NETWORK ANALYSIS OF REVIEWS ON THE AIRBNB'S IN LOS ANGELES, CALIFORNIA, UNITED STATES

EM-623 DATA SCIENCE & KNOWLEDGE DISCOVERY
FINAL PROJECT

ANALYSED AND SUBMITTED BY:

SMIT MEHTA

TOOLS USED: MS EXCEL, WORDij and GEPHI

DATASET SOURCE: http://insideairbnb.com/get-the-data.html (Just Comment File)

STOPWORDS FILE SOURCE: Same file given by Dr. Carlo Lipizzi for HW-5.

The steps as per CRISP-DM are as below:

1. PROJECT GOALS AND CONDITIONS

- Our goal is to perform analysis on reviews.csv dataset containing reviews about Airbnb's in Boston, USA.
- This dataset contains 6 columns representing listing_id, Id, date, reviewer_id, reviewer_name and comments. This file basically contains reviews from every person who ever stayed at an Airbnb in Los Angeles and their respective reviews. We have a total of 651,939 records in the dataset.

2. BUSINESS UNDERSTANDING

- We need to perform network and text analysis on the .csv file and find relations and patterns between the issues or topics the people who ever stayed at an Airbnb in Boston are commenting about.
- In order to perform this, I cleaned the dataset and divided all the records into 5 buckets and performed network and text analysis on each of them using WORDij and Gephi. Through this I checked the modularity of clusters in each bucket and figured out that which cluster is more relevant and what does it say.

3. DATA UNDERSTANDING

- The only column that we are taking for our analysis is the 'comments' column.
- As we can see from the above screenshot of the .csv file, the last column doesn't need much cleaning except removing spaces and removing nulls (empty cells).
- When the data is cleaned, it needs to be divided into 7 buckets containing tentative equal number of reviews in them. ie: around 93134 records in each bucket.
- Then we need to analyse each bucket using WORDij and Gephi.

4. DATA PREPARATION

There is not much cleaning to do but still it needs a bit of refinement.

