Analysis and comparison of Twitter networks of different universities

Haripriya Pandya (A20299643)

Shreesh Kumara Bhat (A20358074)

Introduction:

We are approaching the problem of analyzing & comparing different university Twitter networks by looking into their network properties and doing sentiment analysis on the network users' tweets to profile the university. We intend to do sentiment analysis of college students' tweets to classify and predict the different categories for the nature of their conversation.

Performing an analysis of college students's social media network can not only educate a lot about their social and professional activities, but specifically performing sentiment analysis on their conversation can helps us profile their topics of interest.

Data:

We collected followers of alumni association, career services and main university twitter pages for IIT and North Western University. Also, we collected the friend ids of these followers. We used REST and Streaming API to collect the tweets of user accounts to perform sentiment analysis.

Collected 6,565 follower accounts for IIT and 35,984 follower accounts for North Western University. Using those user accounts, we collected over 140,000 tweets inclusive of both universities.

Fields collected: User objects of follower accounts of 3 pages mentioned above, friend ids of these followers and their tweets.

Methods:

Network data:

While collecting data, if the user didn't authorize to collect their data, then we are assigning an error code of -1. For pages that does not exist anymore, we are assigning an error code of -2. For users having more than 50,000 friends, we are assigning an error code of -3.

While creating the graph, we add nodes that doesn't have any error code assigned to it. We also do this check while adding edges and we add edges, representing friend links, to those nodes which already exists in the graph.

Faced a bug in the code, which results in few nodes having zero degree, which cannot be the case since we are collecting followers of 3 accounts related to university and hence, their degree must be at least one. We tried re-running data collection for IIT network and updating the values. It successfully reduced the number of nodes having zero degree, but, the issue still persists. So, we are removing those nodes having zero degree from the network.

The IIT network after the above steps has 5,904 nodes and 62,891 edges. And, NU network has 31,053 nodes and 7,74,700.

Sentiment analysis data:

We chose 6 categories to classify the tweets, which are academics, sports, politics, technology, current_events and campus_events. Any tweet that would not fall under any one of these six categories would be labeled 'irrelevant'.

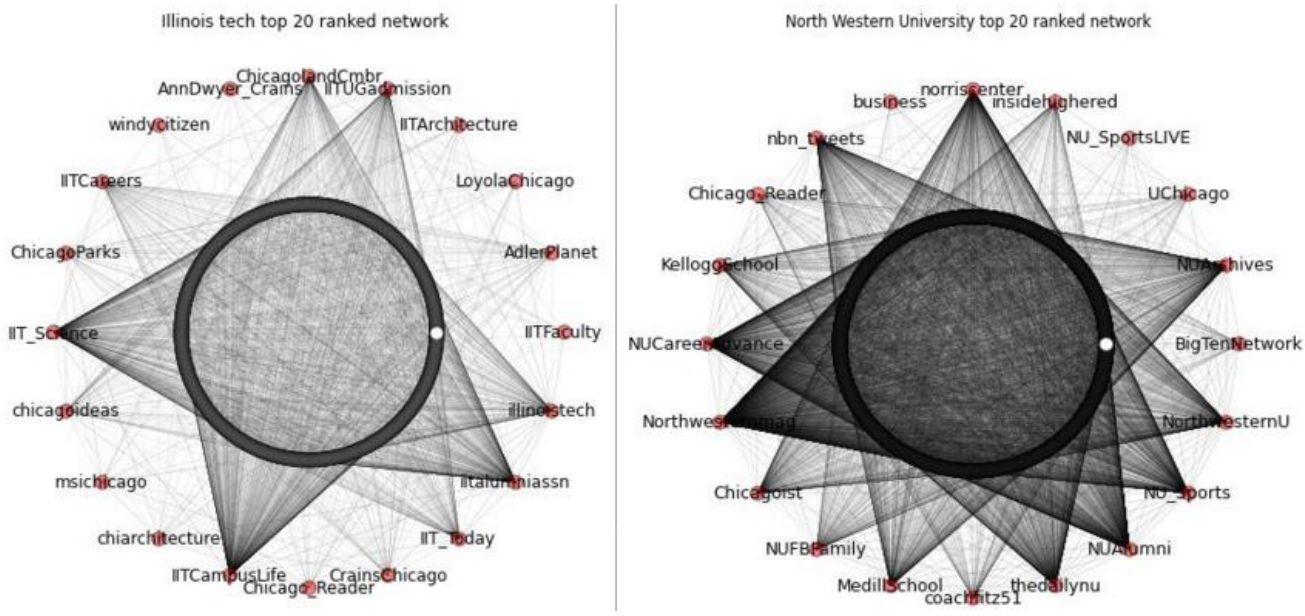
Out of the collected tweets, we labeled 1987 tweets including both universities and trained and fit a logistic regression model on it. A new logistic regression model was created which did not include the tweets labeled 'irrelevant'. The resulting number of tweets was 378. This shows a significantly less amount of relevant data for the study to train and fit the regression model.

