PSOSM MID SEM

ARBAAZ KHAN

2018023

SECTION 1

Q1

a)

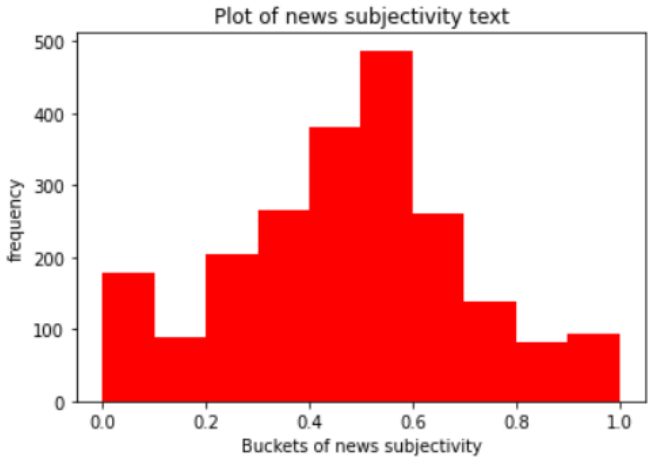
*This observation is for below two graphs.

Observation: Subjectivity of both humor and news are concentrated around 0.5 and the mean and daviation is also similar.

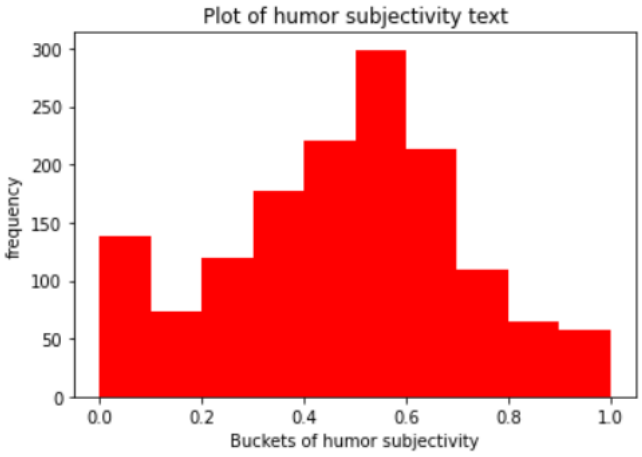
We can see in the below two graphs that the subjectivity of both graphs is positive because it depend on emotions, personal opinion and judgement.

Because both below graphs are distributed so these have more variance as polirity one.

```
Text(0.5, 1.0, 'Plot of news subjectivity text')
```




```
Text(0.5, 1.0, 'Plot of humor subjectivity text')
```

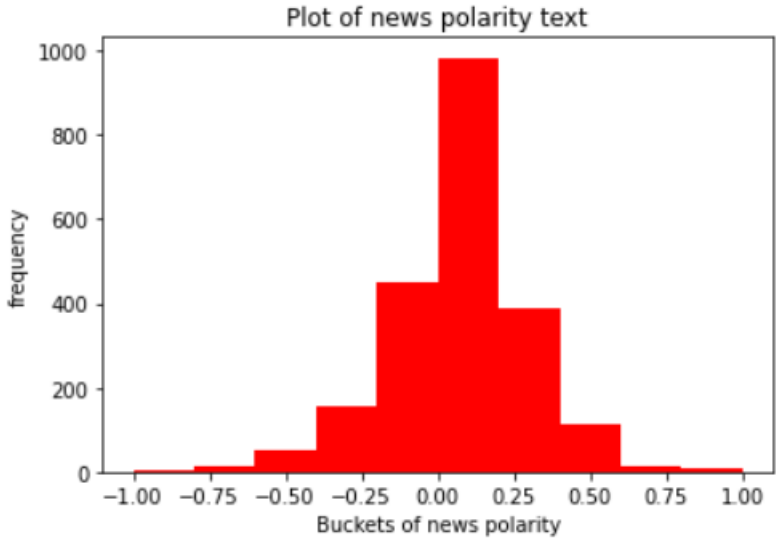


*This observation is for below 2 graphs

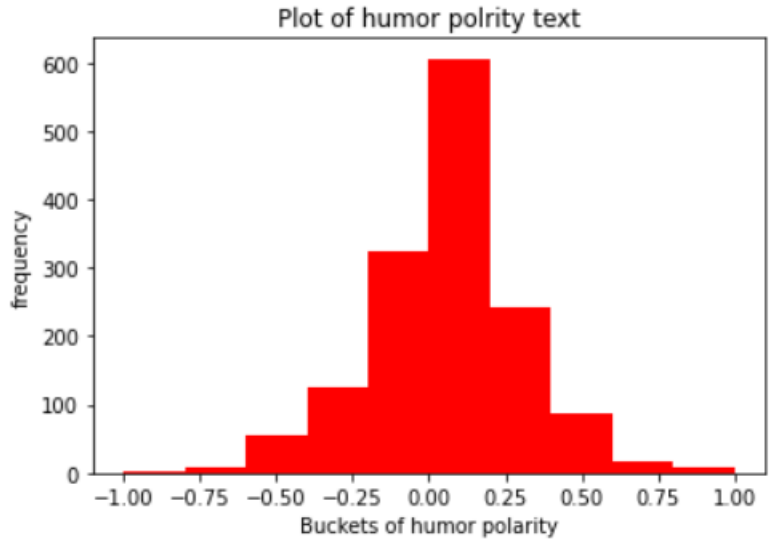
Observation : we can see in the below plots that the peek is at 0.05 or between 0-0.25 and mean and standard daviation for both humor and text are similar.