The steps for cleaning the data is as below:
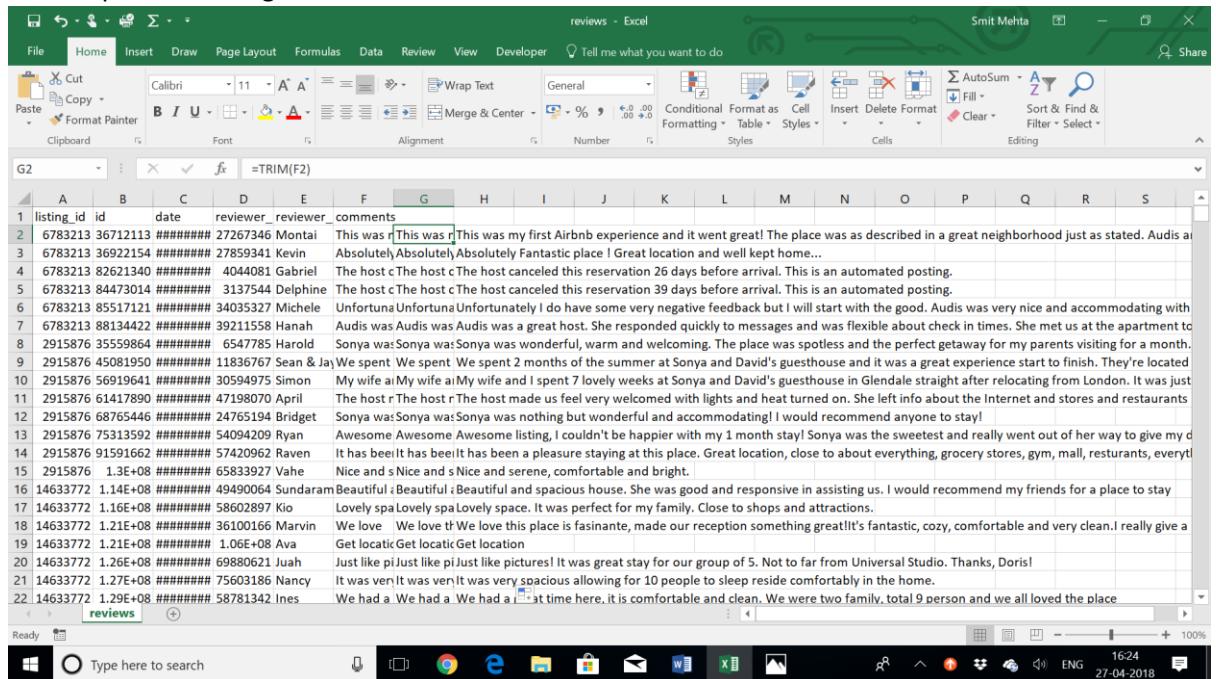
- We use the TRIM function to remove empty spaces from all the reviews in column F.
- So, column G represents all the trimmed comments/reviews.
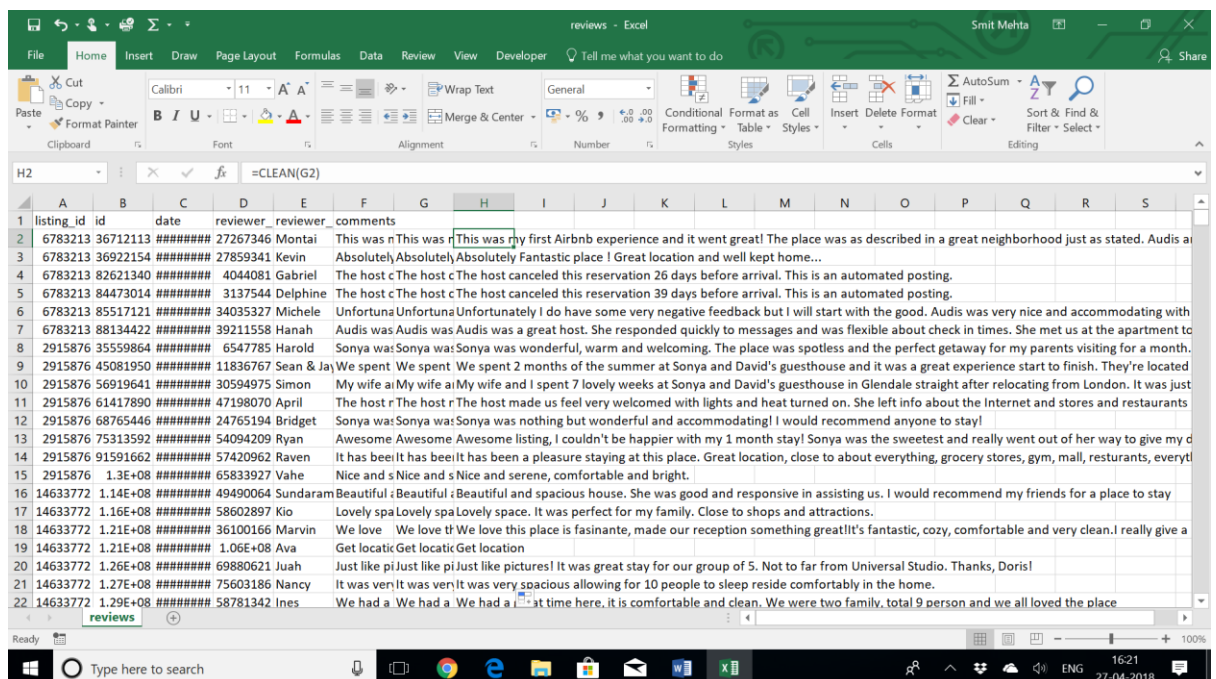
- We use the CLEAN function to remove unwanted symbols from all the reviews in column G.
- So, column H represents all the Cleaned comments/reviews.

Finally We Get Clean Data.

