Cayden Dunn

Data 510

Hw3

Develop an organization scheme for the data used in your final project.

Data should be organized into logical relationships.(2)

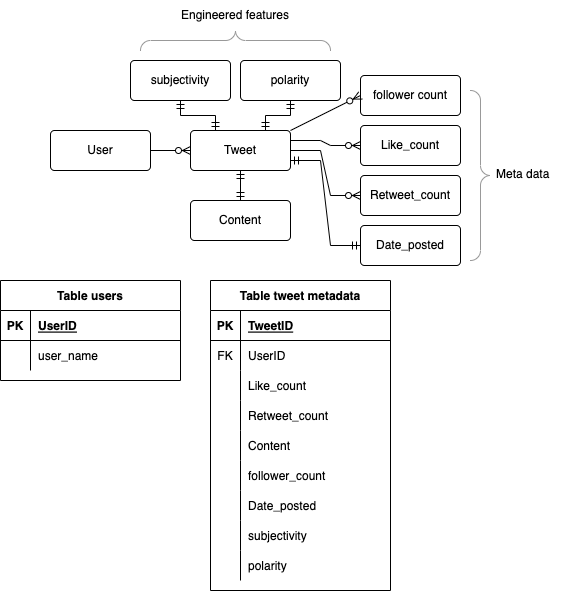


Fig 1

Describe origin data sources, and show where processed data resides in the organization scheme. (2)

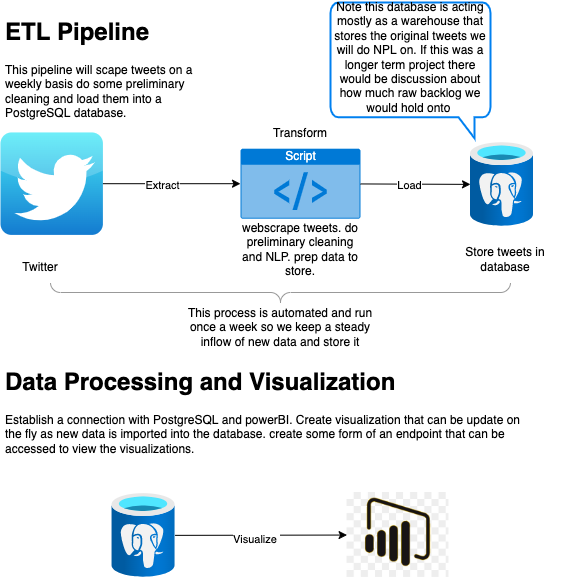


Fig 2

As shown in fig 2 the data will be scapred from twitter pre-processed via a python script and then loaded into sqlite3.

Identify primary keys for individual relationships and show how specific relationships are related to other data in your project. (1)

As shown in fig 1 each user in the users table has a user id and a user name and that is the primary key of that table. Each user can have 0 to many tweets this is the relation ship between the tweets table and users table and user id is a forgien key in tweets table. Therefore we can index a particular user's tweets. Each tweet has 1 sentiment associated with it. Each tweet has 1 count of followers at that time, likes at the time of scrape, retweets at the time of scrape and a date associated with the post.

Estimate the number of records you anticipate for each relationship. (1)

20,000 - 50,000 tweets that will be scraped from twitter all at once. Looking at the number for new tweets each week the numbers really aren't all that substantial. So for now i'm going to start with the 20-50k and if I have time I can tack on automated weekly updating scapres.

Describe the data cleaning, transformation, and feature engineering that will be necessary to populate relationships within your storage scheme. (2)

The fields that will not be preprocces at all before being loaded into the data store are fields like likes, followers, retweets, dates. The Fields that will be preprocessed is tweet content. To process this field we will remove all punctuation, special characters, and stem the words to prep them for sentiment analysis. From here we will run sentiment analysis on the tweet content and generate an engineered feature called sentiment that will store a 0 -> Negative; 1 -> Neutral; 2 -> Positive label to be used in visualizations. There are other enginerreed features that will be created but I will handel them in powerBI instead of python. These are feature engineered fields like normalized follower count, normalized retweet count, normalized like count. These fields will be created by doing arithmetic operations on the incoming data to normalize all virality metrics so we can get a clear unbiased view of how well a tweet performed even when comparing tweets from accounts of different sizes. This will be done in effort to see if we can find any correlations or linkage in the sentiment of a tweet and how well it preforms