Some text polarity is negetive it simply mean there are both type of text present humor and news both. Because both polarity graphs are less distributed that's why standard daviation of both graphs is less than the subjectivity one.

```
Text(0.5, 1.0, 'Plot of news polarity text')
```




```
Text(0.5, 1.0, 'Plot of humor polrity text')
```



Statistics

```
[10] hp1 = np.array(hp)
     print("Mean of polarity humor: ",np.mean(hp1))
     print("Standard Deviation of polarity humor: ",np.std(hp1));
```

```
Mean of polarity humor:  0.05158728575619279
Standard Deviation of polarity humor:  0.24866391452327927
```

```
[11] hs1 = np.array(hs)
     print("Mean of subjectivity humor: ",np.mean(hs1))
     print("Standard Deviation of subjectivity humor: ",np.std(hs1));
```

```
Mean of subjectivity humor:  0.47140332998753287
Standard Deviation of subjectivity humor:  0.24269324821277793
```

```
[12] np11 = np.array(np1)
     print("Mean of polarity news: ",np.mean(np11))
     print("Standard Deviation of polarity news: ",np.std(np11));
```

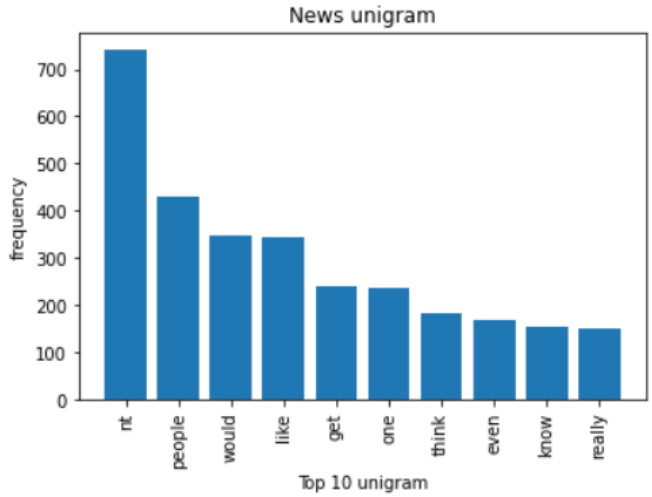
```
Mean of polarity news:  0.064048534801197
Standard Deviation of polarity news:  0.23041556585535447
```

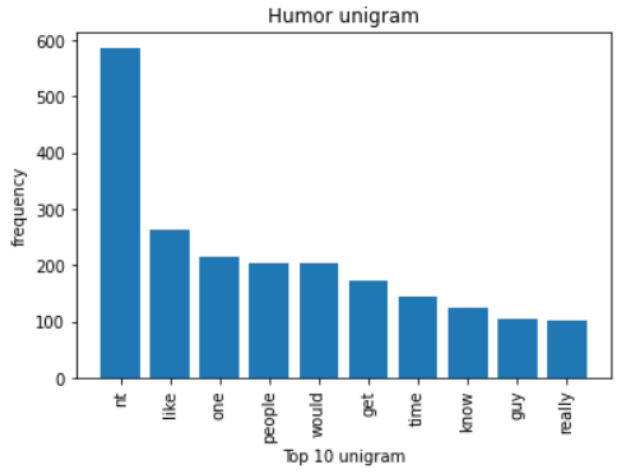
```
[13] ns1 = np.array(ns)
     print("Mean of subjectivity news: ",np.mean(ns1))
     print("Standard Deviation of subjectivity news: ",np.std(ns1));
```

```
Mean of subjectivity humor:  0.4704749753548112
Standard Deviation of subjectivity humor:  0.23274569397299044
```

b)

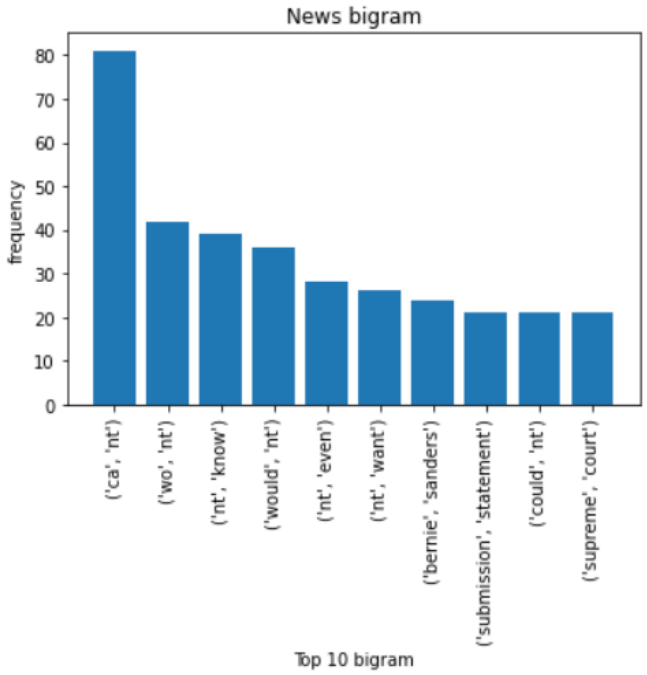
Inferences for unigram → "nt" which means "not" is most occurring word in both news and humor graphs. We can see in both below two graphs that almost all top 10 words are same. These unigrams are occurring in both negative and positive sentiments. You can easily see by graphs that unigrams words are most use in news than humor. Because of nearly same graph of unigram we cannot classify by using single attribute.