5. DATA MODELING

- I have used WORDij to create .net files of the 7 different buckets of data.
- We open WORDij, click on WordLink. Then we upload the .txt file and the stopwords file and click on analyze.
- Then I used Gephi to get the Network graph by loading the .net file.
- The network for all the 7 buckets are as below:

1. BUCKET-1
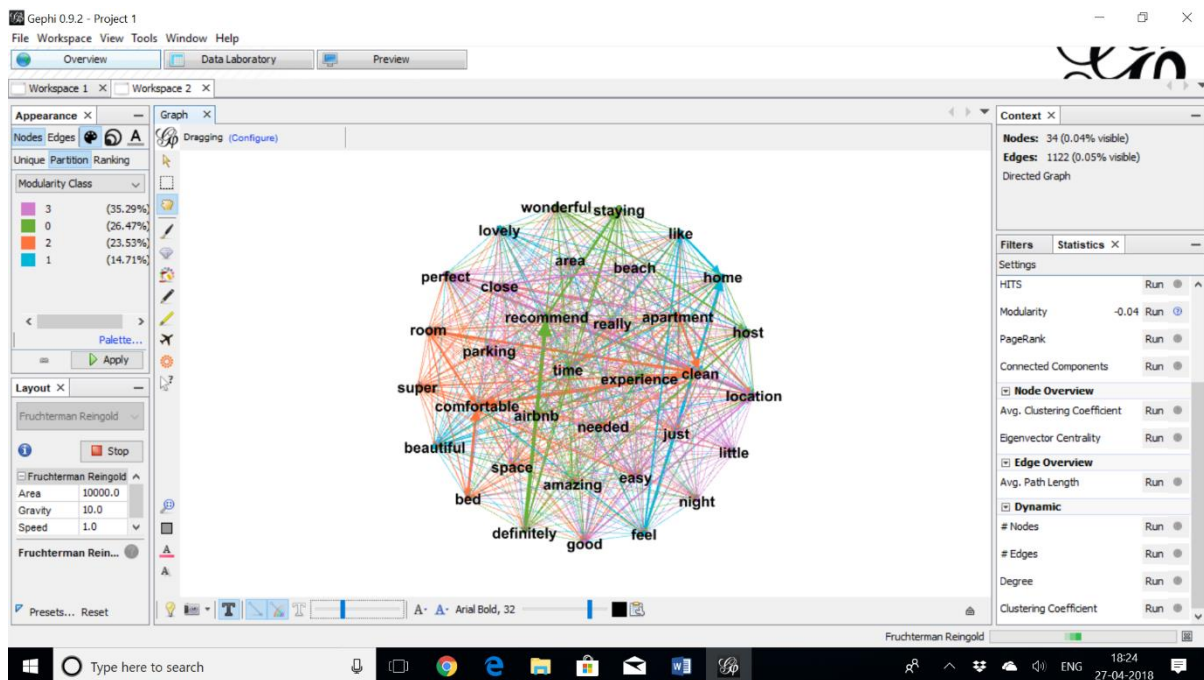
After using the Fruchterman layout, we have 4 modularity classes represented by different colors. Cluster 3 (purple) has the highest modularity of 35.29% and cluster 0 (green) has 20% while Cluster 1(blue) has least with 14.71%. We can see that cluster 3 lays focus on words like 'close, 'good', 'beach', 'night, 'clean', 'perfect' and 'location'. It means that cluster 3 signifies that it is good to stay around the area of the Airbnb especially at night and it has a beach around which is a perfect location.

