Final Project Notebook

DS 5001 Exploratory Text Analytics | Spring 2024

Metadata

- Full Name: Kristian Olsson
- Userid: kno5cac
- GitHub Repo URL: https://github.com/kristianolsson23/DS5001_AAPL_Reddit_Analysis
- UVA Box URL: https://virginia.box.com/s/yhex3pa1hpcdn2x1dniu06lsxqkajlrh

Overview

The goal of the final project is for you to create a **digital analytical edition** of a corpus using the tools, practices, and perspectives you've learning in this course. You will select a corpus that has already been digitized and transcribed, parse that into an F-compliant set of tables, and then generate and visualize the results of a series of fitted models. You will also draw some tentative conclusions regarding the linguistic, cultural, psychological, or historical features represented by your corpus. The point of the exercise is to have you work with a corpus through the entire pipeline from ingestion to interpretation.

Specifically, you will acquire a collection of long-form texts and perform the following operations:

- **Convert** the collection from their source formats (F0) into a set of tables that conform to the Standard Text Analytic Data Model (F2).
- **Annotate** these tables with statistical and linguistic features using NLP libraries such as NLTK (F3).
- **Produce** a vector representation of the corpus to generate TFIDF values to add to the TOKEN (aka CORPUS) and VOCAB tables (F4).
- **Model** the annotated and vectorized model with tables and features derived from the application of unsupervised methods, including PCA, LDA, and word2vec (F5).
- **Explore** your results using statistical and visual methods.
- **Present** conclusions about patterns observed in the corpus by means of these operations.

When you are finished, you will make the results of your work available in GitHub (for code) and UVA Box (for data). You will submit to Gradescope (via Canvas) a PDF version of a Jupyter notebook that contains the information listed below.

Some Details

- Please fill out your answers in each task below by editing the markdown cell.
- Replace text that asks you to insert something with the thing, i.e. replace `(INSERT IMAGE HERE)` with an image element, e.g. `![](image.png)`.

- For URLs, just paste the raw URL directly into the text area. Don't worry about providing link labels using `[label](link)`.
- Please do not alter the structure of the document or cell, i.e. the bulleted lists.
- You may add explanatory paragraphs below the bulleted lists.
- Please name your tables as they are named in each task below.
- Tasks are indicated by headers with point values in parentheses.

Raw Data

Source Description (1)

Provide a brief description of your source material, including its provenance and content. Tell us where you found it and what kind of content it contains.

This dataset was found on Kaggle. It includes Reddit posts from 2016-2021 of posts that are related to AAPL stock and Apple. This dataset was created to allow people to perform NLP and sentiment analysis to find potential relationships with Apple and AAPL stock. At the meta level, it includes subreddit information such as its ID, name, NSFW status. At the post level, we have the post content, the number of upvotes, the title of the post, a link to the post, and the time it was created in UTC.

As a side note, the data states it is from 2005-2010. However, after checking the posts themselves, they are from 2016-2021.

Source Features (1)

Add values for the following items. (Do this for all following bulleted lists.)

- Source URL: https://www.kaggle.com/datasets/thedevastator/aapl-on-reddit-2005-2010/
- UVA Box URL: https://virginia.box.com/s/7vz6uz9cp7gtsjxg9yu6gk1clsp9uvp6
- Number of raw documents: A CSV with 15483 reddit posts.
- Total size of raw documents (e.g. in MB): 7.1 MB
- File format(s), e.g. XML, plaintext, etc.: CSV

Source Document Structure (1)

Provide a brief description of the internal structure of each document. That, describe the typical elements found in document and their relation to each other. For example, a corpus of letters might be described as having a date, an addressee, a salutation, a set of content paragraphs, and closing. If they are various structures, state that.

Each document contains its subreddit ID, a subreddit name, a boolean value for if the subreddit is NSFW, the time it was created, the link to the post, the domain, a link to websites or images the post may include, the text of the post, the title of the post, and the score of the post (the number of upvotes it has).

Parsed and Annotated Data

Parse the raw data into the three core tables of your addition: the `LIB`, `CORPUS`, and `VOCAB` tables.

These tables will be stored as CSV files with header rows.

You may consider using `|` as a delimitter.

Provide the following information for each.

LIB (2)

The source documents the corpus comprises. These may be books, plays, newspaper articles, abstracts, blog posts, etc.

Note that these are *not* documents in the sense used to describe a bag-of-words representation of a text, e.g. chapter.

- UVA Box URL: https://virginia.box.com/s/8lri1rb37m5z61rckbscdbkje24mwakt
- GitHub URL for notebook used to create: https://github.com/kristianolsson23/DS5001_AAPL_Reddit_Analysis/blob/main/FinalProject_Code.ipy
- Delimitter: ,
- Number of observations: 12 subreddits with a total of 1065 posts. I limited the library to subreddits with over 10 posts that had 5 or more upvotes and 50 or more characters.
- List of features, including at least three that may be used for model summarization (e.g. date, author, etc.): Subreddit name, subreddit description, number of posts, mean/median scores of posts.
- Average length of each document in characters: 937.84 characters

CORPUS (2)

The sequence of word tokens in the corpus, indexed by their location in the corpus and document structures.

- UVA Box URL: https://virginia.box.com/s/82kbtne5dcrn3hrkzyv2p2iazswtc8jl
- GitHub URL for notebook used to create: https://github.com/kristianolsson23/DS5001_AAPL_Reddit_Analysis/blob/main/FinalProject_Code.ipy
- Delimitter: ,
- Number of observations Between (should be >= 500,000 and <= 2,000,000 observations.): 168341.
- OHCO Structure (as delimitted column names): 'subreddit_num','post_num','paragraph_num','sentence_num'
- Columns (as delimitted column names, including `token_str`, `term_str`, `pos`, and `pos_group`): 'subreddit_num', 'post_num', 'paragraph_num', 'sentence_num', 'token_num', 'token_str', 'term_str', 'pos', 'pos_group'

VOCAB (2)

The unique word types (terms) in the corpus.

- UVA Box URL: https://virginia.box.com/s/t7l5gmwdl8hgxtet6aezconblrc6rrxn
- GitHub URL for notebook used to create: https://github.com/kristianolsson23/DS5001_AAPL_Reddit_Analysis/blob/main/FinalProject_Code.ipy
- Delimitter: ,
- Number of observations: 10369
- Columns (as delimitted names, including `n`, `p`, `i`, `dfidf`, `porter_stem`, `max_pos` and `max_pos_group`, `stop`): 'term_str', 'n', 'p', 'i', 'max_pos', 'max_pos_group', 'stop', 'porter_stem', 'dfidf'
- List the top 20 significant words in the corpus by DFIDF; 'one', 'average', 'time', 'an', 'price', 'now', 'in', 'there', 'of', 'up', 'would', 'them', 'not', 'if', 'their', 'the', 'or', 'that', 'other', 'was'

Derived Tables

BOW (3)

A bag-of-words representation of the CORPUS.

