > temp_1.tweet,
    > word
    > from temp_1
    > lateral view explode(words) dummy as word;
2021-06-14T19:38:34,220 INFO [main] org.apache.hadoop.hive.conf.HiveConf - Using the default value passed in for log id: 09e05d70-d561-4b0c-b2db-07325f1c27c0
2021-06-14T19:38:34,221 INFO [main] org.apache.hadoop.hive.ql.session.SessionState - Updating thread name to 09e05d70-d561-4b0c-b2db-07325f1c27c0 main
OK
Time taken: 0.558 seconds
2021-06-14T19:38:34,789 INFO [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.hive.conf.HiveConf - Using the default value passed in for log id: 09e05d70-d
561-4b0c-b2db-07325f1c27c0
2021-06-14T19:38:34,791 INFO [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.hive.ql.session.SessionState - Resetting thread name to  main
hive>
    > select * from temp_2 limit 10;
2021-06-14T19:38:43,661 INFO [main] org.apache.hadoop.hive.conf.HiveConf - Using the default value passed in for log id: 09e05d70-d561-4b0c-b2db-07325f1c27c0
2021-06-14T19:38:43,663 INFO [main] org.apache.hadoop.hive.ql.session.SessionState - Updating thread name to 09e05d70-d561-4b0c-b2db-07325f1c27c0 main
2021-06-14T19:38:44,027 INFO [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.hive.common.FileUtils - Creating directory if it doesn't exist: file:/C:/User
s/HP/AppData/Local/Temp/HP/09e05d70-d561-4b0c-b2db-07325f1c27c0/hive_2021-06-14_19-38-43_699_8592587916288302549-1/-mr-10000/.hive-staging_hive_2021-06-14_19-38-43_699_
8592587916288302549-1
OK
NULL    Tweets  tweets
0       real white fungus       real
0       real white fungus       white
0       real white fungus       fungus
1       white black fungus better aid know      white
1       white black fungus better aid know      black
1       white black fungus better aid know      fungus
1       white black fungus better aid know      better
1       white black fungus better aid know      aid
1       white black fungus better aid know      know
Time taken: 0.415 seconds, Fetched: 10 row(s)
2021-06-14T19:38:44,289 INFO [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.hive.conf.HiveConf - Using the default value passed in for log id: 09e05d70-d
561-4b0c-b2db-07325f1c27c0
2021-06-14T19:38:44,291 INFO [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.hive.ql.session.SessionState - Resetting thread name to  main
hive>
```

Figure 19: View temp 2

9. Create view temp 3

create view temp_3 as select

id,

temp_2.tweet,

temp_2.word,

case s_d.polarity

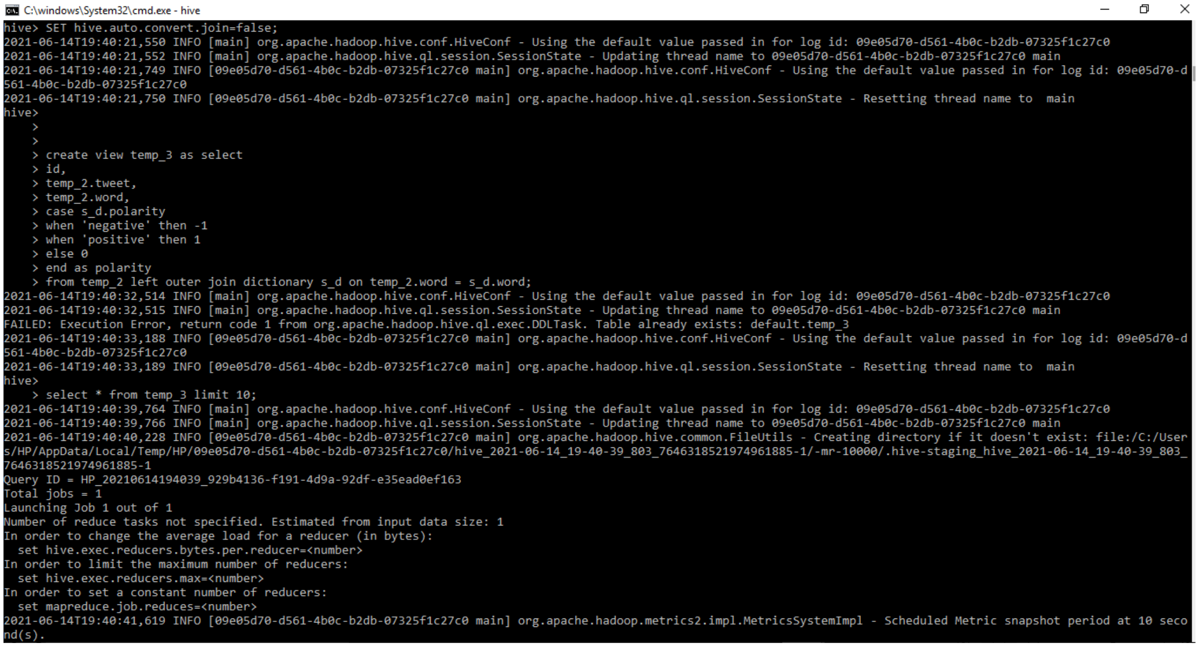
when 'negative' then -1

when 'positive' then 1

else 0

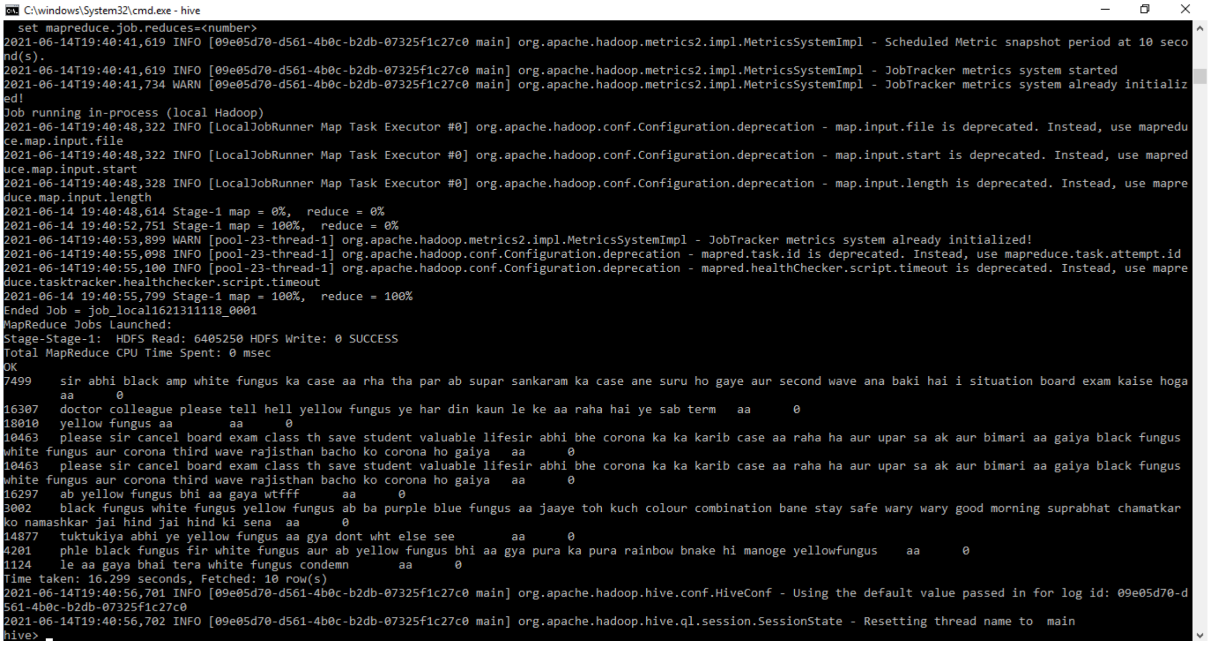
end as polarity

from temp_2 left outer join dictionary s_d on temp_2.word = s_d.word;



```
hive> SET hive.auto.convert.join=false;
2021-06-14T19:40:21,550 INFO [main] org.apache.hadoop.hive.conf.HiveConf - Using the default value passed in for log id: 09e05d70-d561-4b0c-b2db-07325f1c27c0
2021-06-14T19:40:21,552 INFO [main] org.apache.hadoop.hive.ql.session.SessionState - Updating thread name to 09e05d70-d561-4b0c-b2db-07325f1c27c0 main
2021-06-14T19:40:21,749 INFO [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.hive.conf.HiveConf - Using the default value passed in for log id: 09e05d70-d
561-4b0c-b2db-07325f1c27c0
2021-06-14T19:40:21,750 INFO [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.hive.ql.session.SessionState - Resetting thread name to  main
hive>
    >
    >
    > create view temp_3 as select
    > id,
    > temp_2.tweet,
    > temp_2.word,
    > case s_d.polarity
    > when 'negative' then -1
    > when 'positive' then 1
    > else 0
    > end as polarity
    > from temp_2 left outer join dictionary s_d on temp_2.word = s_d.word;
2021-06-14T19:40:32,514 INFO [main] org.apache.hadoop.hive.conf.HiveConf - Using the default value passed in for log id: 09e05d70-d561-4b0c-b2db-07325f1c27c0
2021-06-14T19:40:32,515 INFO [main] org.apache.hadoop.hive.ql.session.SessionState - Updating thread name to 09e05d70-d561-4b0c-b2db-07325f1c27c0 main
