CSEE5590/490: Big Data Programming

Project Increment 2

Instructor - Zeenat Tariq

Team Number :- 6

1.Project Title and Team Members

International Student Data Analysis using Twitter API

- Tanvi Jain
- Saikumar Reddy Papagari
- Amarnadha Reddy Ankireddypalli
- Thotakura Naga Mounika

2. Introduction

This project focuses on extraction of dataset (Twitter Data) from the Twitter API regarding all the information related to international students and using different tools like Hive,Hue pySpark, Solr to show different aspects of the data in a visualization form.

3. Background

Twitter sentiment analysis is done in different ways before, using panda is the most common way. We will be using different hashtags which are related to international students and events related to international students to get a hold of better results based on those hashtags and more information about what events really took place and different aspects they

bring out. Here are some references of projects people have done implementing twitter analysis with big data, we are going to follow some aspects of these projects but our dataset is new and is extracted only for this project, so we could only find references related to those projects which have totally different dataset and completely different objectives.

1. https://github.com/dsuarez993/bigdata-realtime-twitter-analysis
2. https://github.com/6vedant/TwitterAnalyticsHadoop
3. https://www.toptal.com/apache/apache-spark-streaming-twitter

4. Goals and Objectives:

- **Motivation**

Studying abroad is a journey of education and discovery. There are currently over 1 million international students from more than 220 countries, coming to the United States annually. The individuals from this group belong to the international student community and we came to the United States to get a higher education. There are numerous situations where we are relied upon to follow those standards which the overall US citizens are not expected to follow in the long run since we have a place with the foreigner gathering. So, this gave us the establishment for this task and we chose to feature the fundamental information like percentage of students going to the United States for education, the probability of getting a work VISA, Immigration rules change

for F-1 Visa during COVID, Jobs for international students during covid(H1B sponsorship).

- **Significance**

Big data tools help to analyze the huge data which helps to provide efficient results.The sentimental analysis provides a brief understanding of various challenges that international students are currently facing and impact of covid-19 on the visa assurance.

- **Objectives**

The objective of this project is to get twitter data extraction using Twitter Data Analysis and then cleaning the data, performing sentimental analysis and importing files into Hadoop where we used different tools like Hive,Solr and pySpark.

- **Features**

The main feature of the project is to collect the Real timed tweets from the twitter API, also by performing the ETL which means we preprocess the data and extract the necessary data and then we load the extracted data in our HIVE. Performed topic modelling using LDA and gensim model, and done visualization of the top 4 topics by t-SNE visualization. Performed word cloud visualization using pyspark.

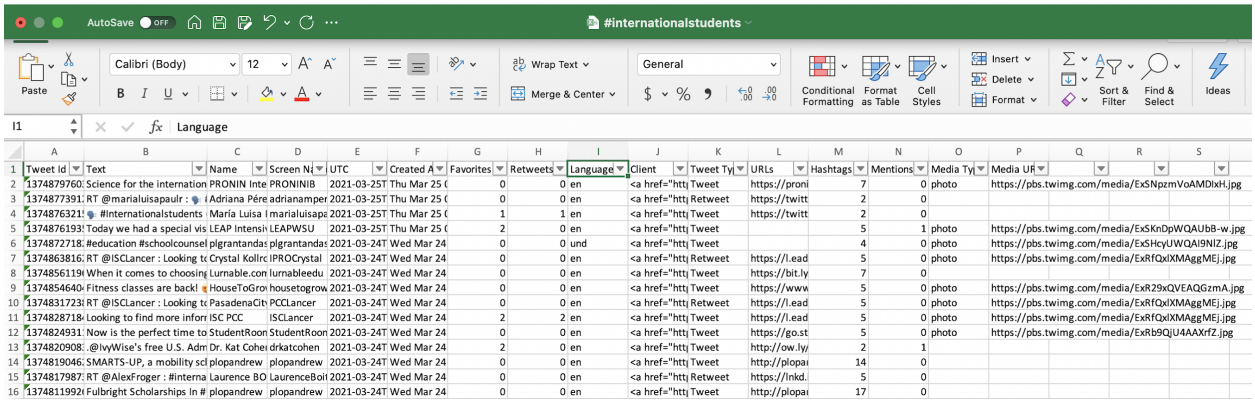
5. For all the features developed for this increment, write a documentation describing the design, implementation, testing, and deployment (including the precise descriptions and screenshots).

•Dataset

We collected the data using twitter API using developer account and API keys. We have used different hashtags to get data related to 3 different scenarios as follows :-

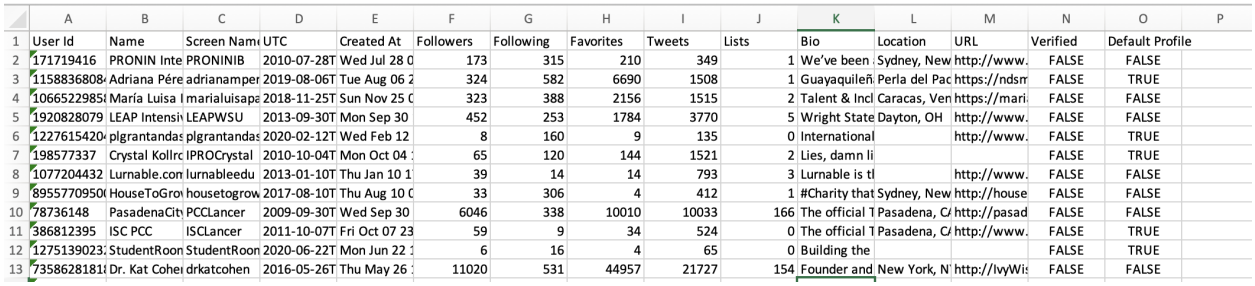
1. Generic International Student Data **(#F1visa #intlstudents #internationalstudents #studyinUSA)**
2. Immigration rules change for F1 Visa during COVID in 2020 **(#AbolishICE)**
3. Jobs for international students during covid (H1B sponsorship) **(#H1B, #h1bjobs)**

   From the extraction we have the information about the tweet itself like this :-



Tweet Id	Text	Name	Screen Na	UTC	Created A	Favorites	Retweets	Language	Client	Tweet Ty	URLs	Hashtags	Mentions	Media Ty	Media UR
1374879760	Science for the internation	PRONIN	Inte	PRONINIB	2021-03-25T	Thu Mar 25 (	0	0	en	<a href="htt	Tweet	https://proni	7	0	photo	https://pbs.twimg.com/media/ExSNpzmVoAMDlxH.jpg
1374877391	RT @marialuisapaulr :	Adriana Pére	adrianamper	2021-03-25T	Thu Mar 25 (	0	0	en	<a href="htt	Tweet	Retweet	https://twitt	2	0		
1374876321	#internationalstudents	María Luisa I	marialuisapa	2021-03-25T	Thu Mar 25 (	1	1	en	<a href="htt	Tweet	https://twitt	2	0		
1374876193	Today we had a special vis	LEAP Intensi	LEAPWSU	2021-03-25T	Thu Mar 25 (	2	0	en	<a href="htt	Tweet		5	1	photo	https://pbs.twimg.com/media/ExSKnDpWQAUbB-w.jpg
1374872718	#education #schoolcounsel	plgrantandas	plgrantandas	2021-03-24T	Wed Mar 24	0	0	und	<a href="htt	Tweet		4	0	photo	https://pbs.twimg.com/media/ExSHcyUWQAI9NIZ.jpg
1374863816	RT @ISCLancer : Looking tc	Crystal Koll	IPROCrystal	2021-03-24T	Wed Mar 24	0	0	en	<a href="htt	Tweet	Retweet	https://l.ead	5	0	photo	https://pbs.twimg.com/media/ExRfQxlXMAggMEj.jpg
1374856119	When it comes to choosing	Lurnable.com	lurnableedu	2021-03-24T	Wed Mar 24	0	0	en	<a href="htt	Tweet	https://bit.ly	7	0		
1374854640	Fitness classes are back!	HouseToGro	housetogrow	2021-03-24T	Wed Mar 24	0	0	en	<a href="htt	Tweet	https://www	5	0	photo	https://pbs.twimg.com/media/ExR29xQVEAQGzmA.jpg
1374831723	RT @ISCLancer : Looking tc	PasadenaCit	PCCLancer	2021-03-24T	Wed Mar 24	0	0	en	<a href="htt	Tweet	Retweet	https://l.ead	5	0	photo	https://pbs.twimg.com/media/ExRfQxlXMAggMEj.jpg
1374828718	Looking to find more infor	ISC PCC	ISCLancer	2021-03-24T	Wed Mar 24	2	2	en	<a href="htt	Tweet	https://l.ead	5	0	photo	https://pbs.twimg.com/media/ExRfQxlXMAggMEj.jpg
1374824931	Now is the perfect time to	StudentRoon	StudentRoon	2021-03-24T	Wed Mar 24	0	0	en	<a href="htt	Tweet	https://go.st	5	0	photo	https://pbs.twimg.com/media/ExRb9QjU4AAXrfZ.jpg
1374820908	.@IvyWise's free U.S. Adm	Dr. Kat Cohe	drkatcohen	2021-03-24T	Wed Mar 24	2	0	en	<a href="htt	Tweet	http://ow.ly/	2	1		
1374819046	SMARTS-UP, a mobility scl	plopandrew	plopandrew	2021-03-24T	Wed Mar 24	0	0	en	<a href="htt	Tweet	http://plopa	14	0		
1374817987	RT @AlexFroger : #interna	Laurence BO	LaurenceBoi	2021-03-24T	Wed Mar 24	0	0	en	<a href="htt	Tweet	Retweet	https://lnkd.	5	0		
1374811992	Fulbright Scholarships In #	plopandrew	plopandrew	2021-03-24T	Wed Mar 24	0	0	en	<a href="htt	Tweet	http://plopa	17	0		

