Project Report for IAL621: Content Analysis for Social Network Data

Extract & Convert Social Network (Twitter) content for sentiment analysis

Bala Mallampati   Thursday, December 9, 2021
Master's in informatics and Analytics
University of North Carolina, Greensboro, USA
b_mallampat@uncg.edu

Abstract:

I understand the Social Network (Twitter, Facebook) and analyzed using NodeXL software. And then extracted twitter text information and did sentiment analysis of topics/tweets by following all privacy, regulatory & access policies.

In this project, I answered the following data research questions:

A. Twitter network analysis to identify their networks as part of exploratory analytics.

B. Extract information for UNCG top tweets & topics from Twitter site and convert to numbers as part of Data wrangling.

C. Classify UNCG tweets/hashtags using Machine Learning Models(Linear, KNN, Decision Tree, Random Forest..)

D. Do sentiment analysis of UNCG Tweets/Topics and present extracted insights to corresponding organization.

Introduction and Motivation:

Now a days everyone using social media for News, follow Trends and share their feelings and get in touch with their friends & family/Employers/Learning. Social media can create impacts on every organization if management is not up to date on latest trends/news.

Social media include all the ways people connect to people through computation. Mobile devices, social networks, email, texting, micro-blogging and location sharing are just a few of the many ways people engage in computer-mediated collective action. As people link, like, follow, friend, reply, retweet, comment, tag, rate, review, edit, update, and text one another (among other channels) they form collections of connections. These collections contain network structures that can be extracted, analyzed, and visualized. The result can be insights into the structure, size, and key positions in these networks.

I love to play with data, now “The world’s most valuable resource is no longer oil, but data” as per The Economist article (https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data).

If we can understand how these social network works and can read & extract useful information, then we can predict how it will impact our organization by doing sentiment analysis on Tweets/topics.

In this project, I used MassMine framework for twitter data extraction and analysis to predict impacts on organization and take preventive actions.

Data Source and description:

Tweets are the basic atomic building block of all things Twitter. Tweets are also known as “status updates.” The Tweet object has a long list of ‘root-level’ attributes, including fundamental attributes such as id, created_at, and text. Tweet objects are also the ‘parent’ object to several child objects. Tweet child objects include user, entities, and extended_entities. Tweets that are geo-tagged will have a place child object.

Data extracted from below social media site. I had extracted 50K UNCG tweets from past one week data using Massmine and summarized to 2165 tweets/topics with below attributes.

https://twitter.com/

https://developer.twitter.com/en/portal/dashboard

Attribute	Alias in Project	Type	Description
created_at	tweet.datetime	String	UTC time when this Tweet was created. Example:"created_at": "Wed Oct 10 20:19:24 +0000 2018"
text	text	String	The actual UTF-8 text of the status update. See twitter-text for details on what characters are currently considered valid. Example:"text":"To make room for more expression, we will now count all emojis as equal—including those with gender and skin t… https://t.co/MkGjXf9aXm"
user.description	user.description	String	User Object: "description": "The Real Twitter API. Tweets about API changes, service issues and our Developer Platform. Don't get an answer? It's on my website.",
user.favourites_count	total.favorites	Int64	User Object: "favourites_count": 31,
user.followers_count	total.followers	Int64	User Object: "followers_count": 6129794,
user.friends_count	total.friends	Int64	User Object: "friends_count": 12,
user.location	user.location	String	User Object: location: "San Francisco, CA",
user.name	user.name	String	User Object: "name": "Twitter API",
user.screen_name	screen.name	String	User Object: "screen_name": "Twitter API",
user.statuses_count	total.statuses	Int64	User Object: "statuses_count": 3658,

entities.user_mentions.0.screen_name	tweet.mentions	String	Entities:"user_mentions":[],
retweeted_status.user.screen_name	retweet.mentions	String	Retweets can be distinguished from typical Tweets by the existence of a retweeted_status attribute. This attribute contains a representation of the original Tweet that was retweeted. Note that retweets of retweets do not show representations of the intermediary retweet, but only the original Tweet. (Users can also unretweet a retweet they created by deleting their retweet.)
entities.urls.0.expanded_url	url	String	Entities: "urls":[],
retweeted_status.user.entities.url.urls.0.expanded_url	retweet.url	String	Retweets can be distinguished from typical Tweets by the existence of a retweeted_status attribute. This attribute contains a representation of the original Tweet that was retweeted. Note that retweets of retweets do not show representations of the intermediary retweet, but only the original Tweet. (Users can also unretweet a retweet they created by deleting their retweet.)
retweet_count	total.retweets	Int64	Number of times this Tweet has been retweeted. Example:"retweet_count":160

Methodology:

Data processing is an important step for in the data analysis. Data science involves methods of analyzing massive amounts of data for the purposes of knowledge extraction. It evolved from statistics and traditional data management. Data comes in many shapes and forms, and many times we need to get it ready to be able to analyze it. The phrase “garbage-in and garbage-out” is particularly applicable to text mining to Train and Test Data.

In this project I have used the following languages, frameworks, tools, libraries, packages from Twitter data extraction to Sentiment analysis and Tweets data classification/regression.

Technologies & Libraries: R(ndjson, tidyverse, tm, lubridate, fliptime, stringr,ggmap, maps, quantenda, readtxt, textplots, widyr), Python(numpy, pandas, matplotlib, seaborn, sklearn, google.colab,)

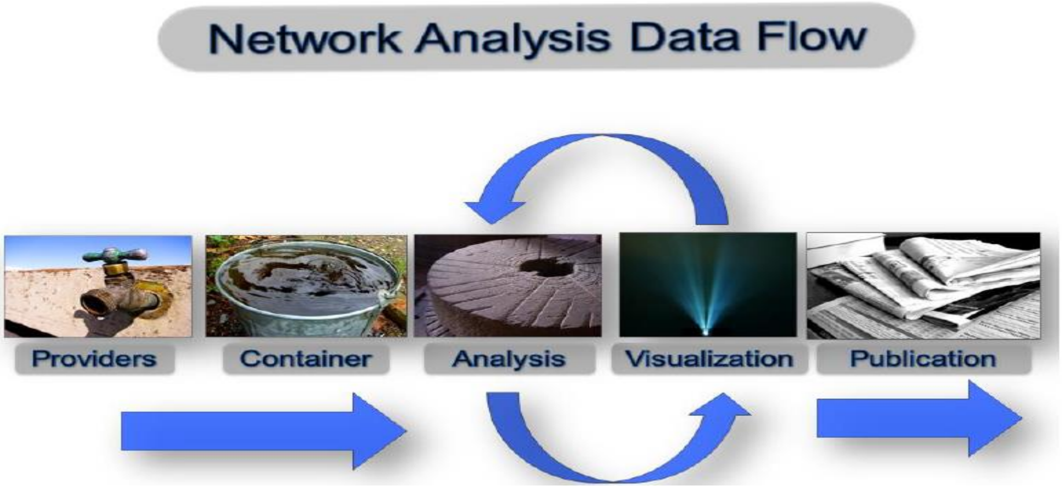
Framework/Softwares: Massmine, NodeXL, Oracle VM virtual box 6.1.26, Ubuntu 20.04

Tools: Rstudio, Colab for python, Power BI, Tableau

Methods: R Selenium, Machine Learning(Linear, Logistic, SVM, RandomForest, KNN, SVM)

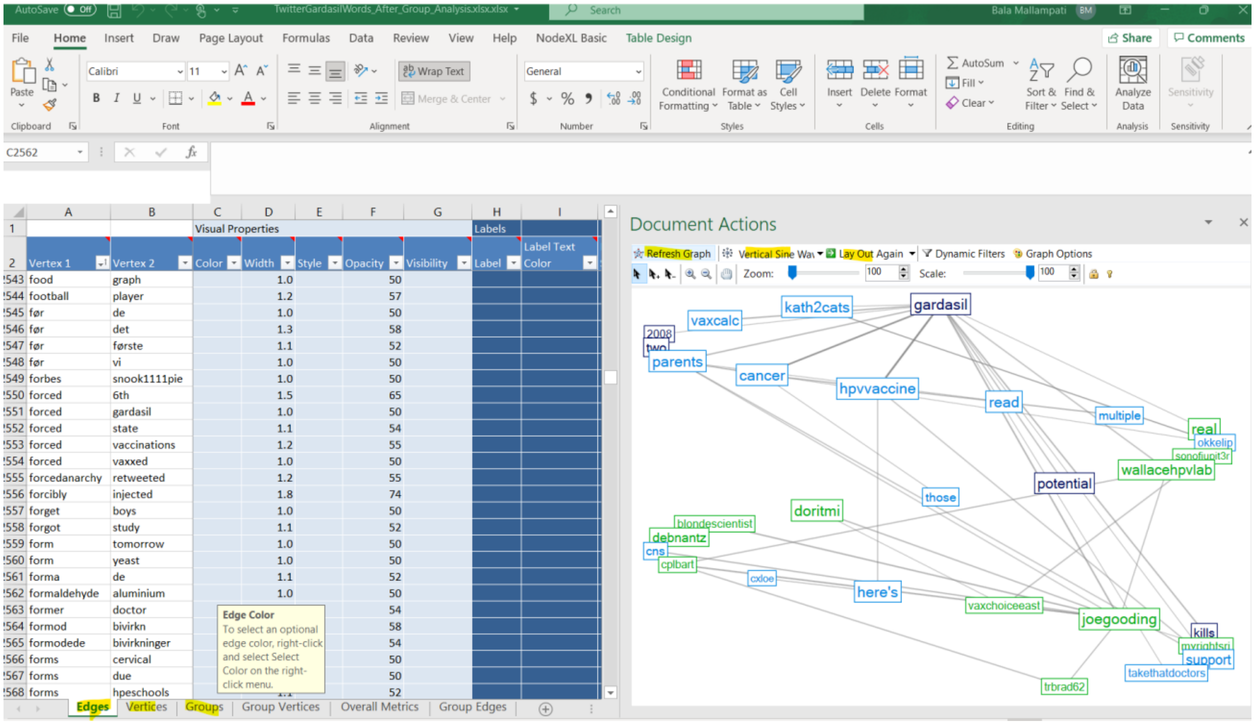
Data Analysis:

Social media networks form in Twitter around a wide range of terms. People talk about the news of the day, celebrities, companies, technology, entertainment, and more. As each person uses Twitter they form networks as they follow, reply and mention one another. These connections are visible in the text of each tweet or by requesting lists of the users that follow the author of each tweet from Twitter.



I came to learn below NodeXL software to analyze Social Networks by input below values. NodeXL enables the automatic execution of a five-step data workflow that starts with data collection from a variety of network data sources, through storage, analysis, visualization and finally publication.