FAILED: Execution Error, return code 1 from org.apache.hadoop.hive.ql.exec.DDLTask. Table already exists: default.temp_3
2021-06-14T19:40:33,188 INFO [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.hive.conf.HiveConf - Using the default value passed in for log id: 09e05d70-d
561-4b0c-b2db-07325f1c27c0
2021-06-14T19:40:33,189 INFO [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.hive.ql.session.SessionState - Resetting thread name to  main
hive>
    > select * from temp_3 limit 10;
2021-06-14T19:40:39,764 INFO [main] org.apache.hadoop.hive.conf.HiveConf - Using the default value passed in for log id: 09e05d70-d561-4b0c-b2db-07325f1c27c0
2021-06-14T19:40:39,766 INFO [main] org.apache.hadoop.hive.ql.session.SessionState - Updating thread name to 09e05d70-d561-4b0c-b2db-07325f1c27c0 main
2021-06-14T19:40:40,228 INFO [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.hive.common.FileUtils - Creating directory if it doesn't exist: file:/C:/User
s/HP/AppData/Local/Temp/HP/09e05d70-d561-4b0c-b2db-07325f1c27c0/hive_2021-06-14_19-40-39_803_7646318521974961885-1/-mr-10000/.hive-staging_hive_2021-06-14_19-40-39_803_
7646318521974961885-1
Query ID = HP_20210614194039_929b4136-f191-4d9a-92df-e35ead0ef163
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 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>
2021-06-14T19:40:41,619 INFO [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.metrics2.impl.MetricsSystemImpl - Scheduled Metric snapshot period at 10 seco
nd(s).
```

Figure 20: View temp 3-1



```
  set mapreduce.job.reduces=<number>
2021-06-14T19:40:41,619 INFO [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.metrics2.impl.MetricsSystemImpl - Scheduled Metric snapshot period at 10 seco
nd(s).
2021-06-14T19:40:41,619 INFO [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system started
2021-06-14T19:40:41,734 WARN [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initializ
ed!
Job running in-process (local Hadoop)
2021-06-14T19:40:48,322 INFO [LocalJobRunner Map Task Executor #0] org.apache.hadoop.conf.Configuration.deprecation - map.input.file is deprecated. Instead, use mapredu
ce.map.input.file
2021-06-14T19:40:48,322 INFO [LocalJobRunner Map Task Executor #0] org.apache.hadoop.conf.Configuration.deprecation - map.input.start is deprecated. Instead, use mapred
uce.map.input.start
2021-06-14T19:40:48,328 INFO [LocalJobRunner Map Task Executor #0] org.apache.hadoop.conf.Configuration.deprecation - map.input.length is deprecated. Instead, use mapre
duce.map.input.length
2021-06-14 19:40:48,614 Stage-1 map = 0%,  reduce = 0%
2021-06-14 19:40:52,751 Stage-1 map = 100%,  reduce = 0%
2021-06-14T19:40:53,899 WARN [pool-23-thread-1] org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-06-14T19:40:55,098 INFO [pool-23-thread-1] org.apache.hadoop.conf.Configuration.deprecation - mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
2021-06-14T19:40:55,100 INFO [pool-23-thread-1] org.apache.hadoop.conf.Configuration.deprecation - mapred.healthChecker.script.timeout is deprecated. Instead, use mapre
duce.tasktracker.healthchecker.script.timeout
2021-06-14 19:40:55,799 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local1621311118_0001
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 6405250 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
7499    sir abhi black amp white fungus ka case aa rha tha par ab supar sankaram ka case ane suru ho gaye aur second wave ana baki hai i situation board exam kaise hoga
        aa      0
16307   doctor colleague please tell hell yellow fungus ye har din kaun le ke aa raha hai ye sab term   aa      0
18010   yellow fungus aa        aa      0
10463   please sir cancel board exam class th save student valuable lifesir abhi bhe corona ka ka karib case aa raha ha aur upar sa ak aur bimari aa gaiya black fungus