The information about the user as well -



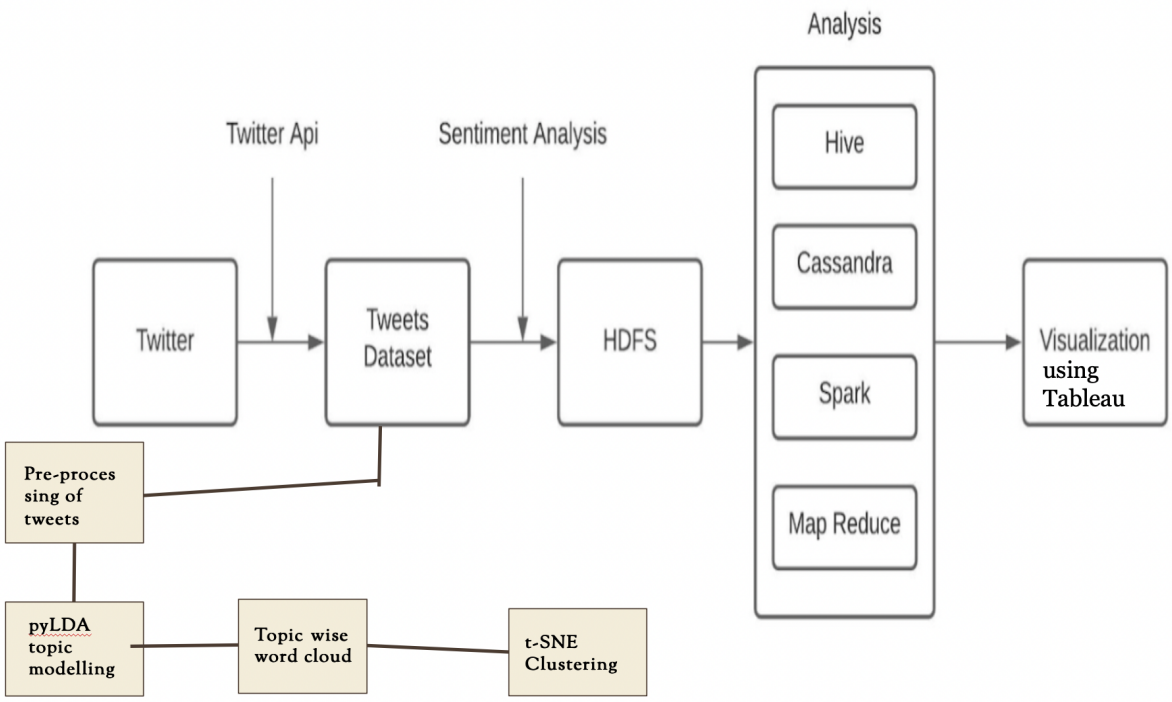
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	User Id	Name	Screen Nam	UTC	Created At	Followers	Following	Favorites	Tweets	Lists	Bio	Location	URL	Verified	Default Profile	
2	171719416	PRONIN Inte	PRONINIB	2010-07-28T	Wed Jul 28 0	173	315	210	349	1	We've been	Sydney, New	http://www.	FALSE	FALSE	
3	1158836808	Adriana Pére	adrianamper	2019-08-06T	Tue Aug 06 2	324	582	6690	1508	1	Guayaquileñ	Perla del Pac	https://ndsm	FALSE	TRUE	
4	1066522985	María Luisa I	marialuisapa	2018-11-25T	Sun Nov 25 0	323	388	2156	1515	2	Talent & Incl	Caracas, Ven	https://mari	FALSE	FALSE	
5	1920828079	LEAP Intensi	LEAPWSU	2013-09-30T	Mon Sep 30	452	253	1784	3770	5	Wright State	Dayton, OH	http://www.	FALSE	FALSE	
6	1227615420	plgrantandas	plgrantandas	2020-02-12T	Wed Feb 12	8	160	9	135	0	International		http://www.	FALSE	TRUE	
7	198577337	Crystal Kollrc	IPROCrystal	2010-10-04T	Mon Oct 04 1	65	120	144	1521	2	Lies, damn li			FALSE	TRUE	
8	1077204432	Lurnable.com	lurnableedu	2013-01-10T	Thu Jan 10 1	39	14	14	793	3	Lurnable is tl		http://www.	FALSE	FALSE	
9	8955770950	HouseToGro	housetogrow	2017-08-10T	Thu Aug 10 0	33	306	4	412	1	#Charity that	Sydney, New	http://house	FALSE	FALSE	
10	78736148	PasadenaCity	PCCLancer	2009-09-30T	Wed Sep 30	6046	338	10010	10033	166	The official T	Pasadena, C/	http://pasad	FALSE	FALSE	
11	386812395	ISC PCC	ISCLancer	2011-10-07T	Fri Oct 07 23	59	9	34	524	0	The official T	Pasadena, C/	http://www.	FALSE	TRUE	
12	1275139023	StudentRoon	StudentRoon	2020-06-22T	Mon Jun 22 1	6	16	4	65	0	Building the			FALSE	TRUE	
13	7358628181	Dr. Kat Cohe	drkatcohen	2016-05-26T	Thu May 26 1	11020	531	44957	21727	154	Founder and	New York, N	http://IvyWis	FALSE	FALSE	

It has all the different fields, which we will filter and use later using big data tools to perform queries and visualization.

These are some of the features and the description of those features :-

Feature	Description
id	Unique id of user
created at	Tweet created time stamp in UTC
text	UTF-8 text Tweet data
source	Type of device used to post the tweet
Name	Name of the user
Place	Geo location of user at time of tweet posted
Screen_name	Profile or Screen name of the user
Lang	Language opted by user
user location	Location of the user
user_followers_count	Count of followers of user
user_friends_count	Count of friends of user
user_favourites_count	Count of people favorited the tweet
reply_count	Count of replies to the tweet
retweet_count	Count of retweets of the tweet

- **Detail design of Features with Project Workflow**



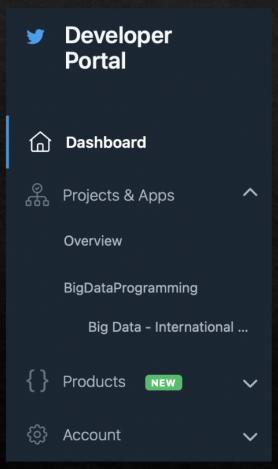
•Analysis (Details about data)

We have extracted dataset using twitter API, these datasets are extracted using different hashtags focusing on different events related to international students, dividing the events and dataset is crucial since we want this project to be informational and focus on different aspects through different datasets. We will use sentiment analysis on the text (tweet) and use different queries to extract important features and visualize them in Tableau.

•Implementation (Using Hadoop and Cassandra for Analysis)

Data extraction :-

Extracting data using Twitter API keys




```python
# Python Script to Extract tweets of a
# particular Hashtag using Tweepy and Pandas


# import modules
import pandas as pd
import tweepy


# function to display data of each tweet
def printtweetdata(n, ith_tweet):
  print()
  print(f"Tweet {n}:")
  print(f"Username:{ith_tweet[0]}")
  print(f"Description:{ith_tweet[1]}")
  print(f"Location:{ith_tweet[2]}")
  print(f"Following Count:{ith_tweet[3]}")
  print(f"Follower Count:{ith_tweet[4]}")
  print(f"Total Tweets:{ith_tweet[5]}")
  print(f"Retweet Count:{ith_tweet[6]}")
  print(f"Tweet Text:{ith_tweet[7]}")
  print(f"Hashtags Used:{ith_tweet[8]}")


# function to perform data extraction
def scrape(words, date_since, numtweet):

  # Creating DataFrame using pandas
  db = pd.DataFrame(columns=['username', 'description', 'location', 'following',
             'followers', 'totaltweets', 'retweetcount', 'text', 'hashtags'])
```

Preprocessing of the data

Preprocessing

```python
def remove_pattern(input_txt, pattern):
    r = re.findall(pattern, input_txt)
    for i in r:
        input_txt = re.sub(i, '', input_txt)

    return input_txt

# remove twitter handles (@user)
train['tidy_tweet'] = np.vectorize(remove_pattern)(train['tweets'], "@[\w]*")

# remove special characters, numbers, punctuations
train['tidy_tweet'] = train['tidy_tweet'].str.replace("[^a-zA-Z#]", " ")

#removing words less than length 2
train['tidy_tweet'] = train['tidy_tweet'].apply(lambda x: ' '.join([w for w in x.split() if len(w)>2]))

tokenized_tweet = train['tidy_tweet'].apply(lambda x: x.split())
tokenized_tweet.head()
```

```
0    [any, hope, for, international, students, who,...
1    [Canada, and, the, are, among, the, choicest, ...
2    [Vaccine, approved, Ready, pay, for, quarantin...
3    [Email, sharpwriters, com, your, orders, for, ...
4    [Scotty, promised, Australians, cat, the, fron...
Name: tidy_tweet, dtype: object
```

```python
#stemming
from nltk.stem.porter import *
stemmer = PorterStemmer()

tokenized_tweet = tokenized_tweet.apply(lambda x: [stemmer.stem(i) for i in x]) # stemming
tokenized_tweet.head()
```