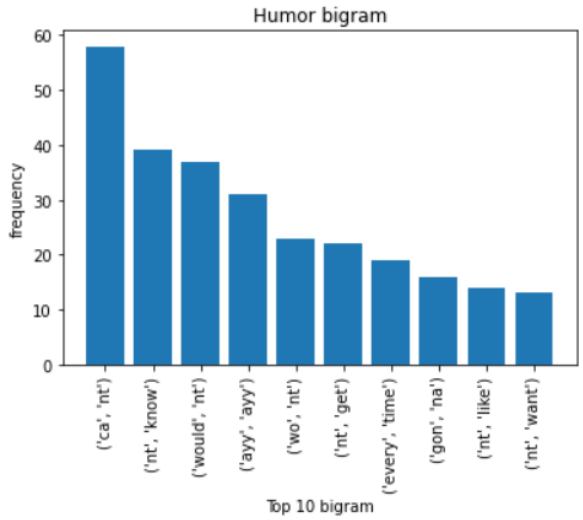




Inferances for bigram → “ca,nt” means “cannot” is the most occuring word in both bigrams word and words like court, supreme, submission,could etc. are very factual.

The bigram used in both humor and news or positive and negetive sentiments and these bigrams are mostly used in news because mean of news bigrams is greater than humor bigrams.





Q2)

a)

Preprocessing Steps:

- **Label Encoding**: It convert the label into integer because ML algorithms only process the integer data.
- **Shuffling of data**: By shuffling the data we will get uniformly variation of labels in our train and test data.
- **Bag of Words**: By ML algorithm we can convert a text into vector form which is necessary for further processing of data because ML algo only work on integer values.
- **Removing STOPWORDS**: It is important for us because by these we reduce unnecessary attributes.
- **Word Lemmatization**: it is possible that there are some words in different form like play, playing which only increase our attribute so by word Lemmatization we take root of these types of words.
- **Counting Frequency**: by using dictionary in python we count the frequency of every root of words.
- **Creating Xtrain, Ytrian, Xtest, Ytest**: By using ML library I split the data into 70 to 30 ratio means 70% data for training set and 30% for test set.

So , I use vectorization to make bag of word which convert our text into vector form and we use those vector to do further processing for our ML model which take xtest in vector form and predict ytest.

b)

I used multinomial naïve bayes model for prediction.

https://drive.google.com/file/d/1pax3Si7MrrCv4wxVNdB3oQyuffpHDQxM/view?usp=sharing

c)

```
  precision    recall  f1-score   support

       humor       0.83      0.71      0.77       467
        news       0.81      0.90      0.85       629

    accuracy                           0.82      1096
   macro avg       0.82      0.80      0.81      1096
weighted avg       0.82      0.82      0.81      1096

Confusion Matrix:
[[331 136]
 [ 66 563]]
```

d)

below are the starting 5 performed well and failed words:

5 examples where model performed well

i) generally speaking i m an asshole that being said i am profoundly nice to customer service reps whenever i have to call a call center for anything my primary objective is to get them to have an actual conversation with me the reason i called is secondary it ll be handled anyway: humor

Reason: words like asshole are comes in negative sentiments that's why it is humor.

ii) my husband would have done the exact same thing you keep pulling shit like that your wife wo nt try to do anything sexy anymore: humor

Reason: words like sexy ,shit comes in negative sentiments so it is humor

iii) i will go out of my way to transfer you to make a second 40 minute car journey to come back for a refund: humor

Reason: because it contains some words which make it personal decision other than a news so it is humor.

iv) maybe the bird is a really tough ski coach you call that skiing smack smack beakpoke get up get up and work those poles i do nt actually ski so i m having a hard time coming up with proper sounding ski tips so i ll just keep trying swish those goddamn hips like you mean it smash that snow i think i m not very good at this: humor

Reason: words like smash hip good etc are comes in negative sentiments so it is humor.

v) this war on cash is such bullshit people carried cash until 12 years ago and it was perfectly fine now they are trying to restrict flying with cash purchasing with cash seizing cash the only problem is that the money is due on one day they need to have the tax office take payment on any day of the week so it s stacked: news

Reason: words like restricting, seizing, problem, tax office and car purchasing make it news text.