white fungus aur corona third wave rajisthan bacho ko corona ho gaiya   aa      0
10463   please sir cancel board exam class th save student valuable lifesir abhi bhe corona ka ka karib case aa raha ha aur upar sa ak aur bimari aa gaiya black fungus
white fungus aur corona third wave rajisthan bacho ko corona ho gaiya   aa      0
16297   ab yellow fungus bhi aa gaya wtfff      aa      0
3002    black fungus white fungus yellow fungus ab ba purple blue fungus aa jaaye toh kuch colour combination bane stay safe wary wary good morning suprabhat chamatkar
ko namashkar jai hind jai hind ki sena  aa      0
14877   tuktukiya abhi ye yellow fungus aa gya dont wht else see        aa      0
4201    phle black fungus fir white fungus aur ab yellow fungus bhi aa gya pura ka pura rainbow bnake hi manoge yellowfungus    aa      0
1124    le aa gaya bhai tera white fungus condemn       aa      0
Time taken: 16.299 seconds, Fetched: 10 row(s)
2021-06-14T19:40:56,701 INFO [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.hive.conf.HiveConf - Using the default value passed in for log id: 09e05d70-d
561-4b0c-b2db-07325f1c27c0
2021-06-14T19:40:56,702 INFO [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.hive.ql.session.SessionState - Resetting thread name to  main
hive> _
```

Figure 21: View temp 3 -2

10. Create table sentiment

create table tweets_sentiment as select

id,

case

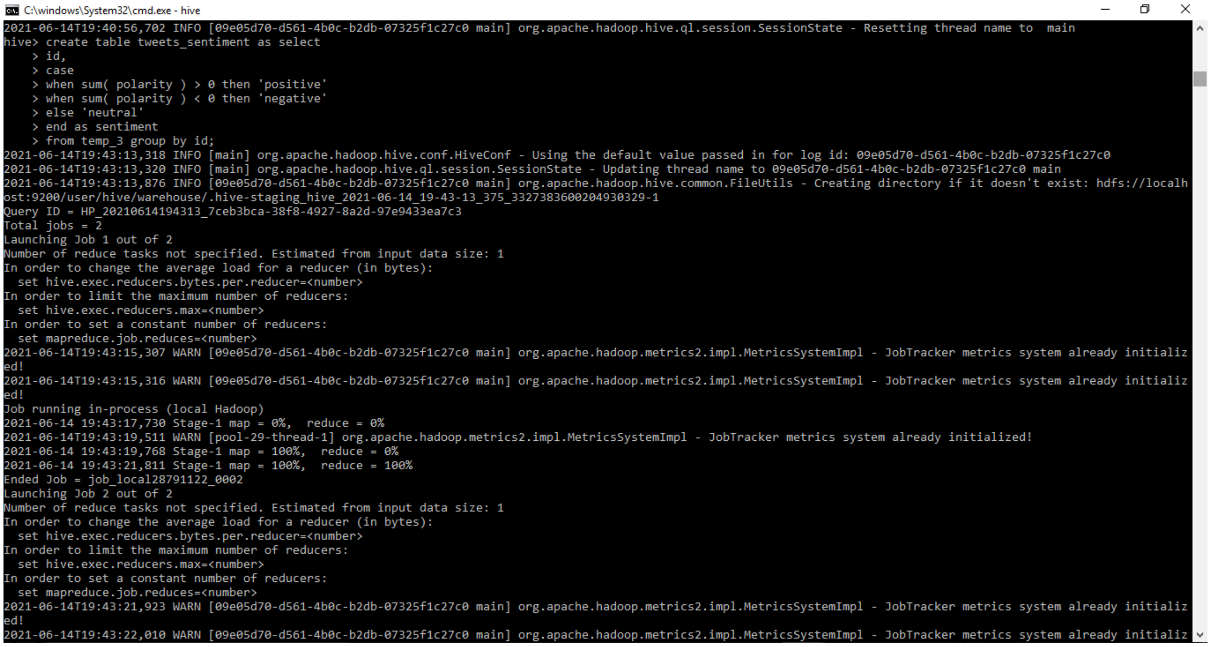
when sum(polarity) > 0 then 'positive'

when sum(polarity) < 0 then 'negative'

else 'neutral'

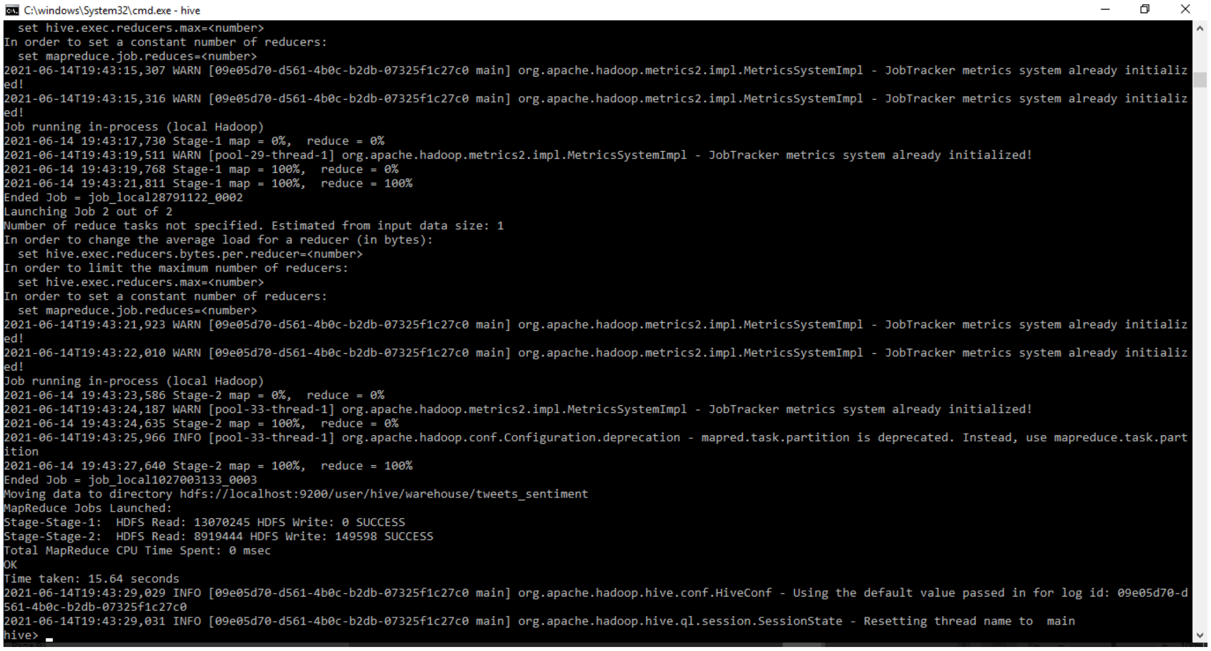
end as sentiment

from temp_3 group by id;



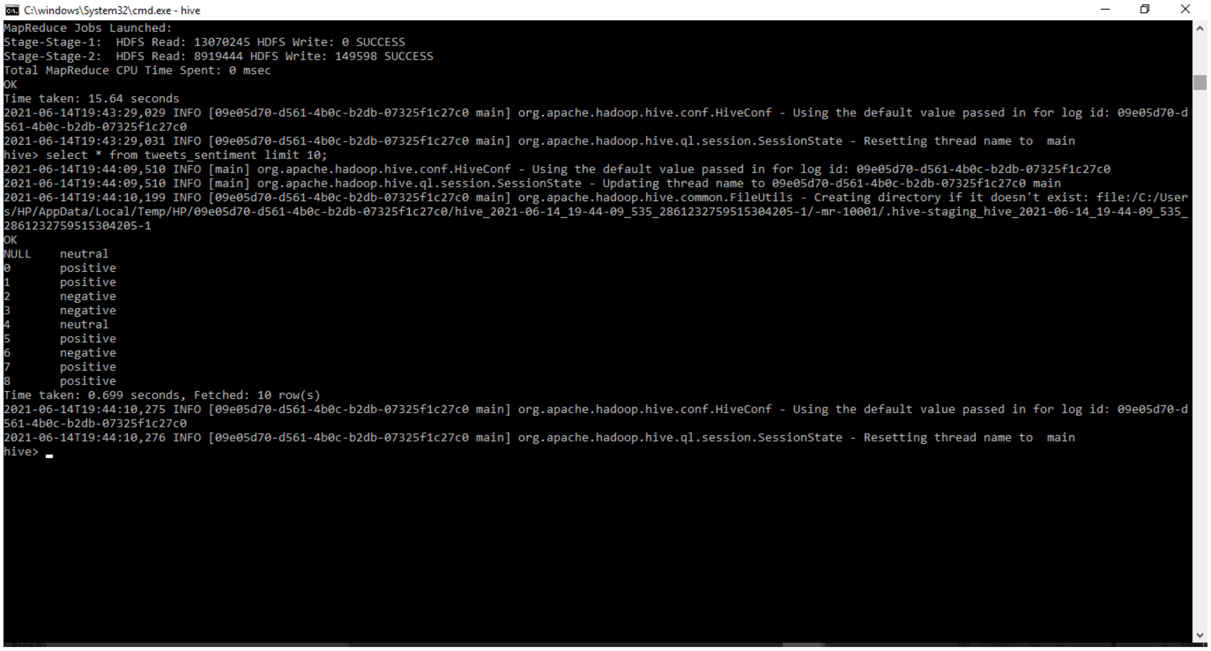
```
2021-06-14T19:40:56,702 INFO [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.hive.ql.session.SessionState - Resetting thread name to  main
hive> create table tweets_sentiment as select
    > id,
    > case
    > when sum( polarity ) > 0 then 'positive'
    > when sum( polarity ) < 0 then 'negative'
    > else 'neutral'
    > end as sentiment
    > from temp_3 group by id;
2021-06-14T19:43:13,318 INFO [main] org.apache.hadoop.hive.conf.HiveConf - Using the default value passed in for log id: 09e05d70-d561-4b0c-b2db-07325f1c27c0
2021-06-14T19:43:13,320 INFO [main] org.apache.hadoop.hive.ql.session.SessionState - Updating thread name to 09e05d70-d561-4b0c-b2db-07325f1c27c0 main
2021-06-14T19:43:13,876 INFO [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.hive.common.FileUtils - Creating directory if it doesn't exist: hdfs://localh
ost:9200/user/hive/warehouse/.hive-staging_hive_2021-06-14_19-43-13_375_3327383600204930329-1
Query ID = HP_20210614194313_7ceb3bca-38f8-4927-8a2d-97e9433ea7c3
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 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>
2021-06-14T19:43:15,307 WARN [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initializ
ed!
2021-06-14T19:43:15,316 WARN [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initializ
ed!
Job running in-process (local Hadoop)
2021-06-14 19:43:17,730 Stage-1 map = 0%,  reduce = 0%
2021-06-14T19:43:19,511 WARN [pool-29-thread-1] org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-06-14 19:43:19,768 Stage-1 map = 100%,  reduce = 0%
2021-06-14 19:43:21,811 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local28791122_0002
Launching Job 2 out of 2
Number of reduce tasks not specified. Estimated from input data size: 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>
2021-06-14T19:43:21,923 WARN [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initializ
ed!
2021-06-14T19:43:22,010 WARN [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initializ
```

Figure 22: Creating sentiments table -1



```
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
2021-06-14T19:43:15,307 WARN [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initializ
ed!
2021-06-14T19:43:15,316 WARN [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initializ
ed!
