**DS620 Final Project Proposal**

**Team KJW – Ken Popkin, Jagdish Chhabria, Willie Smalls**

**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 create a social network graph showing connectedness between students of the CUNY DATA 620 – Summer 2020 course. This graph is expected to be a Multi-Graph with different types of edges (links) between the nodes (students). The motivation for performing this analysis is twofold – we expect that it would throw up data-driven insights on links between students based on common interests that they themselves may not be aware of, and secondly it’s interesting to look at recent data generated by our own class as opposed to some abstract data collected by someone else.

**Methodology**

The DS620 student graph would comprise of the following 3 types of edges:

1. Based on project group membership.
2. Based on initial posts and comments/responses thereon in the weekly discussion forum on Blackboard.
3. Based on lexical similarity between the initial posts and comments thereon. For determining the lexical similarity, we propose to use Natural Language Processing (NLP). We are considering modelling this as a text network and using the textnet library. We’re also considering using NLTK and Gensim libraries for this, and calculating cosine similarity between the discussion threads.

**Data Collection and Preparation**

The data would be manually scraped from the Blackboard site. It would need to be cleaned to remove hyperlinks and stop-words etc.

**Roles and Responsibilities**

Assuming we get approval to proceed on this proposal, we expect to share responsibilities more or less equally, for data collection & cleaning, reading the external research on lexical similarity, doing the coding to derive insights and presenting to the class.

**Result**

We propose to create a network visualization of all the students with the 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/>