- UVA Box URL: https://virginia.box.com/s/8x3vvlwe7ge7irnaki9cvlglki8z1sx5
- GitHub URL for notebook used to create: https://github.com/kristianolsson23/DS5001_AAPL_Reddit_Analysis/blob/main/FinalProject_Code.ipy
- Delimitter: ,
- Bag (expressed in terms of OHCO levels): 'subreddit_num' or OHCO[:1]
- Number of observations: 26413
- Columns (as delimitted names, including `n`, `tfidf`): 'subreddit_num', 'term_str','n', 'tfidf'

DTM (3)

A represenation of the BOW as a sparse count matrix.

- UVA Box URL: https://virginia.box.com/s/tbi48q7r9xic8hnyixaawu768d4k99qn
- UVA Box URL of BOW used to generate (if applicable): https://virginia.box.com/s/8x3vvlwe7ge7irnaki9cvlglki8z1sx5
- GitHub URL for notebook used to create: https://github.com/kristianolsson23/DS5001_AAPL_Reddit_Analysis/blob/main/FinalProject_Code.ipy
- Delimitter: ,
- Bag (expressed in terms of OHCO levels): 'subreddit_num' or OHCO[:1]

TFIDF (3)

A Document-Term matrix with TFIDF values.

- UVA Box URL: https://virginia.box.com/s/ue0a4cab6cigggqtehoakeju9rbnj8vk
- UVA Box URL of DTM or BOW used to create: https://virginia.box.com/s/8x3vvlwe7ge7irnaki9cvlglki8z1sx5
- GitHub URL for notebook used to create: https://github.com/kristianolsson23/DS5001_AAPL_Reddit_Analysis/blob/main/FinalProject_Code.ipy

- Delimitter: ,
- Description of TFIDIF formula ($LaTeX$ OK): I used 'subbredit_num' as the bag of words and then used the sum term factor method. This calculates TF the sum of the term's occurrences in a document, and IDF is the logarithm of the total number of documents divided by the document frequency of the term. Multiplying these two together, you get TFIDF.

Reduced and Normalized TFIDF_L2 (3)

A Document-Term matrix with L2 normalized TFIDF values.

- UVA Box URL: https://virginia.box.com/s/iochznoljujb90cmby1kps5k5xrdopl3
- UVA Box URL of source TFIDF table: https://virginia.box.com/s/ue0a4cab6cigggqtehoakeju9rbnj8vk
- GitHub URL for notebook used to create: https://github.com/kristianolsson23/DS5001_AAPL_Reddit_Analysis/blob/main/FinalProject_Code.ipy
- Delimitter: ,
- Number of features (i.e. significant words): 3791
- Principle of significant word selection: I chose words that had a DFIDF value over the mean DFIDF value of across the entire vocabulary which was 4.70. I also subset words that are adjectives, verbs, and nouns.

Models

PCA Components (4)

- UVA Box URL: https://virginia.box.com/s/brfpyvxvut18yw32p4k5c1n365wkse98
- UVA Box URL of the source TFIDF_L2 table: https://virginia.box.com/s/iochznoljujb90cmby1kps5k5xrdopl3
- GitHub URL for notebook used to create: https://github.com/kristianolsson23/DS5001_AAPL_Reddit_Analysis/blob/main/FinalProject_Code.ipy
- Delimitter: ,
- Number of components: 5
- Library used to generate: Code from module 7.01
- Top 5 positive terms for first component: 'ampnbsp', 'spreadsheet', 'iv', 'calls', 'call'
- Top 5 negative terms for second component: 'iphone', 'amp', 'calls', 'sales', 'saw'

PCA DCM (4)

The document-component matrix generated.

- UVA Box URL: https://virginia.box.com/s/kcc27zee9xignyni7d2zkygp4s49h8ev
- GitHub URL for notebook used to create: https://github.com/kristianolsson23/DS5001_AAPL_Reddit_Analysis/blob/main/FinalProject_Code.ipy
- Delimitter: ,

PCA Loadings (4)

The component-term matrix generated.

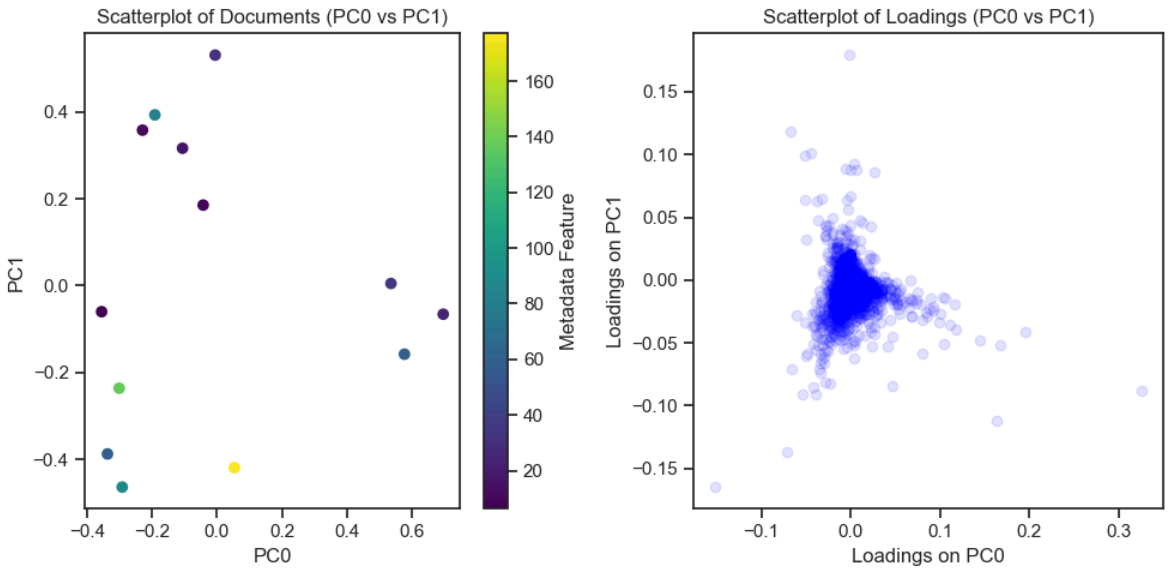
- UVA Box URL: https://virginia.box.com/s/dty4x1t184ju0o6d2k3qkkheq2pbmx2r
- GitHub URL for notebook used to create: https://github.com/kristianolsson23/DS5001_AAPL_Reddit_Analysis/blob/main/FinalProject_Code.ipy
- Delimitter: ,

PCA Visualization 1 (4)

Include a scatterplot of documents in the space created by the first two components.

Color the points based on a metadata feature associated with the documents.

Also include a scatterplot of the loadings for the same two components. (This does not need a feature mapped onto color.)



