PSOSM Midsem Assignment

Section 1

Q1 (A)

Observations For polarity

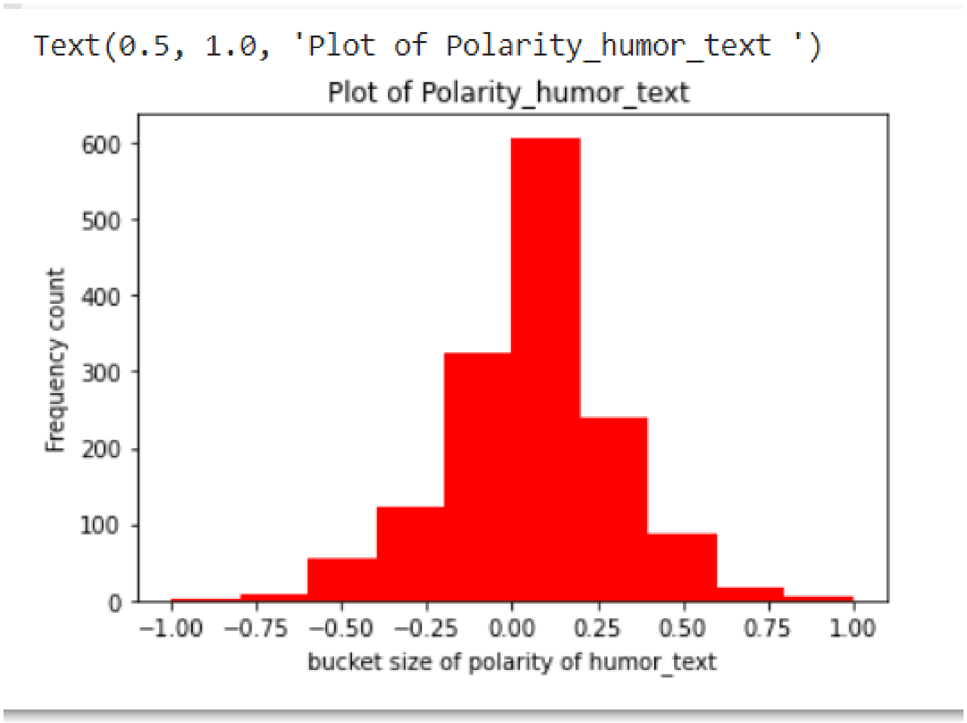
Perceptions or observations: as we can see in the plots (below 2 plots for polarity) peek are at between 0-0.25 and their mean and standard deviation for both humor and news text are similar(approx.)

Polarity of some text is negative it's means there are both type of text present in humor and news (positive and negative text).

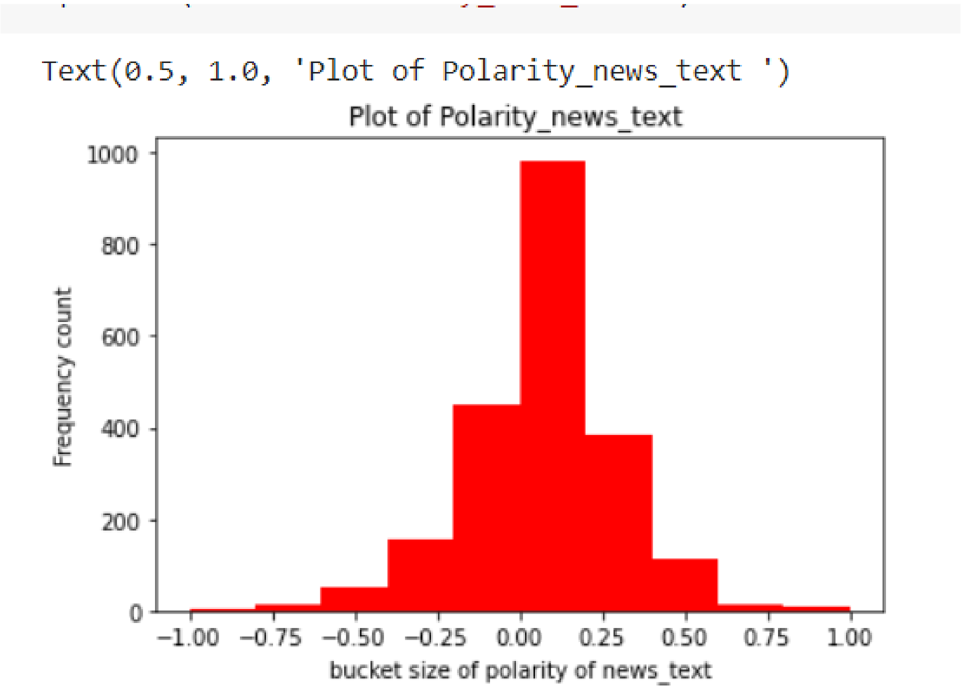
Observations from stats

Polarity_news_text is concentrated at 0.06, and mean of 0.06 and comparatively less standard deviation.

Polarity_humor_text is concentrated at 0.05, and mean of 0.05 and comparatively less standard deviation.



```
Text(0.5, 1.0, 'Plot of Polarity_humor_text ')
```



```
Text(0.5, 1.0, 'Plot of Polarity_news_text ')
```