Experiments

For network data:

We tried visualizing the entire graphs for both university but it wasn't providing fruitful results. Hence, we ran PageRank algorithm to select top 20 ranked nodes of the network and visualized their follower network using shell layout.

Here's a figure visualizing the top 20 ranked networks for both universities:

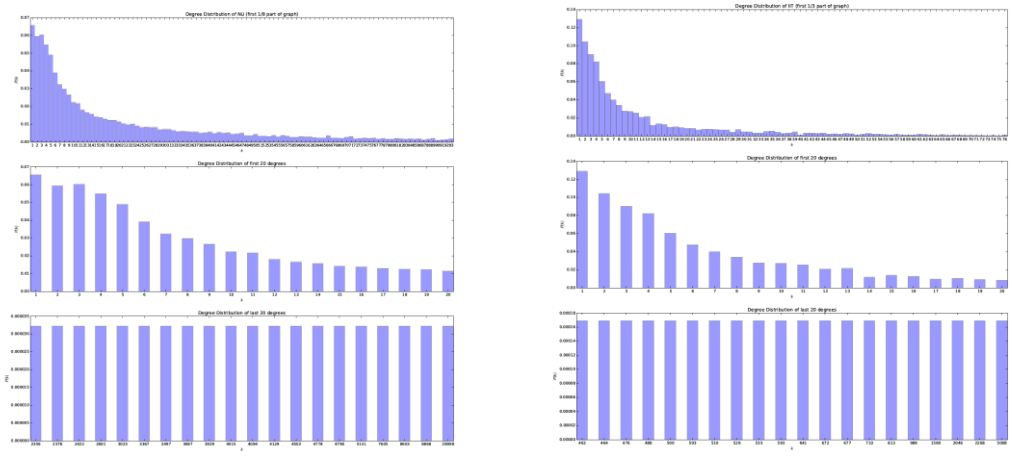


For IIT network, IIT Campus life, IIT Science, IIT Alumni assn, IIT univ page have high number of in-links.

And for NU network, it is NU Magazine, TheDailyNU, NU CareerAdvance, NU Alumni, NU Sports, NU univ page, NU Archives.

Next, we plot the degree distribution (indicating the probability that a randomly chosen node has a degree k) of both networks.

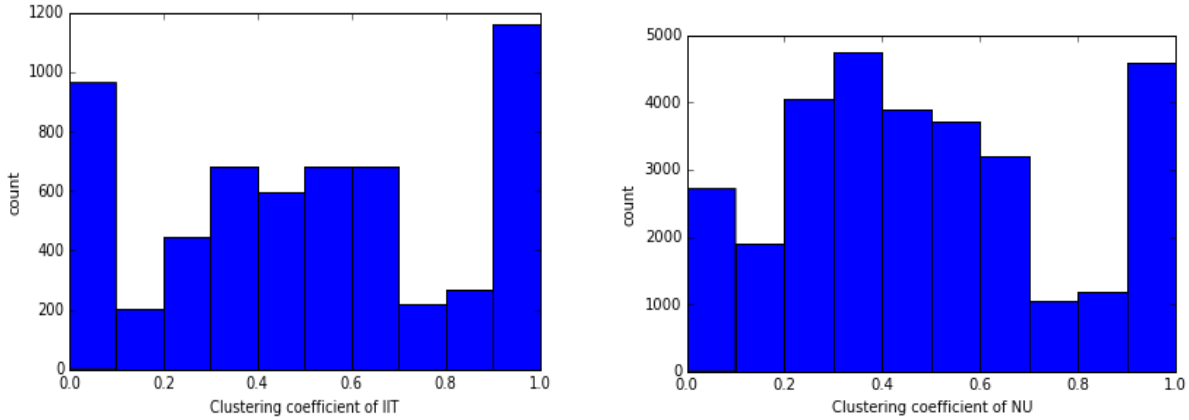
Note: The below 2 figures are PDF documents, representing the degree distribution of both networks, which can be double-clicked to view in actual size:



We can see long tailed graphs for both college networks, which should be the case for real social networks. It works as a sanity check too. Here, we see that NU network has more number of nodes having less degree distribution as compared to IIT network. And the last two values for NU network jumps from 8868 to 28098, indicating that there is this one account which has celebrity status.

