Midpoint Report

Project Objective:

Wikipedia is a free encyclopedia written collaboratively by volunteers around the world. A small part of Wikipedia contributors are administrators (users with access to additional technical features that aid in maintenance). In order for a user to become an administrator, a Request for adminship (RfA) is issued and the Wikipedia community via a public discussion or a vote decides who to promote to adminship.

This project aims to study the complexity of social interactions by utilizing weighted networks and graph scoring.

Project Approach:

The project will be implemented in 4 main steps:

1. Data Mutation in 2 steps:
    a. Flatten data in 2 directions
    b. Construct directed and weighted relationship between data entities
2. Build a Machine Learning model for identifying clustering of connected groups (social clicks)
3. Based on ML groups identification introduce 3rd affecting variable for study of dynamics
4. Analysis & Findings: Identify voting patterns

Team Structure:

The project is being implemented by 1 person, Tatiana Luchian. So I will be completing all of the functions on the project (project management, data cleaning, data transformation, implementation, analysis, and reporting)

Project Milestones & Progress Update:

Task	Description	Due Date	Status
Data Cleaning	Pre-process and clean data for API ingestion and data model	07/10	COMPLETED
Data Model Creation & Mutation	Design data model and relationships in the model for Neo4j. Via python script transforms the data.	07/15	COMPLETED

DB creation in Neo4j	Create a script to populate transformed data into Neo4j, create a graphical presentation of the data.	07/25	COMPLETED
ML for cluster	Create ML program based on K-means algorithm in order to identify clusters (voting clicks) int he dataset	07/25	IN PROGRESS
ML data into Neo4j	Transform data (clusters identified) as a 3rd variable to graph data to Neo4j	08/01	NOT STARTED
Research Hypothesis	Identify at 3 hypotheses based on social network studies that could be tested int his project	08/01	IN PROGRESS
Data analysis & reporting	Perform data visualization and analysis of findings, document, and present the findings via the final presentation.	08/10	NOT STARTED

Currently, I am on track with the project. Although the ML program step is on track, it takes much longer than I initially estimated.
In addition, the research step was added to the project plan, which could slightly delta the final implementation and testing of the data.
Initially, I planned on implementing the project n 2 separate graph database systems: GraphQL and Neo4j, so I can examine their advantages and disadvantages. However, after some research, I realized that systems are very different and it would be difficult to compare them in a meaningful way. Hence, I decided to add research component to my project along with the ML program.

Date / time of scheduled midpoint meeting with Prof.
July 23, 2020, at 3:20 pm.