5 examples where model failed

i) most importantly inventory does nt track preferences almost as important stocktaking only tells you about the past whereas prices tell you about the present and people s beliefs about the future: news

Reason : some words like stock taking and inventory which are technical word makes it news.

ii) a railroad engineer must be sure not to lose his train of thought or he might go down the wrong track: humor

Reason: words like might , and , must are less technical words which make it humor

iii) he is not following an antiquated convention that no one gives two shits about that guy is no gentleman but an asshole: humor

Reason: words like asshole, smash, hips which are very less technical make it humor although it have conventions like high technical words.

iv) i had cox for 3 years service was good and over that 3 years doubled my download speeds twice no extra cost upload was still garbage but hey 10015 internet for 65 a month was still pretty killer i have 7575 fios now: news

Reason: words like download, speed, upload and internet etc. make it news.

v) damnit i misunderstood the joke at first and tried counting the words he could say per year at first i was like pft no punchline here wordcount is different then i backread fail: humor

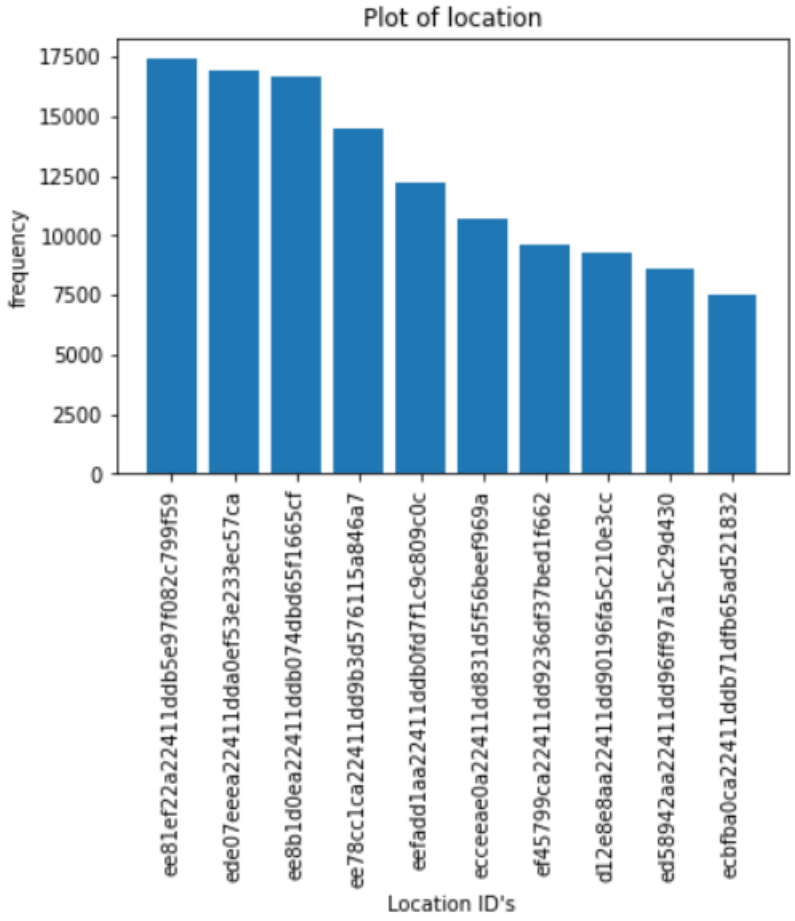
Reason: words like dammit, joked etc are less technical which make it humor although is have also high technical words.

SECTION 2

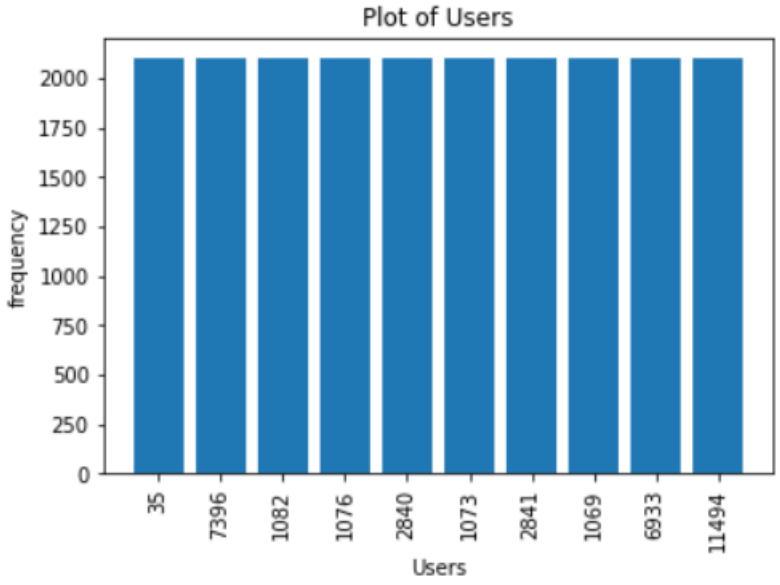
Q1)

a)

Inferences:- we can see below in the bar graph there are some location where are most numbers of check-ins its means some places are famous than others those places may be famous for tourism or may be office works.