Briefly describe the nature of the polarity you see in the first component:

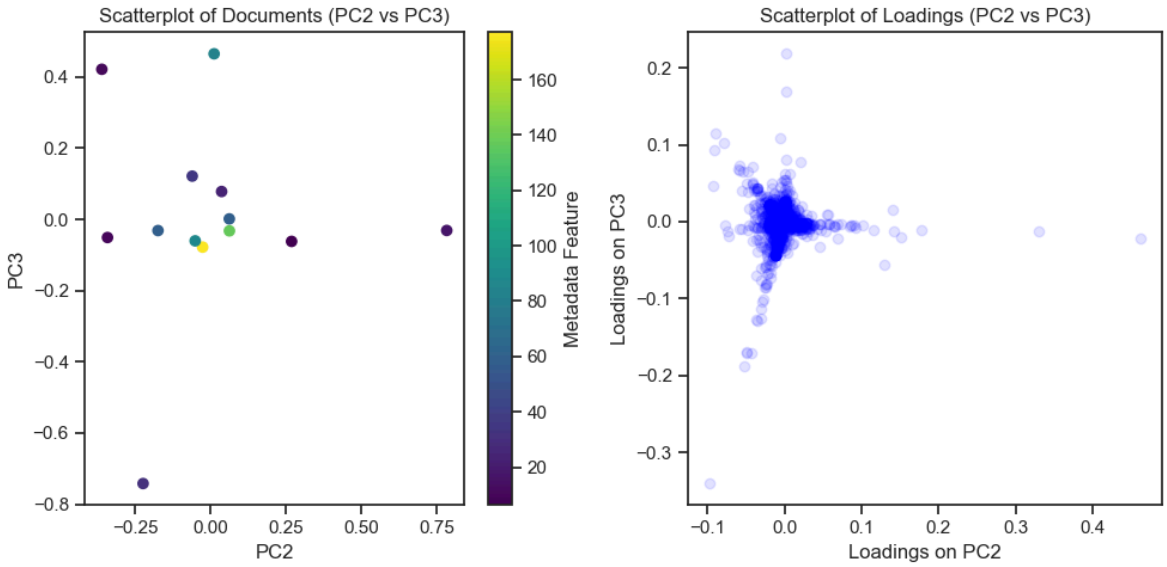
The metadata feature chosen is the mean average number of upvotes in each subreddit. The plot on the left shows that subreddits with higher upvote averages tend to have content that is negative in both principle component 0 and 1. These principle components relate to iphones, China, calls, dividends, and sales, indicating a general notion of sales revenue and AAPL option/stock. The plot on the right shows the contribution of each word to the principal components, which seem to cluster around 0.0 on PC0 and vary much more on PC1.

PCA Visualization 2 (4)

Include a scatterplot of documents in the space created by the second two components.

Color the points based on a metadata feature associated with the documents.

Also include a scatterplot of the loadings for the same two components. (This does not need a feature mapped onto color.)



Briefly describe the nature of the polarity you see in the second component:

The metadata feature chosen is the mean average number of upvotes in each subreddit. The plot on the left shows that subreddits with higher upvote averages tend to have content that is neutral or centered around 0 in both both principle component 2 and 3. These principle components relate to forecasts, alerts, dividends, payouts, and packages, indicating a general notion of forecasted/guaranteed profits. The plot on the right shows the contribution of each word to the principal components, which seem to cluster around 0.0 on both components, with some trails of outliers going in a few directions.

LDA TOPIC (4)

- UVA Box URL: https://virginia.box.com/s/d87kexl0nwczdgrqyvq3p3lt2hgpycrh
- UVA Box URL of count matrix used to create: https://virginia.box.com/s/tbi48q7r9xic8hnyixaawu768d4k99qn
- GitHub URL for notebook used to create: https://github.com/kristianolsson23/DS5001_AAPL_Reddit_Analysis/blob/main/FinalProject_Code.ipy
- Delimitter: ,
- Libary used to compute: LatentDirichletAllocation from sklearn
- A description of any filtering, e.g. POS (Nouns and Verbs only): Nouns
- Number of components: 5
- Any other parameters used: bag = subreddit_num or OHCO[:1], ngram_range = (1, 2), n_top_terms = 5, max_features = 1000, stop_words = 'english', max_iter = 5, learning_offset = 50, random_state = 0 ngram_range
- Top 5 words and best-guess labels for topic five topics by mean document weight:
  - T01: apple stock market year company, Company Performance
  - T02: apple earnings options week stock, Earnings and Options
  - T00: apple iphone market earnings stock, iPhone Market Impact
  - T04: apple stock calls im earnings, Stock and Earnings Calls

- T03: table perfect averages reversal setup, Financial Market Analysis

LDA THETA (4)

- UVA Box URL: https://virginia.box.com/s/3u0qpd0gaewy8n8ycvubrgarq64eayt9
- GitHub URL for notebook used to create: https://github.com/kristianolsson23/DS5001_AAPL_Reddit_Analysis/blob/main/FinalProject_Code.ipy
- Delimitter: ,

LDA PHI (4)

- UVA Box URL: https://virginia.box.com/s/muo2d6xzps4fyjqn2qaqpk6uq3vz6135
- GitHub URL for notebook used to create: https://github.com/kristianolsson23/DS5001_AAPL_Reddit_Analysis/blob/main/FinalProject_Code.ipy
- Delimitter: ,

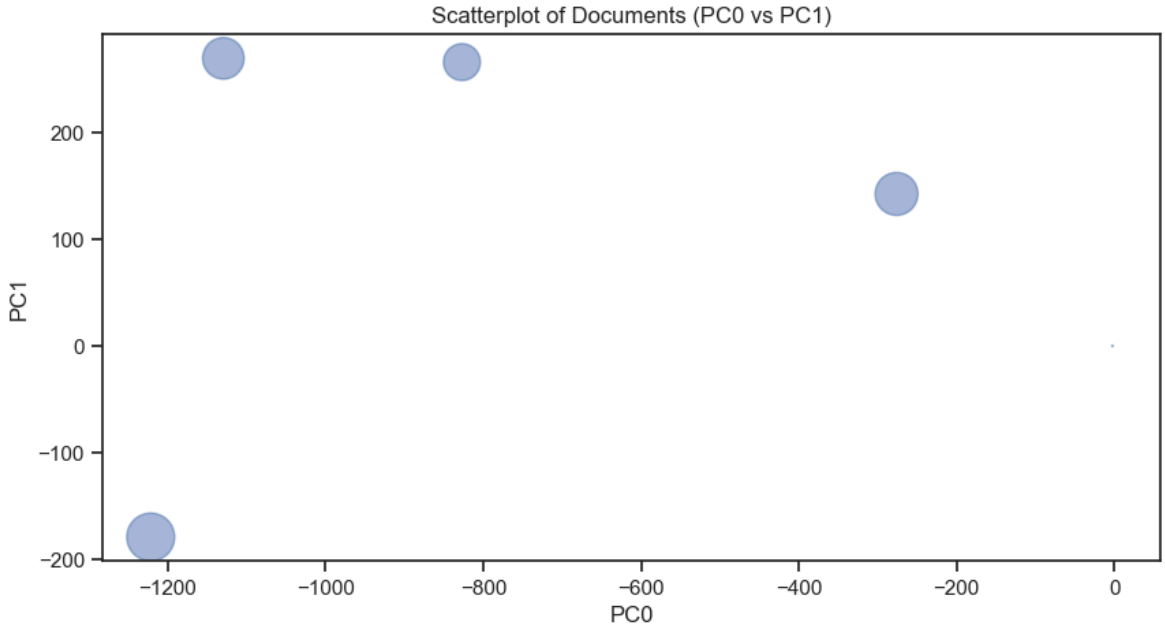
LDA + PCA Visualization (4)

Apply PCA to the PHI table and plot the topics in the space opened by the first two components.