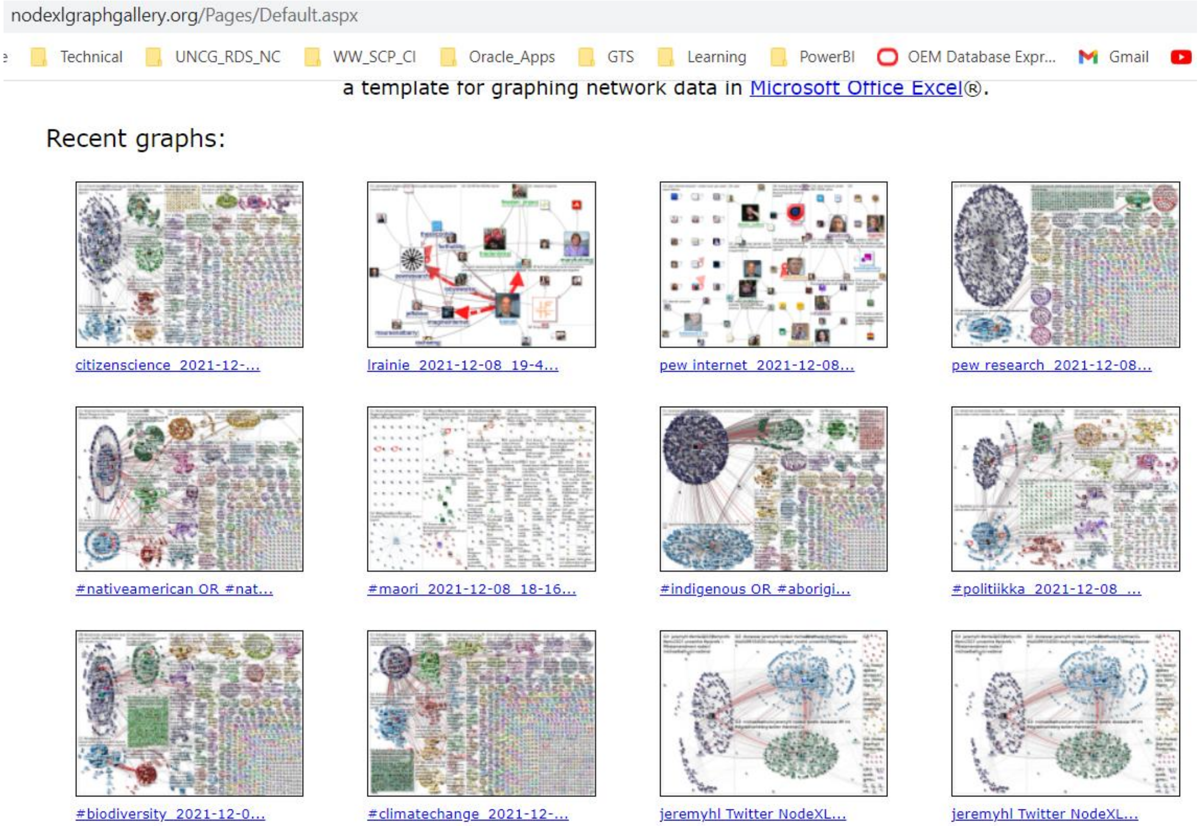
https://nodexl.com/guides/how-to-install-nodexl-pro



	A	B	C	D	E	F	G	H	I
1			Visual Properties					Labels	
2	Vertex 1	Vertex 2	Color	Width	Style	Opacity	Visibility	Label	Label Text Color
2543	food	graph		1.0		50			
2544	football	player		1.2		57			
2545	før	de		1.0		50			
2546	før	det		1.3		58			
2547	før	første		1.1		52			
2548	før	vi		1.0		50			
2549	forbes	snook1111pie		1.0		50			
2550	forced	6th		1.5		65			
2551	forced	gardasil		1.0		50			
2552	forced	state		1.1		54			
2553	forced	vaccinations		1.2		55			
2554	forced	vaxxed		1.0		50			
2555	forcedanarchy	retweeted		1.2		55			
2556	forcibly	injected		1.8		74			
2557	forget	boys		1.0		50			
2558	forgot	study		1.1		52			
2559	form	tomorrow		1.0		50			
2560	form	yeast		1.0		50			
2561	forma	de		1.1		52			
2562	formaldehyde	aluminium		1.0		50			
2563	former	doctor		[illegible]		54			
2564	formod	bivirkn		[illegible]		58			
2565	formodede	bivirkninger		[illegible]		54			
2566	forms	cervical		[illegible]		50			
2567	forms	due		[illegible]		50			
2568	forms	hpeschools		1.1		52			

Edge Color
To select an optional edge color, right-click and select Select Color on the right-click menu.

http://nodexlgraphgallery.org/Pages/Default.aspx – It has different network analysis graphs to analyze data with graphs(ex: Covid19, Twitter networks, Climate changes..etc)



nodexlgraphgallery.org/Pages/Default.aspx

Technical | UNCG_RDS_NC | WW_SCP_CI | Oracle_Apps | GTS | Learning | PowerBI | OEM Database Expr... | Gmail

a template for graphing network data in Microsoft Office Excel®.

Recent graphs:

citizenscience_2021-12-...

lrainie_2021-12-08_19-4...

pew internet_2021-12-08...

pew research_2021-12-08...

#nativeamerican OR #nat...

#maori_2021-12-08_18-16...

#indigenous OR #aborigi...

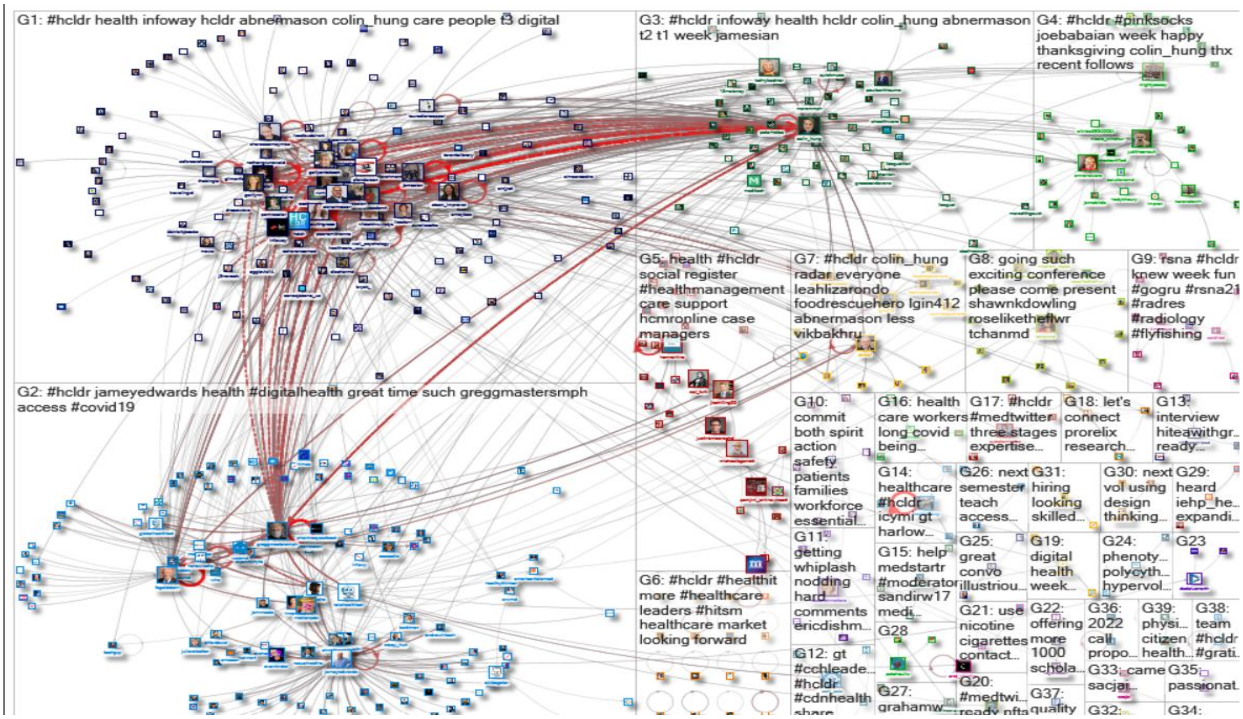
#politiikka_2021-12-08_...

#biodiversity_2021-12-0...

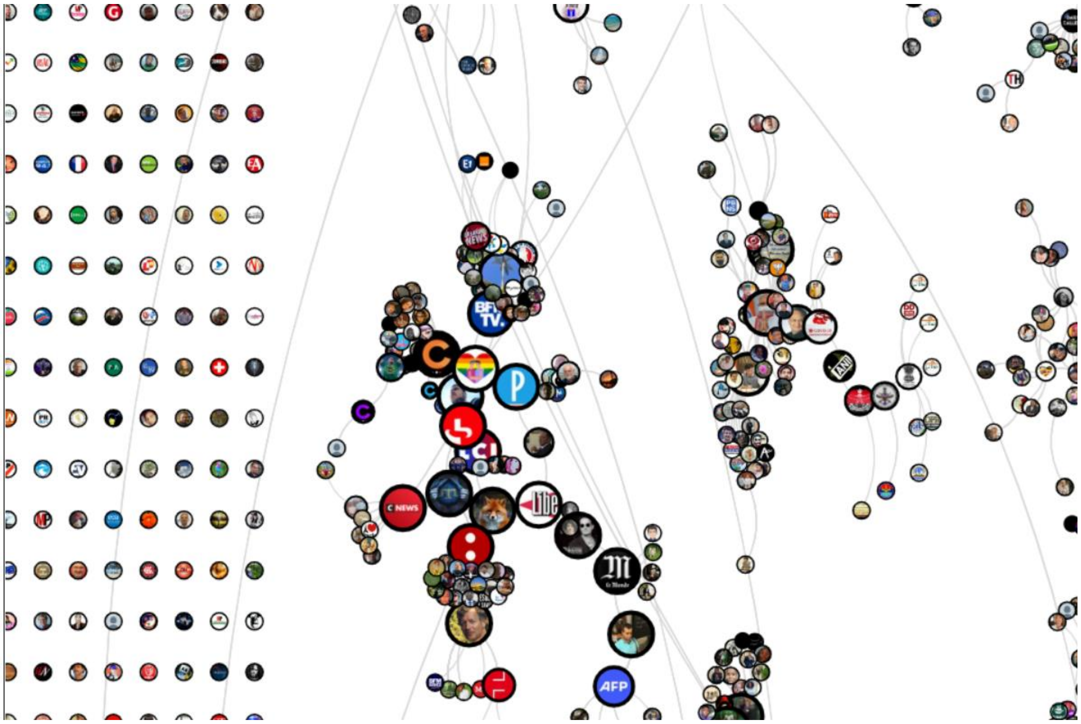
#climatechange_2021-12-...

jeremyhl Twitter NodeXL...