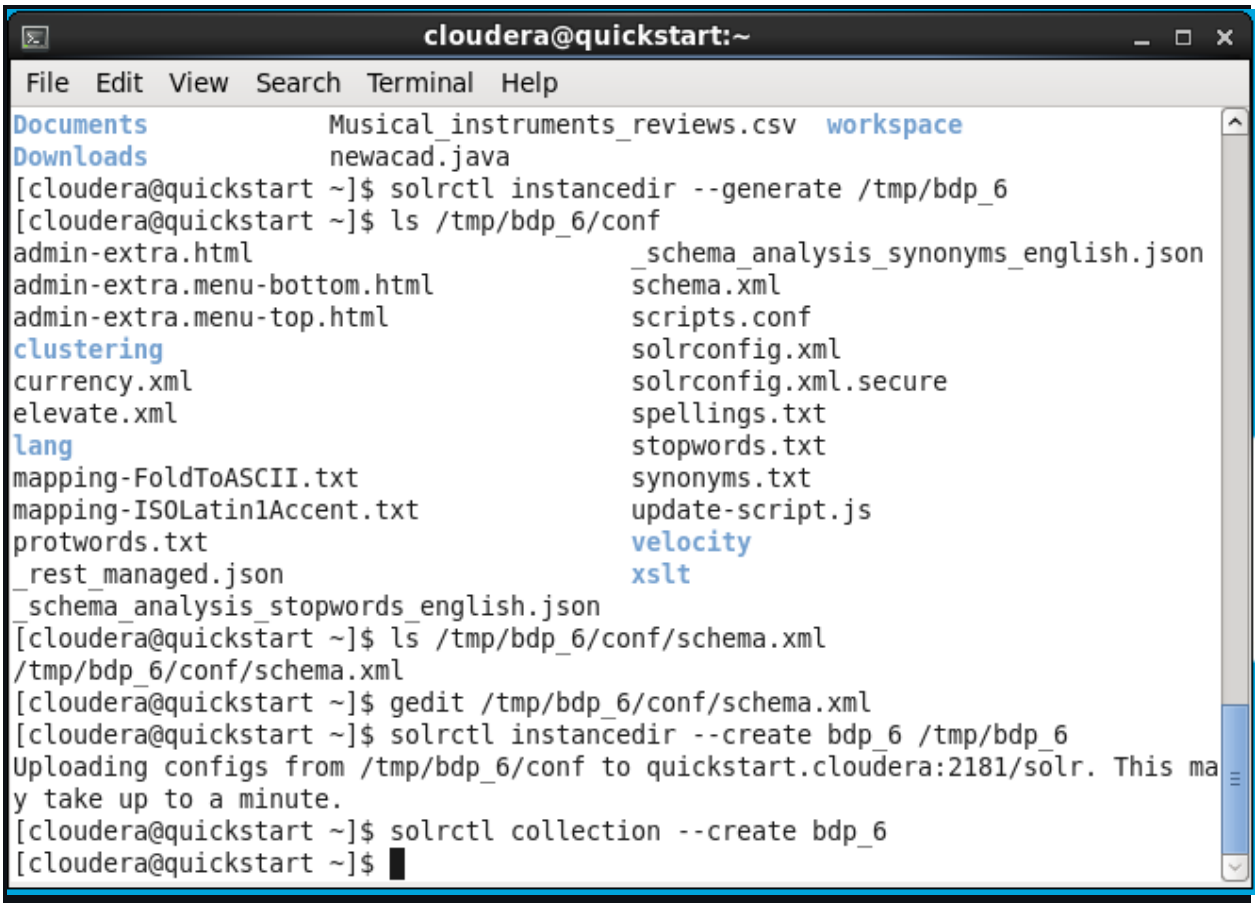
```
[32]  train.head()
```

	tweets	tidy_tweet
0	b'@marcomendicino @marcomendicino any hope for...	ani hope for intern student who miss the histo...
1	b'Canada and the US are among the choicest stu...	canada and the are among the choicest studi ab...
2	b"RT @brarharry685: Vaccine approved. Ready to...	vaccin approv readi pay for quarantin pleas pr...
3	b'DM or Email:sharpwriters94@gmail.com your or...	email sharpwrit com your order for qualiti ass...
4	b'RT @JulianHillMP: Scotty promised Australian...	scotti promis australian cat the front the vac...

```
[41]  tidy_tweets = (train['tidy_tweet'])
```

Solr:

1. Creation/generation of instance & Collection:

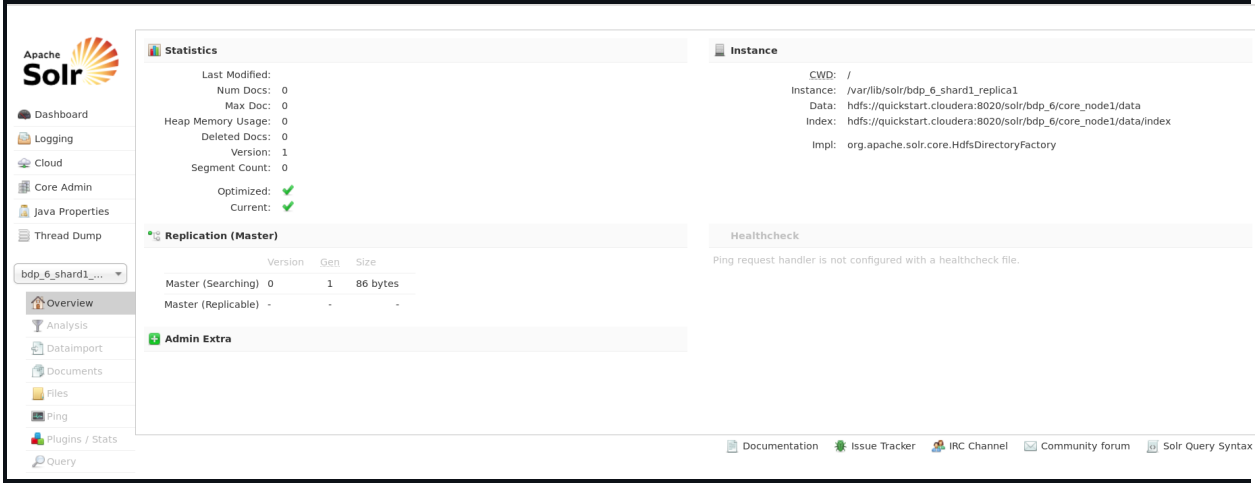


2. Edit the schema.xml created with the instance generation inside the configuration folder to change the attributes based on the dataset given:

schema.xml

```
      <field name="Tweet Id" type="string" indexed="true" stored="true" required="true" multiValued="false" />
      <field name="Text" type="string" indexed="true" stored="true" />
      <field name="Name" type="string" indexed="true" stored="true" />
      <field name="Screen Name" type="string" indexed="true" stored="true" />
      <field name="UTC" type="string" indexed="true" stored="true" />
      <field name="Created At" type="string" indexed="true" stored="true" />
      <field name="Favorites" type="string" indexed="true" stored="true" />
      <field name="Retweets" type="string" indexed="true" stored="true" />
      <field name="Language" type="string" indexed="true" stored="true" />
      <field name="Client" type="string" indexed="true" stored="true" />
      <field name="Tweet Type" type="string" indexed="true" stored="true" />
      <field name="URLs" type="string" indexed="true" stored="true" />
      <field name="Hashtags" type="string" indexed="true" stored="true" />
      <field name="Mentions" type="string" indexed="true" stored="true" />
      <field name="Media Type" type="string" indexed="true" stored="true" />
      <field name="Media URLs" type="string" indexed="true" stored="true" />
       <field name="" type="string" indexed="true" stored="true" />
      <field name="" type="string" indexed="true" stored="true" />
      <field name="" type="string" indexed="true" stored="true" />

 <!-- points to the root document of a block of nested documents. Required for nested
      document support, may be removed otherwise
 -->
 <field name="_root_" type="string" indexed="true" stored="false"/>

 <field name="sku" type="text_en_splitting_tight" indexed="true" stored="true" omitNorms="true"/>
 <field name="name" type="text_general" indexed="true" stored="true"/>
 <field name="manu" type="text_general" indexed="true" stored="true" omitNorms="true"/>
 <field name="cat" type="string" indexed="true" stored="true" multiValued="true"/>
 <field name="features" type="text_general" indexed="true" stored="true" multiValued="true"/>
 <field name="includes" type="text_general" indexed="true" stored="true" termVectors="true" termPositions="true" termOffsets="true" />

 <field name="weight" type="float" indexed="true" stored="true"/>
 <field name="price"  type="float" indexed="true" stored="true"/>
 <field name="popularity" type="int" indexed="true" stored="true" />
 <field name="inStock" type="boolean" indexed="true" stored="true" />

 <field name="store" type="location" indexed="true" stored="true"/>
```