Observations For subjectivity

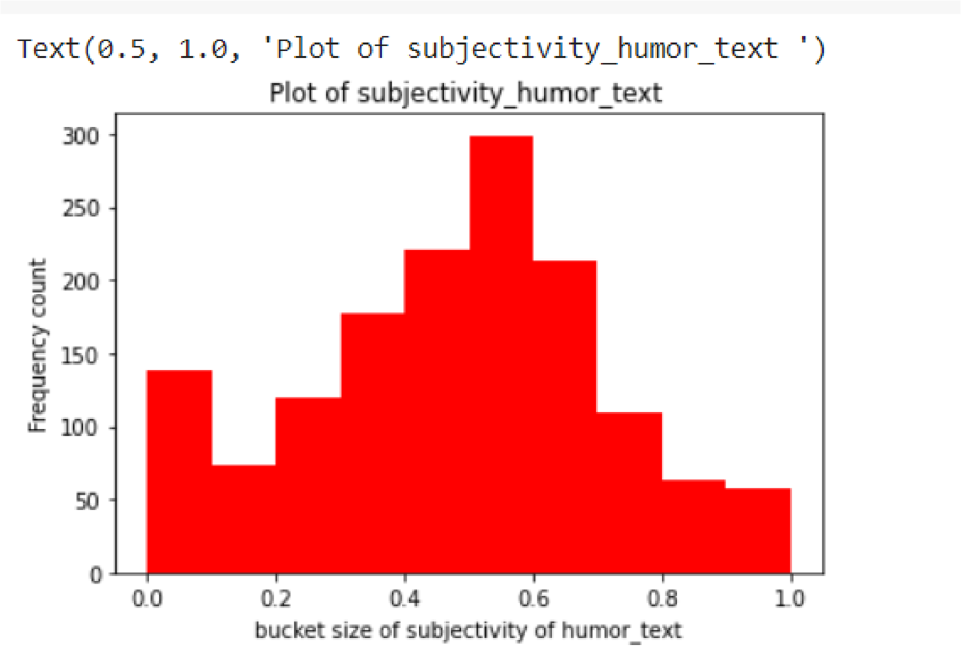
Perceptions or observations: as we can see in the plots (below 2 plots for subjectivity) peek are at between 0.5-0.6 and their mean and standard deviation for both humor and news text are similar(approx.)

Subjectivity of all text is positive (which is default as this is not depend on positive or negative text) and it more depend on emotions, technicality, judgment and personal opinion of text present in text column.

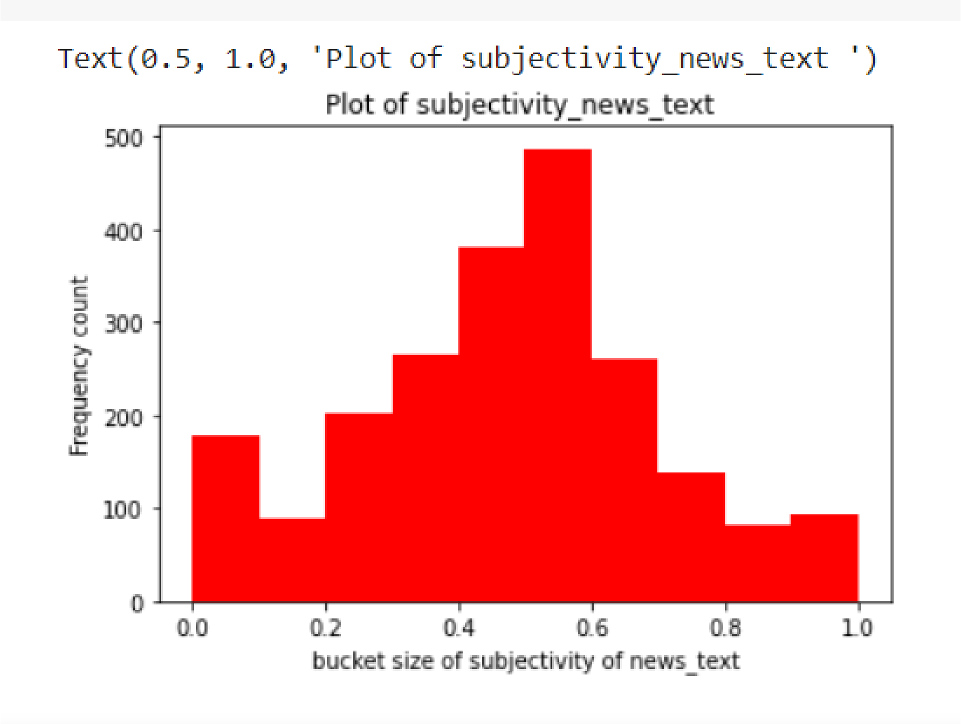
Observations from stats

subjectivity_news_text is concentrated around 0.5. It has comparatively more standard deviation as we can see from the plot and widely distributed.

subjectivity_humor_text is concentrated around 0.5. It has comparatively more standard deviation as we can see from the plot and widely distributed.

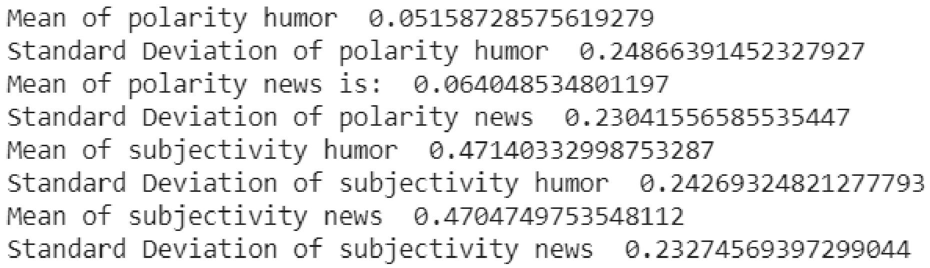


```
Text(0.5, 1.0, 'Plot of subjectivity_humor_text ')
```



```
Text(0.5, 1.0, 'Plot of subjectivity_news_text ')
```

Plot of stats



```
Mean of polarity humor  0.05158728575619279
Standard Deviation of polarity humor  0.24866391452327927
Mean of polarity news is:  0.064048534801197
Standard Deviation of polarity news  0.23041556585535447
Mean of subjectivity humor  0.47140332998753287
Standard Deviation of subjectivity humor  0.24269324821277793
Mean of subjectivity news  0.4704749753548112
Standard Deviation of subjectivity news  0.23274569397299044
```

(B)