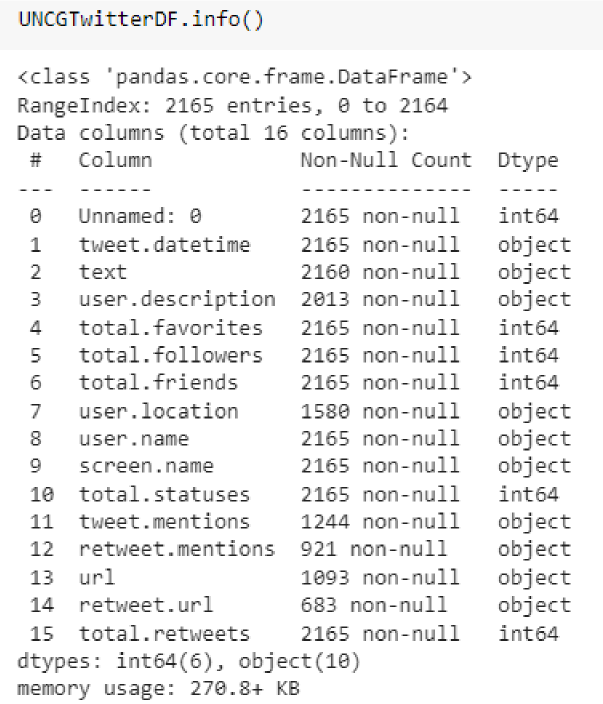
jeremyhl Twitter NodeXL...


G1: #hcldr health infoway hcldr abnermason colin_hung care people t3 digital
G3: #hcldr infoway health hcldr colin_hung abnermason t2 t1 week jamesian
G4: #hcldr #pinksocks joebabaian week happy thanksgiving colin_hung thx recent follows
G5: health #hcldr social register #healthmanagement care support hcmronline case managers
G7: #hcldr colin_hung radar everyone leahlizarondo foodrescuehero lgin412 abnermason less vikbakhru
G8: going such exciting conference please come present shawnkdowling roselikethefiwr tchanmd
G9: rsna #hcldr knew week fun #gogru #rsna21 #radres #radiology #flyfishing
G2: #hcldr jameyedwards health #digitalhealth great time such greggmastersmph access #covid19
G10: commit both spirit action safety patients families workforce essential...
G16: health care workers long covid being...
G17: #hcldr #medtwitter three stages expertise...
G18: let's connect prorelix research...
G13: interview hiteawithgr... ready...
G14: healthcare #hcldr icymi gt harlow...
G26: next semester teach access...
G31: hiring looking skilled...
G30: next vol using design thinking...
G29: heard iehp_he... expandi...
G11: getting whiplash nodding hard comments ericdishm...
G15: help medstartr #moderator sandirw17 medi...
G25: great convo illustriou...
G19: digital health week...
G24: phenoty... polycyth... hypervol...
G23
G6: #hcldr #healthit more #healthcare leaders #hitsm healthcare market looking forward
G21: use nicotine cigarettes contact...
G22: offering more 1000 schola...
G36: 2022 call propo...
G39: physi... citizen health...
G38: team #hcldr #grati...
G12: gt #cchleade... #hcldr #cdnhealth share
G28
G20: #medtwi... ready nfta
G27: grahamw...
G37: quality
G33: came sacjai...
G35: passionat...
G32:
G34:

https://nodexlgraphgallery.org/Pages/InteractiveGraph.aspx?graphID=267938



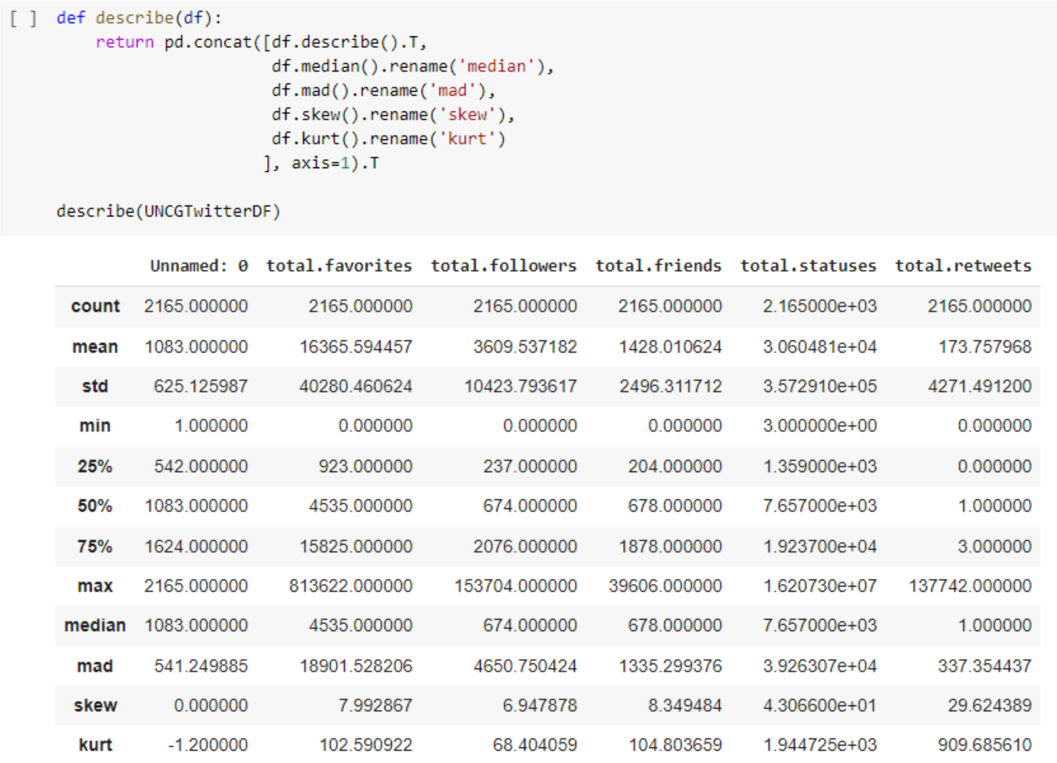
Exploratory Analysis: Below Python code shows 15 features and 2165 observations.It has 5 metrics with integer data type and rest of them are object type(String).



```
UNCGTwitterDF.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2165 entries, 0 to 2164
Data columns (total 16 columns):
 #   Column            Non-Null Count  Dtype 
---  ------            --------------  ----- 
 0   Unnamed: 0        2165 non-null   int64 
 1   tweet.datetime    2165 non-null   object
 2   text              2160 non-null   object
 3   user.description  2013 non-null   object
 4   total.favorites   2165 non-null   int64 
 5   total.followers   2165 non-null   int64 
 6   total.friends     2165 non-null   int64 
 7   user.location     1580 non-null   object
 8   user.name         2165 non-null   object
 9   screen.name       2165 non-null   object
 10  total.statuses    2165 non-null   int64 
 11  tweet.mentions    1244 non-null   object
 12  retweet.mentions  921 non-null    object
 13  url               1093 non-null   object
 14  retweet.url       683 non-null    object
 15  total.retweets    2165 non-null   int64 
dtypes: int64(6), object(10)
memory usage: 270.8+ KB
```

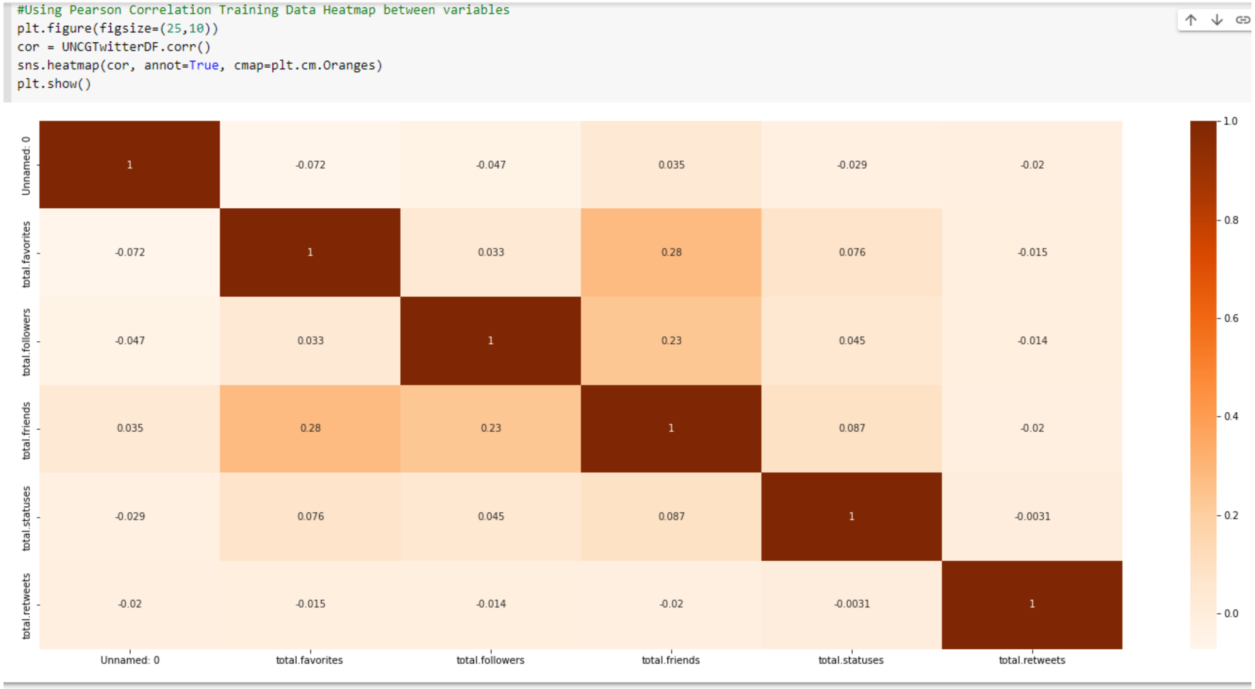
Following are the stats about twitter data set metrics with min, max, std, count, mean, mad, skew and kurt.



```
[ ]  def describe(df):
         return pd.concat([df.describe().T,
                           df.median().rename('median'),
                           df.mad().rename('mad'),
                           df.skew().rename('skew'),
                           df.kurt().rename('kurt')
                          ], axis=1).T

     describe(UNCGTwitterDF)
```

