**DS620 Final Project – Team KJW**

**Draft Proposal**

**Final Project**

Your project should incorporate one or both of the two main themes of this course: network analysis and text processing. You need to show all of your work in a coherent workflow, and in a reproducible format, such as an IPython Notebook or an R Markdown document. If you are building a model or models, explain how you evaluate the “goodness” of the chosen model and parameters.

**Objective**

To come up with a social network graph showing connectedness between students of the CUNY DATA 620 – Summer 2020 course. This graph would be a MultiGraph (multiple edges between nodes) where each student would be a node and there would be 3 types of edges:

1. Edges based on project group membership: Simple unweighted groups of nodes not connected to other group of nodes. In the graph visualization, these edges would be color-coded with the same color to differentiate them from the other type of edges described below. We could collect this information ourselves by noting down the group member names during the class meetup on July 9.
2. Uni-directional edges based on comments posted by other students in response to “original” threads posted on weekly discussion forums. The idea would be to link students to each other based on interaction in the weekly discussion forums across all weeks i.e. aggregate across all weeks. Before aggregation, each edge could be weighted based on Newman’s method i.e. if only 1 student has commented then weight = 1, but if multiple students have commented, then value of each response is lower. Having computed the weights of each edge, sum up the weights for each pair of student nodes. The visualization would depict the thickness of the edge based on this aggregated weight and the color for this set of edges would be different from the ones in #1 above. We may need to create an edge attribute called thread topic i.e. threads are uniquely identified by the description (subject line) and not by the week number / part number combination. We can decide later if this is required or not.
3. Edges based on lexical similarity. We can do this in 2 different ways as described in the 2 documents in the references section.
   1. The first is the Text Network approach which relies on the textnet library to link nodes (students) with each other based on usage of common words. The strength of the linkage is determined by the similarity in terms used in the threads. The built-in methods can be used to come up with the network graph.
   2. The second is the document similarity approach which relies on the NLTK and the Gensim libraries to calculate the cosine similarity measure between sentences attributed to a pair of nodes (students).

**Data Prep**

Before employing both of the approaches mentioned above, we’d need to clean up the threads – for example remove hyperlinks, stop words, punctuations etc. We may even need to remove all words except nouns i.e. remove verbs, adjectives etc. leaving in only nouns as a “true” indicator of lexical or topic similarity.

**Data Collection**

We’d need to collect all the threads from all the weekly submissions and tag each post by a) author b) original post or response comment c) thread title or description. We should save each post as its own text file. This will take time, but if we divide the work between the 3 of us, it may not take that long. We could possibly ignore week 1 (only introductions were made) and go only from week 2 to week 6. Assuming 20 students in the course, we’d probably end up with between 100-300 small text files ranging from 1 to 3 paragraphs of text. Alternatively we could just copy and paste the posts into an Excel file (pre-specified 4 columns) – each one of us can collect the data for 2 weeks only and we can combine the 3 Excel files into one and push the data into a csv file which can be directly loaded into a pandas dataframe.

**Methodology / Analysis**

Using a combination of NLP and SNA, we could show how the students in this class are linked to each other in 3 different ways:

* Explicit project group membership
* Links based on comments posted in response to weekly discussion threads. These are potentially uni-directional edges (from responder to original author).
* Links modelled using lexical similarity analysis to tie students with similar content (words overlap)

Final output would be a visualization of all students with 3 different type of edges. We could potentially compute centrality measures based on these edges to show “influential” nodes.

**References**

<https://compsocialscience.github.io/summer-institute/2019/materials/day3-text-analysis/text-networks/rmarkdown/Text_Networks.html>

<https://pypi.org/project/textnets/>

<https://dev.to/coderasha/compare-documents-similarity-using-python-nlp-4odp>