Nitin Ranjan, 18BCE0272

CSE3021: Social and Information Network

Digital Assignment

Q1. Read through the following papers and summarise the work carried out in those papers.

a. Say It with Colors: Language-Independent Gender Classification on Twitter
b. TUCAN: Twitter User Centric ANalyzer.
c. A Case Study in Text Mining: Interpreting Twitter Data From World Cup Tweets

Q2. Using the following visualization softwares for social network

a. R Tools for Social Network Analysis or Gephi
b. Social Networks Visualiser (SocNetV)
c. Pajek

Visualize your own social network from Facebook.

Q1.

"Say it with colours: Language-Independent Gender Classification on Twitter" by Alowibdi et al.

The authors propose that colour used by a Twitter user on one's Twitter handle might be used to identify the gender of the user.

The colours are extracted from 5 parts of a user's twitter handle –

1. Background colour in profile picture and cover picture
2. Text colour
3. Link colour
4. Sidebar colour
5. Sidebar border colour

Algorithm:

The algorithm involved is –

Assumption: RGB is a 256 value paradigm (0-255).

1. Extraction of colours from the account
2. Reduction and quantization to reduce the number of colours i.e. to normalize all colours to a certain scale.
3. Conversion of the RGB values of the colours to HSV values (Hue, Saturation and Value).
4. Sorting the colours based on hue and value.
5. Labelling the sorted colours into groups of similar colours. It is assumed that colours with consecutive numbers when sorted are similar.
6. Reconversion of all colours in step 5 to their RGB values.
7. The RGB scale is reduced from 0-255 to 0-7. While this compression will change the colours themselves, it is assumed that the colours still maintain a minimum Euclidean distance from a set zero in the same order as obtained in step 4.
8. The RGB is again converted to HSV. This time, hue is the parameter used for sorting. If two given colours have the same hue, values is used as the parameter for sorting them.

A model was trained under the models of Probabilistic Neural Network, Decision Tree, Naïve Bayes and Naïve Bayes-Decision Tree hybrid.

The following results were obtained after the model was run –

1. Quantization of colours in algorithm step 7 helped to improve accuracy.
2. The Naïve Bayes-Decision Tree hybrid yielded the best results as compared to the other algorithms used in the training model.

3. However, given the fact that users have the ability to alter all the five components that were chosen for extracting colours, the model achieved a maximum accuracy of about 71%.

Achievements of the model –

1. An accuracy of about 71% was achieved in correctly predicting the gender of the user in the sample space.
2. The model is completely language-independent. As a matter of fact, the researchers used a users from about 36 different language pools to form their population of study.
3. The quantization is an effective method to reduce the overloading of the machine being trained as the number of possible colour combinations which is about 255*255*255 for a generic colour expressed in RGB and that value raised to the 5^{th} power for this algorithm (5 features, each with 255*255*255 colour combinations) to 8^{3+5} (5 features each with 8*8*8 colour combinations).

Drawbacks:

1. While the initial increase in the dataset increased accuracy, the researchers reached a bottle-neck at about 53,000 profiles.
2. The genders of the twitter users that were compared to the results predicted by the algorithm have been declared by the research team as Twitter has no official policy to ask the users to declare their gender.

18BCE0272

"TUCAN: Twitter User Centric Analyzer" by Luigi Grimaudo et al.

The authors propose a methodology to trace the pattern in which individual users raise topics through their tweets and then by extension, to study groups of similar users and to try and trace a collective common pattern of tweets from these groups. The researchers have dubbed a set of tweets from a given user over a small timeframe as a 'bird song'. Thus, the aim of the research article is to study the way in which the bird song changes and similarly how bird songs change over time – in order to analyse the pattern in which information was presented, its recurrence and the topics of interest.

***Algorithm*:**

1. Each tweet on twitter is time stamped. So, a target user is selected and a group of tweets retrieved from his/her/their timeline. The tweets are then stored in a repository.
2. A time period T is selected and the tweets extracted in step 1 are categorized into groups of tweets tweeted in a time period T. This is called a bird song.
3. Each bird song is then pre-processed. The pre-processing mostly involves removal of stop words. Stop words are words that do not add any meaning to the sentence for the machine. E.g. articles, conjunctions, interjections etc. Further, Twitter mentions are removed and the resulting songs are then processed through lemmatization (grouping together of similar words or the replacement of similar words by a common term) and ontological lexicon generalization.
4. The bird songs are converted into a bag-of-words, and correlation is established among all tweets in a bird song using TF-IDF approach. (There can be several tweets given the fact that a bird song is group of tweets posted in a given time window.)
5. A vector model is constructed with the terms of the bird song characterized by the score established in step 4.
6. The cosine similarity score between a given pair of bird songs is computed to establish correlation between two bird songs.
7. The results obtained in step 5 and 6 are visualized.

