Problem Statement

How can I build or reengineer a twitter-based recommendation algorithm through pyspark within the next 4 months that feeds users content they are unlikely to interact with and promotes a more balanced information environment.

Context

Access to information is more prevalent today than ever before. However, which information users access is often determined not by the user but by recommendation algorithms built into online systems and platforms. This has caused an effect known as a "filter bubbling" in which users tend to interact with information that conforms to their preconceived beliefs or ideas. In turn, this reinforces those beliefs and pushes users either into isolated information camps or towards political extremism. This issue has been raised by popular films including "The Social Dilemma" [1] and articles published by MIT[2]. I believe that reengineering those recommendation algorithms to feed users not simply what they would "prefer" to interact with, but also what may not conform to their worldview is an important and necessary step to solve this problem.

Criteria for Solution Space

Create a recommendation algorithm or reengineer a recommendation algorithm within the next 4 months that feeds users information they are unlikely to see (based on some recommendation value cutoff, tbd) 50% of the time. At the end of the project, I will have a GitHub repository with a slide deck and a project report to show for it as well.

Constraints

There is a time constraint to complete this project by the end of July. There is also the constraint of limited expertise on my part in knowing how to use pyspark and which steps to take first.

Stakeholders

Outside of my own involvement, my mentor Kevin Glynn will be giving advice and assisting in the project. I also expect to refer to the career coaches at Springboard as I curate my presentation of the project for employers.

Data Sources

I plan to use the dataset found at this link to pull user data and interactions with news articles from twitter to train and test my recommendation algorithm. I may gather additional information from Google Dataset Search as the project moves forward.