Capstone Progress Report (Part III)

Thomas Voreyer

Recap

Over the course of the past weeks, the focus of my capstone project has been to identify whether a news article will cause movement within the stock market. For instance, if Samsung announces they will be recalling the Galaxy Note 7, the model should predict that the corresponding news article will be important towards stock price movement. Contrarily, if Disney World announces they will be selling a new kind of cotton candy, the model should predict that the corresponding news article will not be important towards stock price movement. My research and conversations with Alex have led me to believe that this sort of project does not have an achievable end. My findings have been that news relevant to the markets is realized before, after, and/or at the time of the news release, and the implications to the stock price can take place over a very long period.

A stock price, in essence, is calculated by the equity plus the expected discounted value of future profits. Many smart investors, especially institutional investors, will have market research from a variety of sources, both private and non-private. As such, the market is extremely complex, and should not move according to a single front page article. Instead, market movement begins when an institutional investor believes there is an undervaluation or overvaluation based upon an accumulation of information from multiple sources over time. As more investors realize the new valuation, the markets will shift, but this takes place over time, and not suddenly. Alas, this proves the market is unresponsive to a single piece of evidence, even if that evidence indicates that stock prices should move. As my Capstone project can no longer be seen as achievable, I propose a change in topics.

Project Proposal

Topic

I propose changing my topic to creating a subreddit recommendation engine. I believe this topic to be a much stronger choice, as no subreddit recommendation engine currently exists, which makes this a highly creative idea. Furthermore, the concepts are more straightforward, as there is no right or wrong answer in this project. Lastly, the skills I will be using in this project are the skills I will be using in employment.

Thought Process

In looking at the data I had pulled for my previous project attempt, I thought it would be easiest to reuse my scraping code for Reddit data (For one, I have already seen the data; secondly, I have the scraping code written.) I began to map out a plan of action based upon the data I could possess, which would be subreddit (sub) top posts, sub hot posts, sub controversial posts, post comments, user posts, user comments, user up-voted, and user down-voted. With all of this information varying from user to user, and subreddit to subreddit, I thought it would be an interesting idea to identify personalities of top commenters, and to identify latent personalities and ideologies of subreddits. With the personality profiles of users, one could create a similarity matrix, and recommend subreddits based upon currently subscribed subreddits. Alternatively, one could create personality profiles of subreddits based upon user data and hot/controversial posts. From these subreddit personality profiles, one could create an inferred personality of a user based upon their subreddit subscriptions. With this inferred personality information, the recommendation engine would suggest the most similar subreddit that the user is not currently subscribed to. Lastly, these two systems could be used in conjunction to generate a subreddit recommendation.

Skills Used

- NLP
- KMeans Clustering (Potentially PCA, and Feature Selection)
- Matrix Methods
- APIs/Web Scraping
- AWS/Flask
- (Potentially Bayesian Statistics)

Current Progress

There are only two weeks left in this course, so I will have to be efficient to complete this project. Despite the magnitude of this task, I am confident that it is achievable. To ensure I do not fall behind, I have outlined a plan of action below:

Plan of Action

Phase 1 [Due Dec 5]

1. Scrape all of the Reddit data for this project [Current; Finish by Sun. Afternoon]
    a. There is a lot of data, and the reddit API is fairly slow, so this is non-negotiable completion date
    b. On the flip side, I can use this time to write my code for other parts of this project or brainstorm further steps
2. Clean Data [Finish on/by Sunday Evening]
3. Create Ideology and Personality word dictionaries [Finish on/by Monday]

Phase 2 [Due Dec 8]

4. Compile comments by subreddit [Finish on/by Tuesday]
5. Compile comments by user [Finish on/by Tuesday]
6. Compile post types by subreddit (gif, video, text, article,...) [Finish on/by Tuesday]
7. Compile post types by user [Finish on/by Tuesday]
8. Create subreddit-personality matrix [Finish on/by Thursday]
9. Create subreddit-ideology matrix [Finish on/by Thursday]
10. Create user-personality matrix [Finish on/by Thursday]
11. Create inferred user-ideology matrix [Finish on/by Thursday]
12. Check on own profile to see if it makes sense [Ability to extend to on/by Friday]
    a. If not, redo step 3 and restart from step 8

Phase 3 [Due Dec 11]

13. Use Clustering or Matrix Methods (whichever is appropriate for model) to decide similar users or subreddits [Finish on/by Sunday]
14. Test recommendation engines [Able to extend to Monday]
    a. If not redo step 3 and restart from step 8

Phase 4 [Due Dec 17]

15. Create input ability via AWS and flask [Finish on/by Wednesday]
16. Polish everything [Finish on/by Saturday]

Personality/Ideology Abstraction and Theory

Ideology is the bedrock of a person. It dictates what a person does on a daily basis, and can predict a from how a person will vote to how a person will spend money.

Personality is the substance to an individual. It is the unique identifier of an individual from a sea of people. It can be used to predict how one will act on a daily basis, to how one will interact in a social environment.

My theory is that people are highly complex, and it is hard to model them without a complex model and an abundance of data. The user data in itself is limited, so why not infer traits based upon subscribed subreddits? If one were to compile and extract information upon what was popular for a subreddit, what was unpopular for a subreddit, how people interact within a subreddit, and how people think within a subreddit, one could begin to create a subreddit profile strong enough to be used in personality inference.

User Personalities

- User data is limited
- There is user data on comments, posts, likes, and dislikes; Can gather:
    - Language style -> simple, crude, intelligent, technical, plain, etc
    - Entertainment/Ideology Preferences [Issue of separating entertainment from more objective subreddits/comments]
    - Sociability -> How often does one post/comment
    - Popularity -> How up-voted is someone
- Personalities will only be partially complete with user data; must infer traits from subreddit personalities

Subreddit Latent Personalities

- Regard Hot Posts and Controversial Posts
    - These will not necessarily make it to the front page, but they demonstrate a common personality in mass
    - Hot posts are personality positive
    - Controversial posts are personality negative
- Regard comments for:
    - Language styles
    - Trending topics
    - Sanity level
    - Intelligence level

Subreddit Ideologies

- Regard comments and posts for ideology words
- Can mark subreddits as:
    - Conservative or liberal
    - Soft or hardcore
    - Upbeat or Depressed
    - Narcissistic or self-less
    - Girly or Boyish
    - Et cetera

Recommendation Models

Crude Model

- Using gathered user data, extract subreddits for each user by regarding the subreddit a comment or post was created in
- From this user data, create a matrix of users and subreddits
- From this user-subreddit matrix, use cosine similarity to recommend the most highly rated subreddit of all non subscribed subreddits

User Personality Model

- Similar to the crude model, except one would infer personalities of users from subreddit and user data
- From these inferred personalities, an individual could complete a personality form and see what subreddits are recommended

Subreddit Personality Model

- Rather than referring to a user personality, one could refer to a network of subreddit personalities, and have the closest neighbors be selected as the recommended subreddits