Conclusion:

The model established is a good method to understand the patterns in which user behavior varies on Twitter and similarly, to study the way in which public opinion varies on the platform. It also helps to establish semantic relationships between users with similar opinions and users with dissimilar opinions on Twitter.

The experiments that the researchers conducted is not very generic as the experimental sample was tweets on Obama and the White House, however, the results that the experiment yields enforce the conclusions thus presented.

18BCE0272

"A Case Study in Text Mining: Interpreting Twitter Data From World Cup Tweets" by Daniel Godfrey et al.

Text tweets are obviously the largest group of tweets. This particular article aimed to study cluster analysis on text tweets on the 2014 World Cup. Identifying that linguistic noise is inherent to text tweets, the model proposed uses a DBSCAN (a clustering algorithm with the basic assumption that high density regions in a graph are clusters while lower density regions are simply to separate these clusters) followed by a consensus matrix (a matrix representing the fraction of times the two elements in question are clustered) to reduce the noise. Finally, the article compares between k-means clustering and the non-negative matrix factorization clustering.

Algorithm:

1. Extract tweets that contain the word, "World Cup"
2. Remove the retweets.
3. Remove noise by removing stop words (e.g. a, an, the, he, she etc.) and carrying out stemming (removal of suffixes and prefixes). The authors further propose 4 algorithms for noise removal –
   i) Carrying out multiple k-means with varying values of k to establish outliers and removing them as 'noisy' points.
   ii) Carrying out multiple DBSCANs with varying minimum radius distance ε to establish points that are outliers to the clusters thus established and removing the points as noise.
   iii) Run DBSCAN on a consensus matrix. Remove outliers as noise.
   iv) Run the above three algorithms separately. If two of the three algorithms classify a noise.
4. Choose the number of topics the algorithm must extract from the tweets processed till now. The authors propose the number of topics to be represented by a matrix L that is related to the Consensus matrix C and the diagonal matrix D where each diagonal element is the sum total of row values of corresponding rows in D as
   L = D – C.
   The number of gaps between the Eigen values of L are the number of topics that lie in that range of consensus matrix. The user can choose a number of topics to suit their purpose.
5. Cluster the tweets using the consensus matrix as input and k-means as the algorithm.
6. Cluster the tweets using Non-negative matrix factorization with the same value of k as in step 5.

The authors conclude that NFM is a more efficient model to create clusters in their use case.

Advantages and achievements:

1. NMF is established to be better than k-means clustering as in k-means clustering, the client or the end-user has no reference to the cluster itself but only to the fact that a certain tweet/topic belongs to a certain cluster.
2. A very clear visualization of the concept of tweets grouping to form clusters, grouping to form topics is evident.

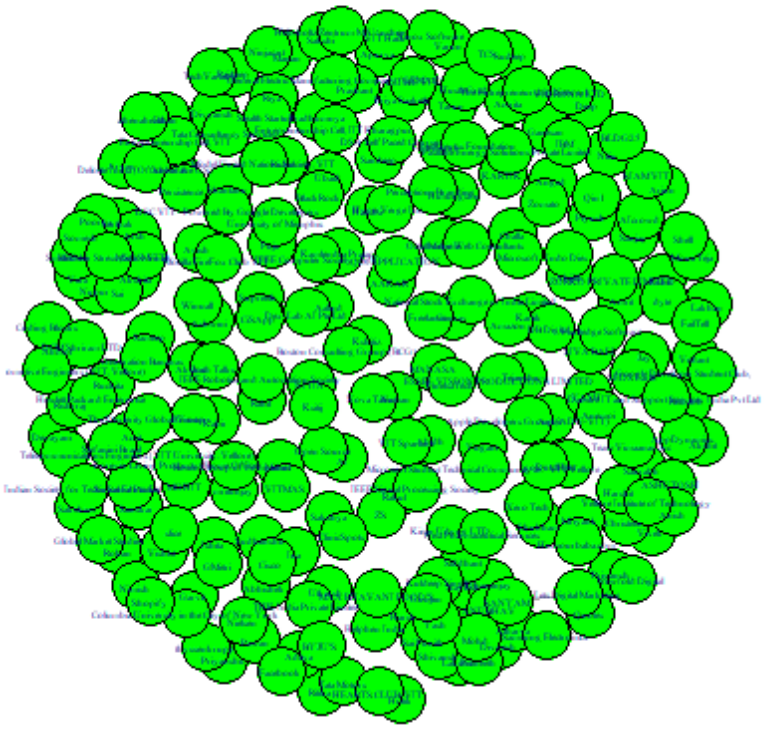
Drawbacks:

In algorithm step 3(ii) and 3(iii), there is an evident drawback on using DBSCAN with varying values of radius and varying density of data points. While the researchers have assumed that the radius is the only varying parameter, it is not the case when the analysis is of live tweets i.e. real - time tweets where the use of topics and words is fairly dynamic. It makes the algorithm unsuitable for modern use cases of live tweets analysis. The algorithm needs a set of Tweets pre-extracted to work.