Job running in-process (local Hadoop)
2021-06-14 19:43:17,730 Stage-1 map = 0%,  reduce = 0%
2021-06-14T19:43:19,511 WARN [pool-29-thread-1] org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-06-14 19:43:19,768 Stage-1 map = 100%,  reduce = 0%
2021-06-14 19:43:21,811 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local28791122_0002
Launching Job 2 out of 2
Number of reduce tasks not specified. Estimated from input data size: 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>
2021-06-14T19:43:21,923 WARN [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initializ
ed!
2021-06-14T19:43:22,010 WARN [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initializ
ed!
Job running in-process (local Hadoop)
2021-06-14 19:43:23,586 Stage-2 map = 0%,  reduce = 0%
2021-06-14T19:43:24,187 WARN [pool-33-thread-1] org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-06-14 19:43:24,635 Stage-2 map = 100%,  reduce = 0%
2021-06-14T19:43:25,966 INFO [pool-33-thread-1] org.apache.hadoop.conf.Configuration.deprecation - mapred.task.partition is deprecated. Instead, use mapreduce.task.part
ition
2021-06-14 19:43:27,640 Stage-2 map = 100%,  reduce = 100%
Ended Job = job_local1027003133_0003
Moving data to directory hdfs://localhost:9200/user/hive/warehouse/tweets_sentiment
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 13070245 HDFS Write: 0 SUCCESS
Stage-Stage-2:  HDFS Read: 8919444 HDFS Write: 149598 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
Time taken: 15.64 seconds
2021-06-14T19:43:29,029 INFO [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.hive.conf.HiveConf - Using the default value passed in for log id: 09e05d70-d
561-4b0c-b2db-07325f1c27c0
2021-06-14T19:43:29,031 INFO [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.hive.ql.session.SessionState - Resetting thread name to  main
hive> _
```

Figure 23: Creating sentiments table -2



```
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 13070245 HDFS Write: 0 SUCCESS
Stage-Stage-2:  HDFS Read: 8919444 HDFS Write: 149598 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
Time taken: 15.64 seconds
2021-06-14T19:43:29,029 INFO [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.hive.conf.HiveConf - Using the default value passed in for log id: 09e05d70-d
561-4b0c-b2db-07325f1c27c0
2021-06-14T19:43:29,031 INFO [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.hive.ql.session.SessionState - Resetting thread name to  main
hive> select * from tweets_sentiment limit 10;
2021-06-14T19:44:09,510 INFO [main] org.apache.hadoop.hive.conf.HiveConf - Using the default value passed in for log id: 09e05d70-d561-4b0c-b2db-07325f1c27c0
2021-06-14T19:44:09,510 INFO [main] org.apache.hadoop.hive.ql.session.SessionState - Updating thread name to 09e05d70-d561-4b0c-b2db-07325f1c27c0 main
2021-06-14T19:44:10,199 INFO [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.hive.common.FileUtils - Creating directory if it doesn't exist: file:/C:/User
s/HP/AppData/Local/Temp/HP/09e05d70-d561-4b0c-b2db-07325f1c27c0/hive_2021-06-14_19-44-09_535_2861232759515304205-1/-mr-10001/.hive-staging_hive_2021-06-14_19-44-09_535_
2861232759515304205-1
OK
NULL    neutral
0       positive
1       positive
2       negative
3       negative
4       neutral
5       positive
6       negative
7       positive
8       positive
Time taken: 0.699 seconds, Fetched: 10 row(s)
2021-06-14T19:44:10,275 INFO [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.hive.conf.HiveConf - Using the default value passed in for log id: 09e05d70-d
561-4b0c-b2db-07325f1c27c0
2021-06-14T19:44:10,276 INFO [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.hive.ql.session.SessionState - Resetting thread name to  main
hive> _
```

Figure 24: Creating sentiments table - 3

11. Output with tweets and id

create table tweet_sentiment as select

id,tweet,

case

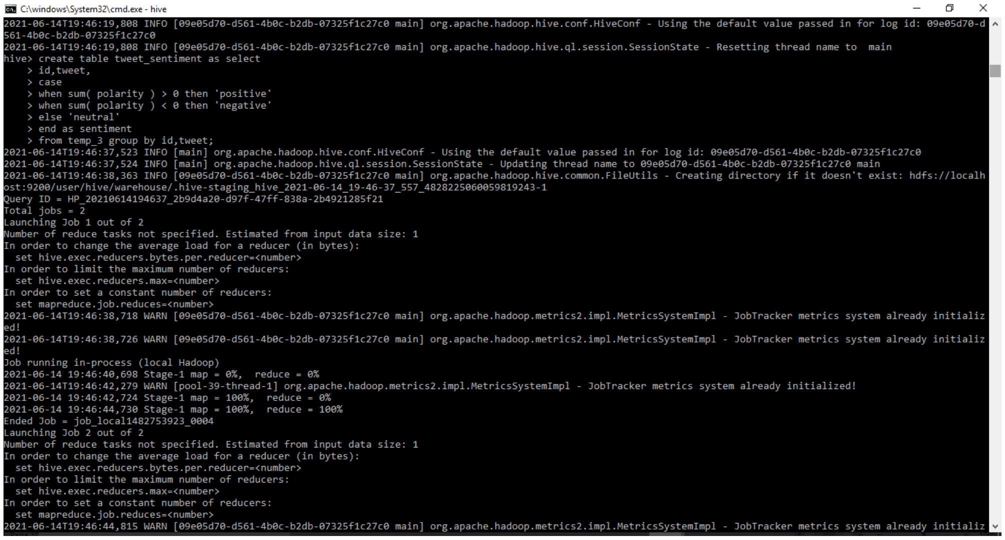
when sum(polarity) > 0 then 'positive'

when sum(polarity) < 0 then 'negative'

else 'neutral'

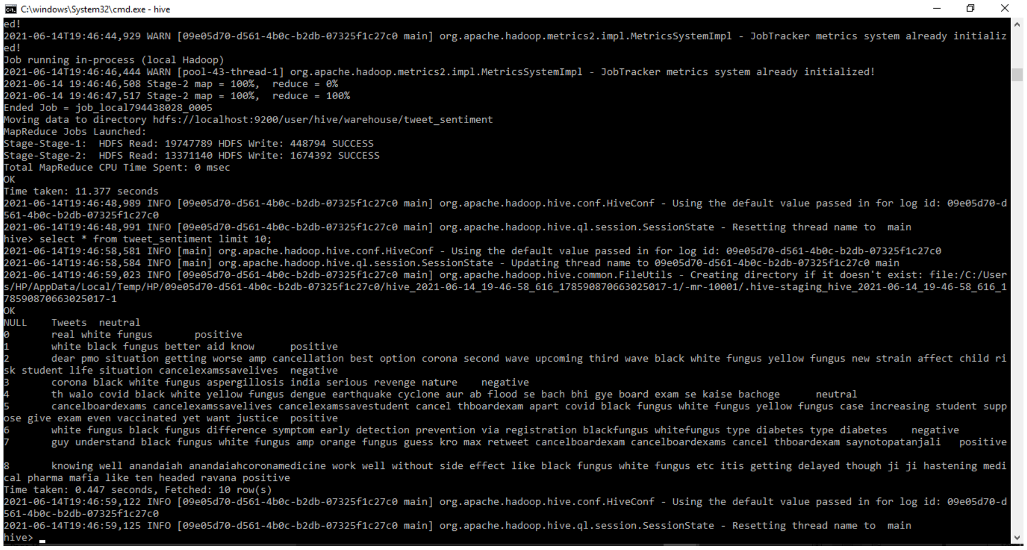
end as sentiment

from temp_3 group by id,tweet;



```
2021-06-14T19:46:19,808 INFO [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.hive.conf.HiveConf - Using the default value passed in for log id: 09e05d70-d
561-4b0c-b2db-07325f1c27c0
2021-06-14T19:46:19,808 INFO [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.hive.ql.session.SessionState - Resetting thread name to  main
hive> create table tweet_sentiment as select
    > id,tweet,
    > case
    > when sum( polarity ) > 0 then 'positive'
    > when sum( polarity ) < 0 then 'negative'
    > else 'neutral'
    > end as sentiment
    > from temp_3 group by id,tweet;
2021-06-14T19:46:37,523 INFO [main] org.apache.hadoop.hive.conf.HiveConf - Using the default value passed in for log id: 09e05d70-d561-4b0c-b2db-07325f1c27c0
2021-06-14T19:46:37,524 INFO [main] org.apache.hadoop.hive.ql.session.SessionState - Updating thread name to 09e05d70-d561-4b0c-b2db-07325f1c27c0 main
2021-06-14T19:46:38,363 INFO [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.hive.common.FileUtils - Creating directory if it doesn't exist: hdfs://localh
ost:9200/user/hive/warehouse/.hive-staging_hive_2021-06-14_19-46-37_557_4828225060059819243-1
Query ID = HP_20210614194637_2b9d4a20-d97f-47ff-838a-2b4921285f21
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 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>
2021-06-14T19:46:38,718 WARN [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initializ
ed!