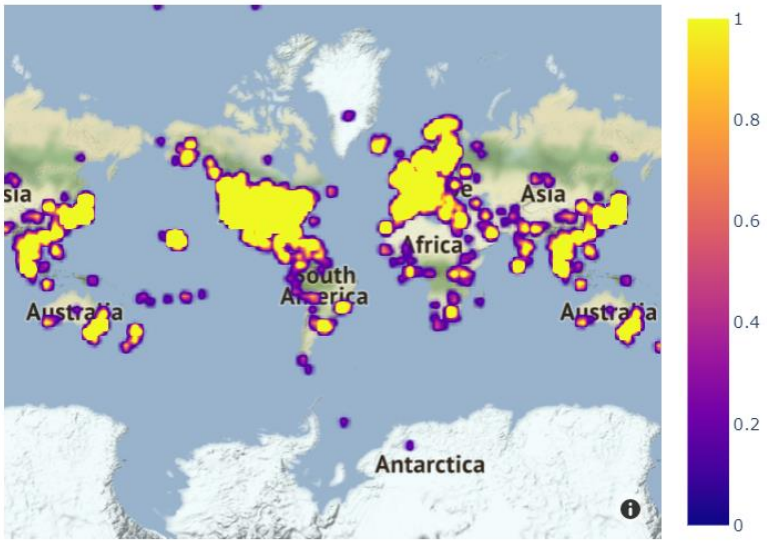


Inferences:- All the top 10 users have same number of check-ins (you can see in the below bar graph) its means may be they are tourist or company employs who were travel across the world for company works.



b)

The yellow colour on map indicate the heat and more heat on map show the more number of check-ins on that place.
you can see in the map that North America and Europe and Singapore and Myanmar sides heat is very high which show these are the most visited countries because these are famous tourist places that's why the number of check-ins in these places is so high.



c)

This time Series graph is for the user with most number of check-ins

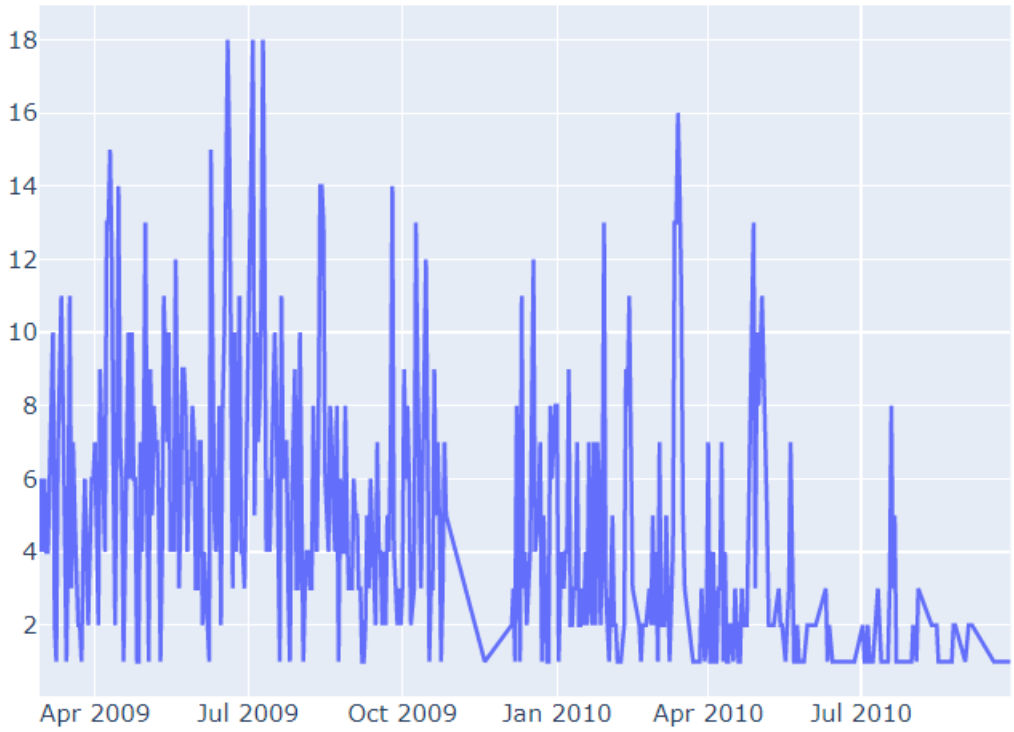
Inferences:- you can see in the graph the person whose user id is 35 do most number of check-ins in July 2009 he/she do 18 check-ins in a single day in July 2009.

He/she may be a traveler or may be some terrorist group use a same ID for doing some attack or meeting in the hotels.