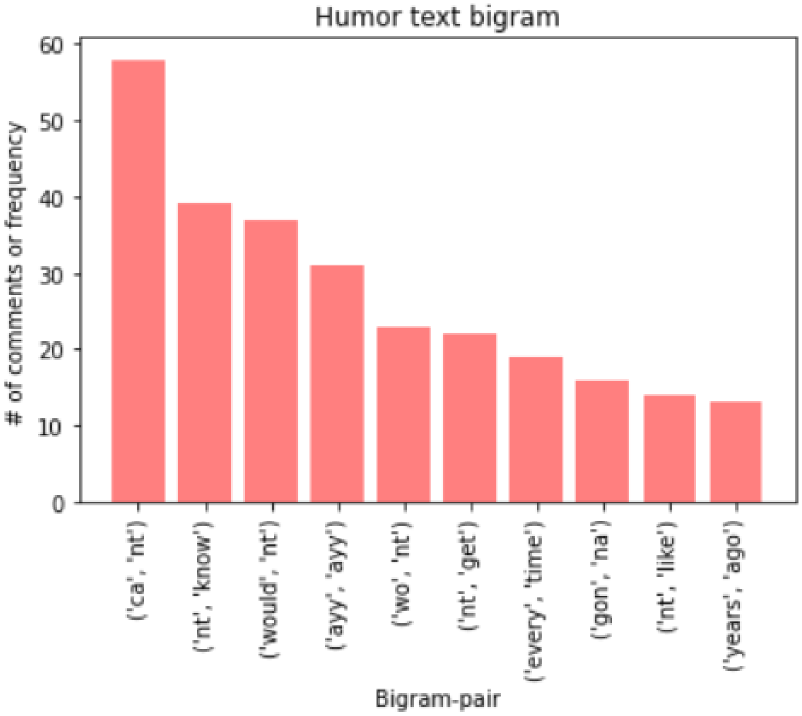
Inferences: - for bigram plot

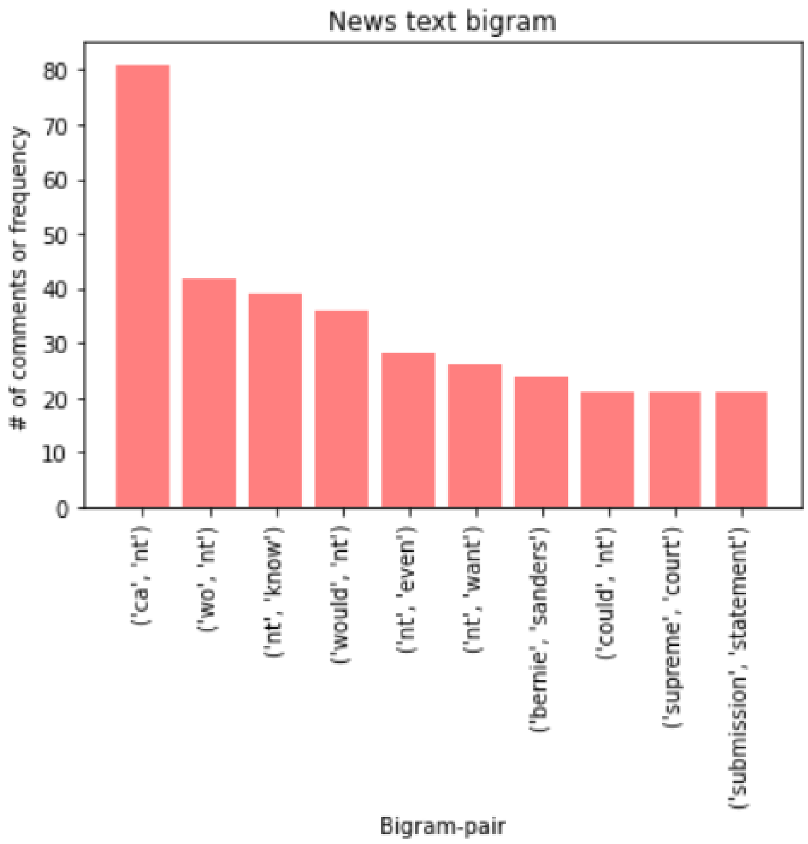
“can’t” is the most frequent bigram

Word like supreme, court, submission, could etc. are more factual.

These bigrams have both positive and negative sentiments.

Bigram word pair are most subjective in news_text_bigram than humor_text_bigram.


Humor text bigram
# of comments or frequency
60
50
40
30
20
10
0
('ca', 'nt')
('nt', 'know')
('would', 'nt')
('ayy', 'ayy')
('wo', 'nt')
('nt', 'get')
('every', 'time')
('gon', 'na')
('nt', 'like')
('years', 'ago')
Bigram-pair



Inferences: - for unigram plots

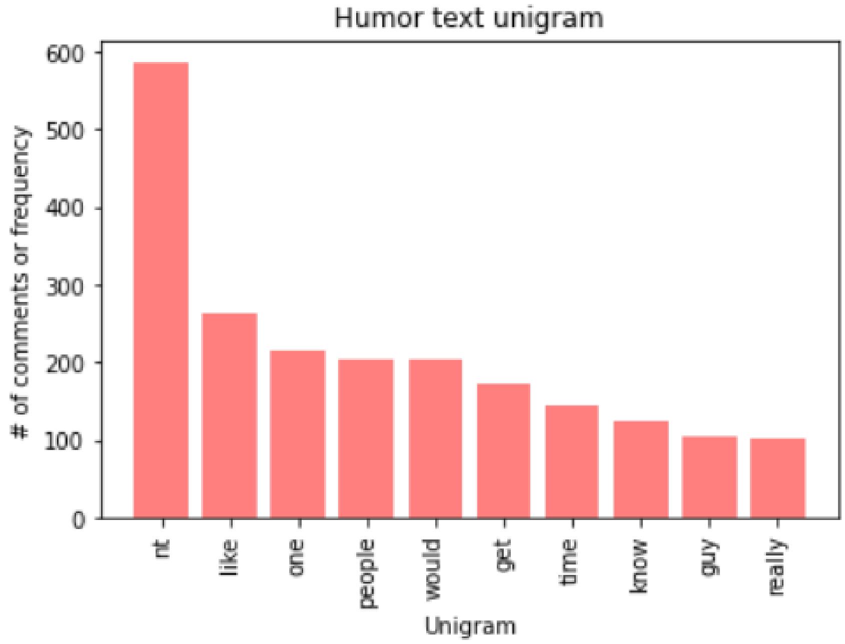
“nt” comes more time than others unigram word in both unigram plots.

