MIE 451/1513 Decision Support Systems
Lab and Assignment 5:
Social Network Analysis

November 14, 2019

This assignment involves social network analysis based on twitter data. Through this assignment you will have better understanding of graph analysis methods, as well as different centrality measures in the graph.

- Programming language: Python (Google Colab Environment)
- Due Date: Posted in Syllabus

Marking scheme and requirements: Full marks will be given for (1) working, readable, reasonably efficient, documented code that achieves the assignment goals, (2) for providing appropriate answers to the questions in a Jupyter notebook (named `sna-assignment.ipynb`) committed to the student's assignment repository, and (3) attendance lab code review session, running your solution notebook for instructors, and providing clear and succinct answers in response to instructor questions regarding your solution.

Please adhere to the collaboration policy on the course website – people you discussed the assignment solution with, or websites with source code you used should be listed in the submitted Jupyter notebook.

What/how to submit your work:

1. All your code should be included in a notebook named `sna-assignment.ipynb` that is provided in the cloned assignment repository.
2. Commit and push your work to your github repository in order to submit it. Your last commit and push before the assignment deadline will be considered to be your submission. You can check your repository online to make sure that all required files have actually been committed and pushed to your repository.
3. A link to create a personal repository for this assignment is posted on QUERCUS.

1 The Twitter Data

We will work on the data of tweets collected in one month of 2009.

The data is stored in a csv files. Each tweet is a line with three fields: time, user and text of the tweet. Here is a snippet:

```
   2009-06-11 16:59:45, ibbored, amberback #squarespace does?  Hot damn.  Now I want
to win more.
```

Twitter users abbreviate topics with hash-tags (#squarespace) and mention others with @ sign (answerback). In this assignment, you will need to extract records like user names and hash-tags from the text, and analyze different patterns of communication in the given dataset.

2 In the Introductory lab

In this introductory lab, we will build a mention graph for a given hash-tag, and explore different ways to visualize graphs, and incorporate additional information.

We will cover graph creation, manipulation, and analysis using `NetworkX`, and visualization using `Plotly`.

3 Main Assignment (Assessed in Code Review Session)

Q1. Choose a hash-tag

You are required to choose a hash-tag to perform the social network analysis on. The chosen hash-tag should be unique and not shared with the any of your classmates. In order to keep track on the hash-tags being used by the other students, please refer to the discussion board on Piazza. Once you choose a hash-tag, please make sure it is still available and then post your chosen hash-tag on the discussion board.

- The first one to post a hash-tag will be the one to perform the analysis on this hash-tag.
- It is your responsibility to make sure the hash-tag is still available when you post it to Piazza.
- Make sure the hash-tag will be suitable for the analysis required in this assignment.
- If you want to change the hash-tag, please post another message declaring the old hash-tag (now available), and the new one (now allocated to you). As before, it is your responsibility to make sure the new hash-tag is available.

Q2. Build a Mention Graph

In this question you are required to build a mention graph for your chosen hash-tag. The mention graph is the mention relations between users. In this graph, each user is viewed as a node. If a user `Alice` mentions another user `Bob` in her tweet(s) with the @ sign, then an undirected edge connects `Alice` and `Bob`. The edge weight is the number of mentions in the tweets.

(a) How many nodes and how many edges in your mention graph?

(b) Build a histogram of the graph nodes' degree (i.e., the degree distribution of the graph). What can you learn from the degree distribution?

(c) Provide a list of top 5 edges with highest weights (edges are identified be the two nodes they connects, e.g., $\langle node1, node2 \rangle$).

(d) Provide a visualization of the mention graph in which the edge color reflects its weight (i.e., the number of mentions).

Q3. Content Analysis

In the this question you are asked to perform a basic content analysis on your chosen hash-tag.

(a) Analyze the most common words in all the tweets with the chosen hash-tag, and provide a basic description of the main themes.

(b) In the visualization of the mention graph, add hover information for the nodes, describing the 3 most common words for this user. Add any other hover information that may help you understand the social network better.

Q4. Centrality Analysis

In this question, your need to analyze centrality of users in the mention graph.

Here is a list of `networkx` functions that calculate the different centrality measures: `https://networkx.github.io/documentation/stable/reference/algorithms/centrality.html`

Note that PageRank lives in a different place in `networkx`: `https://networkx.github.io/documentation/stable/reference/algorithms/link_analysis.html`

(a) Choose two centrality measures and calculate the centrality of the nodes on your graph based on each of the measures.

(b) Provide a visualization that demonstrates the centrality of each node using a visual property (size, color, etc) for each of the centrality measures.

(c) Identify the key players in the mentioned graph based on the centrality measures.

1. Are the results similar or different? Explain what can be the reason for the observed similarity or difference.
2. What centrality measure produced a more meaningful interpretation? Why?

Q5. Connectivity Patterns

In this question, you will analyze the cliques in a graph. A clique in an undirected graph is a subset of the nodes, such that every two different nodes are adjacent (directly connected with an edge).

Here is a list of `networkx` functions that calculate the different measures related to cliques in graph: `https://networkx.github.io/documentation/stable/reference/algorithms/clique.html`

(a) Calculate two or more of the following measures to analyze cliques in your mention graph:

1. Number of maximal cliques in your graph.
2. The graph's clique number (size of the largest clique in the graph).

3. Number of maximal cliques for each node
4. Size of the largest maximal clique containing each given node.

(b) Provide some insights on the connectivity patterns of your mention graph based on the information calculated in (a). What was the largest clique? How large was it? What else do your clique properties and cliques themselves tell you about the social network?

4 Helpful Links

4.1 NetworkX Documentation

`https://networkx.github.io/documentation/stable/index.html`

4.2 Plotly

`https://plot.ly/python/`