In the graph

x-axis is date

y-axis is number of check-ins by user

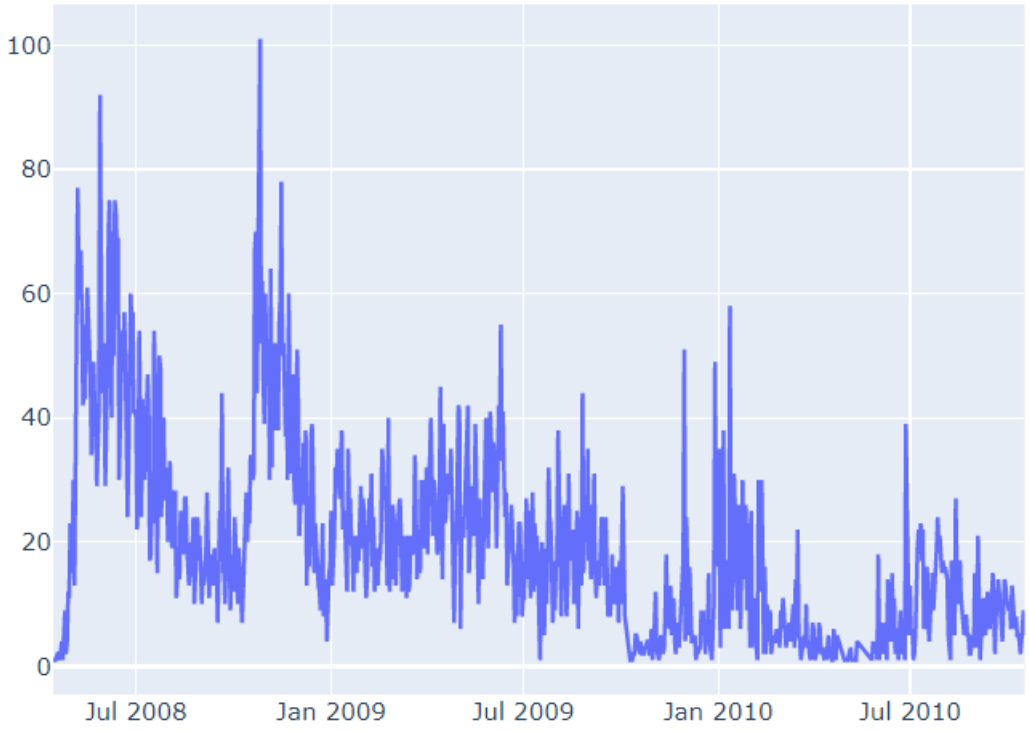


The below graph is showing the number of check-ins in location ID of ee81ef22a22411ddb5e97f082c799f59.

You can see in the graph the most number of check-ins happened on around Oct of 2008 So it may possible it is famous place it may be possible that it is famous 5 star hotel or some tourist place hotel.

x-axis = date

y-axis = No. of check-ins on the given location.



Comment:-

Time Series graph show periodicity it may be possible the person come after interval of time or come occasionally and because of period = 1day it shows the user do check-ins 18 times in a single day.

Q2)

Promoting it to other parties: we can elevate this record to different gatherings for acquiring a couple of benefits. Expect a couple of transport associations wants to perceive the current locale in their worker.

Location recommendation: by understanding the example of the development of the individual, we will propose to them a couple of areas to go to. For example: on the off chance that we comprehend a couple of clients is now and again visiting to the participation, we can exhort extra close to their district.

Showing favorable ads to the user: On the off chance that we have some business venture close wherein the individual is by and by at, we can promote the one's things to the client, and afterward, they can procure benefit from those advertisements.

Advising for a favorable place of an event: In the event that a couple of tolls or enormous event is happening close to the territory, we can encourage those exercises to our area

Security and privacy concers:

Threating the consumers: records of area will not do any physical damage to the person or this record will not complete to threat the person.

Bodily harming the user: information about any individual can be utilized to har and seize any individual for a various explanation. Furthermore, this thought can be utilized by hoodlums and so on

Stalking of person: On the off chance that somebody knows about my place, it's miles a protection subject for me, as an individual can easily utilize it to follow us. this will develop wrongdoing nearby, it particularly will be perilous for young ladies in the public arena

Prediction of present location of user : areas of clients can be expected the utilization of their previous styles of development, that permits you to be awful for privateness and security for clients.

Showing them focused commercials : a few gatherings can convey focused on promotions to control the individual and furthermore can trap them in explorer trick.

SECTION 3

Q1)

a)

There are many matrices which can find the distance between two string but I use Masi and Jaccard

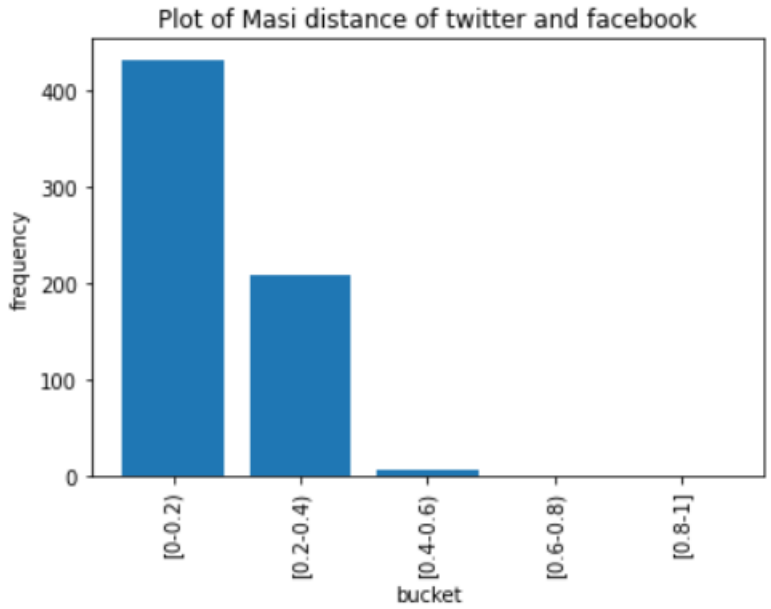
following screenshot is the result of distance matrix between Masi and Jaccard.