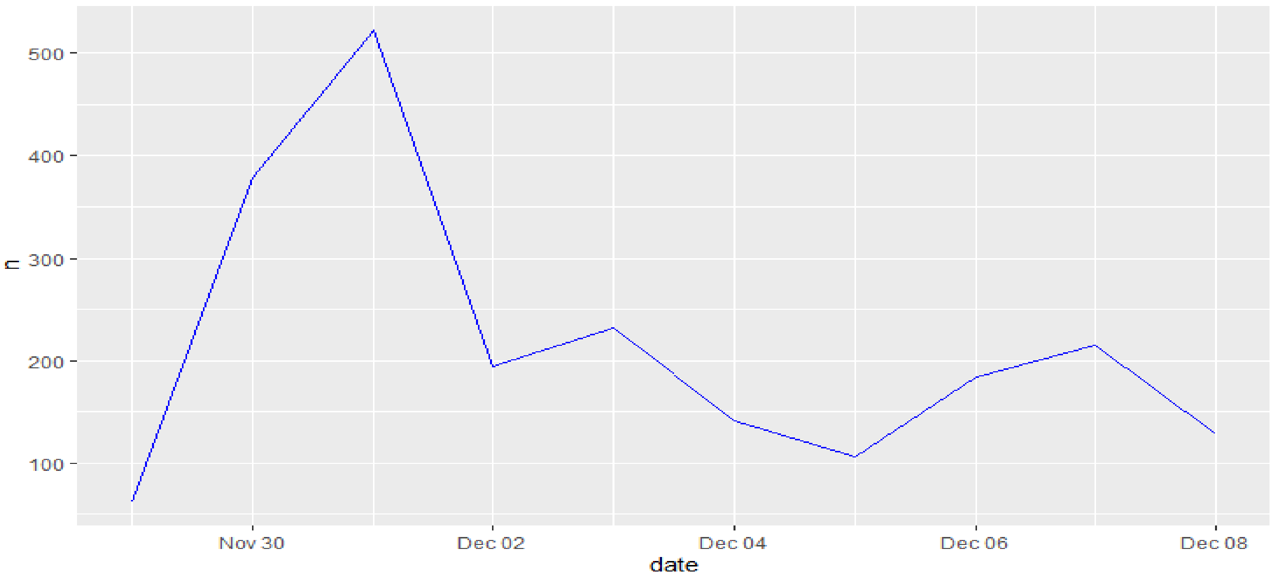
	Unnamed: 0	total.favorites	total.followers	total.friends	total.statuses	total.retweets
count	2165.000000	2165.000000	2165.000000	2165.000000	2.165000e+03	2165.000000
mean	1083.000000	16365.594457	3609.537182	1428.010624	3.060481e+04	173.757968
std	625.125987	40280.460624	10423.793617	2496.311712	3.572910e+05	4271.491200
min	1.000000	0.000000	0.000000	0.000000	3.000000e+00	0.000000
25%	542.000000	923.000000	237.000000	204.000000	1.359000e+03	0.000000
50%	1083.000000	4535.000000	674.000000	678.000000	7.657000e+03	1.000000
75%	1624.000000	15825.000000	2076.000000	1878.000000	1.923700e+04	3.000000
max	2165.000000	813622.000000	153704.000000	39606.000000	1.620730e+07	137742.000000
median	1083.000000	4535.000000	674.000000	678.000000	7.657000e+03	1.000000
mad	541.249885	18901.528206	4650.750424	1335.299376	3.926307e+04	337.354437
skew	0.000000	7.992867	6.947878	8.349484	4.306600e+01	29.624389
kurt	-1.200000	102.590922	68.404059	104.803659	1.944725e+03	909.685610

I also ran correlation plot between Total Retweets with other features. Its highly correlated with Statuses, friends, followers.

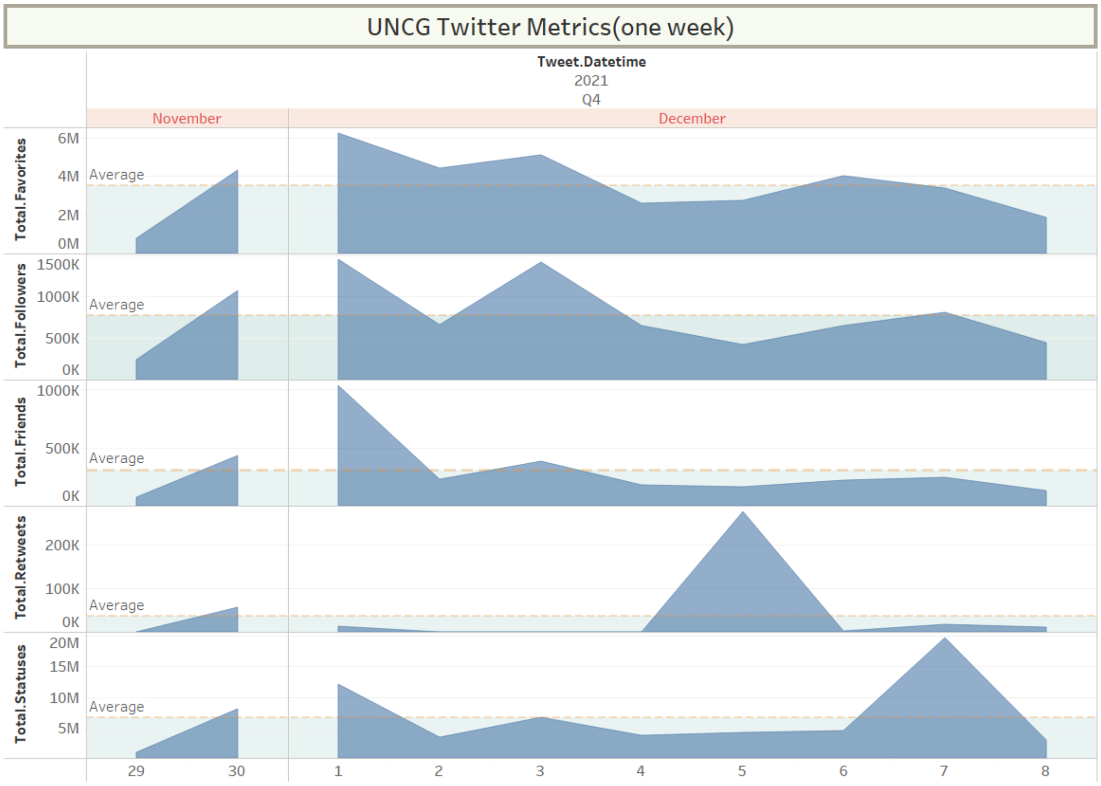


```
#Using Pearson Correlation Training Data Heatmap between variables
plt.figure(figsize=(25,10))
cor = UNCGTwitterDF.corr()
sns.heatmap(cor, annot=True, cmap=plt.cm.Oranges)
plt.show()
```

Below R Plot shows trend of UNCG Tweets for past one week. Dec 1st 2021 shows high tweets than other dates.



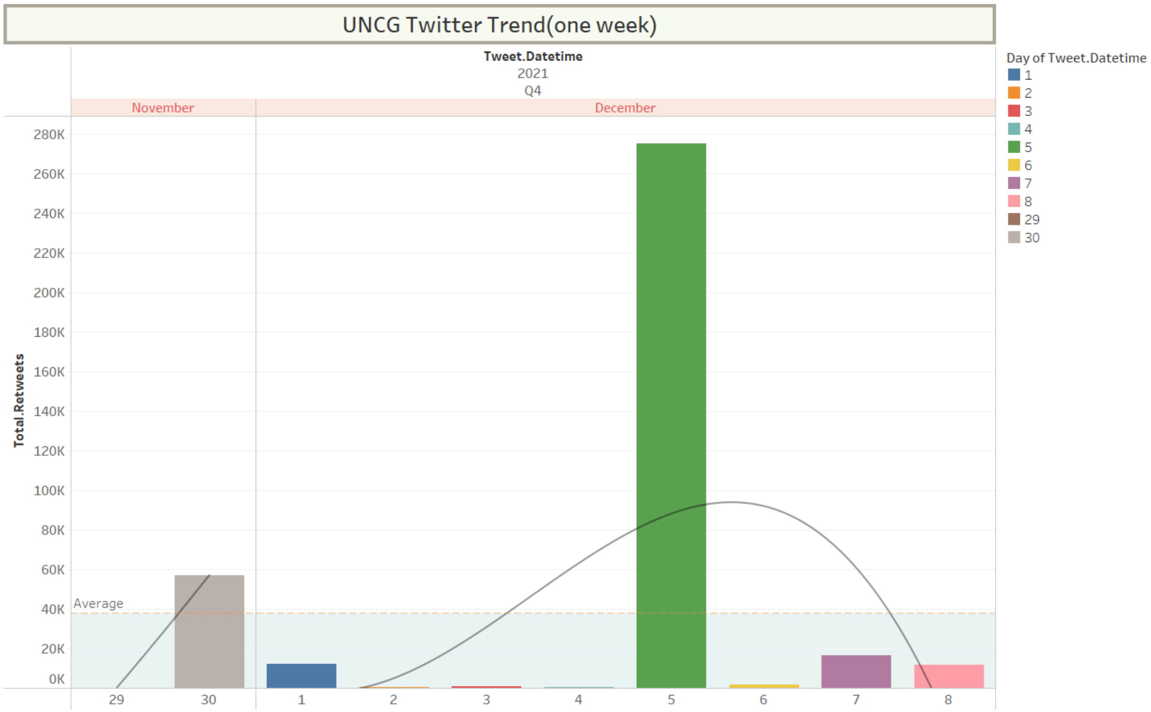
I also verified all metrics comparison using tableau tool, below shows comparison between UNCG Total Tweets with Status, Friends Followers & favorites.



Sum of Total.Favorites, sum of Total.Followers, sum of Total.Friends, sum of Total.Retweets and sum of Total.Statuses for each Tweet.Datetime Day broken down by Tweet.Datetime Year, Tweet.Datetime Quarter and Tweet.Datetime Month. The data is filtered on User.Name, which keeps 1,058 of 1,058 members.

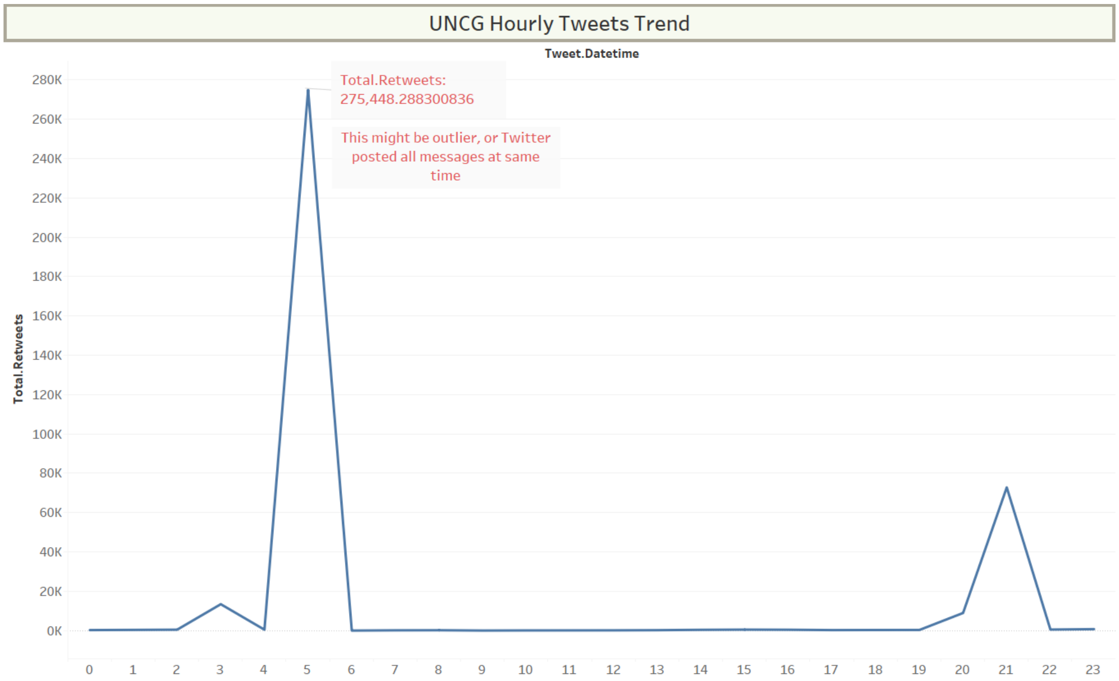
s

Day wise comparison: Below bar graph shows Tweets per day for past one-week (Nov/Dec 2021) UNCG twitter data. I see Dec 5th has many tweets posted, it might be due to System generated and outliers.