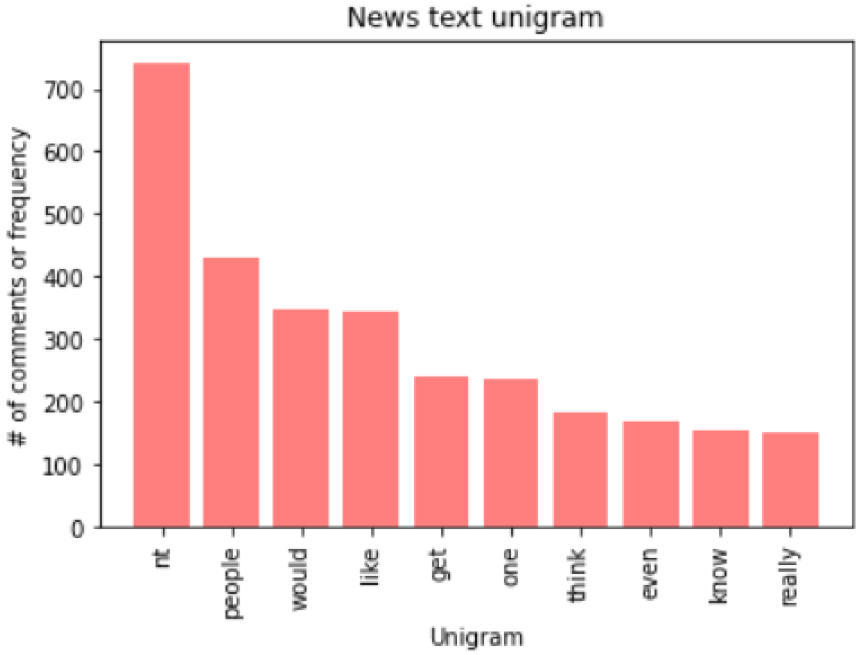
Both plots have same word almost (top 10 word) in unigram plot.

These unigrams have both positive and negative sentiments.

Unigram word most subjective in news_text_unigram than humor_text_unigram.

As top 10 unigram are same we can not able to classify correctly by using single attribute.





Q2

(A) following code and comments for preprocessing, splitting and transforming using bag of words approach with justification

```
pkl=pd.read_pickle('redditDataset.pkl') #reading file

dis_pkl=pkl.drop_duplicates() #droping duplicates from data frame

  dis_pkl.isnull().sum()        #checking null value in data frame

bi_pkl=pd.get_dummies(dis_pkl['subreddit'], drop_first=True) #categoriszing humor =0 and news =1

con_pkl=pd.concat([dis_pkl,bi_pkl],axis=1) #concating news column (binary column) with prev_data frame

con_pkl.drop(['subreddit'],axis=1, inplace=True) #deleting subreddit column now we have data frame with #text and news column

x=con_pkl['text']

y=con_pkl['news']

x_tr, x_te, y_tr, y_te = train_test_split(x,y,test_size=.3, random_state=1) #splitting data into 70-30 train-test data

xt=x_te

yt=y_te

vect = CountVectorizer() # basically by using it we do bag of words approach (collecting unique word into a vector), word lemmatization, removing stop words, counting frequency, transforming and creating matrix of train data.

x_tr= vect.fit_transform(x_tr)  #transforming data into suitable form and using bag of words approach

x_te = vect.transform(x_te)

feat_names = vect.get_feature_names() #extractng features from train data

naive_base = MultinomialNB() #creating a model for training data

naive_base.fit(x_tr, y_tr) #training the data
```

(B)

Using Multinomial naive Bayes model

Pickle file link

https://drive.google.com/file/d/1Sdb20A5DjUkF3pbFBF8W5Maqib129arq/view?usp=sharing

(C)