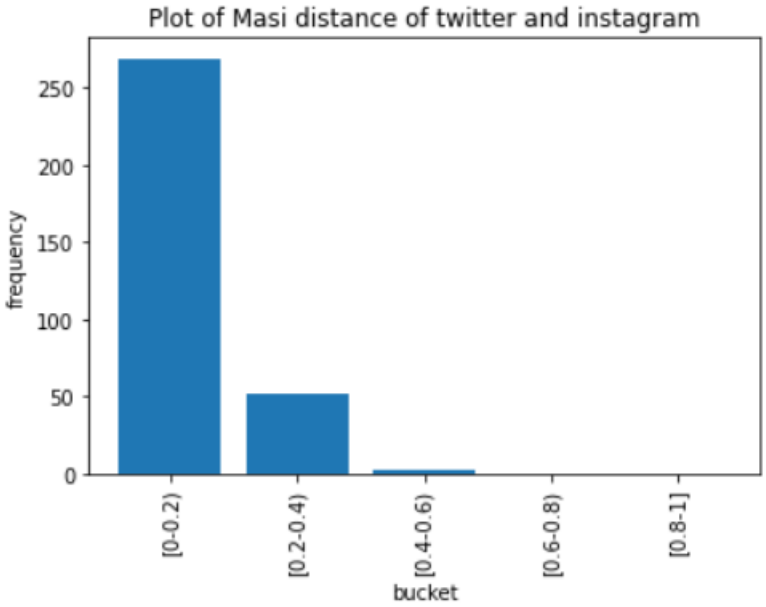
```
masi for twitter-facebook
masi for twitter-instagram
masi for facebook-instagram
```

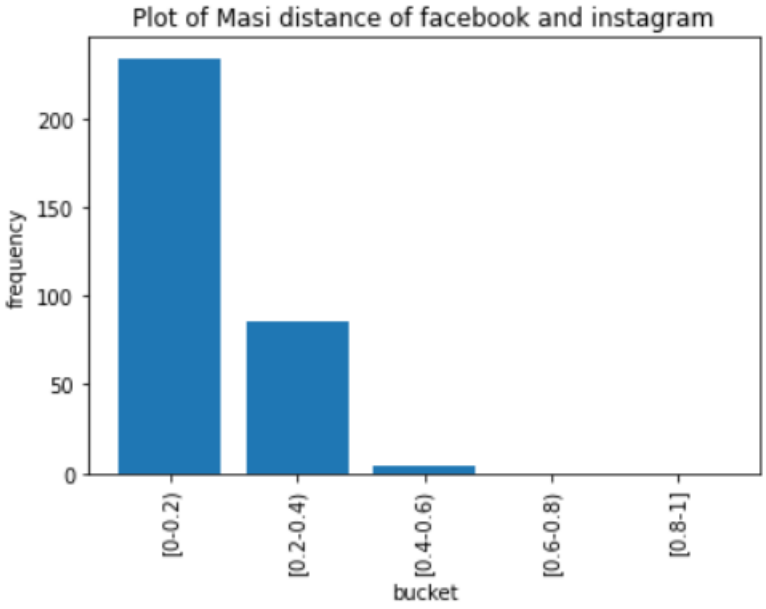
b)

Inference for below 3 bar graphs because all are almost same.

Inference for Masi distance graphs → The below 3 graphs are Masi distance graph and we can see most number of masi distance is in the bucket of 0.0-0.2 and 0.2-0.4 this shows that the masi distance gives us the better result than Jaccard distance.