18BCE0272

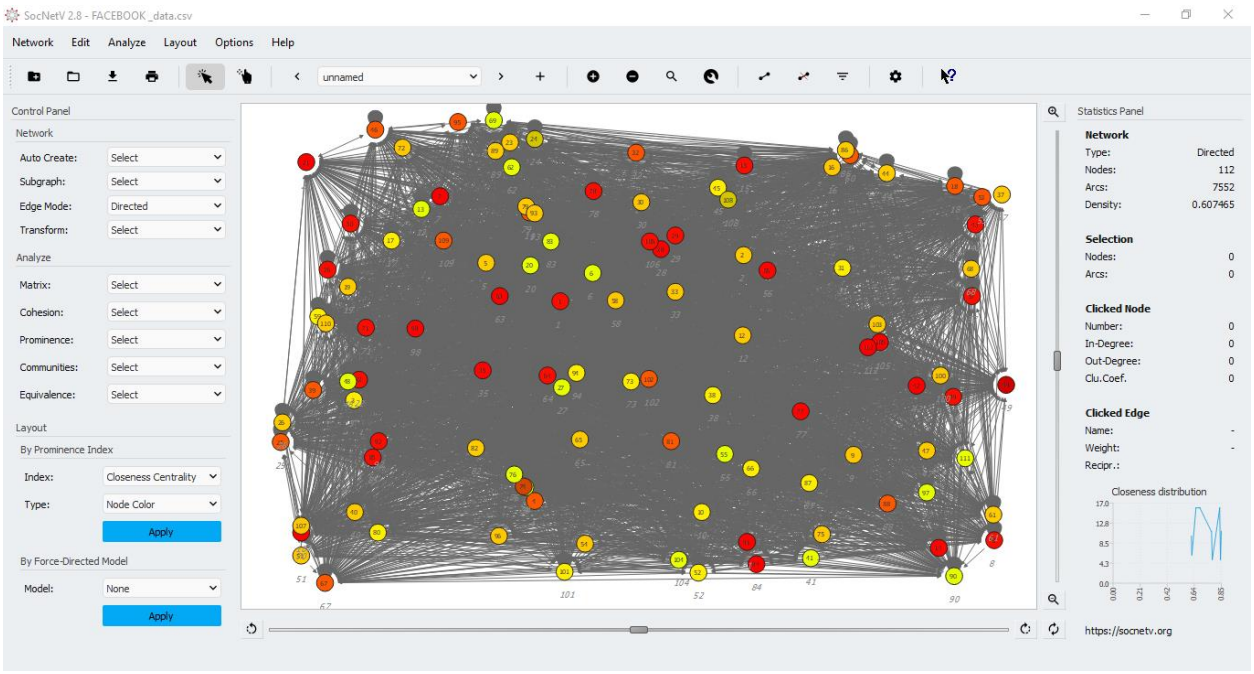
Q2.