Accuracy = 0.8156934306569343

Classification Report

```
            precision    recall  f1-score   support

         0       0.83      0.71      0.77       467
         1       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)

0 = humor

1=news

Well predicted text

1. 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. 0

Reasons: - because it contains word (singular subjective word) like I and abusive word like asshole that has a negative sentiment make this text humor.

2. 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. 0

Reasons: - because it contains word (singular subjective word) like my, you and abusive word like shit that has a negative sentiment and word like sexy less technical make this text humor.

3. i will go out of my way to transfer you to make a second 40 minute car journey to come back for a refund. 0

Reasons: - because it contains word (singular subjective word) like I ,means talking about themselves not like a news text talking about others make this a humor text.

4. 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. 0

Reasons: - because it contains word (singular subjective word) like I, you, I am ,and word like smash, hips, good etc. less technical words make this text humor.

5. 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. 1

Reasons: - because it contains word some serious word like problem, restricting, seizing and some technical word payment, car purchasing, tax office etc. make this a news text.

Wrong predicted text

1. 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. 1

Reasons: - because it contains word some serious word like important, preferences, and some technical word stock taking, inventory etc. make this a news text.

2. a railroad engineer must be sure not to lose his train of thought or he might go down the wrong track. 0

Reasons: - because it contains word (singular subjective word) like he ,and word might, must, etc. less technical words make this text humor.

3. he is not following an antiquated convention that no one gives two shits about that guy is no gentleman but an asshole. 0

Reasons: - because it contains word (singular subjective word) like I, you, I am ,and word like smash, hips, good etc. less technical words and less subjectivity make this text humor.

4. 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 . 1

Reasons: - because it contains word some technical word services, download, speed, upload, internet etc. make this a news text.

5. 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. 0

Reasons: - because it contains word (singular subjective word) like I, he ,and word like dammit, joked etc. less technical words and less subjectivity make this text humor

Section 2

Q1

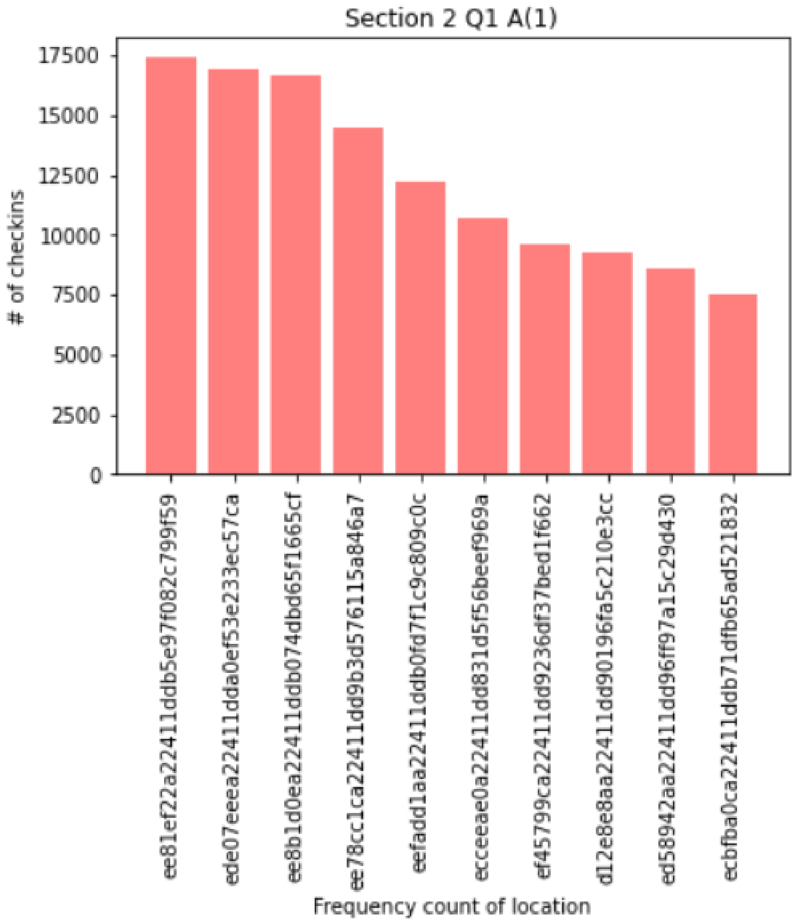
(A)

Inferences: -

As we can see in the plot the number of check-ins decreases as we change the location (no two different locations have same number of check-ins).

It means some places are very famous than other that’s why number of check-ins are very large their (First three location mainly)

Location ids with same or large number of check-ins (approx.) May lie in a same country. Or nearest to each other.