UNCG Twitter Trend(one week)

Sum of Total.Retweets for each Tweet.Datetime Day broken down by Tweet.Datetime Year, Tweet.Datetime Quarter and Tweet.Datetime Month. Color shows details about Tweet.Datetime Day.

Hourly comparison: At 5AM ET, it shows high tweets but strongly suspects its outlier. Next most tweets happened between 7PM to 11PM ET.

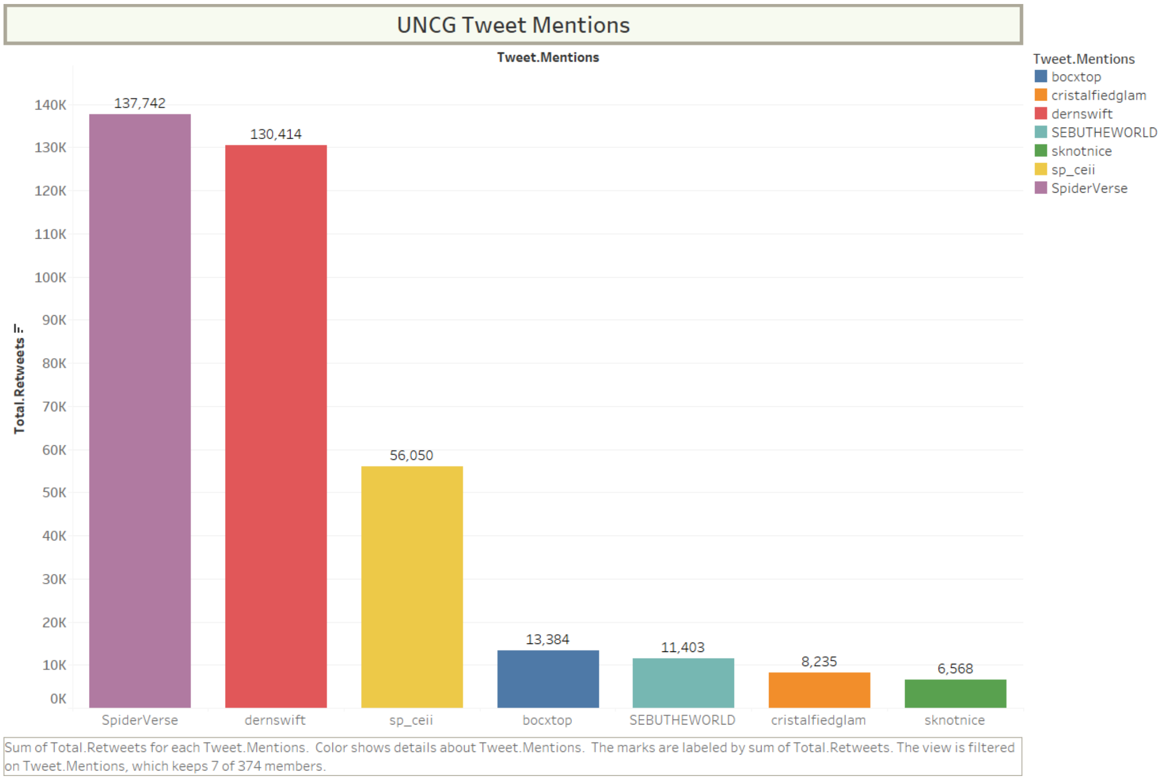

UNCG Hourly Tweets Trend

Total.Retweets: 275,448.288300836

This might be outlier, or Twitter posted all messages at same time

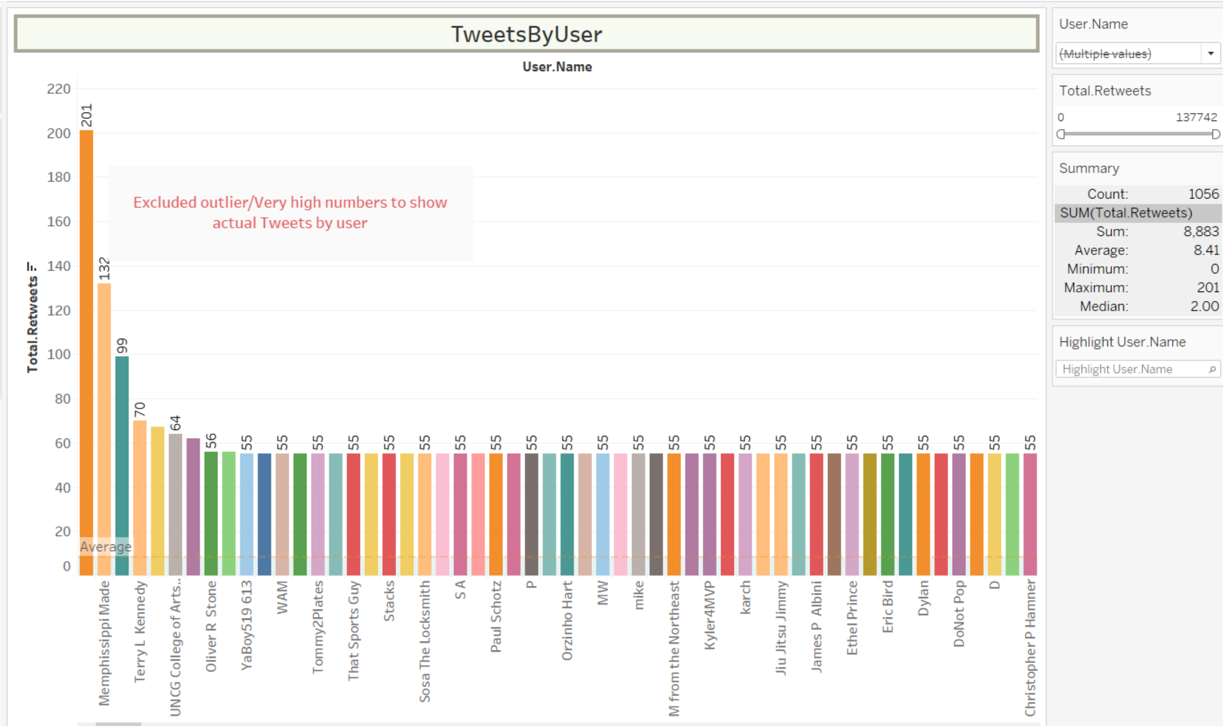
The trend of sum of Total.Retweets for Tweet.Datetime Hour.

Tweets Mentions vs count Comparison: SpiderVerse & Dernswift has highest tweet counts than others.



Sum of Total.Retweets for each Tweet.Mentions. Color shows details about Tweet.Mentions. The marks are labeled by sum of Total.Retweets. The view is filtered on Tweet.Mentions, which keeps 7 of 374 members.

Tweets By User Comparison**:** Below Tableau reports shows Nabil & Memphis, UNCG Athletics tweeted more than other users.



User.Name

~~(Multiple values)~~

Total.Retweets

0 137742

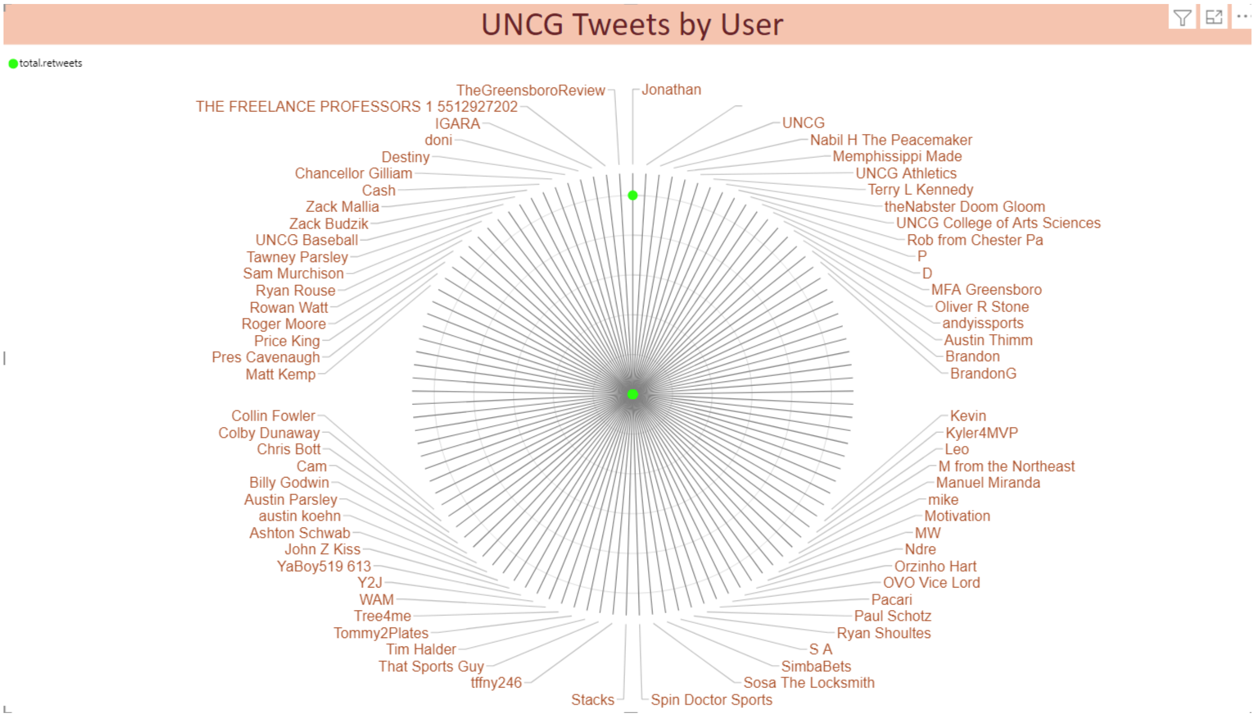
Summary

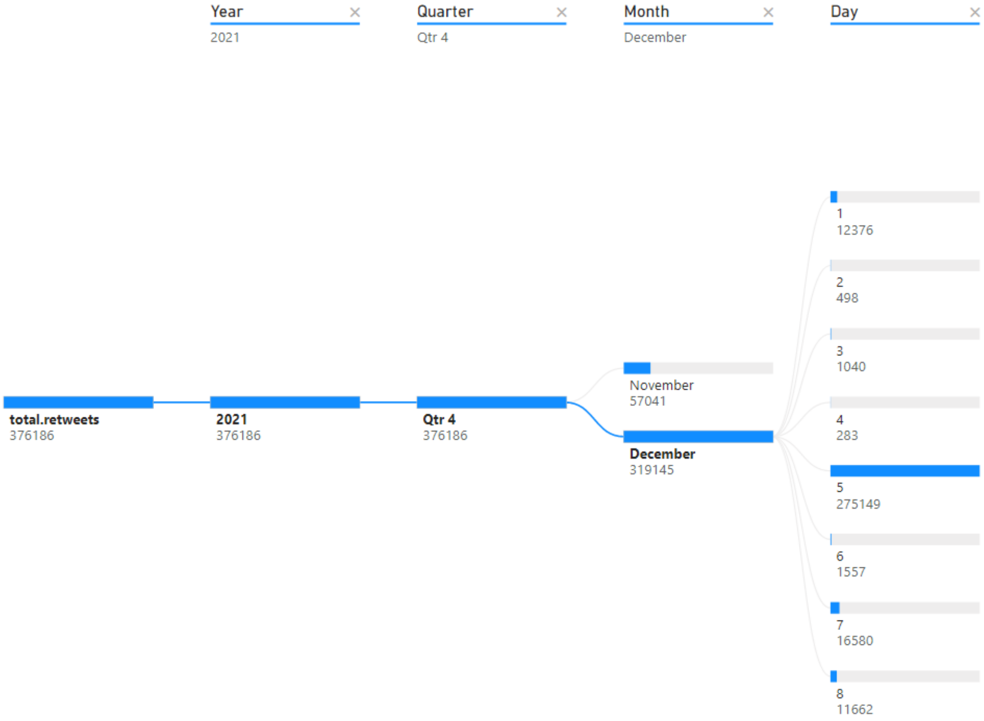
Count:	1056
SUM(Total.Retweets)	
Sum:	8,883
Average:	8.41
Minimum:	0
Maximum:	201
Median:	2.00

Highlight User.Name

Highlight User.Name

Similar Tweets by user in different view using Power BI radar view, Jonathan & UNCG are seems outliers.