2. BUCKET-2
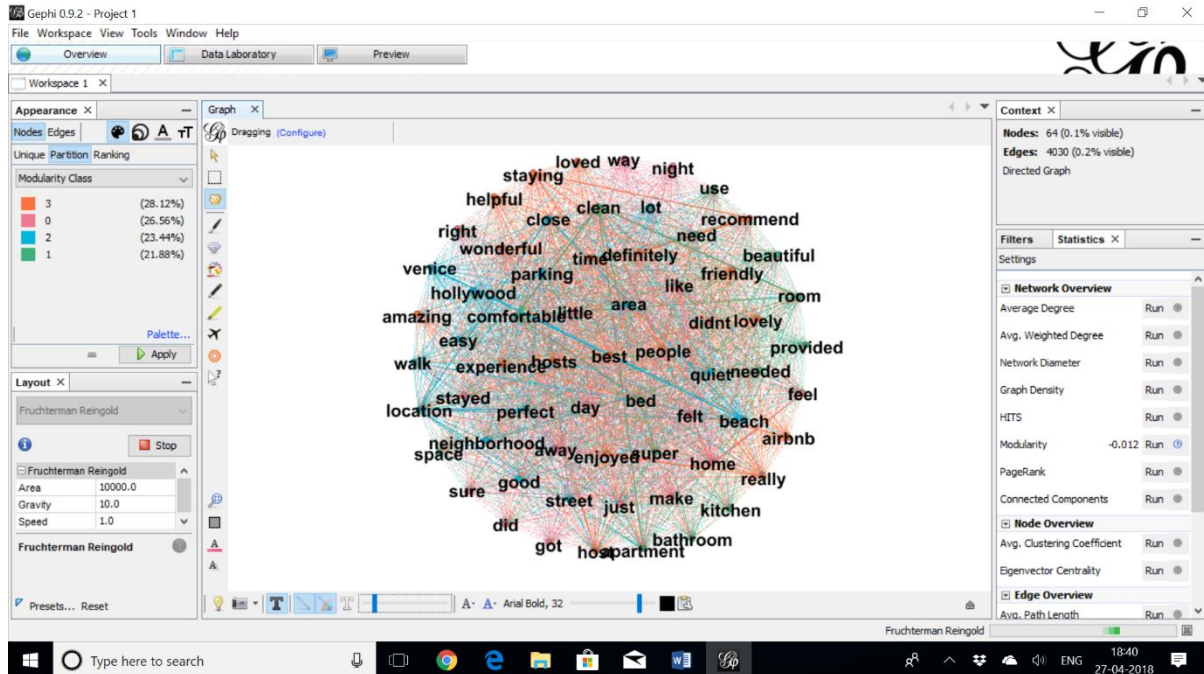
After using the Fruchterman layout, we have 4 modularity classes represented by different colors. Cluster 3 (orange) has the highest modularity of 28.12% and cluster 0 (purple) has 26.56% while Cluster 1(green) has least with 21.88%. We can see that cluster 3 lays focus on words like 'helpful', 'staying', 'loved', 'night, 'felt', 'home', 'definitely', 'recommended' and 'wonderful'. It means that cluster 3 signifies that it feels like home to stay at AIRBNB property. Customers enjoyed the night at their stay and was a wonderful time.

3. BUCKET-3

After using the Fruchterman layout, we have 4 modularity classes represented by different colors. Cluster 2 (purple) has the highest modularity of 33.96% and cluster 1 (orange) has 20.75% while Cluster 0(blue) has least with 18.87%. We can see that cluster 2 lays focus on words like 'lovely', 'space', 'need', 'kitchen', 'clean', 'comfortable', 'best', 'quiet' and 'house'. It means that cluster 2 signifies customer were in need of a house and they felt a bit quiet and had lovely space and a clean room and had good space.