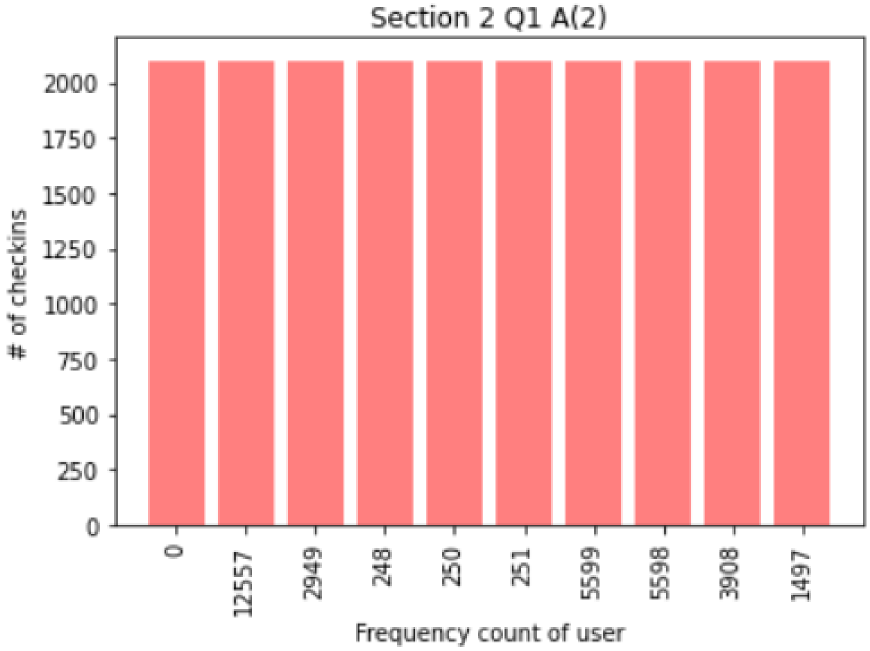


Inferences: -

As we can see in the plot the number of check-ins same for all top 10 users (no two different users have different number of check-ins).

It means most of the users using bright kite or a social network for check-ins.

Maybe they have some business meeting or they are travelers.



(B)

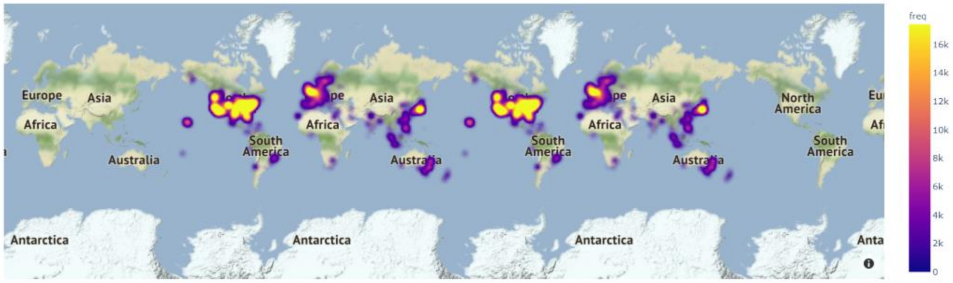
The more heat on the map shows high number of check-ins on that place.

North America has the greatest number of check-ins.

No one goes to Antartica as there is no heat on the map near Antarctica.

After North America Europe is the second most place number of check-ins

So, north America and Europe are the most famous and tourist area as compare to others as the greatest number of people check-ins in these country .



(C)

1.The user with the greatest number of check-ins

Inferences: -

The data is from 2009-05-25 to 2010-10-17 (for user id =0)

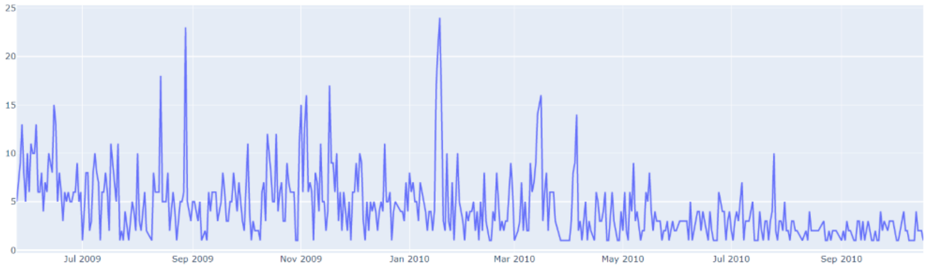
And on 18 January 2010 the greatest number of check-in (24 check-ins) occurs.

It means this person check-ins 24 times in a single day.

Maybe he/she is a traveler or a suspicious person or a doing some kind of meetings.

X-axis= date

Y-axis=no. Of check-ins on a given date for a user



2. The location with the most number of check-ins

Inferences: -

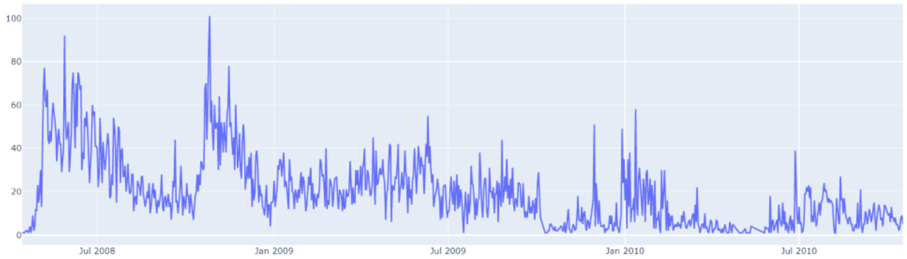
The data is from 2008-04-15 to 2010-10-17 (for location id = ee81ef22a22411ddb5e97f082c799f59)

And on 26 Oct 2008 the greatest number of check-in (101 check-ins) occurs.

So, this location may be a famous or a tourism area.

X-axis= date

Y-axis=no. Of check-ins on a given date for a location



Comment: -

Temporal facts or data of the longitude and latitude for any users may be used to threatening the user and maybe to physically harm it.

Yes, it’s show periodicity because in time series graph for user may a user check-in at a given place daily (period = 1 day) and in location time series graph it’s also shows periodicity may be that place is visited by someone at regular interval or occasionally.

Q2

Some ways by which we can target user using this data:

Location recommendation: by understanding the pattern of movement of the person, we will propose them a few locations to go to. for instance: if we understand a few users is every so often touring to the membership, we can advise extra near their region.

showing favorable ads to user: If we've got some business enterprise near wherein the person is presently at, we can advertise the ones things to the user and then they can earn profit from those ads.

promoting it to other party: we can promote this record to other parties, for earning a few profits. assume a few transport organizations desires to recognize the modern-day region in their employee.

Advising a proper place for an event: If a few fare or big occasion is occurring near the area, then we are able to advise those activities to our region.

Suspecting or tracking a suspicious person by govt: govt can track suspected person using their location data. In case of criminals or other suspicious activity.