Size the points based on the mean document weight of each topic (using the THETA table).

Color the points basd on a metadata feature from the LIB table.

Provide a brief interpretation of what you see.



This plot of PCA of LDA shows that PC0 does well to capture the variation in the topics while PC1 mainly splits T01 (Company Performance, bottom left) from the other 4 topics. PC1 also shows the similarities between T02 (Earnings and Options top right), T00 (iPhone Market Impact, top left), T04 (Stock and Earnings Calls, top middle), which is to be expected. There is a small point representing

T03 at PC0=0 and PC1=0 that is small and hard to see, representing its low topic weight and importance.

Sentiment VOCAB_SENT (4)

Sentiment values associated with a subset of the VOCAB from a curated sentiment lexicon.

- UVA Box URL: https://virginia.box.com/s/grlqn5plxbderi7zru619vk24huyktef
- UVA Box URL for source lexicon: https://virginia.box.com/s/kgk3b0q07tmcvz94zmawfb4jh843nnpa
- GitHub URL for notebook used to create: https://github.com/kristianolsson23/DS5001_AAPL_Reddit_Analysis/blob/main/FinalProject_Code.ipy
- Delimitter: ,

Sentiment BOW_SENT (4)

Sentiment values from VOCAB_SENT mapped onto BOW.

- UVA Box URL: https://virginia.box.com/s/bw9u0jf3m7lb4mb4v4islpvtulpshxdq
- GitHub URL for notebook used to create: https://github.com/kristianolsson23/DS5001_AAPL_Reddit_Analysis/blob/main/FinalProject_Code.ipy
- Delimitter: ,

Sentiment DOC_SENT (4)

Computed sentiment per bag computed from BOW_SENT.

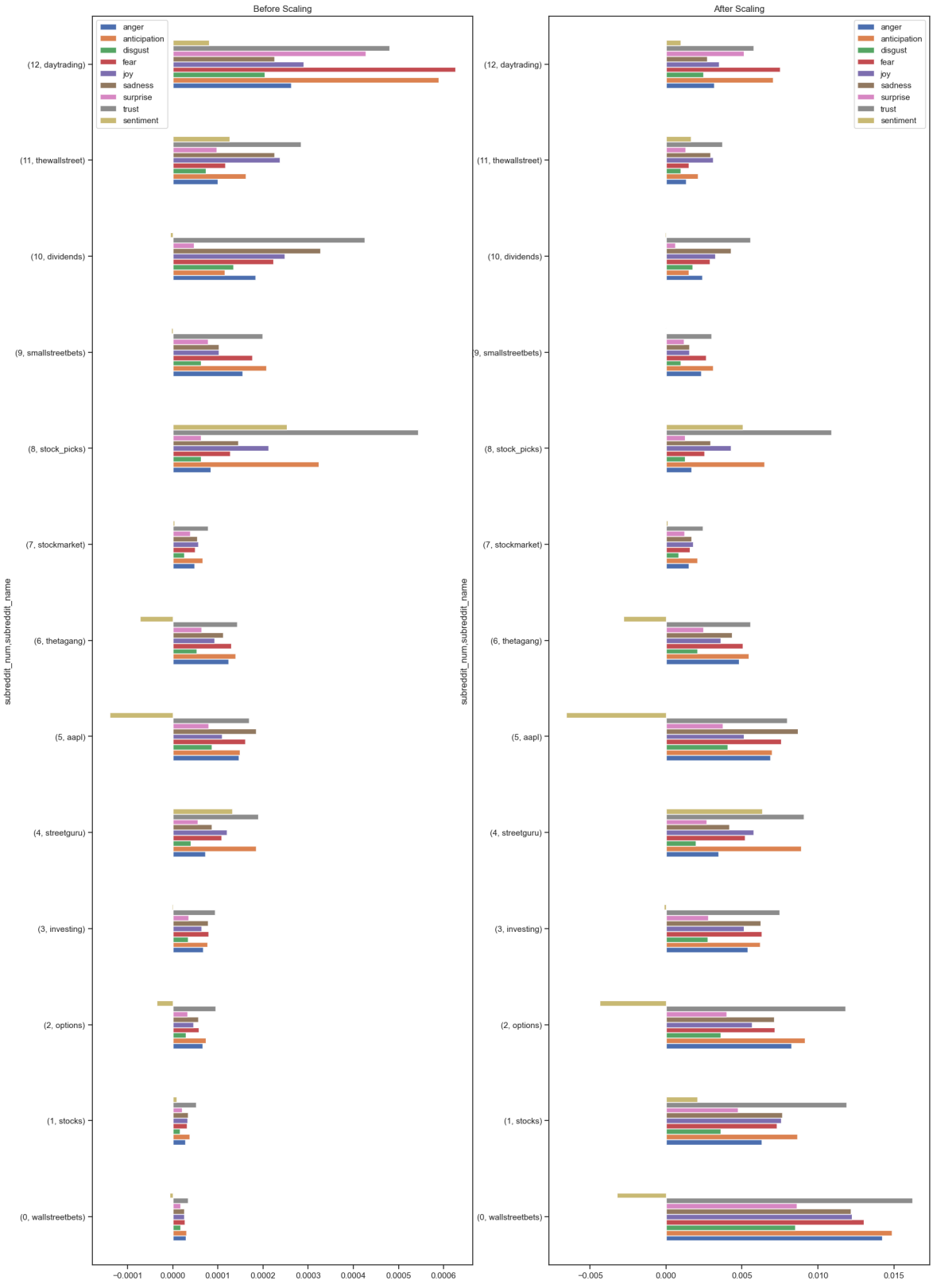
- UVA Box URL: https://virginia.box.com/s/j574cusogcznda82rcsjowk023bd98vo
- GitHub URL for notebook used to create: https://github.com/kristianolsson23/DS5001_AAPL_Reddit_Analysis/blob/main/FinalProject_Code.ipy
- Delimitter: ,
- Document bag expressed in terms of OHCO levels: subreddit_num or OHCO[:1]

Sentiment Plot (4)

Plot sentiment over some metric space, such as time.