4. **BUCKET-4**

After using the Fruchterman layout, we have 4 modularity classes represented by different colors. Cluster 0 (pink) has the highest modularity of 32.08% and cluster 1 (green) has 24.53% while Cluster 3(blue) has least with 20.75%. We can see that cluster 0 lays focus on words like 'Clean, 'Comfortable' , 'Host', 'Bed', 'lovely' , 'bathroom' and 'super' . It means that cluster 0 signifies customer were very happy with the clean and comfortable bedroom as well as super big bathroom.
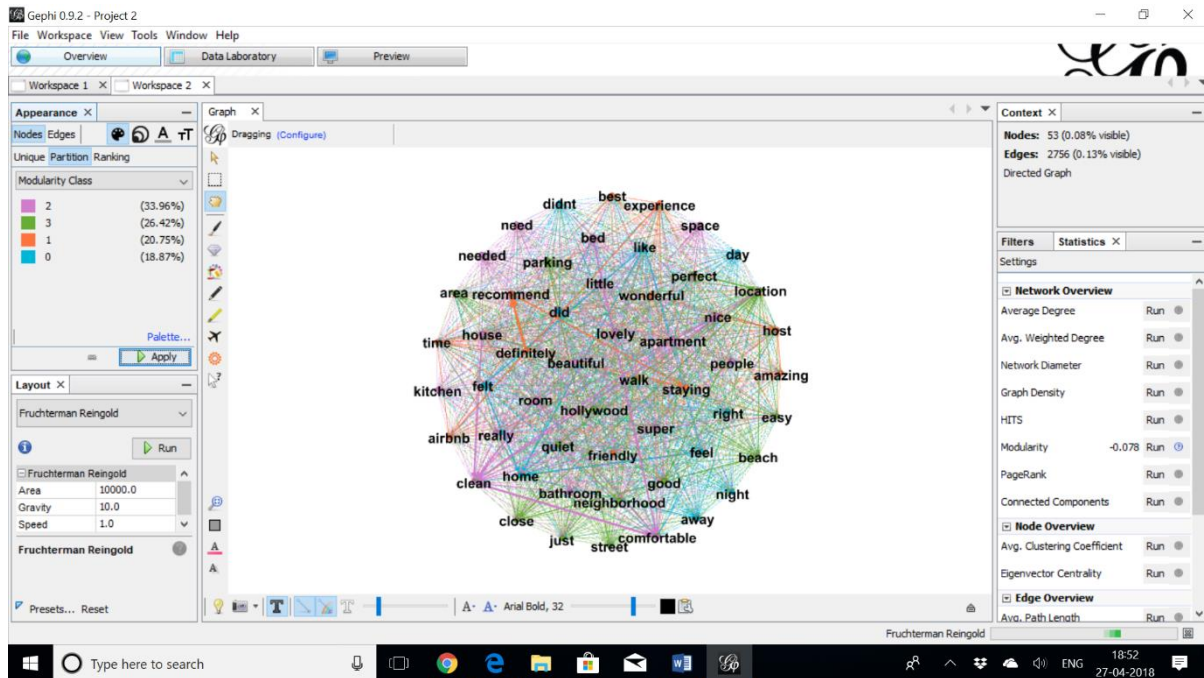
5. BUCKET-5

After using the Fruchterman layout, we have 4 modularity classes represented by different colors. Cluster 2 (purple) has the highest modularity of 29.17% and cluster 0 (green) has 27.08% while Cluster 1(blue) has least with 18.75%. We can see that cluster 2 lays focus on words like 'Hollywood' , ' close' , 'parking' , 'perfect', 'good' , 'walk' and ' location' . It means that cluster 2 signifies customer were very happy with the location. It is close to Hollywood and has good parking space.