2021-06-14T19:46:38,726 WARN [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initializ
ed!
Job running in-process (local Hadoop)
2021-06-14 19:46:40,698 Stage-1 map = 0%,  reduce = 0%
2021-06-14T19:46:42,279 WARN [pool-39-thread-1] org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-06-14 19:46:42,724 Stage-1 map = 100%,  reduce = 0%
2021-06-14 19:46:44,730 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local1482753923_0004
Launching Job 2 out of 2
Number of reduce tasks not specified. Estimated from input data size: 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>
2021-06-14T19:46:44,815 WARN [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initializ
```

Figure 25: Output with tweets & ID - 1

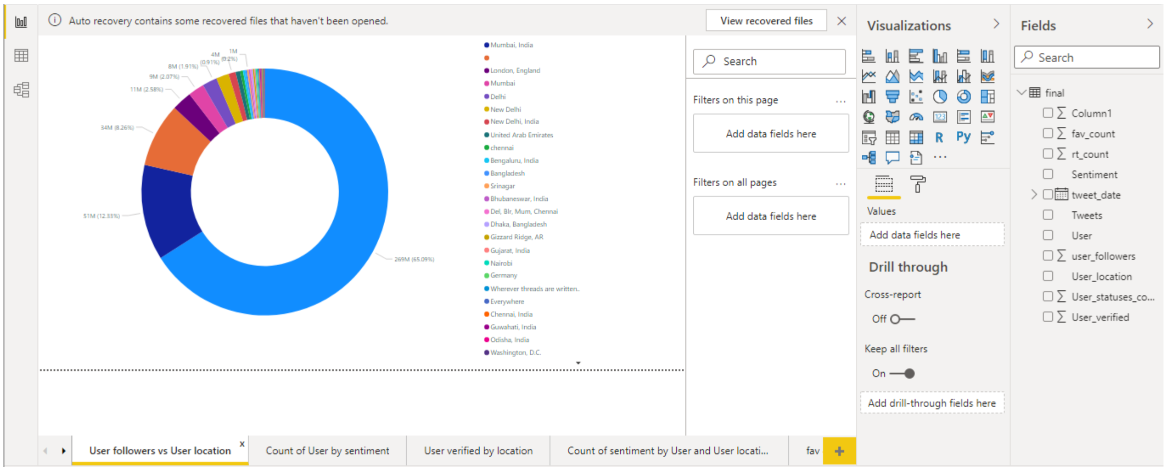


```
ed!
2021-06-14T19:46:44,929 WARN [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initializ
ed!
Job running in-process (local Hadoop)
2021-06-14T19:46:46,444 WARN [pool-43-thread-1] org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-06-14 19:46:46,508 Stage-2 map = 100%,  reduce = 0%
2021-06-14 19:46:47,517 Stage-2 map = 100%,  reduce = 100%
Ended Job = job_local794438028_0005
Moving data to directory hdfs://localhost:9200/user/hive/warehouse/tweet_sentiment
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 19747789 HDFS Write: 448794 SUCCESS
Stage-Stage-2:  HDFS Read: 13371140 HDFS Write: 1674392 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
Time taken: 11.377 seconds
2021-06-14T19:46:48,989 INFO [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.hive.conf.HiveConf - Using the default value passed in for log id: 09e05d70-d
561-4b0c-b2db-07325f1c27c0
2021-06-14T19:46:48,991 INFO [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.hive.ql.session.SessionState - Resetting thread name to  main
hive> select * from tweet_sentiment limit 10;
2021-06-14T19:46:58,581 INFO [main] org.apache.hadoop.hive.conf.HiveConf - Using the default value passed in for log id: 09e05d70-d561-4b0c-b2db-07325f1c27c0
2021-06-14T19:46:58,584 INFO [main] org.apache.hadoop.hive.ql.session.SessionState - Updating thread name to 09e05d70-d561-4b0c-b2db-07325f1c27c0 main
2021-06-14T19:46:59,023 INFO [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.hive.common.FileUtils - Creating directory if it doesn't exist: file:/C:/User
s/HP/AppData/Local/Temp/HP/09e05d70-d561-4b0c-b2db-07325f1c27c0/hive_2021-06-14_19-46-58_616_178590870663025017-1/-mr-10001/.hive-staging_hive_2021-06-14_19-46-58_616_1
78590870663025017-1
OK
NULL    Tweets  neutral
0       real white fungus       positive
1       white black fungus better aid know      positive
2       dear pmo situation getting worse amp cancellation best option corona second wave upcoming third wave black white fungus yellow fungus new strain affect child ri
sk student life situation cancelexamssavelives  negative
3       corona black white fungus aspergillosis india serious revenge nature    negative
4       th walo covid black white yellow fungus dengue earthquake cyclone aur ab flood se bach bhi gye board exam se kaise bachoge      neutral
5       cancelboardexams cancelexamssavelives cancelexamssavestudent cancel thboardexam apart covid black fungus white fungus yellow fungus case increasing student supp
ose give exam even vaccinated yet want justice  positive
6       white fungus black fungus difference symptom early detection prevention via registration blackfungus whitefungus type diabetes type diabetes    negative
7       guy understand black fungus white fungus amp orange fungus guess kro max retweet cancelboardexam cancelboardexams cancel thboardexam saynotopatanjali   positive

8       knowing well anandaiah anandaiahcoronamedicine work well without side effect like black fungus white fungus etc itis getting delayed though ji ji hastening medi
cal pharma mafia like ten headed ravana positive
Time taken: 0.447 seconds, Fetched: 10 row(s)
2021-06-14T19:46:59,122 INFO [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.hive.conf.HiveConf - Using the default value passed in for log id: 09e05d70-d
561-4b0c-b2db-07325f1c27c0
2021-06-14T19:46:59,125 INFO [09e05d70-d561-4b0c-b2db-07325f1c27c0 main] org.apache.hadoop.hive.ql.session.SessionState - Resetting thread name to  main
hive> _
```

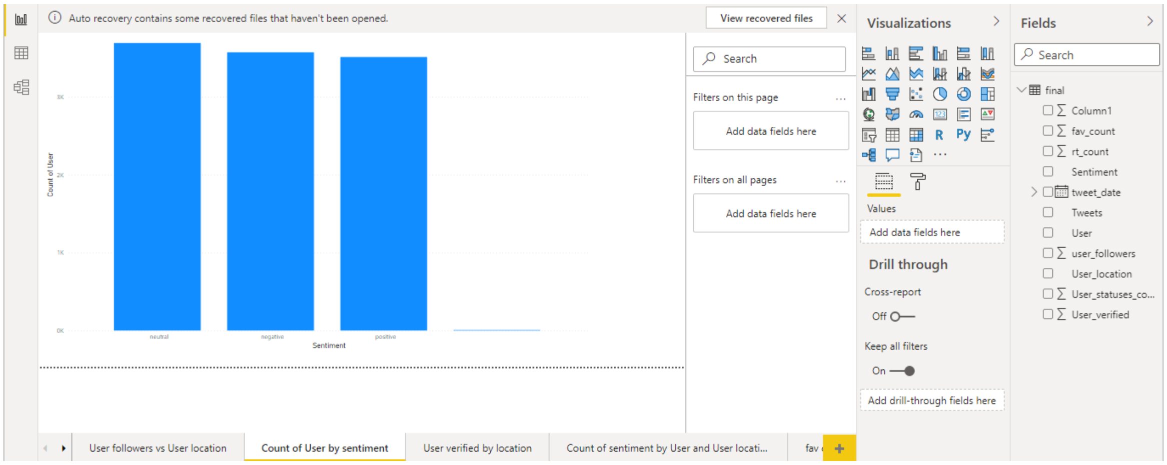
Figure 26: Output with tweets & ID - 2

3.3. Visualization screenshots



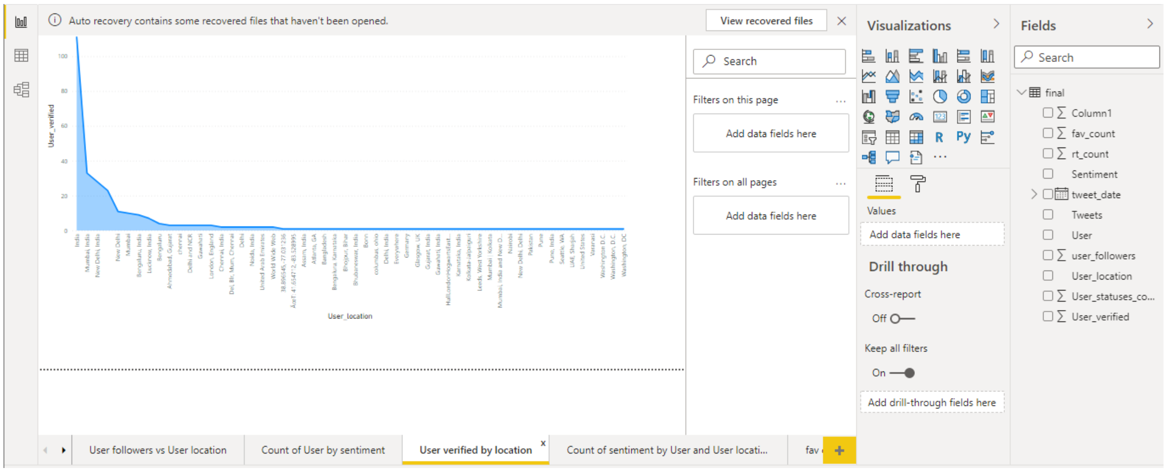
Auto recovery contains some recovered files that haven't been opened.