If you don't have a metric metadata features, plot sentiment over a feature of your choice.

You may use a bar chart or a line graph.



The plot on the right shows initial sentiment scores for each subreddit. However, subreddits at the top have significantly less posts than those at the bottom, which may lead to skewed results as there could be a few posts with more extreme sentiment, so I scaled the emotion values by the number of posts as shown on the right.

VOCAB_W2V (4)

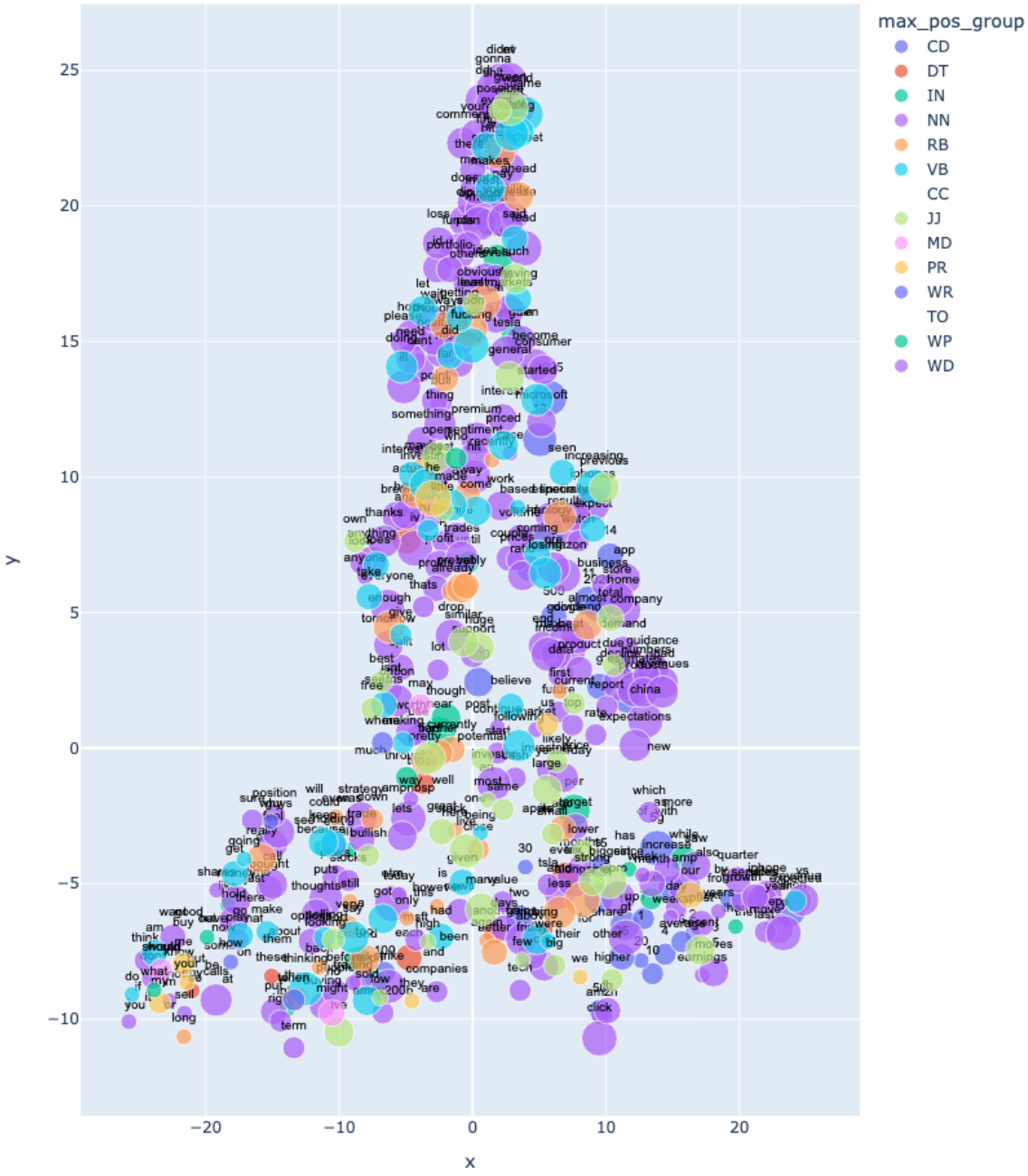
A table of word2vec features associated with terms in the VOCAB table.

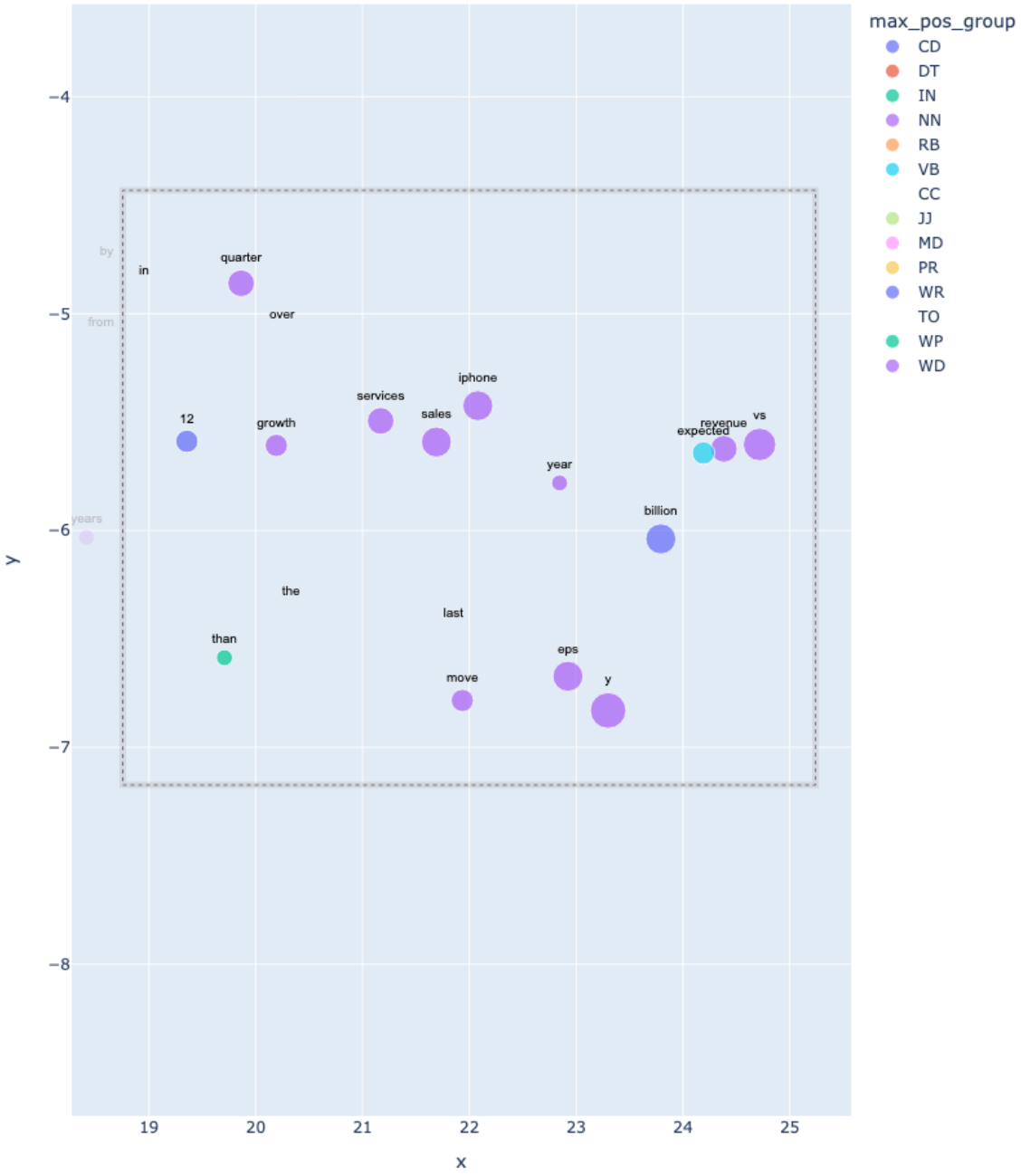
- UVA Box URL: https://virginia.box.com/s/dc0hks6esd7wsbns89bs24a60a80llli
- GitHub URL for notebook used to create: https://github.com/kristianolsson23/DS5001_AAPL_Reddit_Analysis/blob/main/FinalProject_Code.ipy
- Delimitter: ,
- Document bag expressed in terms of OHCO levels: subreddit_num or OHCO[:1]
- Number of features generated: 246
- The library used to generate the embeddings: gensim.models word2vec