6. BUCKET-6

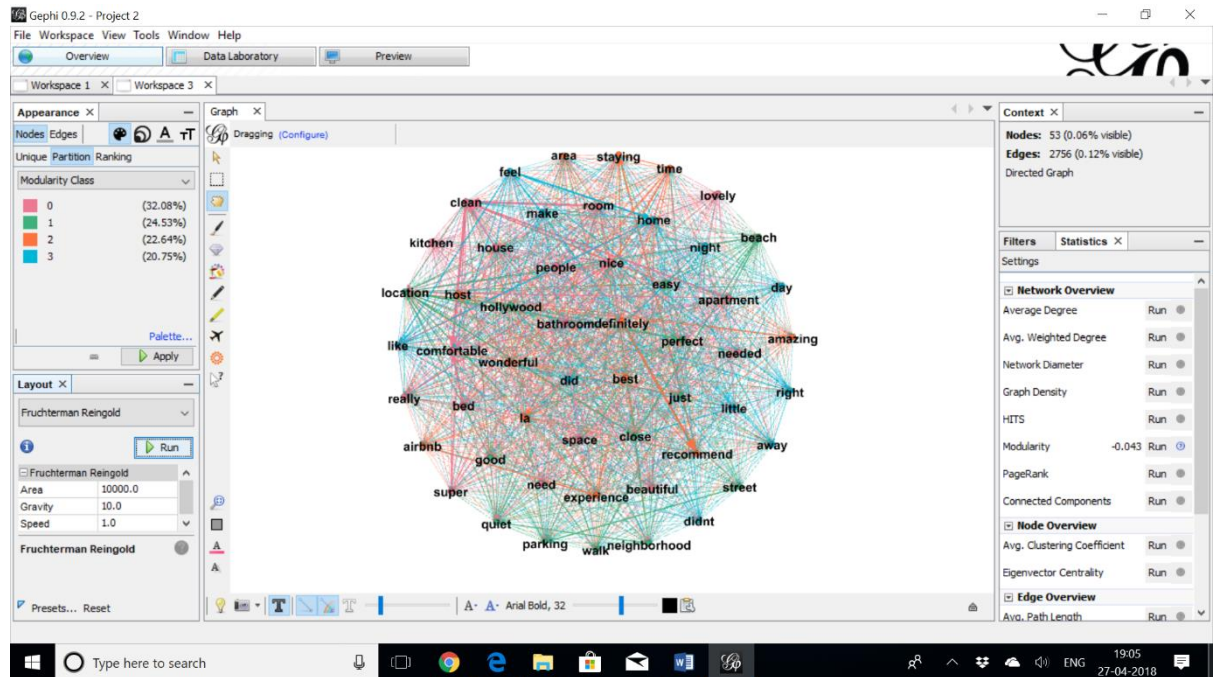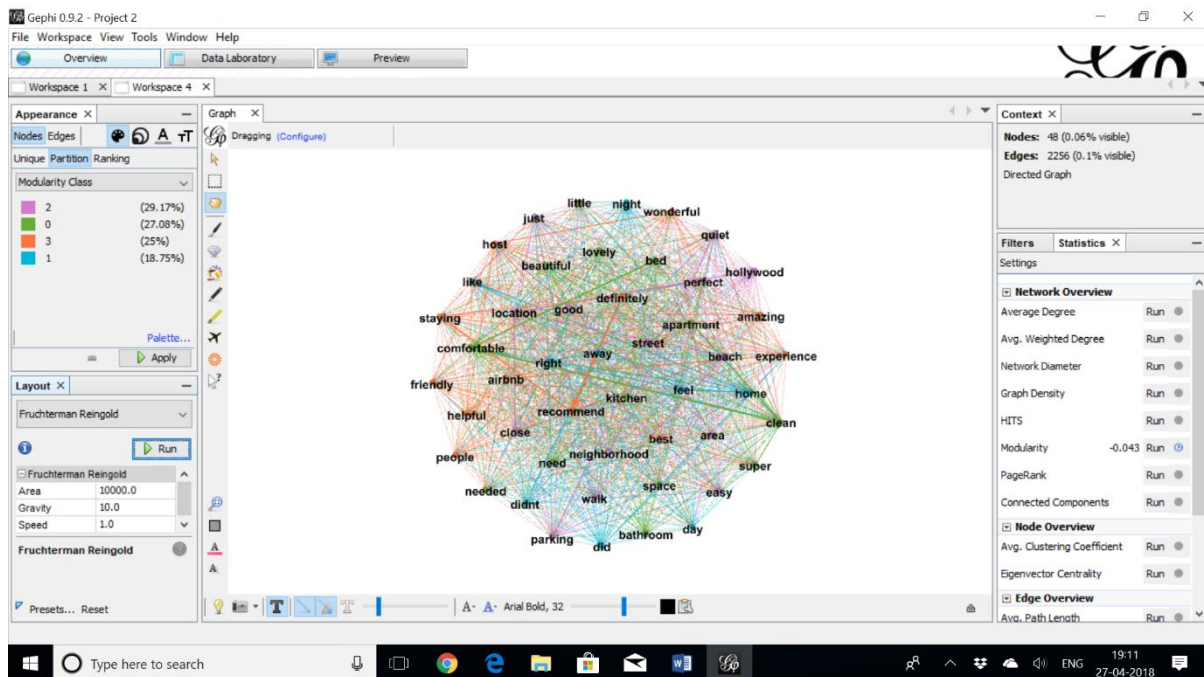

After using the Fruchterman layout, we have 4 modularity classes represented by different colors. Cluster 2 (purple) has the highest modularity of 36.54% and cluster 3 (green) has 25.00% while Cluster 0(blue) has least with 17.31%. We can see that cluster 2 lays focus on words like 'location', 'bathroom ', 'neighbourhood' , 'lovely', 'apartment' 'Hollywood' , ' close' , 'parking' , 'perfect', 'good' , 'walk' and ' location' . It means that cluster 2 signifies customer were very happy with the location. It is close to Hollywood and has good parking space. It are happy with their apartments.
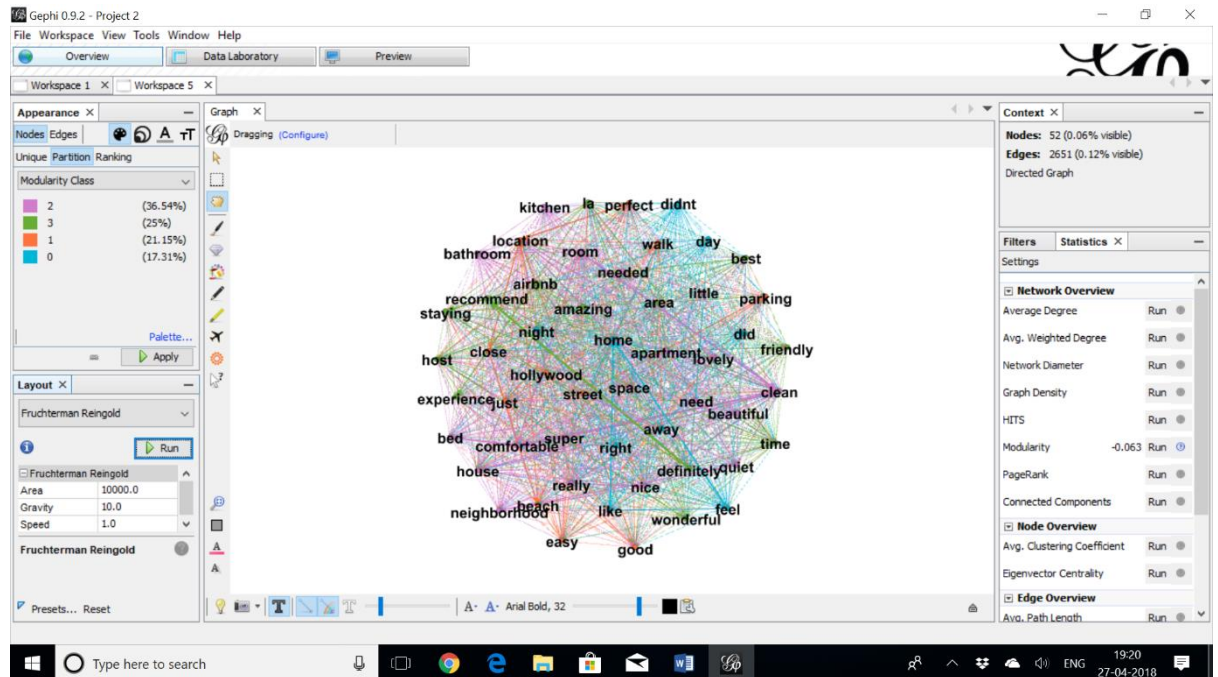
7. BUCKET-7

After using the Fruchterman layout, we have 4 modularity classes represented by different colors. Cluster 2 (purple) has the highest modularity of 36.67% and cluster 3 (orange) has 20.00% while Cluster 1(blue) has least with 10.00%. We can see that cluster 2 lays focus on words like 'wonderful', 'super', 'little' , 'close', 'needed', 'lovely' , 'beautiful' , 'friendly', 'definitely' , easy' and 'parking' . It means that cluster 2 signifies customer were very happy with the location. It is close to what they needed. Parking was easily available too. The place would be recommended to others.

6. EVALUATION PHASE

- We need to check the quality of our models before we deploy them in to the market.
- We need to see whether our analysis answered all the questions raised by our business understanding needs.
- We need to make sure that the cleaning of data was at par and finally make a decision on how to use the relevant data.

7. DEPLOYMENT PHASE

- At this phase, we are ready to use all our network models.
- Simultaneously, we need to test our models and make a report of it.
- Or we can simply deploy the model depending on the urgency, demand and situation the customer is in.

CONCLUSION:

I came to this conclusion that:

- First of all, the reviews focused on the night spent and the beach around the property.
- Then it shifted to that it feels like home and people were of helping nature.
- Then, they talk about the room space in AIRBNB which was comfortable and clean.
- The major point was the apartment was close to Hollywood and it was easy to park and had great location. Their was discussion about street also.
- It talked about a house with kitchen and good parking and clean house.
- Last, Airbnb got best reviews in LOS ANGELES and last buckets refer to their clean and comfortable room which must be recommended to others.