Next, we plot the clustering coefficient (indicating the fraction of node's neighbors who are also neighbors) of both graphs.

Here's a figure visualizing the clustering coefficient for both universities:



We can see that IIT's network has higher ratio of, number of low clustering coefficient nodes to number of high clustering coefficient nodes, as compared to Northwestern University. This can mean that IIT has more number of diverse groups of users or NU users network better among themselves. The right side of the clustering coefficient graphs indicates that the accounts having high clustering coefficient have similar pattern across universities.

Note: The error code -3 assigned to the user, having more than 50,000 friends, while collecting data are not considered for this. It could have been a genuine student/faculty following many

accounts. Due to rate limits in Twitter API and time constraints for project, we had to ignore them. This could have affected NU clustering coefficient values drastically.

For sentiment analysis:

The 5-fold and 10-fold cross-validation accuracy for the model with irrelevant tweets was found 0.8153 and 0.8178 respectively. The 5-fold accuracy and 10-fold accuracy for the new model without irrelevant tweets was found 0.3889 and 0.4576 respectively.

This shows a significant lack of reliability from this training model, however the statistical data clearly shows that the amount of relevant data found was only almost 20% of the total labeled data. Hence The rest of the 80% of data is deemed irrelevant to this study.
Hence the old training model would have a very high accuracy for predicting relevant or irrelevant data, considering almost 80% of labeled data is irrelevant. This can be improved by increasing the amount of data collected and balancing the proportions of data labeled for each category.

We used rest of the unlabeled tweets for both universities to make predictions for those tweets based on the new model. Each university had almost 40,000 unlabeled tweets that were used to create a sparse matrix (CSR matrix) and prediction were made based on the newly trained model. The percentage of number of tweets predicted for each label category for individual university are as follows.
Hence according to above table the highest percentages of total tweets belong to category sports and politics and other categories have almost 10% of total number of tweets each. Moreover it shows that for an unlabeled tweet, it is most likely to be classified to a sports category. Among the two universities, Northwestern has greater chances of sports tweets.

This analysis is very biased towards the irrelevant data that was labeled and hence the feature vector created for the new regression model does not have proper feature elements representing individual category.

This model can be greatly improved by collecting a large amounts of data over a course of time, which would include data of different types and would improve the amount of relevant data collected to train and fit the regression model. Also the amount of labels and the selection of the labels also play an important role in gathering relevant data. For example during the labeling process, due to lack of career related category a great deal of tweets were considered irrelevant.

An explanation for skewed results for sports and political tweets could be found in the time period in which the tweets were collected. These tweets were collected when there were hockey and football games being played on weekends, hence there is a likely increased amount of internet chatter before and after such game events. In addition presidential debates were conducted when the tweets were collected. The labeled tweets also include a fair amount of tweets related to terrorist attack in Paris which can spike the political discussions and debates in social media. Due to holiday season, there was a fair amount of holiday related tweets collected and were labeled ‘irrelevant’ for this study which shows a great amount of overall irrelevant tweets for training purposes.

All this strengthens the point of lack of substantial tweets required for a good analysis and conclusion from the data. Although the widely popular nature of both sport and political categories show that the model is not completely wrong. It does show a commonly found interest in young adults. Especially the prediction of high percentage of sports tweets from Northwestern University, is aligned with the fact that it has a very popular sports team.

Related Work:
A sentiment analysis Depressive Moods of Users Portrayed in Twitter [1] conducted to make similar type of prediction included making prediction of depressed mood of a user based on their depressive nature of tweets. This study showed successful correlation between a user's depressive mood and their tweet.

Conclusions:
According to the trained and fit regression model with 45% accuracy, we can say that for any incoming unlabeled tweet from either IIT or NU, it's most likely to belong to sports category and if not then to politics category.

References:
1. Depressive Moods of Users Portrayed in Twitter. by Minsu Park, Chiyoung Cha and Meeyoung Cha. Jul 2012. National Research Foundation, Korea