Figure 27: User Followers vs User Location



Auto recovery contains some recovered files that haven't been opened.

Figure 28: Count of User by Sentiment



Auto recovery contains some recovered files that haven't been opened.

Figure 29: User verified by Location

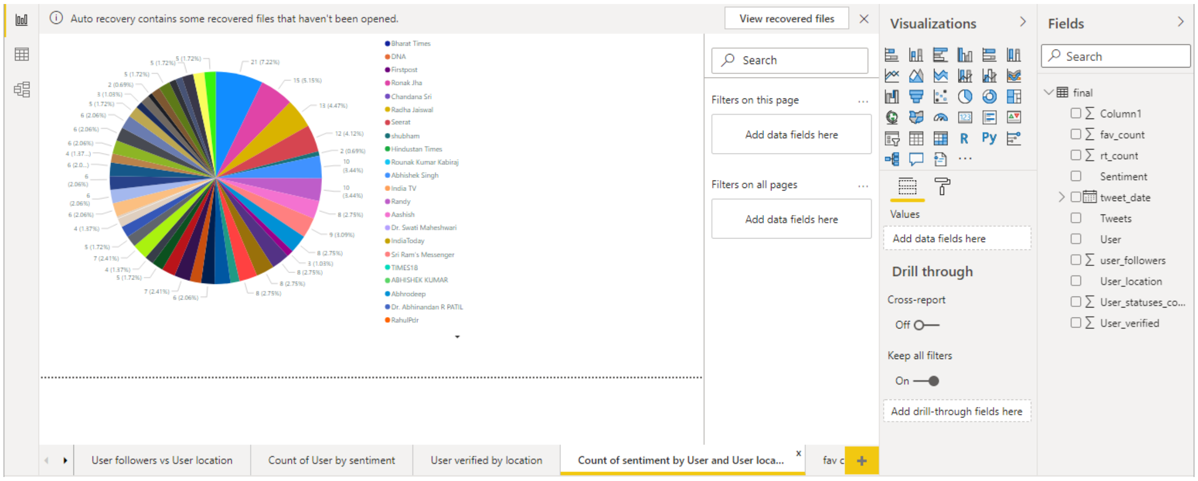


Figure 30: Count of Sentiment by User and User Location

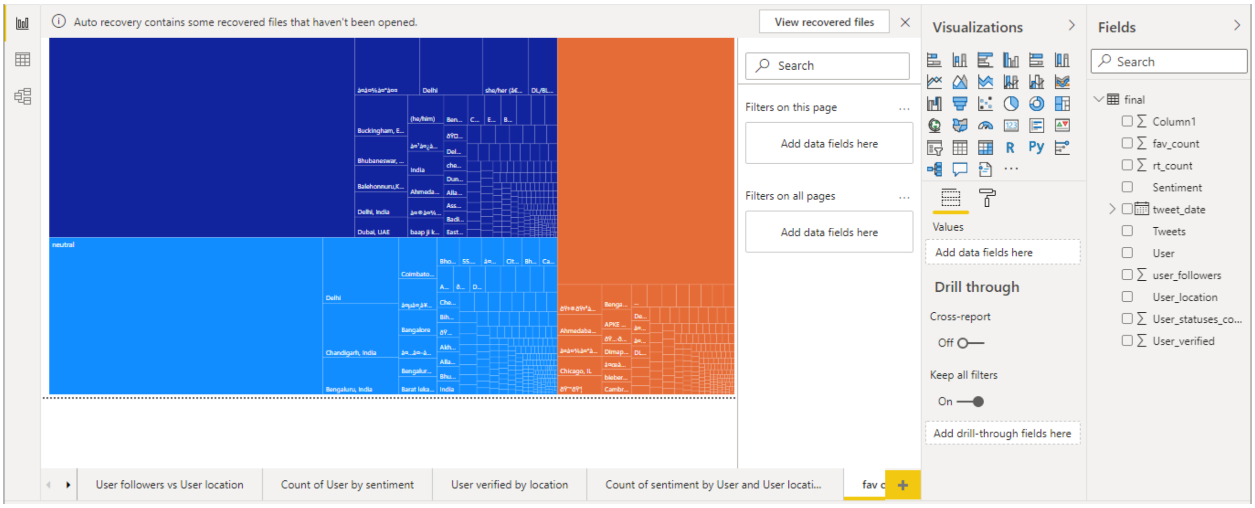


Figure 31: Favorite Count by sentiment and user location

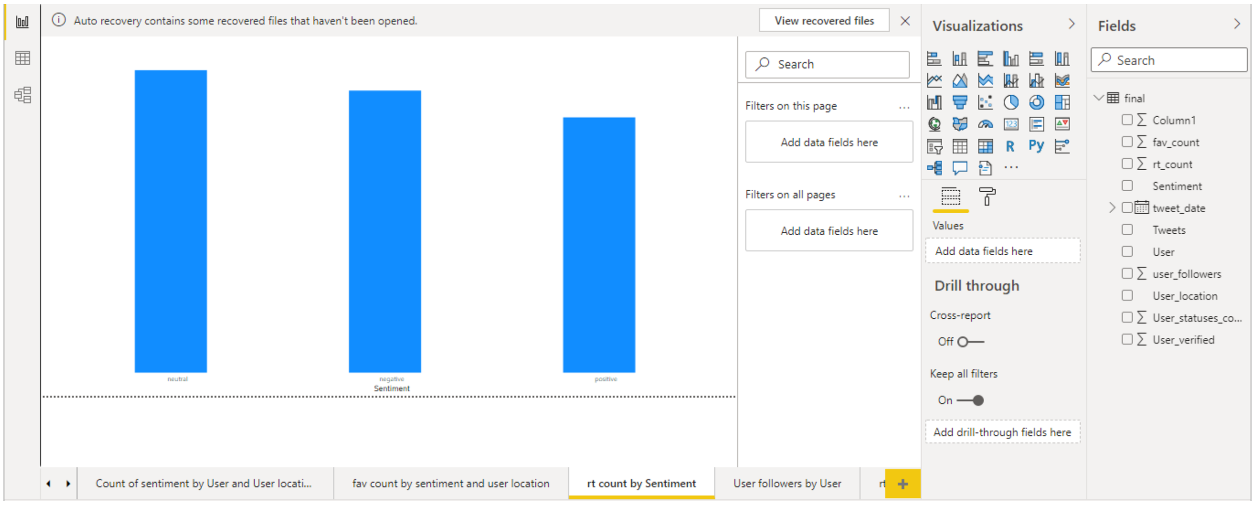


Figure 32: Retweet Count by Sentiment

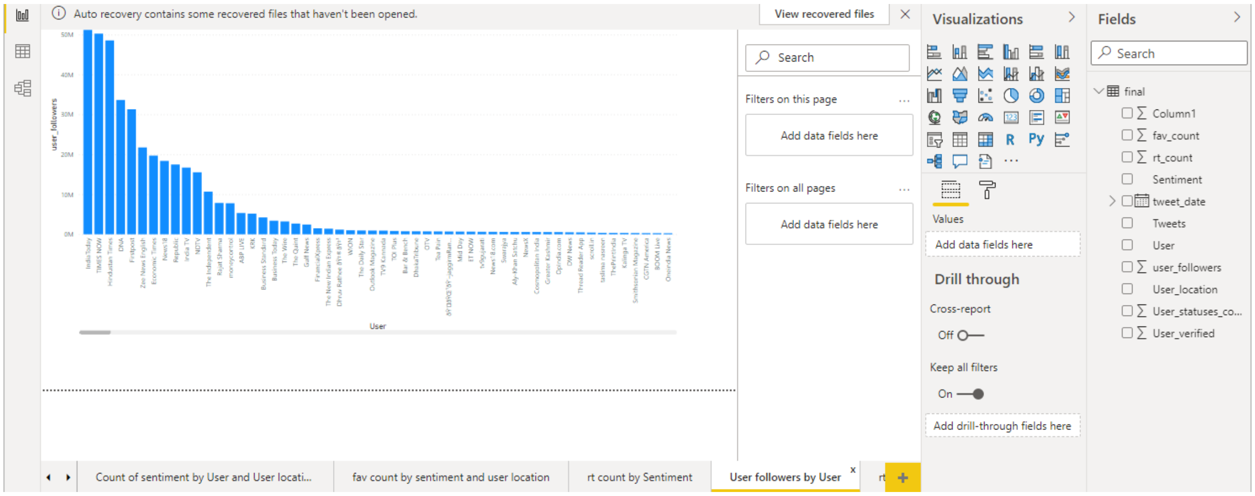


Figure 33: User Followers by User

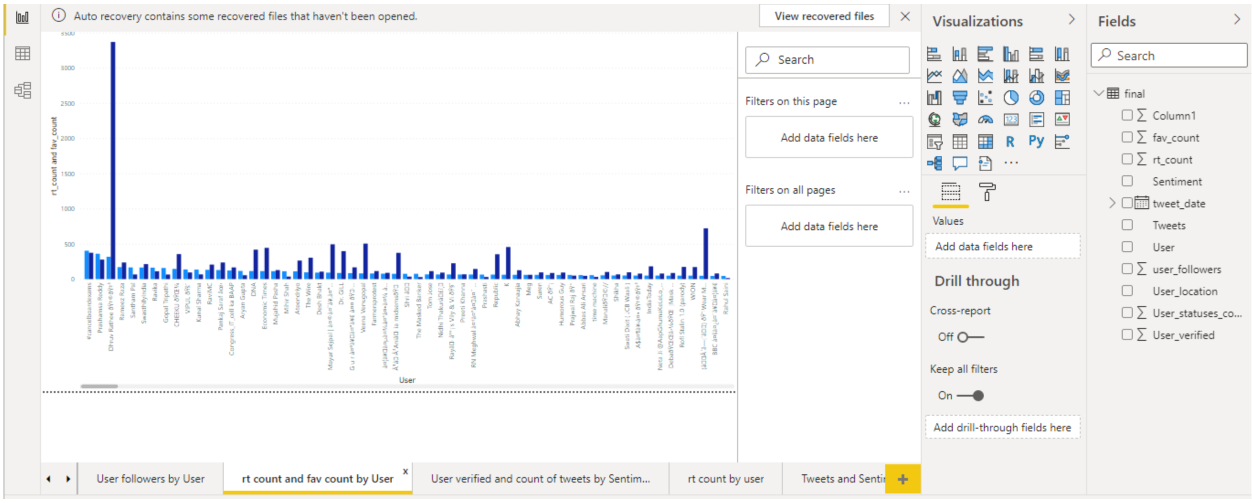


Figure 34: Retweet count and Favorite count by User

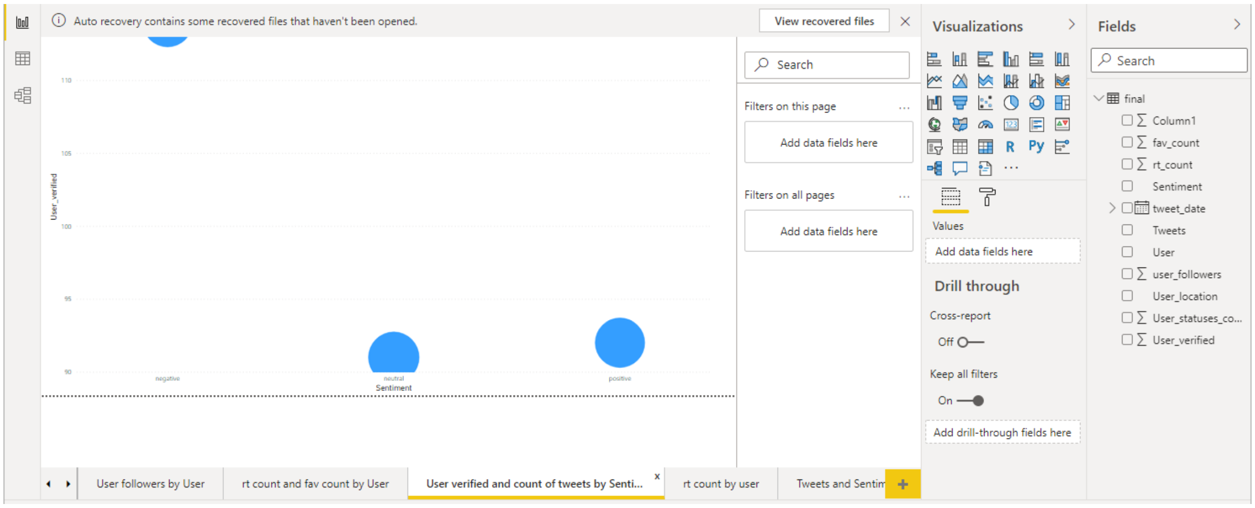
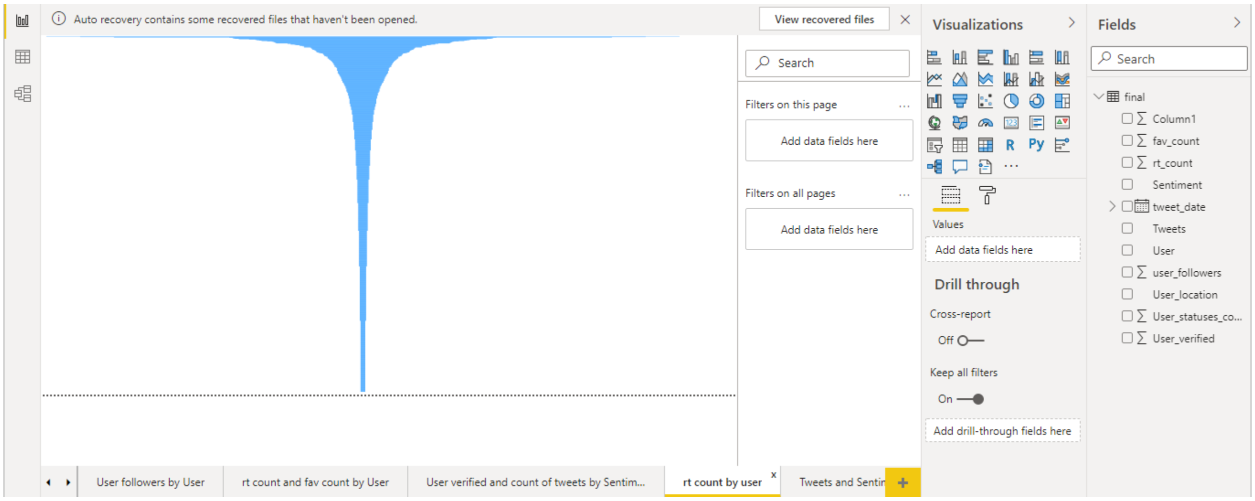


Figure 35: User verified and count of tweets by Sentiment

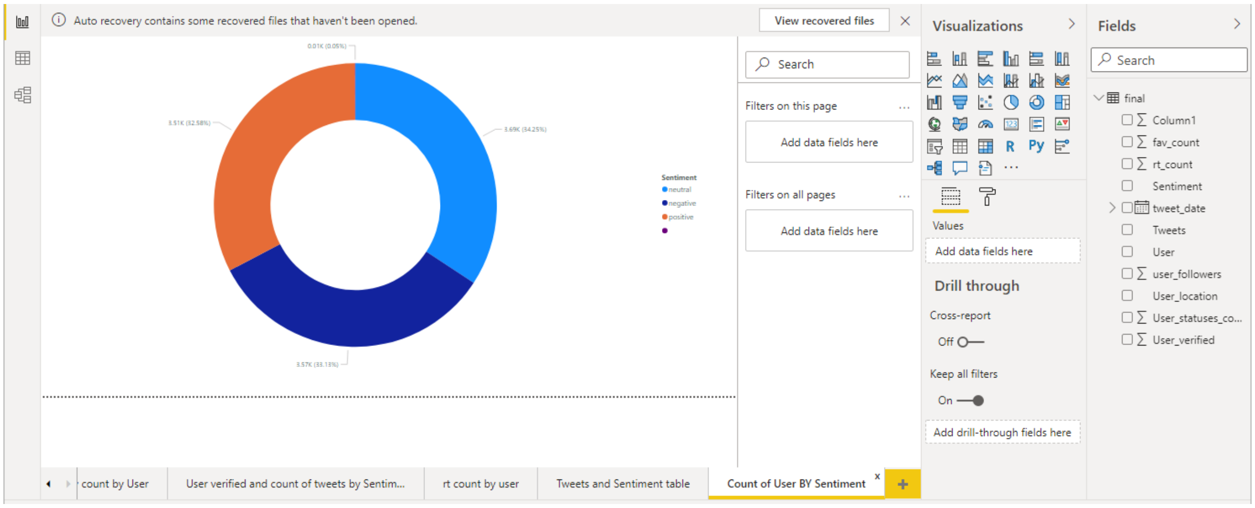


Auto recovery contains some recovered files that haven't been opened.

Cross-report: Off

Keep all filters: On

Figure 36: Retweet Count by User



Auto recovery contains some recovered files that haven't been opened.

Cross-report: Off

Keep all filters: On

Figure 37: Count of User by Sentiment

CHAPTER 4
FUTURE SCOPE

4.1. Conclusion

This study focuses on the application of big data analytics technology in the prevention and control of major public health incidents. First of all, we clarified the definition, characteristics, and prevention difficulties of major public health incidents. To cope with these difficulties, the use of big data is an important means to assist prevention and control in major public health incidents. Governments may make full use of the application of big data in an epidemic situation in all aspects of prevention and control, and they can further improve the epidemic prevention mechanism based on big data analytics. In terms of information collection, data collection platforms for the Internet of Things, mobile devices, navigation and search engines, social media, and large-scale gene banks can be fully established. On the basis of information collection, it is necessary to establish an early warning detection mechanism for big data analytics, e.g., using visual analysis, deep learning, and forecast line analysis techniques. This may be used as a basis for early warning and forecasting, formulating plans, rapid decision-making, and starting emergency mechanisms.