Word2vec tSNE Plot (4)

Plot word embedding featues in two-dimensions using t-SNE.

Describe a cluster in the plot that captures your attention.


max_pos_group
CD
DT
IN
NN
RB
VB
CC
JJ
MD
PR
WR
TO
WP
WD
y
x
25
20
15
10
5
0
−5
−10
−20
−10
0
10
20



I found a cluster between x=(19,25), y=(-7,-5) that captured my attention. There are many financial terms such as eps (earnings per share), revenue, and expected, potentially referring to expected revenue. There are also terms relating to periods in finance such as quarter and year. Lastly, terms that relate to the performance of the company such as services, sales, growth, and most importantly, the iPhone. These represent many of the main forces in considering the performance of Apple and AAPL stock

Riffs

Provde at least three visualizations that combine the preceding model data in interesting ways.

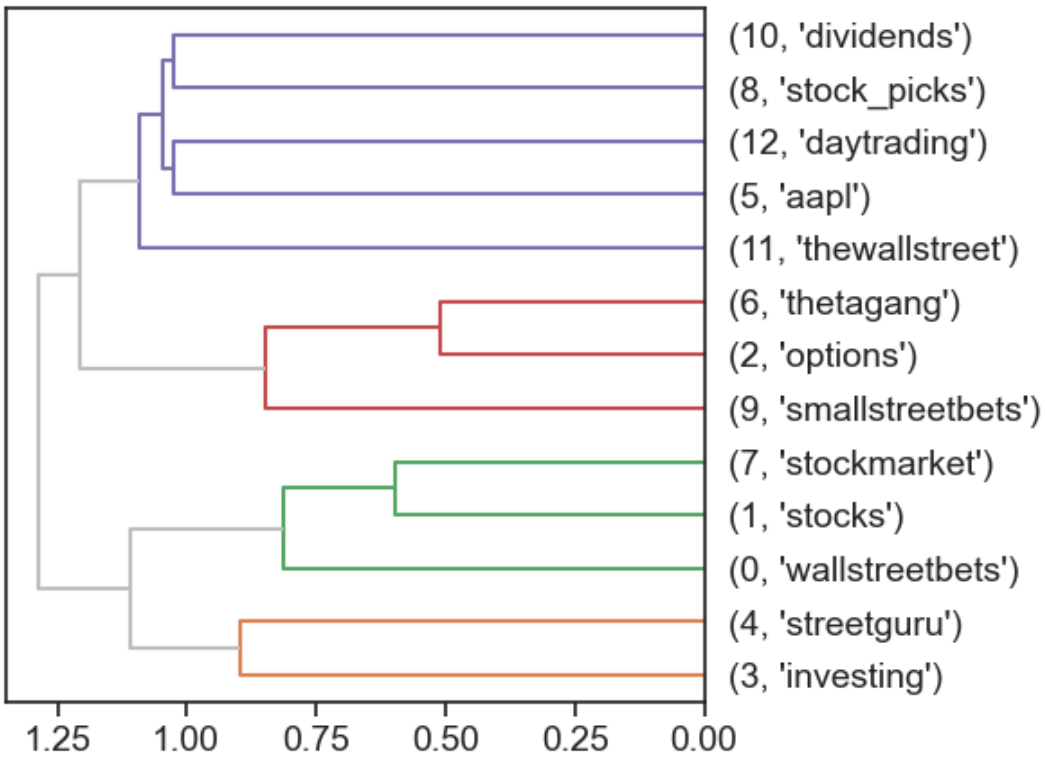
These should provide insight into how features in the LIB table are related.

The nature of this relationship is left open to you -- it may be correlation, or mutual information, or something less well defined.

In doing so, consider the following visualization types:

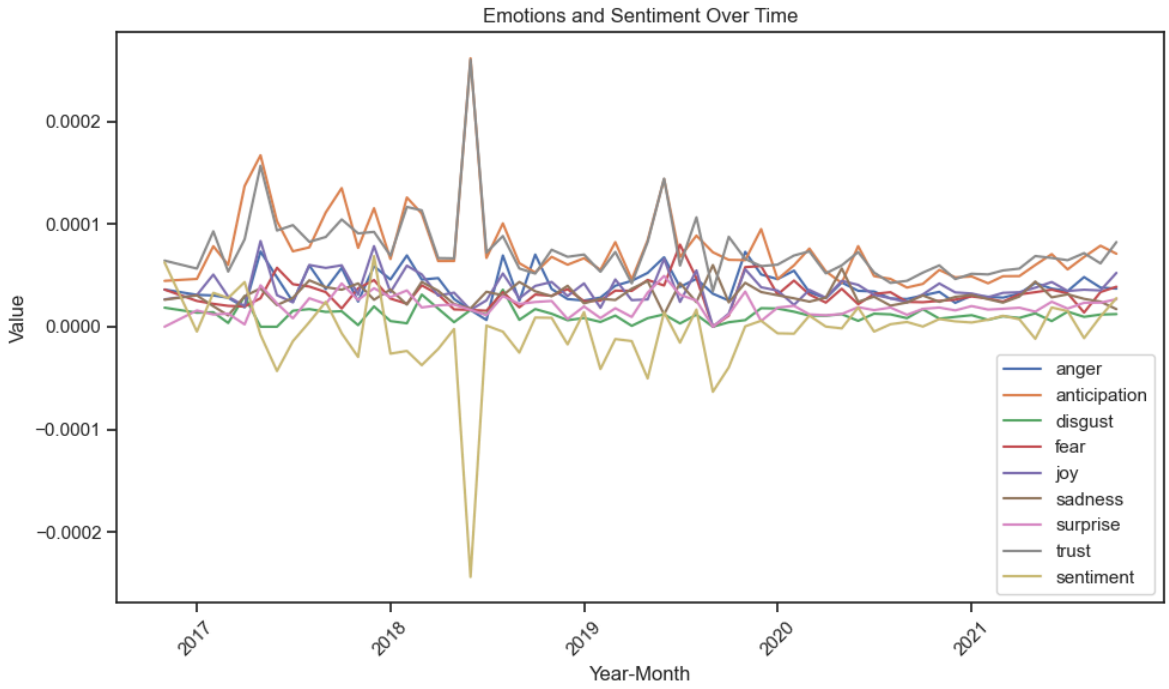
- Hierarchical cluster diagrams
- Heatmaps
- Scatter plots
- KDE plots
- Dispersion plots
- t-SNE plots
- etc.