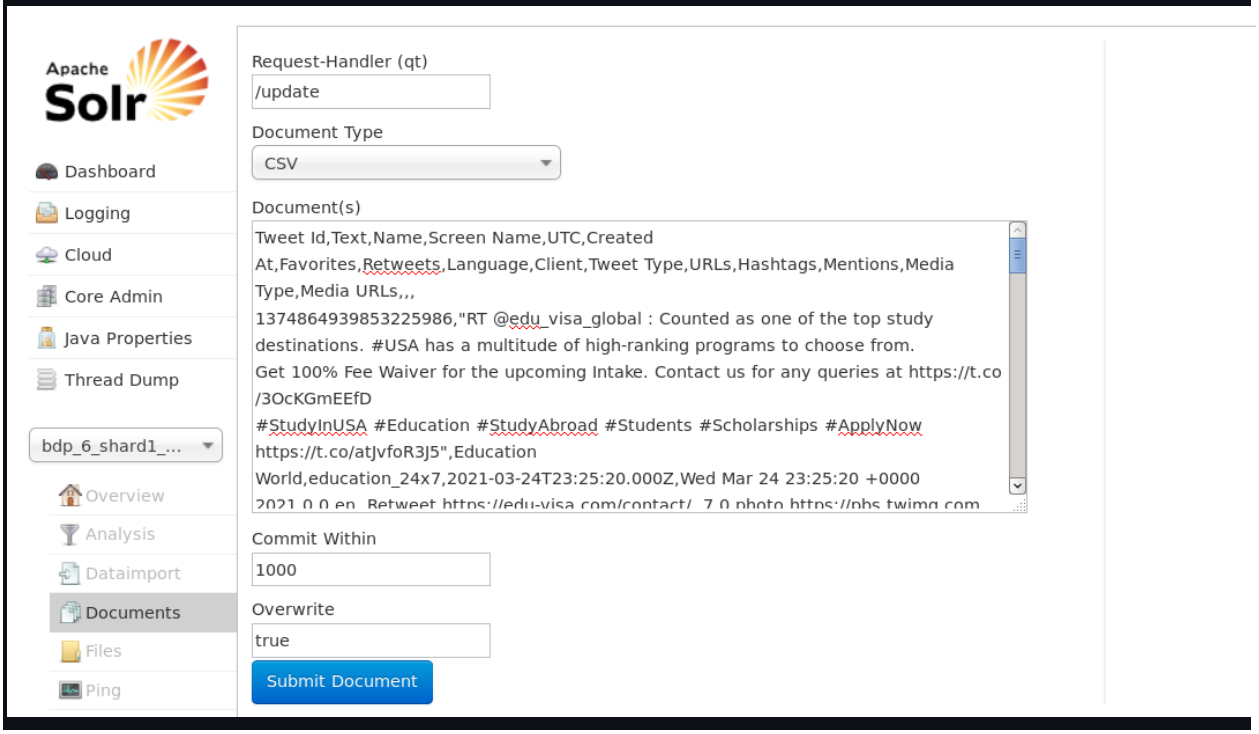
3. Open the solr browser in the web browser and select the create collection on the left side dropdown.



4. Set "Tweet Id" as the primary key.

```
<!-- Field to use to determine and enforce document uniqueness.
     Unless this field is marked with required="false", it will be a required field
  -->
  <uniqueKey> Tweet Id </uniqueKey>
```

5. Select the document type to csv. Then copy paste all the data inside the dataset into the documents field and submit the document



HIVE:

Created the database and database table in hive and loaded the data into hive table for AbolishICE dataset:

```
hive> create table AbolishICE (Tweet_Id string,Text string,Name string,Screen_Name string,UTC string,Created_At string,Favorites string,Retweets string,Language string,Client string,Tweet_Type string,URLs string,Hashtags string,Mentions
string,Media_Type string,Media_URLs string) row format delimited fields terminated by ',' stored as textfile;
OK
Time taken: 0.353 seconds
```

```
hive> load data local inpath '/home/cloudera/Downloads/bdphive/AbolishICE.csv' into table AbolishICE;
Loading data to table default.abolishice
Table default.abolishice stats: [numFiles=1, totalSize=645903]
OK
Time taken: 0.598 seconds
hive> select * from AbolishICE limit 2;
OK
Tweet Id        Text    Name    Screen Name     UTC     Created At      Favorites       Retweets        Language        Client  Tweet Type      URLs    Hashtags        Mentions        Media Type      Media URLs
1374885870491197441     "RT @abolishICE_nynj : After living in the US for nearly his whole life  Hieu and 32 other Vietnamese refugees were deported by @POTUS and @VP last week. Please support his re-entry by donating to this fundraiser
set up by @vietlead and @MekongNYC. If you don't have Venmo let us know. #AbolishICE https://t.co/T71JNy7d8O"    carb    carbnim 2021-03-25T00:48:30.000Z        Thu Mar 25 00:48:30 +0000 2021  0       0       en      "<a href=""https://mo
bile.twitter.com"" rel=""nofollow"">Twitter Web App</a>"         Retweet         1       4       photo
Time taken: 0.382 seconds, Fetched: 2 row(s)
hive>
```

Created the database and database table in hive and loaded the data into hive table for F1visa dataset: Visualized the output:

```
hive> create table F1visa (Tweet_Id string,Text string,Name string,Screen_Name string,UTC string,Created_At string,Favorites string,Retweets string,Language string,Client string,Tweet_Type string,URLs string,Hashtags string,Mentions stri
ng,Media_Type string,Media_URLss string) row format delimited fields terminated by ',' stored as textfile;
OK
Time taken: 0.071 seconds
hive> load data local inpath '/home/cloudera/Downloads/bdphive/F1visa.csv' into table F1visa;
Loading data to table default.f1visa
Table default.f1visa stats: [numFiles=1, totalSize=40034]
OK
Time taken: 0.257 seconds
hive> select * from F1visa limit 3;
OK
Tweet_Id        Text    Name    Screen_Name     UTC     Created_At      Favorites       Retweets        Language        Client  Tweet_Type      URLs    Hashtags        Mentions        Media_Type      Media_URLs
13748342506677116608    "RT @Maiquel59232858 : All categories of immigrant visas in Cuba #IR5Visa #F1Visa #F3Visa #F4Visa #f2AVisa #f2BVisa #IR1Visa #IR2Visa #IR3Visa #IR4Visa #CR1Visa #CR2Visa #K1Visa #K2Visa #K3Visa #K4Visa #DV2020 #DV
2021 &amp;amp; #CFRP should be processed in our country     until when?     NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL
 #ConsularServisesInCuba"       Neibis Aguilar  AguilarNeibis   2021-03-24T21:23:23.000Z        Wed Mar 24 21:23:23 +0000 2021  0       0       en      "<a href=""https://mobile.twitter.com"" rel=""nofollow"">Twitter Web App</a>"   Retwe
et      20      0       NULL
Time taken: 0.064 seconds, Fetched: 3 row(s)
hive>
```

Created the database and database table in hive and loaded the data into hive table for h1b dataset: Visualized the output:

```
hive> create table h1b (Tweet_Id string,Text string,Name string,Screen_Name string,UTC string,Created_At string,Favorites string,Retweets string,Language string,Client string,Tweet_Type string,URLs string,Hashtags string,Mentions string,
Media_Type string,Media_URLs string) row format delimited fields terminated by ',' stored as textfile;
OK
Time taken: 0.106 seconds
hive> load data local inpath '/home/cloudera/Downloads/bdphive/h1b.csv' into table h1b;
Loading data to table default.h1b
Table default.h1b stats: [numFiles=1, totalSize=591164]
OK
Time taken: 0.339 seconds
hive> select * from h1b limit 2;
OK
Tweet_Id        Text    Name    Screen_Name     UTC     Created_At      Favorites       Retweets        Language        Client  Tweet_Type      URLs    Hashtags        Mentions        Media_Type      Media_URLs
13748875418962165 76    "Mohammed Raheem's drive back home ?? home turned soar   please don't for brothers surgery https://t.co/Vhfd01vElB #Sanjose #muslim #muslimah #MuslimLivesMatter #veinbust #Indian #H1B #f1 visa"       Al Hindi ????
indimuslim101   2021-03-25T00:55:08.000Z        Thu Mar 25 00:55:08 +0000 2021  0       0       en      "<a href=""http://twitter.com/download/iphone"" rel=""nofollow"">Twitter for iPhone</a>"         Tweet   https://www.gofundme.com/f/pl
ease-help-for-the-surgery-of-mohammed-raheem?utm_source=twitter&utm_medium=social&utm_campaign=m_pd+share-sheet         8       0
Time taken: 0.107 seconds, Fetched: 2 row(s)
hive>
```


Created the database and database table in hive and loaded the data into hive table for intlstudents dataset: Visualized the output:

```
hive> create table intlstudents (Tweet_Id string,Text string,Name string,Screen_Name string,UTC string,Created_At string,Favorites string,Retweets string,Language string,Client string,Tweet_Type string,URLs string,Hashtags string,Mention
s string,Media_Type string,Media_URLs string) row format delimited fields terminated by ',' stored as textfile;
OK
Time taken: 0.093 seconds
hive> load data local inpath '/home/cloudera/Downloads/bdphive/intlstudents.csv' into table intlstudents;
Loading data to table default.intlstudents
Table default.intlstudents stats: [numFiles=1, totalSize=128138]
OK
Time taken: 0.211 seconds
hive> select * from intlstudents limit 3;
OK
Tweet_Id        Text    Name    Screen_Name     UTC     Created_At      Favorites       Retweets        Language        Client  Tweet_Type      URLs    Hashtags        Mentions        Media_Type      Media_URLs
13748891726752768 03    "RT @rus_students_cn : This China policy         silence towards #intlstudents and the way universities treat us make me feel like the whole COVID19 pandemic is our fault ?? #takeUsBackToChina"       Mr.Chips?????
I_mZainii       2021-03-25T01:01:37.000Z        Thu Mar 25 01:01:37 +0000 2021  0       0       en      "<a href=""http://twitter.com/download/android"" rel=""nofollow"">Twitter for Android</a>"      Retweet         2       0
13748628565926174 76    "RT @clayhensley : During a riveting #PIELive21 discussion on disruptive #edtech solutions driving #digital #recruitment         @cialfoplatform's @rohanpasari notes ""engagement (w/ #intlstudents &amp;amp; counselors
) is not an event       but a process."" Resonates w/ seasoned #intled folk. The more things change... https://t.co/dC4yVgsXTq"        Rohan Pasari    rohanpasari     2021-03-24T23:17:03.000Z        Wed Mar 24 23:17:03 +0000 2021  0   0
en      "<a href=""http://twitter.com/download/iphone"" rel=""nofollow"">Twitter for iPhone</a>"         Retweet         6       2
Time taken: 0.044 seconds, Fetched: 3 row(s)
hive>
```