Second, governments can further improve their epidemic response mechanisms based on big data analytics. Big data technology can be used for diseases identification, decision support, coordination and communication, and technical support. Diseases identification typically uses predictive analysis of infectious disease dynamic models combined with data to make predictions regarding the criticality of an event, provide support for management decisions, report, and take timely response measures. Graph database analysis and geographic information systems can provide a significant advantage in tracking infected persons and their contacts, thus determining the source of infection. For research into virus sources and the research and development of specific drugs and vaccines, we should fully utilize the potential of big data technology for research support and technical consultation regarding genetic data and real-time patient data transmitted by the Internet of Things.

Third, the government should establish an epidemic repair mechanism based on big data analysis and promoting the sharing of big data in different regions, industries, and platforms. This includes the use of big data to eliminate fear, for recovery, audit assessment, and policy adjustment. Big data analytics can be convenient in ameliorating public fear by revealing the real-time epidemic situation and clarifying rumours. The big data analysis model can also be used to estimate the impact of the epidemic on political, economic, and social development, so as to assist governments to make suitable decisions, make policy adjustments, integrate prevention and resistance measures, and promote a rapid economic recovery.

It is important that big data analytics is merely a supportive method to assist in ex-ante prediction and ex-post prevention and control. Big data analytics has certain limitations and application premises. For instance, in terms of predicting in advance, some methods of early warning with big data may be directly used for diseases that scientists already understand, such as influenza, because relevant massive data have been accumulated. However, faced with the first generation of new viruses & diseases, it may not be possible to directly generate closely related big data, so it may not be possible to use big data technology for immediate analysis. Notwithstanding, researchers may explore the combination of data from other relevant sources for analysis, such as the incidence of internal and external causes (climates, other epidemics, etc.) that could affect how a virus is produced, and the probability of the occurrence of new viruses being observed and analysed in advance. Moreover, the application of big data in epidemic prevention and control must consider the practicability of administrative rights, privacy protection, cost, and so on, and the balance of interests with public epidemic prevention so as to ensure operability.