Security and privacy concerns:

bodily harming the user: data about any person can be used to har and kidnap any person for different reason. And this idea can be used by criminals etc.

Threatening the consumer: records like area may be without difficulty used to threaten the person or even physically damage them.

showing them focused commercials: some groups can deliver targeted ads to manipulate the person and also can trap them in traveler scam.

Stalking of persons: If someone is aware of my place, it's miles a privacy subject for me, as a person can effortlessly use it to stalk us. this will growth crime in the area, it especially will be dangerous for girls in society.

Prediction of present loc of user: locations of customers can be anticipated the use of their preceding styles of movement, that allows you to be terrible for privateness and safety for customers.

Section 3

Q1

(A)

I have used **masi and Jaccard** these two-distance metrics

On the basic of distance metric of two different social handle.

Best metric for twitter and Facebook is masi distance metric.

Best metric for Facebook and insta is masi distance metric.

Best metric for twitter and insta is masi distance metric.

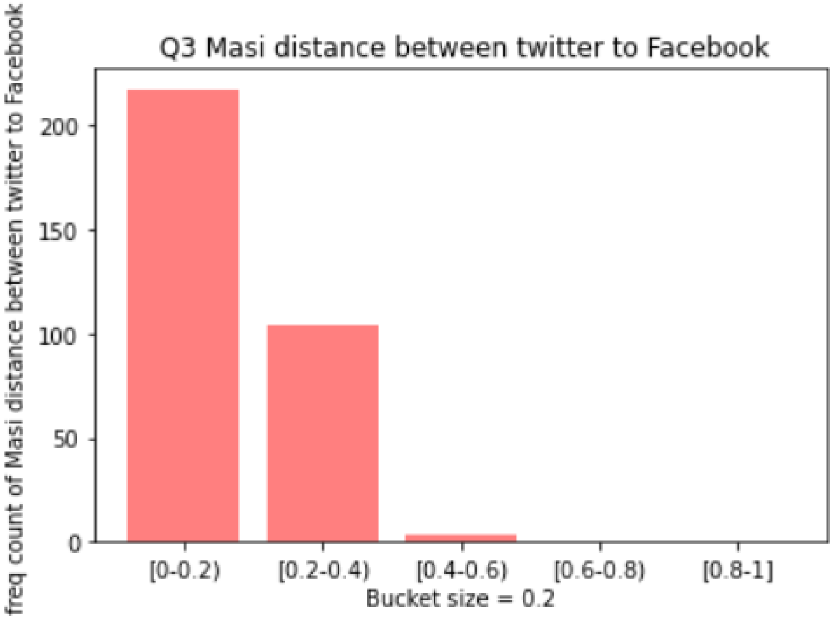
(B)

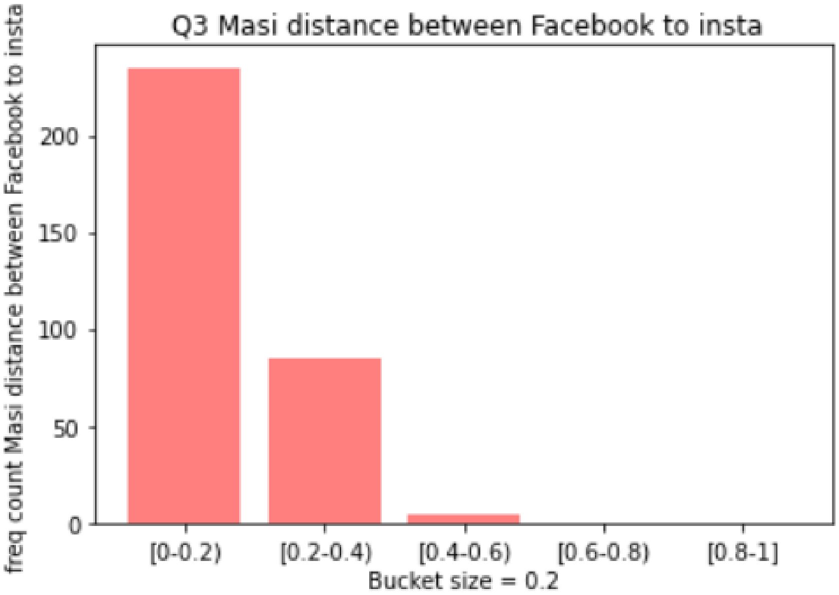
Inferences for masi distance metrics (for below three plot)

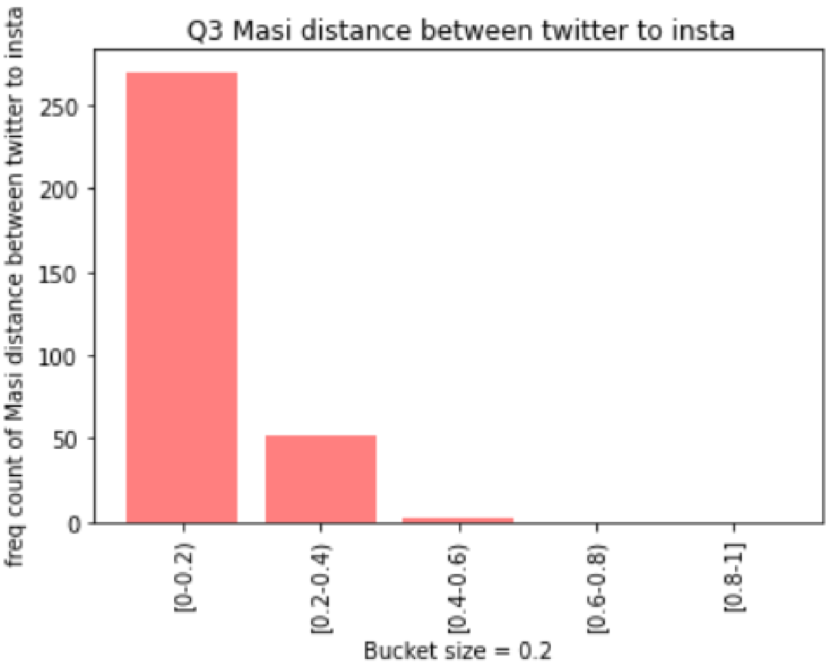
As we can see above masi distance metric is better than Jaccard distance metric.

We can also see this in bar graph (below 3 bar graphs for masi distance metric) as most of the frequency count occurs as bucket 0-0.2 and 0.2-0.4.

So for most of the distance in different social media handle name is less than 0.4 which is very less.


Q3 Masi distance between twitter to Facebook
freq count of Masi distance between twitter to Facebook
0
50
100
150
200
[0-0.2)
[0.2-0.4)
[0.4-0.6)
[0.6-0.8)
[0.8-1]
Bucket size = 0.2


Q3 Masi distance between Facebook to insta
freq count Masi distance between Facebook to insta
0
50
100
150
200
[0-0.2)
[0.2-0.4)
[0.4-0.6)
[0.6-0.8)
[0.8-1]
Bucket size = 0.2



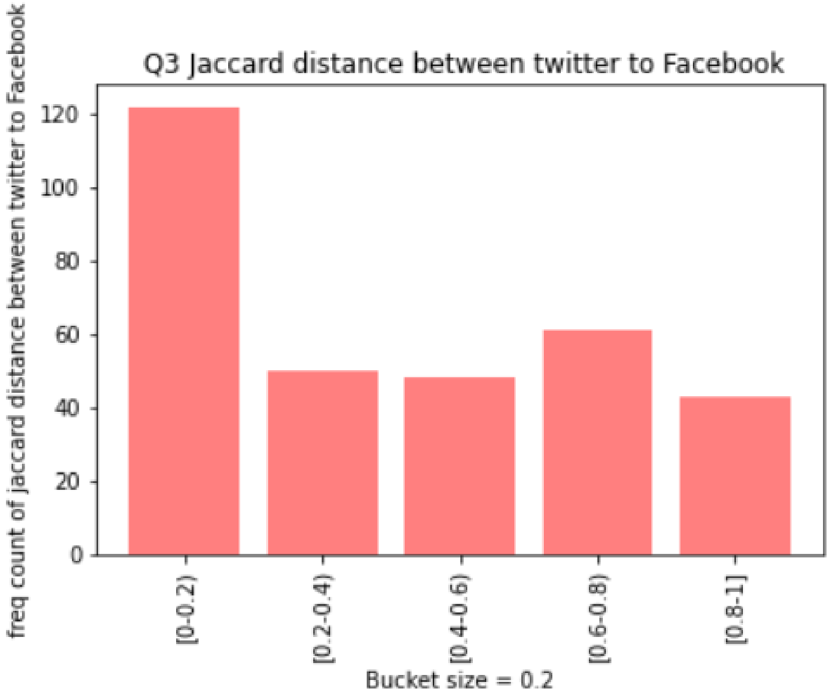
Inference for Jaccard distance metric (for below three plot)

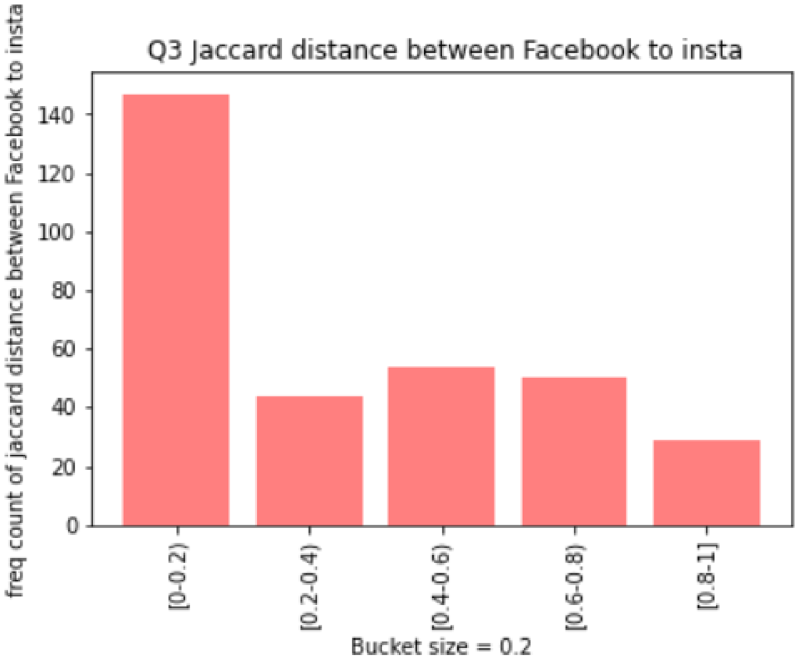
As we can see above Jaccard distance metric not so good for measuring distance between different social handles name.

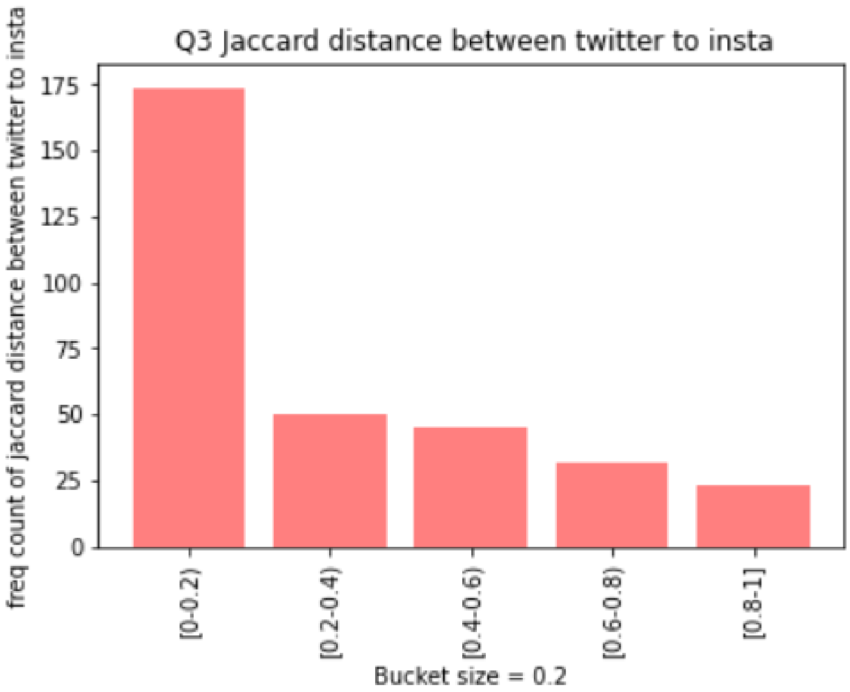
We can also see this in bar graph (below 3 bar graphs for Jaccard distance metric) approx. all frequency count distributed among all bucket size.

That's why it's not better than masi distance metric and after comparing distance metric by both masi and Jaccard (In Ipynb)

Masi distance is always less than or equal to Jaccard distance for every social handle name pair.


Q3 Jaccard distance between twitter to Facebook
freq count of jaccard distance between twitter to Facebook
0
20
40
60
80
100
120
[0-0.2)
[0.2-0.4)
[0.4-0.6)
[0.6-0.8)
[0.8-1]
Bucket size = 0.2


Q3 Jaccard distance between Facebook to insta
freq count of jaccard distance between Facebook to insta
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 size = 0.2