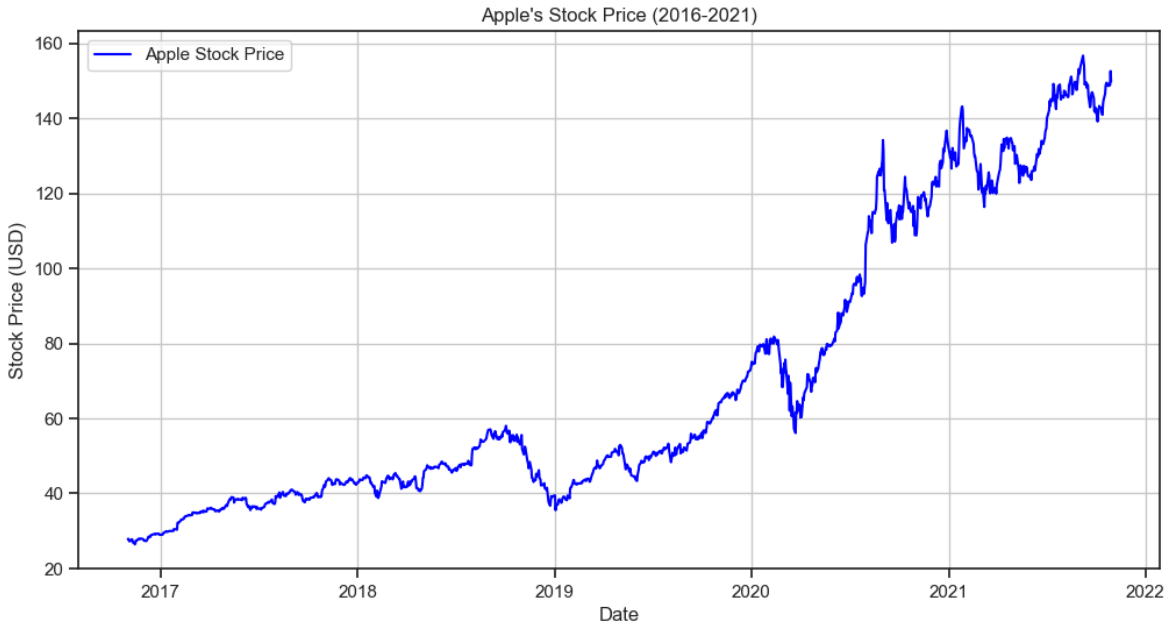
Riff 1 (5)



Using a hierarchical cluster with a cosine similarity score, we are able to see the similarities between specific subreddits. The top purple cluster represents stock trading, dividends, and Apple stock specifically. The red cluster represents options as all three subreddits are described as option focused (where theta measures the rate of time decay of an option's premium). The green cluster, similar to the purple cluster, represents stock trading but with a more general focus on investing and the market. The orange cluster represents a couple of subreddits with a general interest in investing with no specific focus.

Riff 2 (5)