Created the database and database table in hive and loaded the data into hive table for study in USA dataset: Visualized the output:

```
FAILED: Execution Error, return code 1 from org.apache.hadoop.hive.ql.exec.DDLTask. AlreadyExistsException(message:Table intlstudents already exists)
hive> create table studyinusa (Tweet_Id string,Text string,Name string,Screen_Name string,UTC string,Created_At string,Favorites string,Retweets string,Language string,Client string,Tweet_Type string,URLs string,Hashtags string,Mentions string,Media_Type string,Media_URLs string) row format delimited fields terminated by ',' stored as textfile;
OK
Time taken: 0.101 seconds
hive> load data local inpath '/home/cloudera/Downloads/bdphive/studyinusa.csv' into table studyinusa;
Loading data to table default.studyinusa
Table default.studyinusa stats: [numFiles=1, totalSize=123163]
OK
Time taken: 0.283 seconds
hive> select * from intlstudents limit 10;
OK
Tweet Id	Text	Name	Screen Name	UTC	Created At	Favorites	Retweets	Language	Client	Tweet Type	URLs	Hashtags	Mentions	Media Type	Media URLs
1374889172675276803	"RT @rus_students_cn : This China policy	silence towards #intlstudents and the way universities treat us make me feel like the whole COVID19 pandemic is our fault ?? #takeUsBackToChina"	Mr.Chips?????I_mZainii	2021-03-25T01:01:37.000Z	Thu Mar 25 01:01:37 +0000 2021	0	0	en	"<a href=""http://twitter.com/download/android"" rel=""nofollow"">Twitter for Android</a>"	Retweet	2	0
1374862856592617476	"RT @clayhensley : During a riveting #PIELive21 discussion on disruptive #edtech solutions driving #digital #recruitment	@cialfoplatform's @rohanpasari notes ""engagement (w/ #intlstudents &amp; counselors) is not an event	but a process."" Resonates w/ seasoned #intled folk. The more things change... https://t.co/dC4yVgsXTq"	Rohan Pasari	rohanpasari	2021-03-24T23:17:03.000Z	Wed Mar 24 23:17:03 +0000 2021	0	0	en	"<a href=""http://twitter.com/download/iphone"" rel=""nofollow"">Twitter for iPhone</a>"	Retweet	6	2
1374824238192979971	@UCAdvocacy @EdTrust Thank you so much for participating in our chat and for advocating for #immigrant &amp; #intlstudents! #HigherEdImmChat	Presidents' Alliance on Higher Ed & Immigration PresImmAlliance 2021-03-24T20:43:36.000Z	Wed Mar 24 20:43:36 +0000 2021	0	0	en	"<a href=""https://about.twitter.com/products/tweetdeck"" rel=""nofollow"">TweetDeck</a>"	Reply	3	2
1374822478065901573	"RT @clayhensley : During a riveting #PIELive21 discussion on disruptive #edtech solutions driving #digital #recruitment	@cialfoplatform's @rohanpasari notes ""engagement (w/ #intlstudents &amp; counselors) is not an event	but a process."" Resonates w/ seasoned #intled folk. The more things change... https://t.co/dC4yVgsXTq"	Cialfo	cialfoplatform	2021-03-24T20:36:36.000Z	Wed Mar 24 20:36:36 +0000 2021	0	0	en	"<a href=""https://mobile.twitter.com"" rel=""nofollow"">Twitter Web App</a>"	Retweet	6	2
1374805876956086273	"RT @clayhensley : During a riveting #PIELive21 discussion on disruptive #edtech solutions driving #digital #recruitment	@cialfoplatform's @rohanpasari notes ""engagement (w/ #intlstudents &amp; counselors) is not an event	but a process."" Resonates w/ seasoned #intled folk. The more things change... https://t.co/dC4yVgsXTq"	The PIE News	ThePIENews	2021-03-24T19:30:38.000Z	Wed Mar 24 19:30:38 +0000 2021	0	0	en	"<a href=""https://about.twitter.com/products/tweetdeck"" rel=""nofollow"">TweetDeck</a>"	Retweet	6	2
1374799943836299278	"RT @clayhensley : During a riveting #PIELive21 discussion on disruptive #edtech solutions driving #digital #recruitment	@cialfoplatform's @rohanpasari notes ""engagement (w/ #intlstudents &amp; counselors) is not an event	but a process."" Resonates w/ seasoned #intled folk. The more things change... https://t.co/dC4yVgsXTq"	ThePIELive	ThePIELive	2021-03-24T19:07:03.000Z	Wed Mar 24 19:07:03 +0000 2021	0	0	en	"<a href=""http://twitter.com/download/iphone"" rel=""nofollow"">Twitter for iPhone</a>"	Retweet	6	2
1374796754181357576	"During a riveting #PIELive21 discussion on disruptive #edtech solutions driving #digital #recruitment	@cialfoplatform's @rohanpasari notes ""engagement (w/ #intlstudents &amp; counselors) is not an event	but a process."" Resonates w/ seasoned #intled folk. The more things change... https://t.co/dC4yVgsXTq"	Clay Hensley	clayhensley	2021-03-24T18:54:23.000Z	Wed Mar 24 18:54:23 +0000 2021	5	4	en	"<a href=""https://mobile.twitter.com"" rel=""nofollow"">Twitter Web App</a>"	Tweet	6	2
1374754630392020994	10 tips for succeeding at an American university: https://t.co/6tVyU5Ge7W #intlstudents #college https://t.co/7t10kPRzO1	University Language	CampusCommons	2021-03-24T16:07:00.000Z	Wed Mar 24 16:07:00 +0000 2021	0	0	en	"<a href=""https://about.twitter.com/products/tweetdeck"" rel=""nofollow"">TweetDeck</a>"	Tweet	https://bit.ly/2L7djjf	2	0	photo	https://pbs.twimg.com/media/Ew2i7gHWgAA87nb.jpg
```

Imported file into hdfs

```
[cloudera@quickstart Team_6]$ hadoop fs -put /home/cloudera/Downloads/tweets_data_03_21.csv /Team_6
```

```
[cloudera@quickstart ~]$ mkdir Team_6
[cloudera@quickstart ~]$ cd Team_6
```

Created the database and database table in hive and loaded the data into hive table for General dataset:

```
[cloudera@quickstart Team_6]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.properties
```

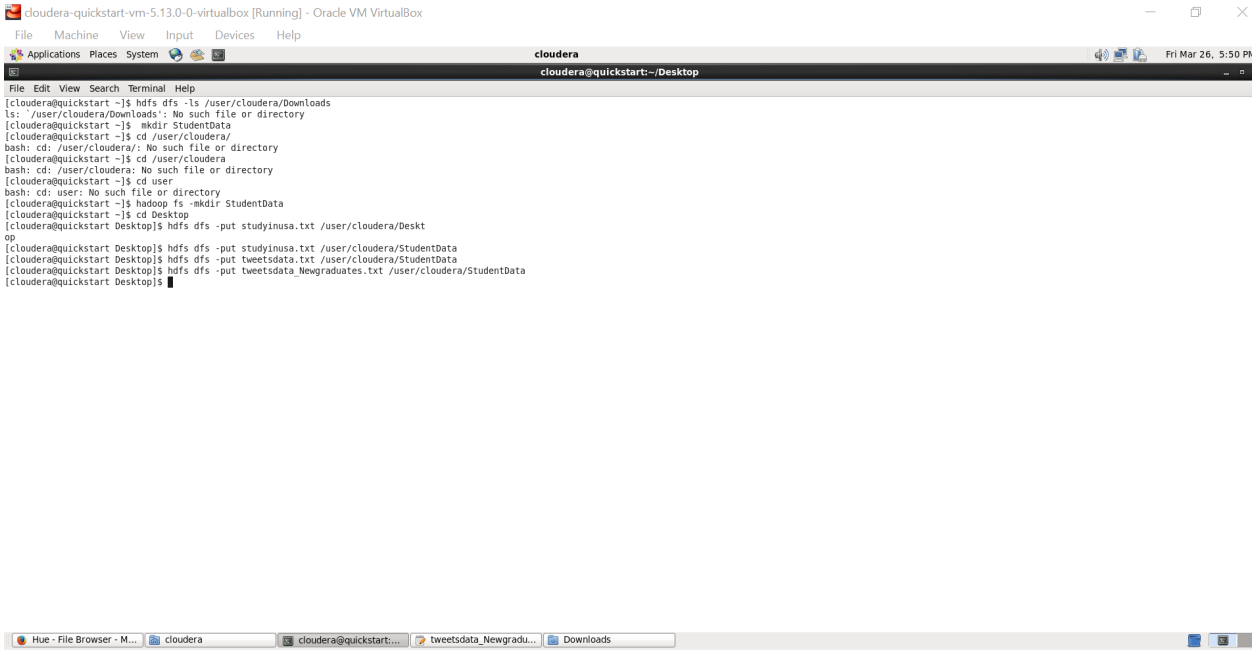
```
hive> CREATE TABLE tweets_data(F1 string, created_at string,id string,id_str string,text string,source string,truncated string,in_reply_to_status_id string,in_reply_to_status_id_str string,in_reply_to_user_id string,in_reply_to_user_id_str string,in_reply_to_screen_name string,user string,geo string,coordinates string,place string,contributors string,retweeted_status string,is_quote_status string,quote_count string,reply_count string,retweet_count string, favorite_count string,entities string,favorited string,retweeted string,filter_level string,lang string,timestamp_ms string,display_text_range string,extended_tweet string,quoted_status_id string,quoted_status_id_str string,quoted_status string,quoted_status_permalink string,extended_entities string,possibly_sensitive string,withheld_in_countries string) row format delimited fields terminated by '\t' stored as textfile;
OK
Time taken: 6.212 seconds
hive> load data local inpath '/home/cloudera/Downloads/tweets_data_03_21.csv' into table tweets_data;
Loading data to table default.tweets_data
Table default.tweets_data stats: [numFiles=1, totalSize=33422781]
OK
Time taken: 3.601 seconds
hive> select * from tweets_data limit 10;
OK
```