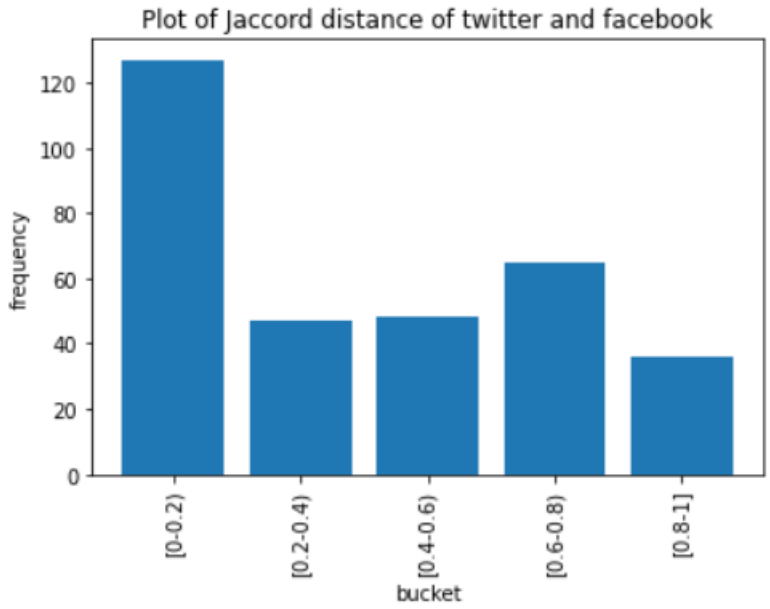


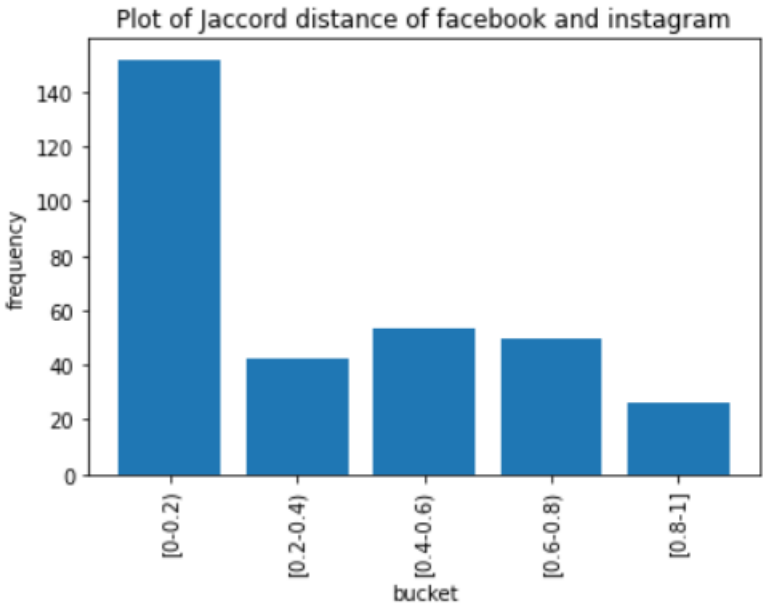

Plot of Masi distance of twitter and instagram
frequency
250
200
150
100
50
0
[0-0.2)
[0.2-0.4)
[0.4-0.6)
[0.6-0.8)
[0.8-1]
bucket

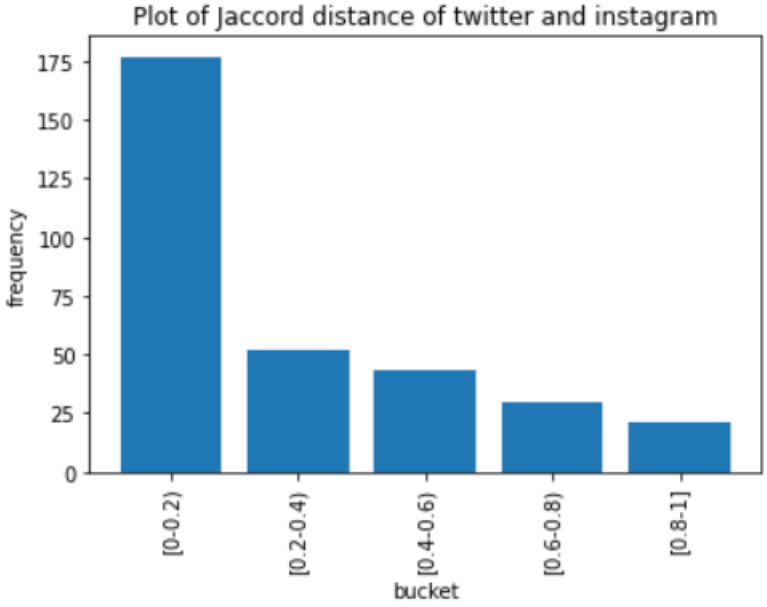

Plot of Masi distance of facebook and instagram
frequency
200
150
100
50
0
[0-0.2)
[0.2-0.4)
[0.4-0.6)
[0.6-0.8)
[0.8-1]
bucket

Below 3 graphs are on the Jaccard distance

Inferences: from above we can say that Jaccard metric is not good for measuring distance between different social media handles. By below three graphs we can easily see that Jaccard distance is distributed among all buckets that's why it is not better than Masi distance and Masi distance is mostly less than the Jaccard distance for social media names.




Plot of Jaccord distance of facebook and instagram
frequency
0
20
40
60
80
100
120
140
[0-0.2)
[0.2-0.4)
[0.4-0.6)
[0.6-0.8)
[0.8-1]
bucket


Plot of Jaccord distance of twitter and instagram
frequency
0
25
50
75
100
125
150
175
[0-0.2)
[0.2-0.4)
[0.4-0.6)
[0.6-0.8)
[0.8-1]
bucket

Section 4

Answer 1)

- We can found out SSN number from date of birth of users from different social media platforms and then after that we can use it to predict SSN numbers.
- We can try similar method to predict the Aadhar number in India. Like if any one will try to get different information from different social media platform and try to predict similarity if there any.
- Aadhar card number is impossible to predict because it generate randomly and it is not related to Date of birth.

Answer 2)

By following reasons it is hard for companies to implement the "Use limitation" principle:

- Hard to regulate 3rd party → Many users share their data to 3rd parties who use their data for different purposes other than companies purpose.
- Slow development process → If companies takes users consent for development data then the development process will be very slow which would affect the companies growth.
- Revenue loss → if user cannot give their consent to company then companies revenue will be decrease because most of the companies revenue comes by selling users information to 3rd party companies.
- Reduce user Experience → if companies start taking consent of user for security purpose for every new update and for any new thing than it would be very annoying for most of the user and it would reduce the usability.
- Misunderstanding or ignoring privacy policy→ Most of the user don't read and don't want to read privacy agreement.

Answer 3)

Following are the reasons that how companies use our information:

- Recommendation: Companies use recommendation algorithm to show us the same or similar information that we are looking for.
- Shows the advertisements relevant to our web server cookies.
- Companies announce reward to users who are using companies platform from long period of time.
- Sharing our information to third party like amazon for delivery details and all.
- By syncing option companies tell us that this person is in network. It will help people to increase their connections.