Year ✕ 2021

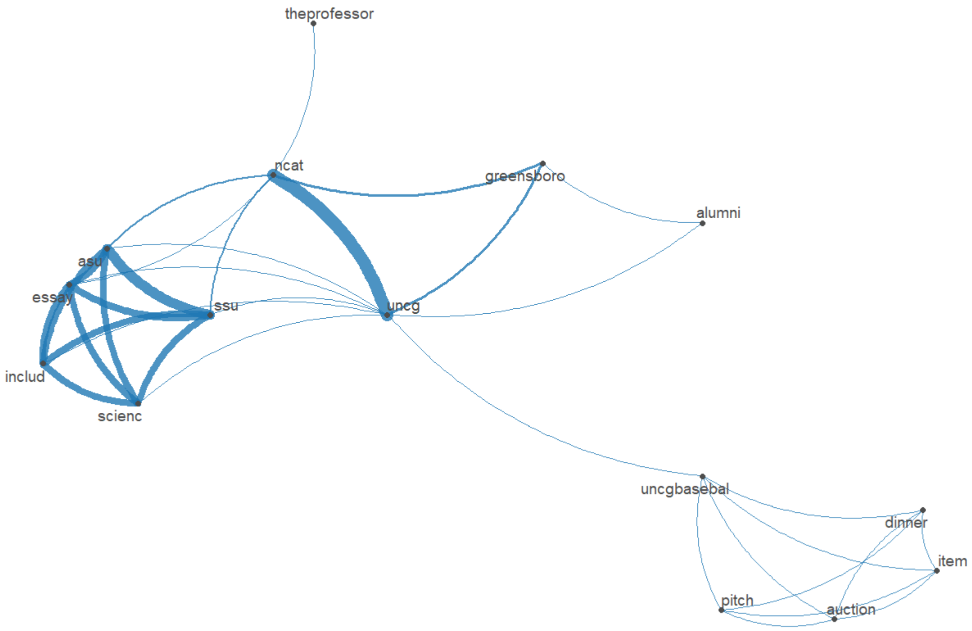
Quarter ✕ Qtr 4

Month ✕ December

Day ✕

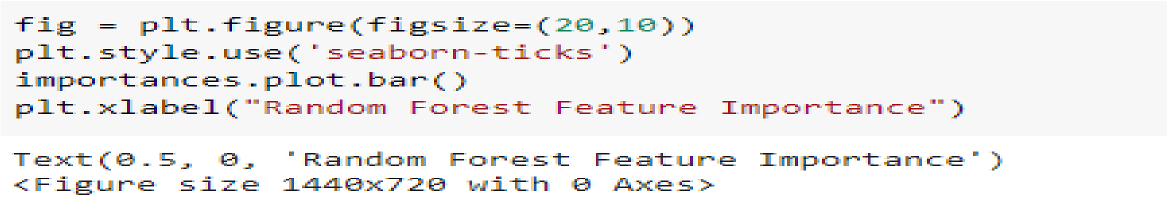
Results:

I have included parameters based on above exploratory analysis, Below R graph shows co-occurrence relationships for mentioned usernames, it’s using quanteda R package text plot.It shows as Decentralized(B) UNCG network.



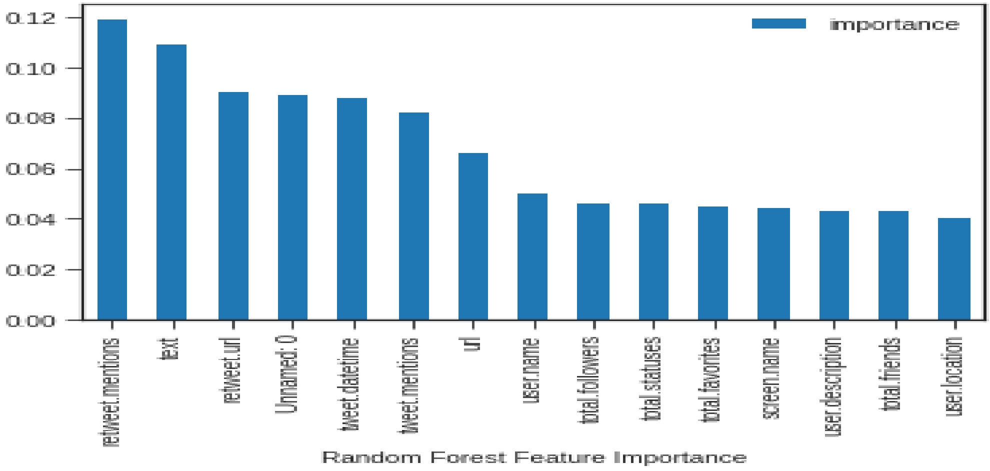
Machine Learning:

I have taken Randomforest Classifier Importance feature extraction for UNCG Tweets,, it shows that Mentions, text, URL, Date & Time are highest importance features to predict Tweets count.

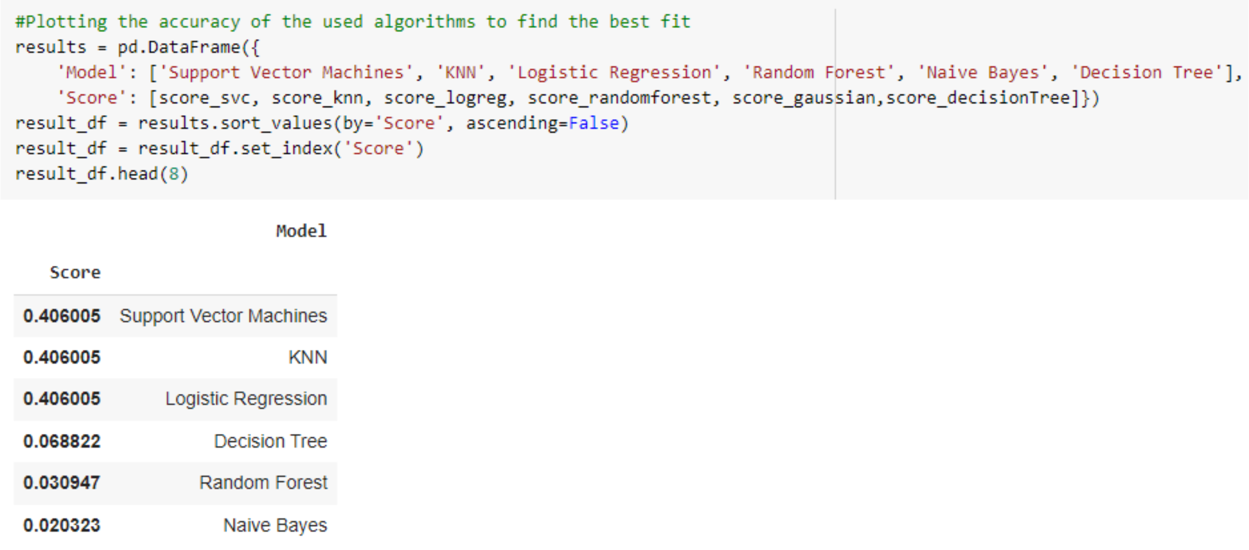


```
fig = plt.figure(figsize=(20,10))
plt.style.use('seaborn-ticks')
importances.plot.bar()
plt.xlabel("Random Forest Feature Importance")
```

```
Text(0.5, 0, 'Random Forest Feature Importance')
<Figure size 1440x720 with 0 Axes>
```



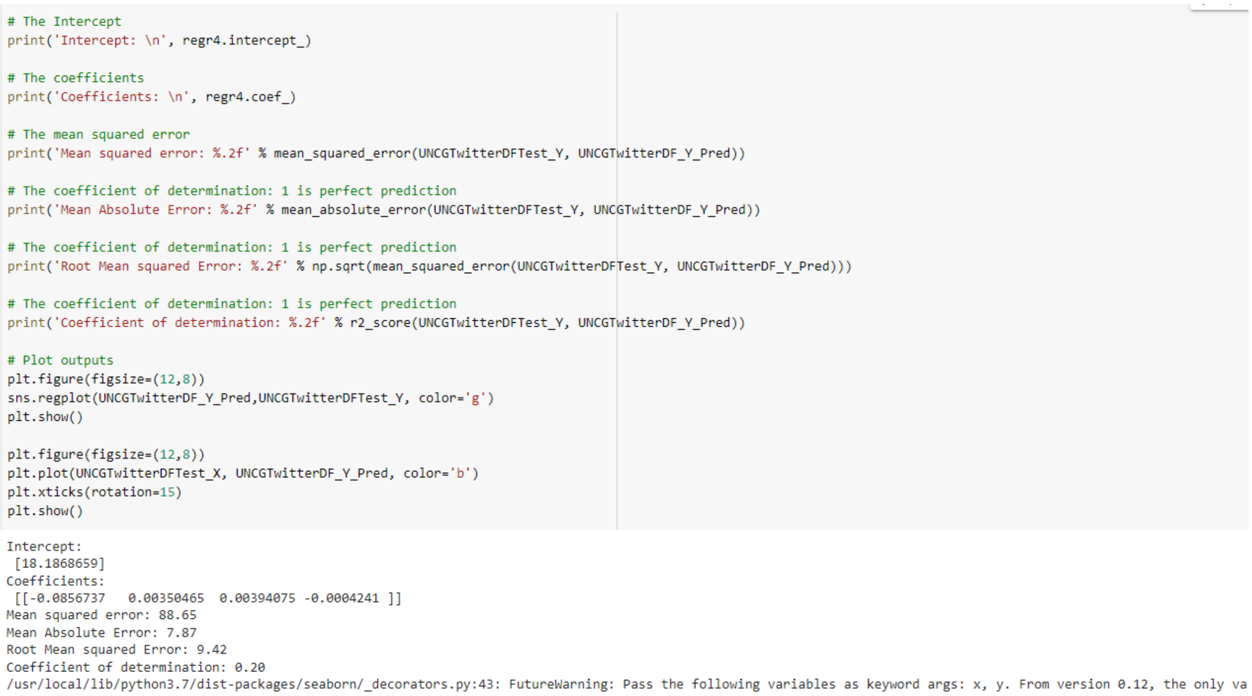
One of my research questions related to Tweets classification, which I ran with below Machine Learning classifiers, which shows 40% accuracy, this might be due to text data.



```
#Plotting the accuracy of the used algorithms to find the best fit
results = pd.DataFrame({
    'Model': ['Support Vector Machines', 'KNN', 'Logistic Regression', 'Random Forest', 'Naive Bayes', 'Decision Tree'],
    'Score': [score_svc, score_knn, score_logreg, score_randomforest, score_gaussian,score_decisionTree]})
result_df = results.sort_values(by='Score', ascending=False)
result_df = result_df.set_index('Score')
result_df.head(8)
```

	Model
Score	
0.406005	Support Vector Machines
0.406005	KNN
0.406005	Logistic Regression
0.068822	Decision Tree
0.030947	Random Forest
0.020323	Naive Bayes

And executed Linear regression with correlated & highly statistically significant variables to predict tweet count. Below model shows RMSE=9.42, MSE=88.65, MAE=7.87.