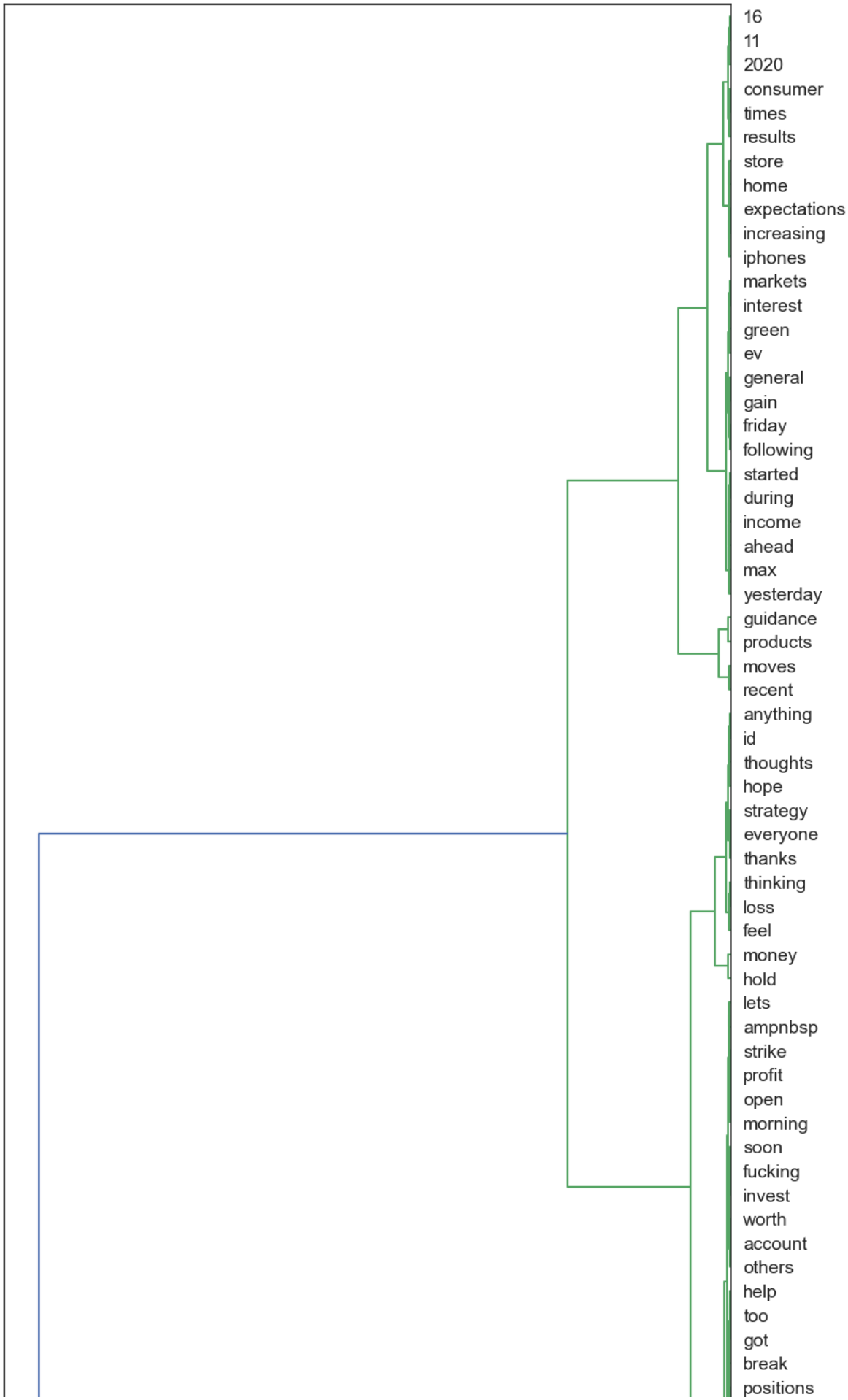


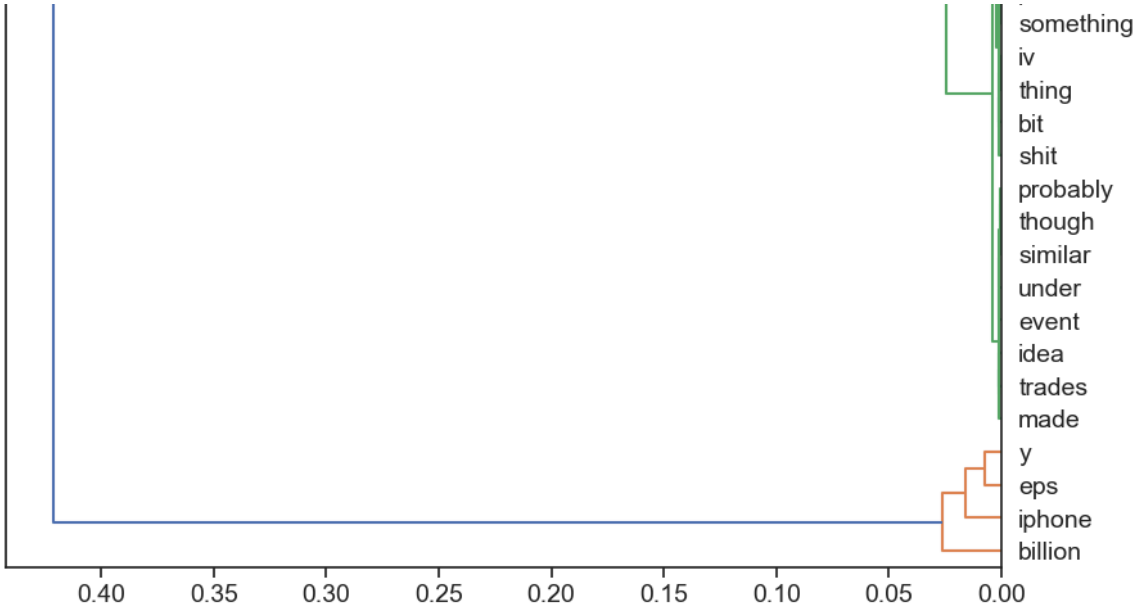
This plot represents emotions and sentiment over time. From the end of 2016 to mid 2018, we see some signal across the emotions, with a high spike mid 2018 although this is mainly due to a lack of posts in that month. The emotions are calmer between mid 2018 and mid 2019, before becoming quite noisy again in mid 2019. From 2020 onwards, it is relatively calm. I also attached a plot of AAPL's performance across the same period. We can see a steady increase in price over time, with a large growth from COVID onwards. There are a couple of drops in price in late 2018 and the early 2020, which do not correlate exactly to the sentiments. However, it feels like there is a time lag from the sentiments to the market, as there are big drops and changes in sentiment and emotions a couple of months before these drops in prices. This may indicate there is some correlaton between public opinion and stock price, although it is not a significant relationship based on these plots.

In terms of specific emotions, it is worth noting trust and anticipation are very highly correlated while sentiment seems to have an inverse relationship with these two. It may indicate that while people's

sentiment is low, there is a lot of hope that it will improve.

Riff 3 (5)


16
11
2020
consumer
times
results
store
home
expectations
increasing
iphones
markets
interest
green
ev
general
gain
friday
following
started
during
income
ahead
max
yesterday
guidance
products
moves
recent
anything
id
thoughts
hope
strategy
everyone
thanks
thinking
loss
feel
money
hold
lets
ampnbsp
strike
profit
open
morning
soon
fucking
invest
worth
account
others
help
too
got
break
positions



Using a hierarchical cluster on Word2Vec inputs, we are able to find relationships between specific words and groups of words. I sampled 75 words that are nouns and have a DFIDF value greater than the mean DFIDF value. There are two main clusters, with just 4 words in the orange cluster representing an array of topics such as iPhones, billion, and earnings. Within the subclusters in the larger green cluster, there are groups of words that closely relate such as thoughts, hope, strategy, thanks, thinking, loss, feel all relating to emotion and coming up with a strategy. We also see a more consumer focused cluster with 2020, consumer, iPhone, results, store, home, expectations, and increasing near the top. This cluster diagram provides a great way to visually group words and find some closely related words within this corpus.

Interpretation (4)

Describe something interesting about your corpus that you discovered during the process of completing this assignment.

At a minumum, use 250 words, but you may use more. You may also add images if you'd like.

To be honest, I was unsure but excited about what relationships and patterns I would be able to find with this dataset, especially considering how small it is. However, I was blown away by some of the insights I came upon. A few that I find to be very interesting are:

- A large change in sentiment over time could be an indicator of decreases to stock price, while a less emotional and calmer sentiment may indicate increases in stock price, as shown in the sentiment plot across time.
- Every subreddit related to options has a negative sentiment, likely due to the higher risk/reward the asset provides, as shown in the sentiment plot across subreddits. On the other hand, subreddits related to day trading and stock trading had heightened levels of fear, which makes sense as you may be constantly watching and reacting to price changes.
- Clustering is a reliable method to find relationships between different values, whether that be in clustering words that relate to consumers or finding subreddits specific to stock versus options

trading.
- t-SNE plots are effective in reducing the dimensionality of words and creating groups of closely related terms, such as words that indicate the performance of Apple.

This type of quantitative analysis into public opinion can provide many insights into what investors are researching and considering, what sentiments they might have based on what community (subreddit) they are a part of or when they post, and changes to stock price that may occur as a result. Overall, this project and this course was very mind-opening into understanding language from a qualitative and quantitative perspective.