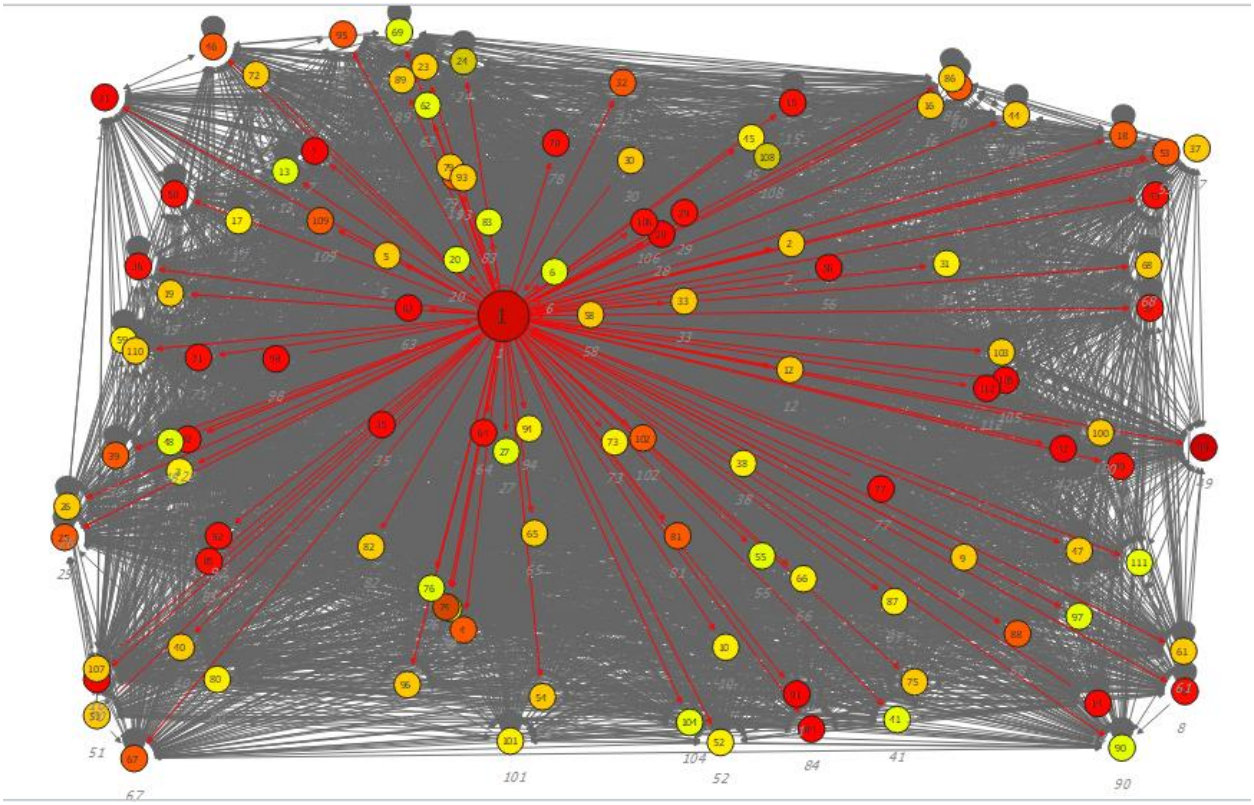
Facebook Network visualized in R using the igraph library. The network was extracted using Selenium crawler.



18BCE0272

The visualization of Facebook network through SocNetV. The network was extracted through Selenium web crawler.





18BCE0272

Pajek software was not compatible with my device. Hence, I did not use Pajek.

I also used Python and R to model my LinkedIn profile.

Just made this one visualization as well of my LinkedIn connections using Python.

Figure: Visualization of my network in terms of the company the people connected to me work in. (made in Python)

18BCE0272

Figure: Visualisation of my network in terms of the jobs that people connected to me are into. (made in Python)

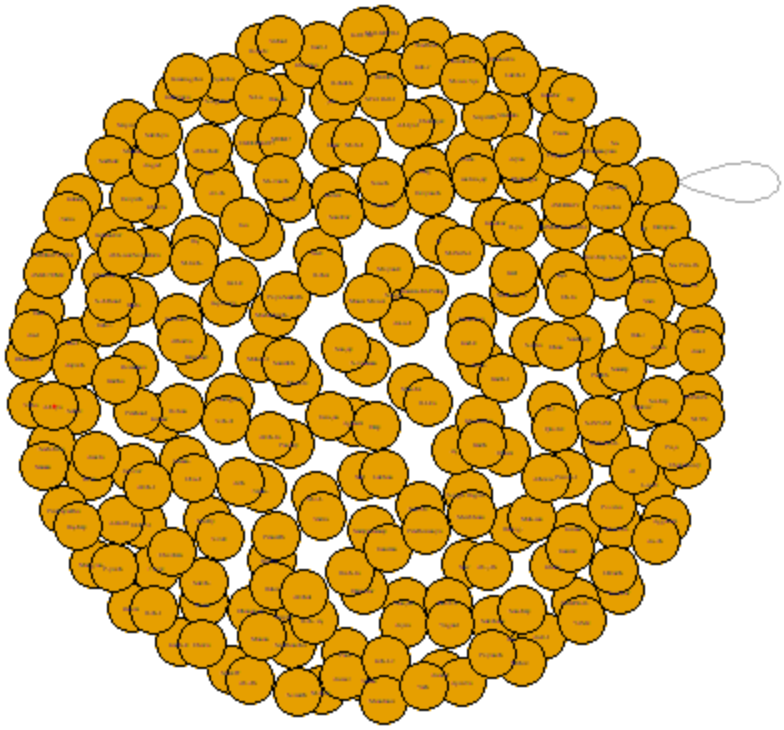


Figure: My LinkedIn Network expressed as names and jobs of connections in R

Submitted by Nitin Ranjan

18BCE0272

On

28th May 2021.