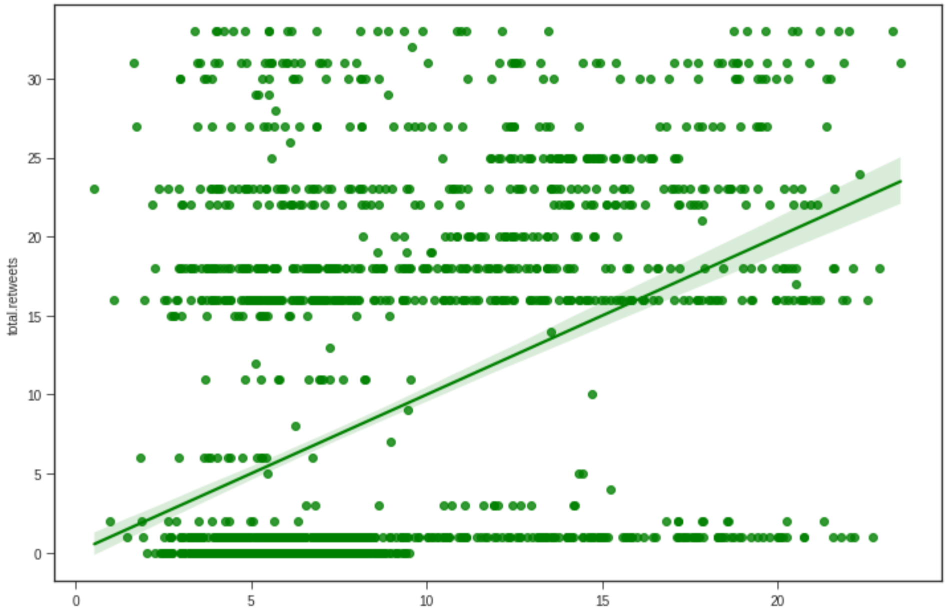
```
# The Intercept
print('Intercept: \n', regr4.intercept_)

# The coefficients
print('Coefficients: \n', regr4.coef_)

# The mean squared error
print('Mean squared error: %.2f' % mean_squared_error(UNCGTwitterDFTest_Y, UNCGTwitterDF_Y_Pred))

# The coefficient of determination: 1 is perfect prediction
print('Mean Absolute Error: %.2f' % mean_absolute_error(UNCGTwitterDFTest_Y, UNCGTwitterDF_Y_Pred))

# The coefficient of determination: 1 is perfect prediction
print('Root Mean squared Error: %.2f' % np.sqrt(mean_squared_error(UNCGTwitterDFTest_Y, UNCGTwitterDF_Y_Pred)))

# The coefficient of determination: 1 is perfect prediction
print('Coefficient of determination: %.2f' % r2_score(UNCGTwitterDFTest_Y, UNCGTwitterDF_Y_Pred))

# Plot outputs
plt.figure(figsize=(12,8))
sns.regplot(UNCGTwitterDF_Y_Pred,UNCGTwitterDFTest_Y, color='g')
plt.show()

plt.figure(figsize=(12,8))
plt.plot(UNCGTwitterDFTest_X, UNCGTwitterDF_Y_Pred, color='b')
plt.xticks(rotation=15)
plt.show()
```

```
Intercept:
 [18.1868659]
Coefficients:
 [[-0.0856737   0.00350465  0.00394075 -0.0004241 ]]
Mean squared error: 88.65
Mean Absolute Error: 7.87
Root Mean squared Error: 9.42
Coefficient of determination: 0.20
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning: Pass the following variables as keyword args: x, y. From version 0.12, the only va
```

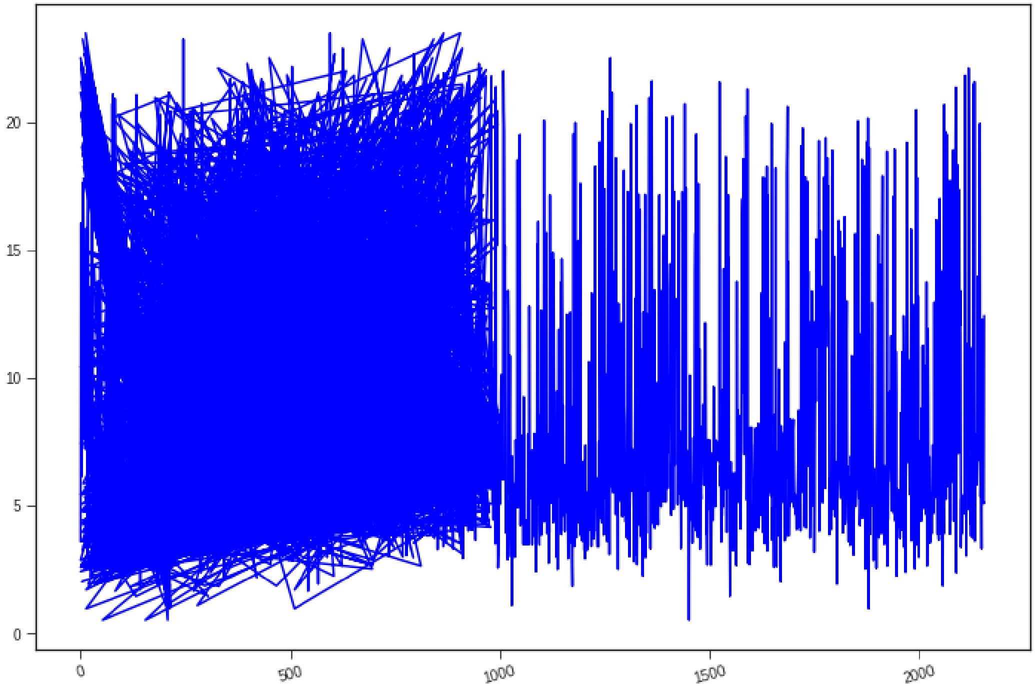
Linear Regression Formula for Tweets prediction:

UNCG Tweet(Y|X)= 18.17-0.085Mentions+0.004Text-0.0005URL

The following graph shows Prediction vs Actual data comparison accuracy, both are linearly correlated.

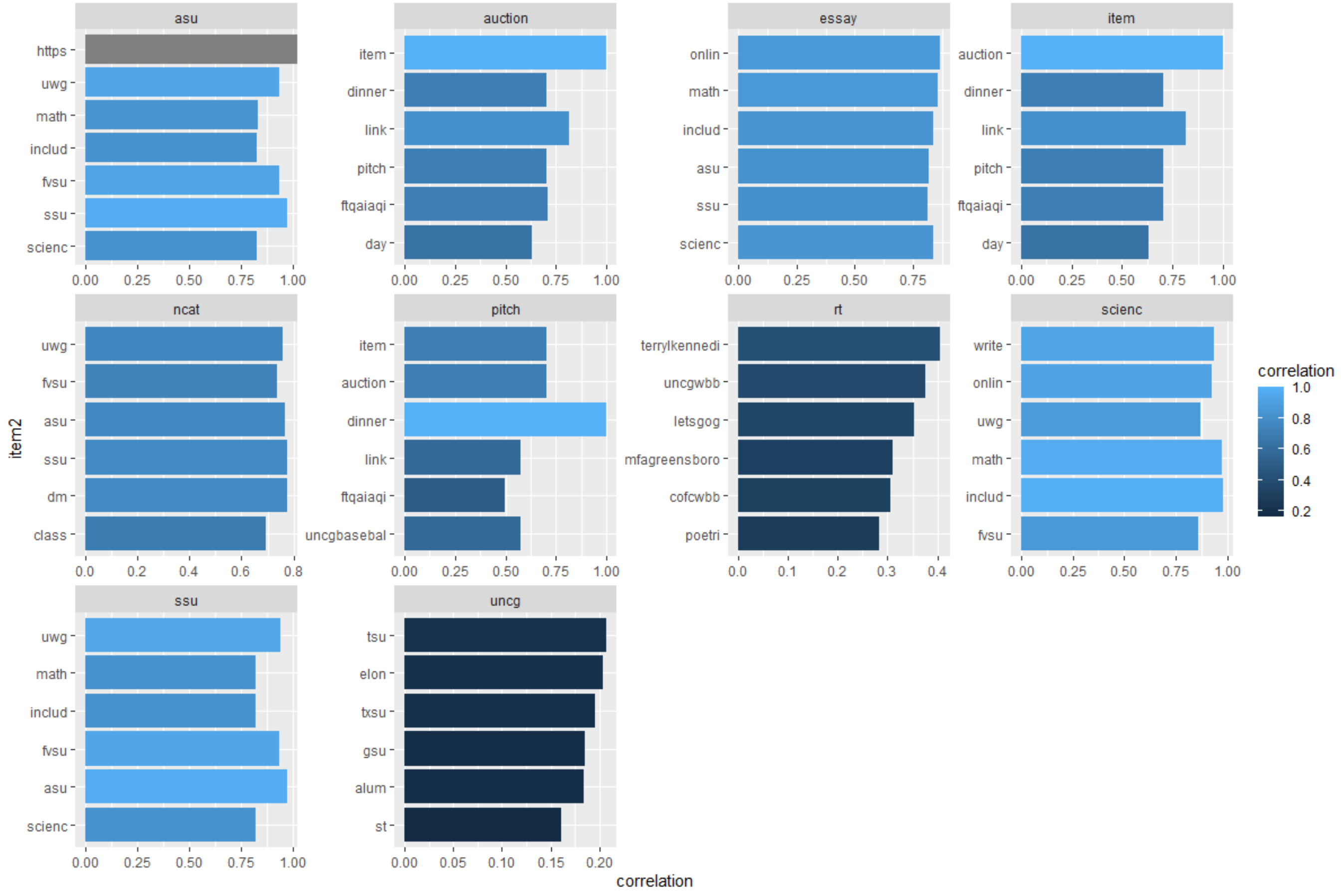


Below graph shows Test vs predicted values by Linear Regression model.