Visualized the Output:

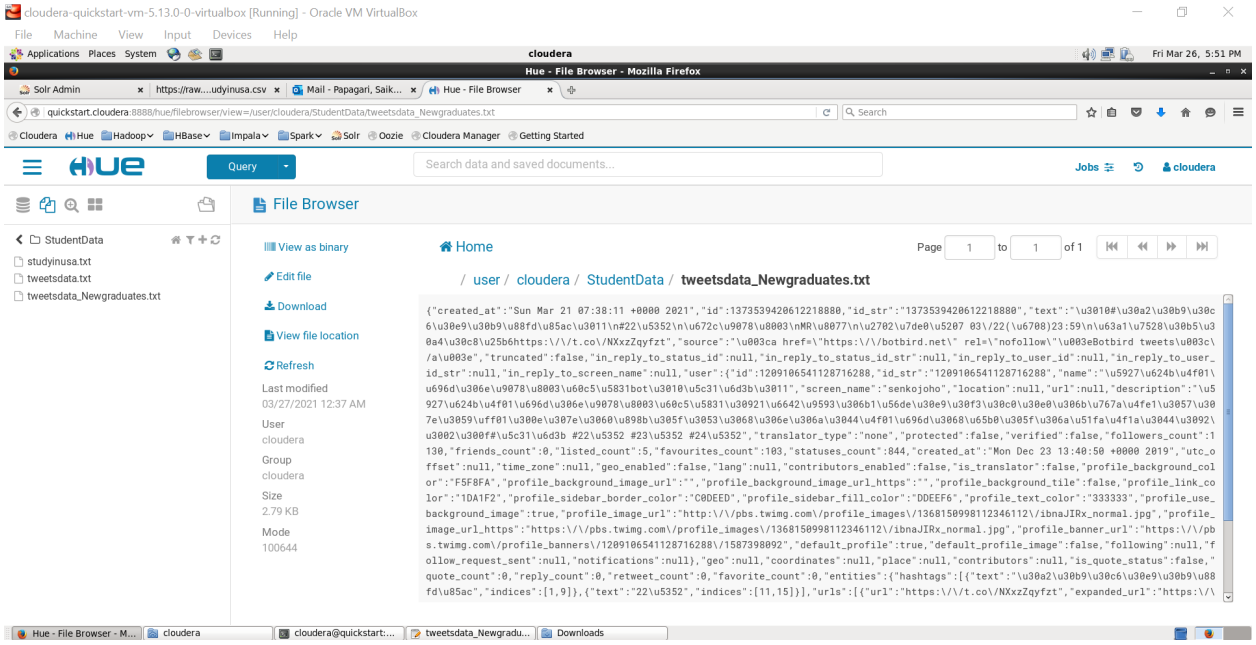

```
Remember when people were thinking that a student was gonna drop out from the war and make room for S…","<a href=""http://twitter.com/download/iphone"" rel=""nofollow"">Twitter for iPhone</a>",FALSE,,,,,,"{'id': 1215252128790450176, 'id_str': '1215252128790450176', 'name': 'Lau ▯eeds Ajax', 'screen_name': 'TOD0ROVS', 'location': 'Tahera's♥ | 22 | BLM', 'url': None, 'description': 'She/Her • I mostly rt fanarts of stuff I love and ramble about them – not spoiler free – Deku stan', 'translator_type': 'none', 'protected': False, 'verified': False, 'followers_count': 243, 'friends_count': 160, 'listed_count': 17, 'favourites_count': 40041, 'statuses_count': 30088, 'created_at': 'Thu Jan 09 12:41:15 +0000 2020', 'utc_offset': None, 'time_zone': None, 'geo_enabled': False, 'lang': None, 'contributors_enabled': False, 'is_translator': False, 'profile_background_color': 'F5F8FA', 'profile_background_image_url': '', 'profile_background_image_url_https': '', 'profile_background_tile': False, 'profile_link_color': '1DA1F2', 'profile_sidebar_border_color': 'C0DEED', 'profile_sidebar_fill_color': 'DDEEF6', 'profile_text_color': '333333', 'profile_use_background_image': True, 'profile_image_url': 'http://pbs.twimg.com/profile_images/1371146374704680960/UKbLPL3__normal.jpg', 'profile_image_url_https': 'https://pbs.twimg.com/profile_images/1371146374704680960/UKbLPL3__normal.jpg', 'profile_banner_url': 'https://pbs.twimg.com/profile_banners/1215252128790450176/1615741672', 'default_profile': True, 'default_profile_image': False, 'following': None, 'follow_request_sent': None, 'notifications': None}",,,,,"{'created_at': 'Thu Mar 18 03:32:23 +0000 2021', 'id': 1372390398316843010, 'id_str': '1372390398316843010', 'text': '#MHA306\n#MHASpoilers\n\nRemember when people were thinking that a student was gonna drop out from the war and make ro… https://t.co/7laxi6znvG', 'source': '<a href=""https://mobile.twitter.com"" rel=""nofollow"">Twitter Web App</a>', 'truncated': True, 'in_reply_to_status_id': None, 'in_reply_to_status_id_str': None, 'in_reply_to_user_id': None, 'in_reply_to_user_id_str': None, 'in_reply_to_screen_name': None, 'user': {'id': 926200464453939200, 'id_str': '926200464453939200', 'name': 'AsarathaHS', 'screen_name': 'AsarathaHS', 'location': 'he/him', 'url': 'https://curiouscat.me/AsarathaHS', 'description': '99% of my tweets are about MHA.', 'translator_type': 'none', 'protected': False, 'verified': False, 'followers_count': 585, 'friends_count': 70, 'listed_count': 2, 'favourites_count': 8172, 'statuses_count': 3596, 'created_at': 'Thu Nov 02 21:32:53 +0000 2017', 'utc_offset': None, 'time_zone': None, 'geo_enabled': False, 'lang': None, 'contributors_enabled': False, 'is_translator': False, 'profile_background_color': 'F5F8FA', 'profile_background_image_url': '', 'profile_background_image_url_https': '', 'profile_background_tile': False, 'profile_link_color': '1DA1F2', 'profile_sidebar_border_color': 'C0DEED', 'profile_sidebar_fill_color': 'DDEEF6', 'profile_text_color': '333333', 'profile_use_background_image': True, 'profile_image_url': 'http://pbs.twimg.com/profile_images/1372791666000556042/md5wt4sD_normal.jpg', 'profile_image_url_https': 'https://pbs.twimg.com/profile_images/1372791666000556042/md5wt4sD_normal.jpg', 'profile_banner_url': 'https://pbs.twimg.com/profile_banners/926200464453939200/1615161438', 'default_profile': True, 'default_profile_image': False, 'following': None, 'follow_request_sent': None, 'notifications': None}, 'geo': None, 'coordinates': None, 'place': None, 'contributors': None, 'is_quote_status': False, 'extended_tweet': {'full_text': ""#MHA306\n#MHASpoilers\n\nRemember when people were thinking that a student was gonna drop out from the war and make room for Shinso?\n\ndamn y'all were right it just wasnt who we expected LMFAOOO"", 'display_text_range': [0, 190], 'entities': {'hashtags': [{'text': 'MHA306', 'indices': [0, 7]}, {'text': 'MHASpoilers', 'indices': [8, 20]}], 'urls': [], 'user_mentions': [], 'symbols': []}}, 'quote_count': 171, 'reply_count': 44, 'retweet_count': 640, 'favorite_count': 4754, 'entities': {'hashtags': [{'text': 'MHA306', 'indices': [0, 7]}, {'text': 'MHASpoilers', 'indices': [8, 20]}], 'urls': [{'url': 'https://t.co/7laxi6znvG', 'expanded_url': 'https://twitter.com/i/web/status/1372390398316843010', 'display_url': 'twitter.com/i/web/status/1…', 'indices': [117, 140]}], 'user_mentions': [], 'symbols': []}, 'favorited': False, 'retweeted': False, 'filter_level': 'low', 'lang': 'en'}",FALSE,0,0,0,0,"{'hashtags': [{'text': 'MHA306', 'indices': [16, 23]}, {'text': 'MHASpoilers', 'indices': [24, 36]}], 'urls': [], 'user_mentions': [{'screen_name': 'AsarathaHS', 'name': 'AsarathaHS', 'id': 926200464453939200, 'id_str': '926200464453939200', 'indices': [3, 14]}], 'symbols': []}",FALSE,FALSE,low,en,1.61634E+12,,,,,,,,	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULLNULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
1,Sun Mar 21 16:20:26 +0000 2021,1.37367E+18,1.37367E+18,"@cephalus en I stand against all forms of racism, ableism, homophobia, transphobia, Islamophobia, hate; and pedophi… https://t.co/htKEN7BOsV","<a href=""http://twitter.com/downloa
```