4.2. Future Work

Future work involves collaborating with "Spark NLP," an open-source library for the processing of natural languages, developed on top of Apache Spark and Spark ML. Integrating with ML Pipelines offers a simple API. The Spark NLP library, which contains Scala and Python APIs for use by Spark, is written in Scala. It has no reliance on any other library of NLP or ML. The library includes the ability to train, modify and store models as a native extension of the Spark ML API so that they can run on a cluster or other machines or save for later.

REFERENCES

[1] The Apache Software Foundation, https://flume.apache.org

[2] The Apache Software Foundation, https://sqoop.apache.org

[3] Dataiku, http://www.dataiku.com/blog/2013/05/01/a-complete-guide-to-writing-hive-udf.html

[4] Pillar, http://www.3pillarglobal.com/insights/how-to-tame-the-machine-learning-beast-with-apache-mahout

[5] Rajurkar G.D., Goudar R.M.: “**A speedy data uploading approach for twitter trend and sentiment analysis using HADOOP”**. *In: 2015 International Conference on Computing Communication Control and Automation, pp. 580–584. IEEE, Pune (2015)* Google Scholar

[6] Mane S.B., Sawnt Y., Kazi S., Shinde V.: “**Real time sentiment analysis of twitter data using Hadoop**”. *In: International Journal of Computer Science and Information Technology, vol. 5(3), pp. 3098–3100, IJCSIT (2014)* Google Scholar

[7] Zarrad A., Aljialoud J.: “**The evaluation of the public opinion a case study: MERS-Cov infection virus in KSA**”. *In: 7th International Conference on Utility and Cloud Computing, pp. 664–607. IEEE, London (2014)* Google Scholar

[8] Hammond K., Varde A.S.: **“Cloud based predictive analytics. In: 13th International Conference on Data Mining Workshops”,** *pp. 607–612. IEEE, Dallas, TX (2013)* Google Scholar

[9] Shang S., Shi M., Shan W., Hong Z.: “**Research on public opinion based on big data**”. *In: 14thInternational Conference on Computer and Information Science, pp. 559–562. IEEE, Las Vegas, NV (2015)* Google Scholar

[10] Lui B., Blasch E., Chen Y., Shen D., Chen G.: “**Scalable sentiment classification for big data analysis using naive bayes classifier.**” *In: International Conference on Big Data, pp. 99–104. IEEE, Silicon Valley, CA (2013)* Google Scholar

[11] Conejero J, Burnap P., Rana O., Morgan J.: “**Scaling archived social media data analysis using a hadoop cloud.”** *In: Sixth International Conference on Cloud Computing, pp. 685–692. IEEE, Santa Clara, CA (2013)* Google Scholar