Project 5 Proposal

Twitter Network Analysis

Motivation

Social media as a whole is the new face of the modern day ad industry. Everyday, platforms are in a constant battle for consumers' attention as their sole focus is to monetize that attention in a cycle dubbed the **"attention economy"**. But what remains unseen is the impact that the constant bombardment of "content" has had on society, especially to young people. In a world that favors the sources that draw the most attention, we've lost what it means to be authentic in favor of flamboyant personas.

The aim for my project is to create healthier digital spaces. I want to promote meaningful interactions and foster real relationships between people. To find connections without the need for a flood of content, but by trying to find a deeper understanding of an individual using other attributes of their digital footprint.

Factoids

- A recent **Cigna study** found that 73% of Gen Z (18–22 year olds, notably the first generation to grow up with technology since day one) report “sometimes” or “always” feeling alone, up from 69% the previous year.
- Every second, approximately 6,000 Tweets are tweeted on Twitter, which corresponds to over 350,000 tweets sent per minute, 500 million tweets per day and around 200 billion tweets per year **(via Brandwatch)**.
- **Pew Research Center** states that currently, 72% of the public uses some type of social media.

Problem Statement: For this project, I am going to use three different data sources using Twitter and create a feature set using NLP, network analysis, and personality classification. I will use these features to create a recommender system to connect Twitter users. My main goal is to use only retweets from the account timeline, tweets that were favorited by the account, and the accounts the person is following. This is to test the hypothesis that connections can be made between people without having to use a catalog of original content.

Description of the Data

The data I will use for this project is Twitter data extracted from the Twitter API via the tweepy/twitterscraper packages. I will pull the data from three sources: tweets from the users' timelines (tweets and retweets), tweets that the user has favorited, and attributes of the user account. Here is a description of the data:

user_timeline

Independent Variable	Type	Description
id	Int	Unique identifier of the tweet.
created_at	Timestamp	Time the tweet was posted.
screen_name	String (Object)	Username of the account that posted the tweet.
user_id	Int	Unique identifier of the account.
in_reply_to_status_id	Int	Unique identifier of the parent tweet to which the tweet is in response to.
in_reply_to_screen_name	String (Object)	Username of the account of the parent tweet to which the tweet is in response to.
in_reply_to_user_id	Int	Unique identifier of the account of the parent tweet to which the tweet is in response to.
favorite_count	Int	Number of favorites the tweet has.
retweet_count	Int	Number of retweets the tweets has.
text	String (Object)	The text of the tweet.

user_favorites

Independent Variable	Type	Description
id	Int	Unique identifier of the tweet.
created_at	Timestamp	Time the tweet was posted.
screen_name	String (Object)	Username of the account that posted the tweet.
user_id	Int	Unique identifier of the account.
favorite_count	Int	Number of favorites the tweet has.
retweet_count	Int	Number of retweets the tweets has.
text	String (Object)	The text of the tweet.

user_attributes

Independent Variable	Type	Description
user_id	Int	Unique identifier of the account.
follower_count	Int	The number of followers the user has.
following_count	Int	The number of accounts the user follows.
following_ids	Int, List	List of IDs corrsponding to the accounts the user follows.
follower_ids	Int, List	List of IDs corresponding to the accounts that follow the parent user.

Known Unknowns/Barriers

- This project requires three unsupervised learning techniques, mainly revolving around messy text data
- Building a network graph may take more time than anticipated. There's also no guarantee that there will be connection between accounts outside of common accounts they follow.

Potential Resources

- nltk, TextBlob, Vader, for topic modeling and sentiment analysis

- pandas, numpy for data wrangling and aggregation
- matplotlib, seaborn, plotly for data visualization
- networkx, graphx, i-graph for network analysis and visualization

Minimum Viable Product / Other Project Ideas

- Create a D3 visualization and/or Flask app of a Twitter ego network.
- Create a text generator for people of different personality types using Tweets, or Reddit data.
- Train a neural network to classify MBTI personality types using Reddit data.