Loading the studentdata into the hadoop

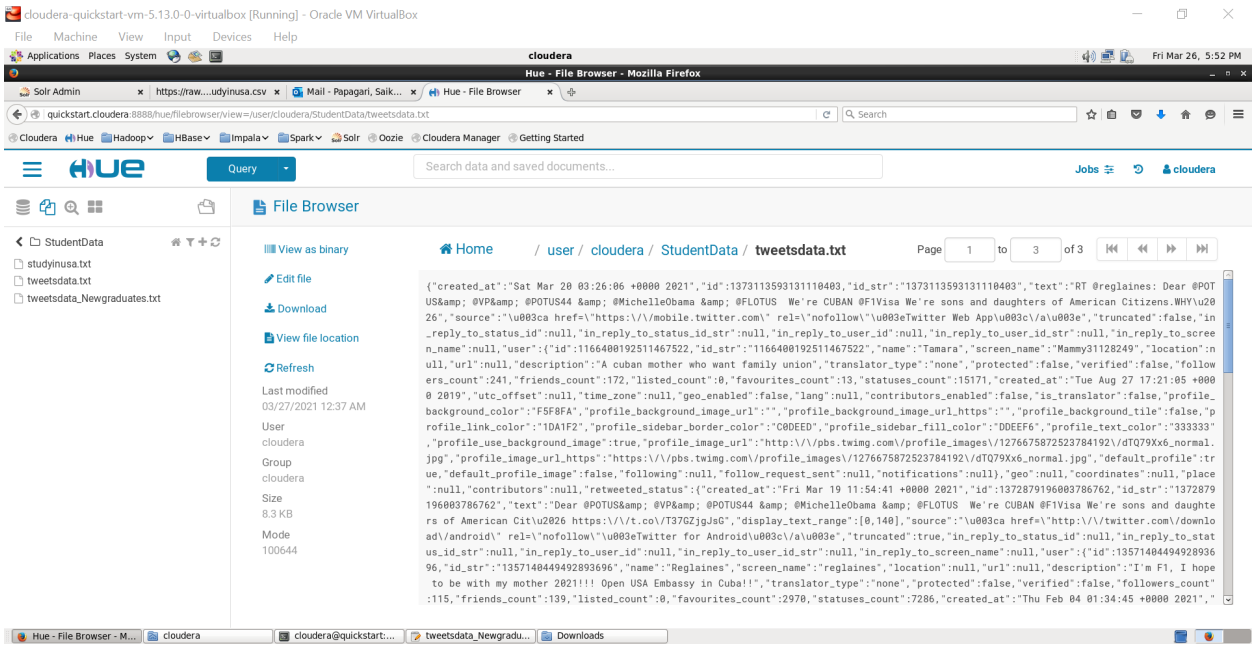
1. First imported the .txt files into cloudera and then created a directory with name “StudentData” and moved the imported files into the directory



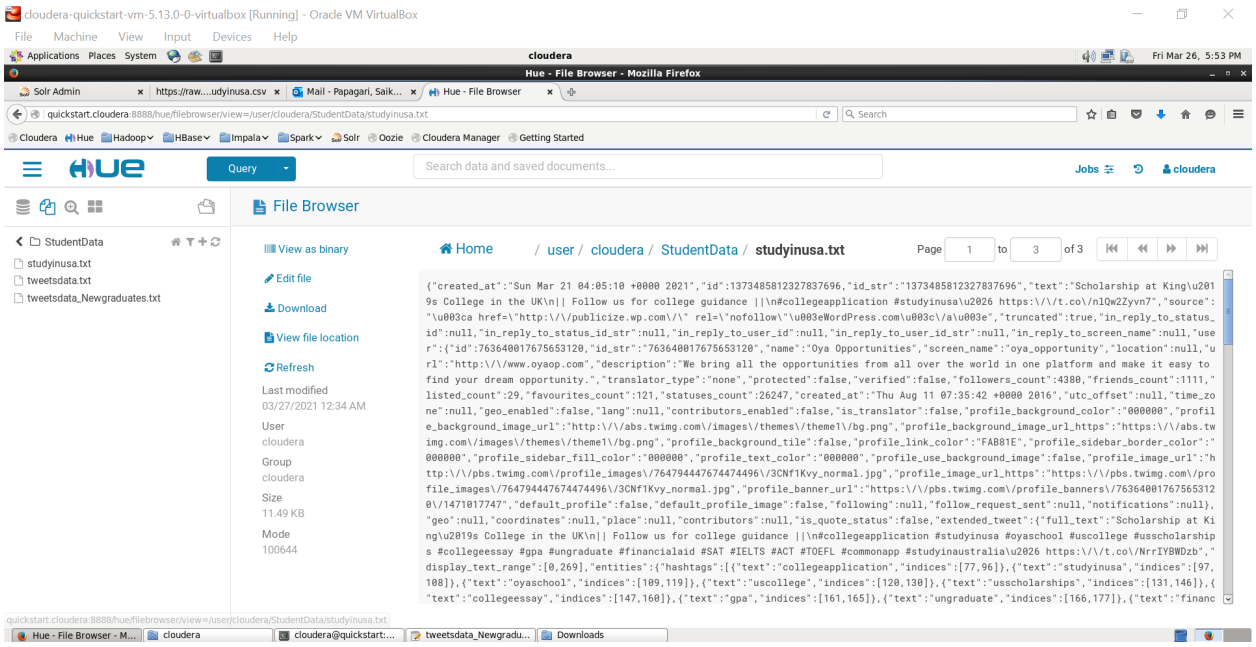
2. The output data which is loaded into Hue for tweetsdata_Newgraduates.txt



3. Similarly, for the tweetsdata.txt



4. The output for file study.txt is as follows:



Working with Spark using Scala:

The number of Tweets from various countries (Query)

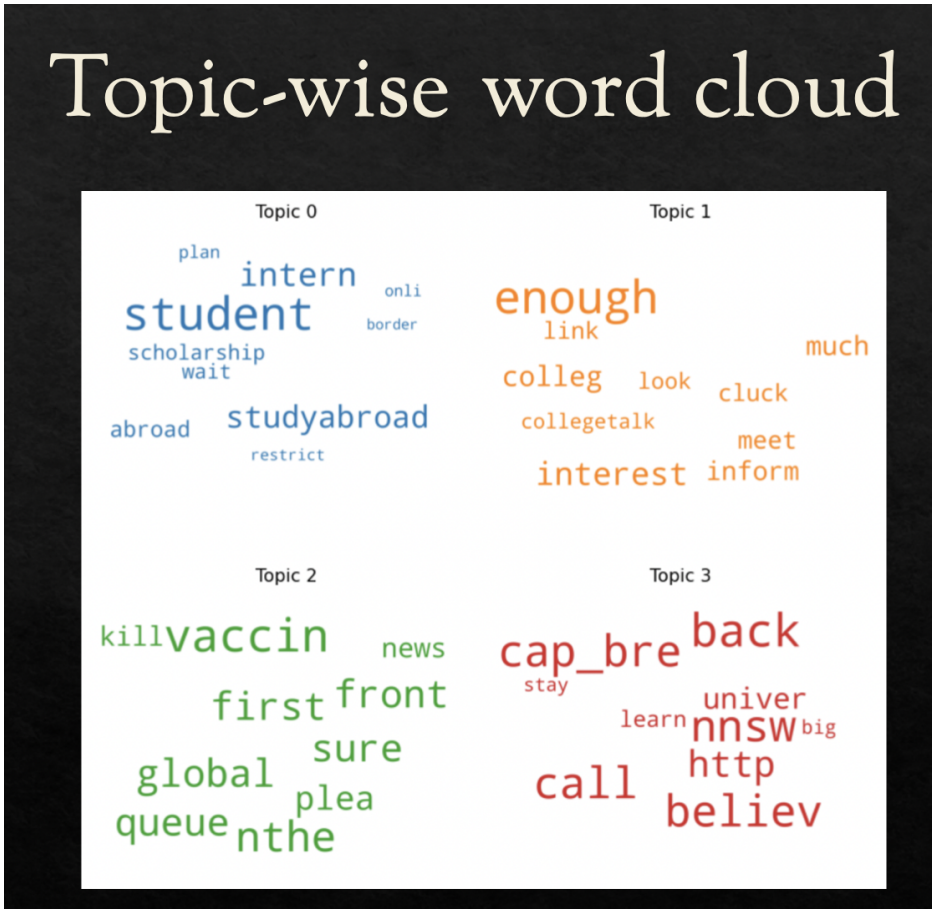
```
scala> val q1 = sqlContext.sql("SELECT place.country,count(*) AS count FROM tweetDatatable GROUP BY place.country ORDER BY count DESC limit 10");
q1: org.apache.spark.sql.DataFrame = [country: string, count: bigint]
```

To retrieve the various languages in which tweets about web series are created.
The number of tweets about the web series in various languages was tallied.

