Recommending 2nd Degree Twitter Users with NLP and Topic Modeling

**Abstract**

The goal of this project was to build a recommendation model that recommends 2nd degree Twitter users to follow based on the each users collective tweets, retweets and likes. I scraped approximately 19,700 from 106 different Twitter accounts using Twitters Tweepy API, and I used nature language processing, topic modeling, and distance calculations to determine the top 10 most similar accounts based on my Twitter preferences.

**Design**

I began the project by first collecting up to 100 tweets / retweets and 100 likes per user using Twitter’s Tweepy API v2. I compiled the data into a data frame and cleaned the data by stripping punctuation, removing html links, converting characters to lower case (using regular expression), and removed certain parts of speech (used NLTK’s pos\_tag to identify unnecessary words for topic modeling). I also created a function to extract crypto-specific terms as the topic modeling algorithms were having issues categorizing these terms. I finally consolidated all 200 tweets per user into one document, resulting in 106 final documents for topic modeling.

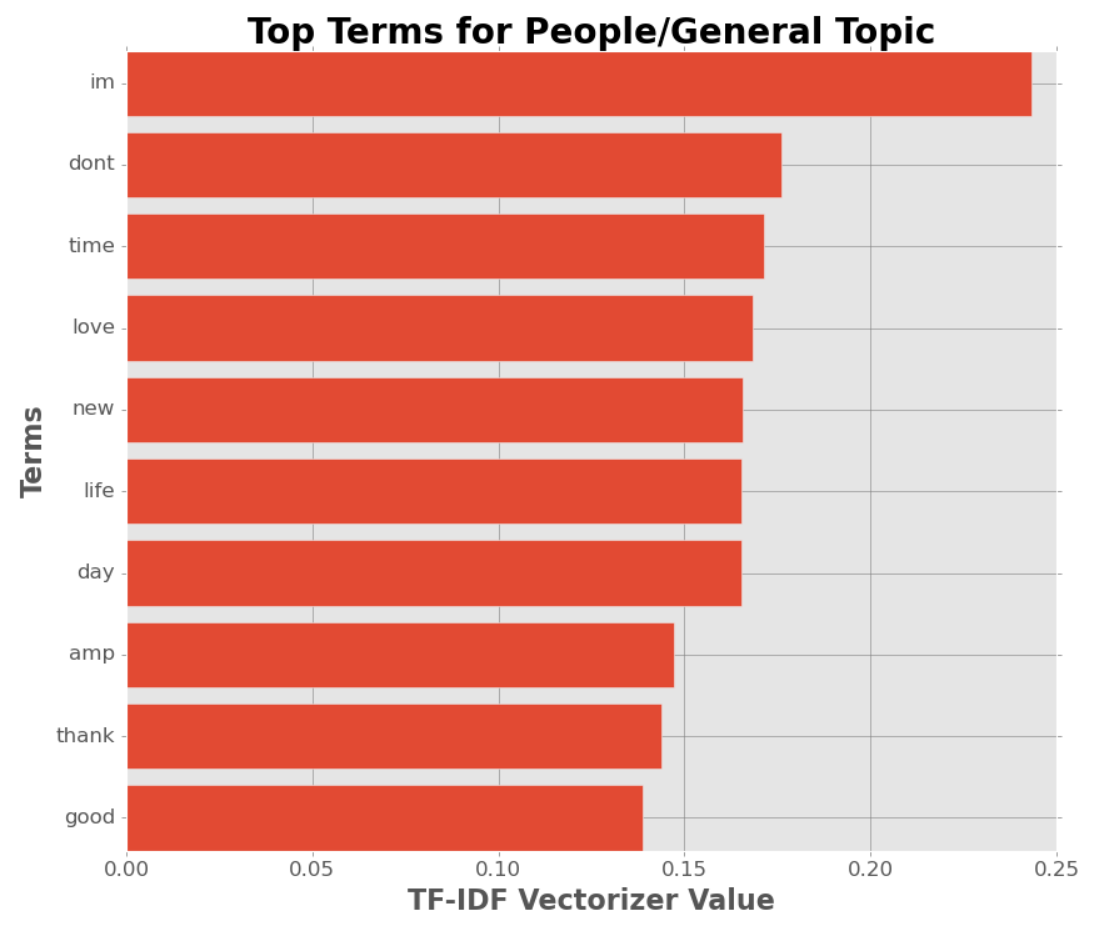
I then used SciKit Learn’s CountVectorizer and TF-IDF Vectorizers to convert each term in the document to a numerical value. I then applied Latent Semantic Analysis (LSA) and Non-Negative Matrix Factorization (also from SciKit Learn), along with Latent Dirichlet Allocation (from Gensm) to compile all terms into 3-5 main topics. This greatly simplified the document-term matrix and reduced dimensionality. I chose the NMF (using the TF-IDF vectorizer) model because was the only model that correctly categorized the two most relevant topics of interest — crypto and sneakers.

With terms in the NMF model reduced to 3 main topics (General, Sneakers, and Crypto) I implemented the cosine-similarity function to compare my personal account’s topic with values with all other 105 documents to determine the top most similar accounts based on my accounts tweets and likes.

**Data**

Using Twitter’s Tweepy API, I scraped 19,759 tweets from 106 different Twitter accounts, consisting of tweets, retweets, and likes. These Twitter accounts are all second degree Twitter followers. For each user I follow, I identified five accounts that that user follows, totaling 958 accounts. I randomly chose 106 of these accounts and extracted 200 tweets each. I would have gathered tweets from more second degree users; however, the scraping process was too long for this project.

**Algorithms**

* Scikit Learn
  + Latent Semantic Analysis (LSA)
  + Non-Negative Matrix Factorization (NMF)
  + Cosine Similarity
  + Count Vectorizer
  + TF-IDF Vectorizer
* Gensim
  + Latent Dirichlet Allocation

**Tools**

* Numpy and Pandas for data manipulation
* Scikit-learn for modeling
* Matplotlib for visualization
* NLTK for Natural Language Processing / Cleaning

**Communication**