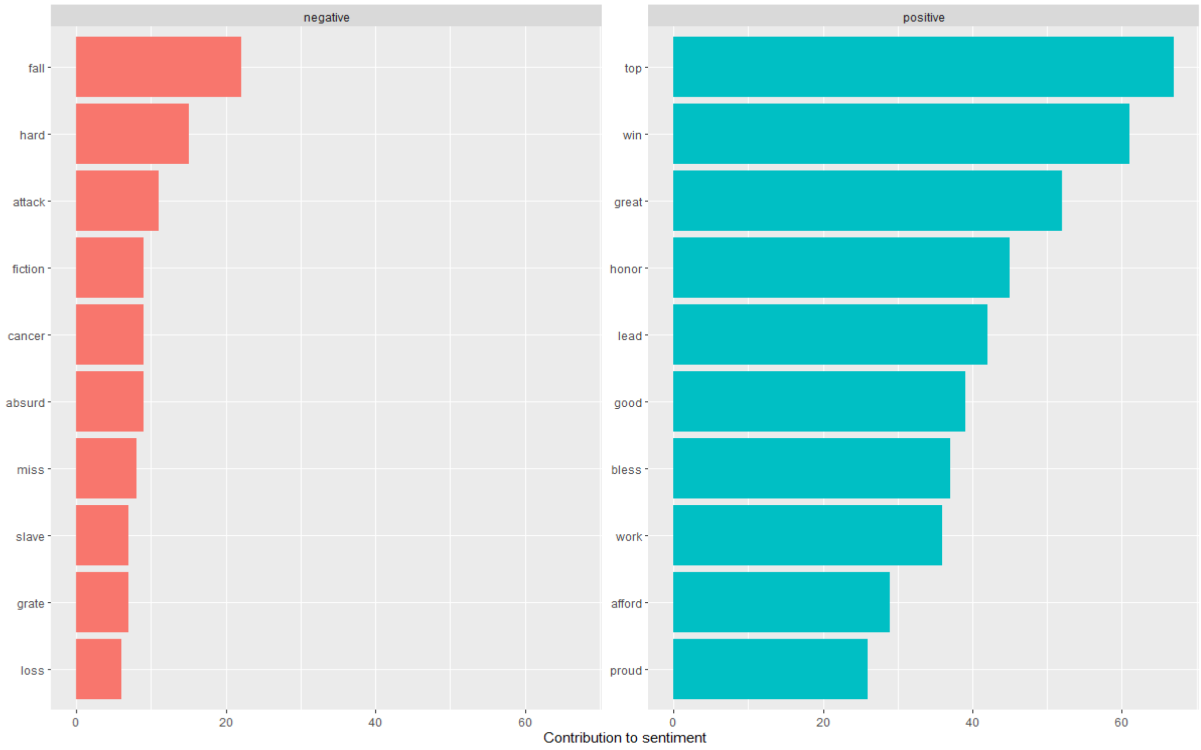


Sentiment Analysis Results:

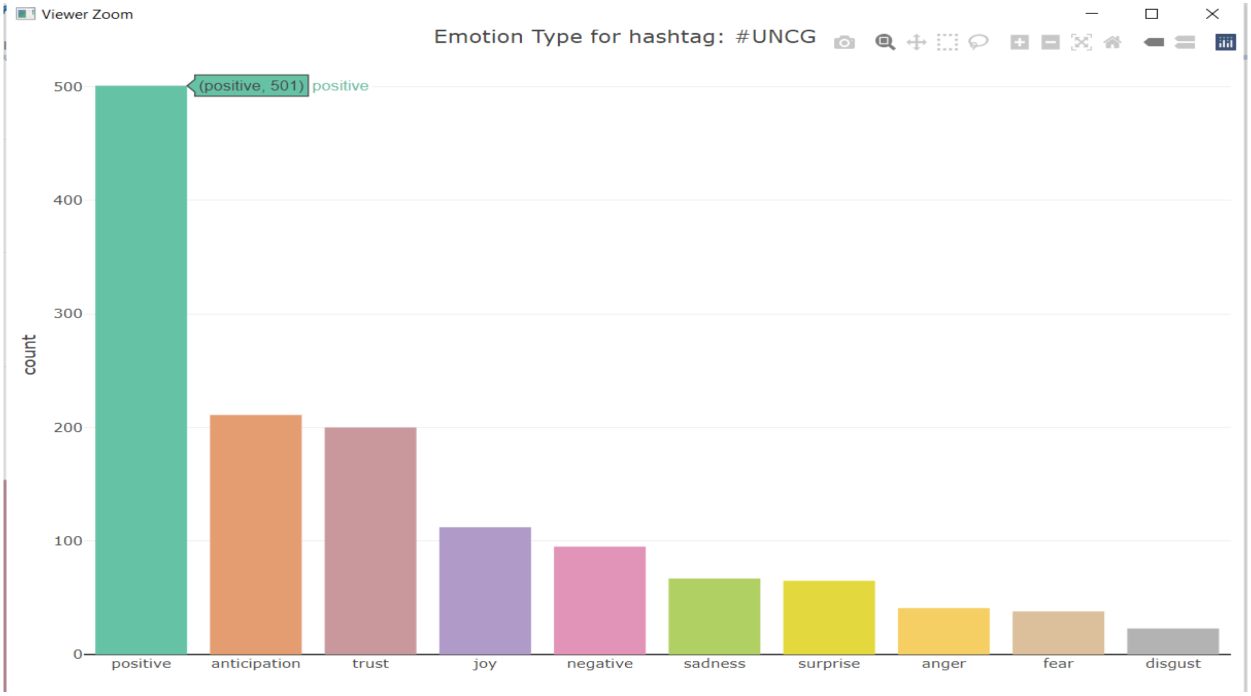
Below R plot shows correlation comparison between each word, example: asu highly correlated to tweet contains ssu, uwg, science words. Similarly Pitch with Dinner, Item with Auction..etc



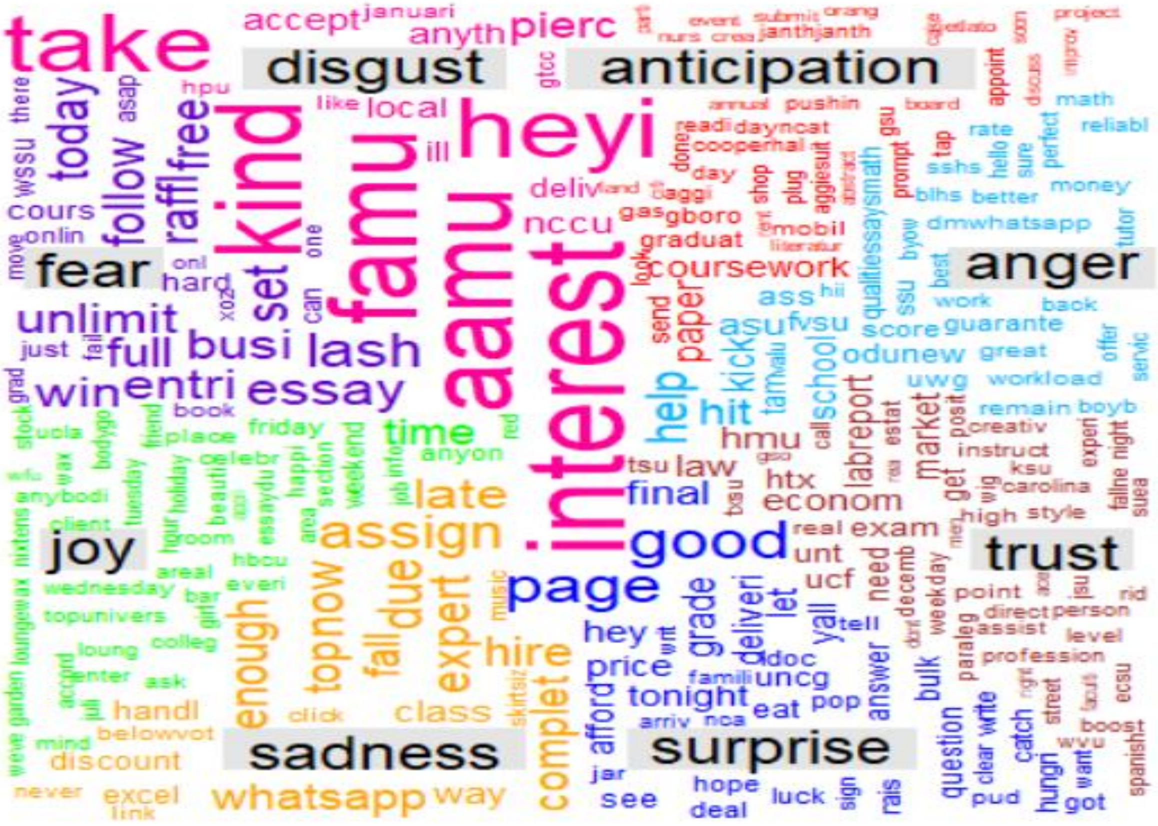
I ran sentiment analysis with UNCG twitter data with R tdm, tm packages. It shows below positive vs negative comparison. Most of words related to soccer game.I observe UNCG tweets are more positiveness than negativity in Tweets.



I ran below Emotion type for hashtags #UNCG tweets using R NRC data dictionary words. It shows Positive with 501 count, anticipation & trust ~200 words. This shows that UNCG tweets are clearer & more perfect.



With word cloud package in R, I have executed UNCG tweets word comparison with max 500 words. Below results shows that UNCG highly uses in tweets based on size of word.



Conclusion:

Github URL for code: https://github.com/BalaMallampatiGIT/IAL621_TwitterProject

I learnt very new subject social media network analysis using NodeXL & how we can utilize twitters tweets & connections contents data for organization (UNCG) growth prediction.

Massmine made my life very easy for twitter data extraction with simple commands, thanks to Dr Aaron.

UNCG tweets have more positivity and most of them related to games (Baseball, Soccer) & ssu(science, asu), uncg.

Though my models ultimately did not result in an effective way to classify and predict UNCG tweets data, the improvements I made to our ML model over time did significantly reduce the MSE loss. Given more time and resources to do further research and feature engineering, I believe I could retrain this model to better success.

I believe that I did not have enough of the contributing features that drive classification of UNCG Twitter dataset & not enough data. Additional data points year over year Tweet growths, seasonal trends of tweets, other modern bigdata techniques may have improved these model results.

References:

- Sklearn, Python library
- https://www.massmine.org/docs/twitter.html
- https://developer.twitter.com/en/portal
- https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/object-model/tweet
- https://towardsdatascience.com/3-super-simple-projects-to-learn-natural-language-processing-using-python-8ef74c757cd9
- https://nodexl.com/guides/how-to-install-nodexl-pro
- http://nodexlgraphgallery.org/Pages/Default.aspx
- How to open NodeXL file : https://alice.endicott.edu/examples/9