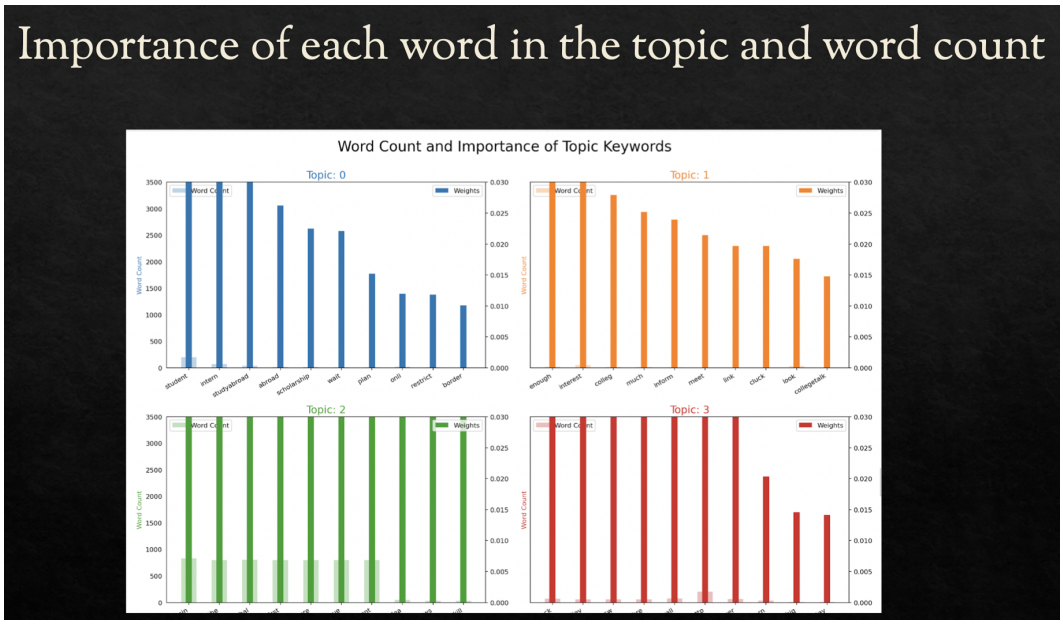
```
scala> val q3=sqlContext.sql("select count(*) as count,lang as language from tweetDatatable where lang is not null group by lang");
q3: org.apache.spark.sql.DataFrame = [count: bigint, language: string]
```

- **Preliminary Results (Visualization of Results)**

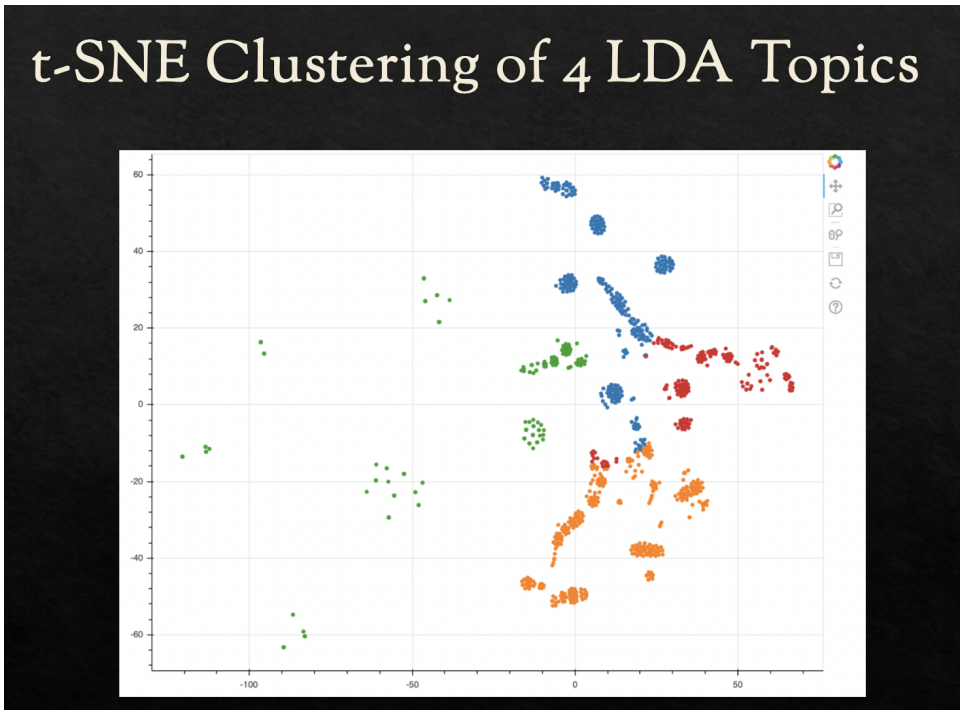
a.



b.

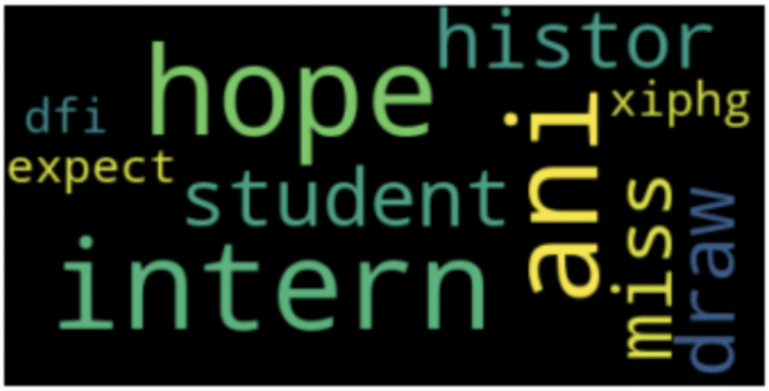


c.



d. Creating word cloud using pySpark

```
1 text = tidy_tweets[0]
2
3 # Create and generate a word cloud image:
4 wordcloud = WordCloud().generate(text)
5
6 # Display the generated image:
7 plt.imshow(wordcloud, interpolation='bilinear')
8 plt.axis("off")
9 plt.show()
```



- **Project Management**
  - **Implementation status report**
    - **Work completed:**
      - **Description :-**

        We have completed the extraction of dataset from twitter using twitter API, completed the tweets preprocessing part, sentiment analysis of those tweets. We have also started working on hadoop and pySpark and completed loading of the dataset in hive and creating tables, and creating pySpark and working on the tweets (text processing) aspect in spark to perform different analysing techniques. Loaded dataset into solr as well to work on the analysing aspect.

- **Responsibility (Task, Person)**

Amar - Data Extraction using tweepy, Creating collection and Instance in Solr, Editing schema.xml in accordance with dataset, Loading the dataset (.csv file) into Solr.
Sai - Importing the text files into Hadoop-Hue and exploring on working with Scala queries.
Mounika - Data extraction , loading table in Hive and documentation.
Tanvi - Twitter data collection from twitter API, preprocessing, (removing numbers, punctuations, words less than length 2, stemming, removing stopwords), using Gensim, Build the bigram and trigram models, LDA model for topic modelling, Word Cloud of Top N words in each topic, plot Word Count and Weights of Topic Keywords, t-SNE clustering of 4LDA topics, using pySpark for word cloud and text analysis/visualization.

- **Contributions (members/percentage)**

Amar - 20%
Sai - 20%
Mounika - 30%
Tanvi - 30%

- **Work to be completed**
- **Description**

  We have to perform ETL process using spark's batch processing, and analysis using different pySpark codes to analyse the text data and different aspects of it, and also visualize them using Tableau. We have to
- **Responsibility (Task,**

Person)

Tanvi - performing more analysis and text visualization and sentiment analysis in pySpark on the other 2 events well (AbolishICE and h1b), visualizing the important data using Tableau

Mounika - Implementing SQL and HIVE queries and using sqoop to transfer data from SQL to HDFS

Amar - Performing various queries in Solr, sentiment analysis using pySpark, visualisation using seaborn.

Sai - Data analysis using Scala for different datasets.

- **Issues/Concerns**

The big data tools and the tasks performed by increment 2 might alter a bit later on in the final phase of the project based on the time and knowledge constraints and what aspects will be important to showcase, if the project demands only hadoop, spark and solr, or addition of another big data tool for better data analysis, this can be only determined after moving forward with current plan and seeing how well it goes or if there is a need to change the plan.

6. Once you are done with the above report, for story telling part of the project, address all the questions given in the following link for Increment 2 storytelling under Assignment 02section:

1. **Who**

The international students who are studying
in the United States Of America.(or overseas).

The dataset is extracted using different hashtags to get data related to 3 different scenarios as follows :-

1. Generic International Student Data **(#F1visa #intlstudents #internationalstudents #studyinUSA)**
2. Immigration rules change for F1 Visa during COVID in 2020 **(#AbolishICE)**
3. Jobs for international students during covid (H1B sponsorship) **(#H1B, #h1bjobs)**

2. What

Yes,the data set records the targeted events, activities, behaviors, etc. in Assignment 1. This is fundamentally about the variables. It records the username, the location, the tweets which tell us about what the users really think about the specific event that happened.

3. When

The events take place on how people reacted to the challenges faced by an International students during COVID like immigration rules for F1visa, jobs for international students during covid(H1B sponsorship).

4. Where means two things:

These challenges can be faced by any international students in a foriegn land, where they are a part of the immigrant community and do not have access to the perks that The citizens of that country get, which makes life a bit harder for even the very basic requirements of living.

5. Why means the possible causes and/origin of the problem.

Due to the increase in the intake of international students over a period of time.

6. How:

It happened during COVID, as a result there was a massive Unavailability of jobs(new graduates), many have lost their jobs due to pandemic, immigration rules change for F1 Visa during COVID.

References:

1. https://docs.cloudera.com/documentation/enterprise/6/6.3/topics/cm_mc_solr_service.html
2. https://docs.cloudera.com/runtime/7.2.7/search-managing/topics/search-updating-the-schema-in-a-solr-collection.html
3. https://github.com/dsuarez993/bigdata-realtime-twitter-analysis
4. https://github.com/6vedant/TwitterAnalyticsHadoop
5. https://www.toptal.com/apache/apache-spark-streaming-twitter
6. **http://spark.apache.org/**
7. https://www.toptal.com/spark/introduction-to-apache-spark
8. **https://hadoop.apache.org/**
9. https://www.infoworld.com/article/3236869/what-is-apache-spark-the-big-data-platform-that-crushed-hadoop.html

10.https://towardsdatascience.com/extracting-data-from-twitter-using-python-5ab67bff553a