Section 4

Ans1

We can find ssn number using D.O.B on different social media then we can predict ssn number using those data.

By using same method as mention above or using different attributes like name, father’s name age, D.O.B etc. for extracting features by which we can predict Aadhaar number.

But it it’s not possible to predict Aadhaar number in India as Aadhar number is different and random for each person and it’s 16 digits (which is a lot) and it is not connected with our D.O.B, name, father’s name etc. Only. It’s an identity of a person that’s why predicting Aadhaar number is not possible.

Ans2

Use limitation principle: - states that individual information ought not be uncovered to anyone or any association with the exception of law specialists and assent of information subject.

Reasons why it’s hard to implement.

1. Client Past uses: If less information is unveiled to the individual client than it will be awful insight for clients as they won't have a clue about their mutual clients (friend or user). For instance, if D.O.B information isn't uncovered or the profile image of the client isn't revealed then it will be an awful encounter for the local area.
2. Data sells by third party: If somebody logs at outsider applications utilizing a similar id, at that point outsider applications can sell client information that the organization has no influence over.
3. Income Loss of companies: The organization will face income loss and consequently they won't be able to overcome their costs.
4. Difference of sensitivity (in rules and regulation of data uses) for companies: - This is one of the reasons depend on different companies mean how much and how they are going to use the data.

5. Misunderstanding or ignoring privacy policy: - because Most of the users mostly don’t read privacy agreement.

Ans3

Reason mention below are suitable to defend by anyone(companies) to using our info.

1. Provide services: - So. They can share our least informative info (which does not go to hurt our privacy completely) to others companies related to their company so they can able to provide use better service. E.g., online shopping companies giving address related details to delivery companies so they can able to provide us service.
2. Recommendation: - They can use our data by which they can recommend their product which are liked by their users. E.g., music recommendation on Spotify or other music platform.
3. Favorable or useful Advertisement to users: - By using user data, they Showing us ads which are liked by that particular user like I searched for ASUS phone on amazon then sometime I can easily find ads related to ASUS phone on my browser.
4. For some special cashbacks or reward: - They use our data to check the criteria for that cashback or reward by app. E.g., like if you do transaction of greater than 2000 you get cashback of 10% etc.

5. Syncing Data from one place to other(platforms): - some companies use our data to syn our data with others to for better engagement with others like recommendation of friend on social media using his contact number saved in your phone etc.