**Top Personal Finance Concerns: the Content Analysis Based on The Reddit Data**

MACSS Jinfei Zhu ([jinfei@uchicago.edu](mailto:jinfei@uchicago.edu))

Introduction

In 2017, a Federal Reserve survey(Federal Reserve, 2018) finds almost 40% of American adults wouldn't be able to cover a $400 unexpected emergency expense with cash, savings or a credit-card charge that they could quickly pay off, saying that they would either not be able to cover it or would cover it by selling something or borrowing money. Why do people in the United States, the most powerful country in the world, so ill-prepared for? What's the heaviest financial burden on people? What are the topics that people who seek financial security talks every day? Do these topic change over time?

In the past, to answer these questions, we may have to turn to the power of survey to ask each interviewee questions individually, but now there are many people discuss their financial concerns online and post their thoughts, questions and suggestions online, in the subreddits from Reddit website. So, we can scrape the text data and do some content analysis including counting, clustering and word embedding to dig the information behind the content data.

In Macroeconomics, there is a formula showing the equilibrium of output and input of economics yields:

Y = C + I + G + NX

(Total economic output = Consumption + Investment + Government spending + Net Export). It shows that the consumption and investment are individual activities that constitute of our society. Therefore, studying the personal finance concerns will give us better insight about macroeconomics about individual’ financial behaviors, including consumptions, debts, loans, investments, insurance and retirement planning. My study will report the most common financial burden on people, and the time trend of the changes most-discussed topics. In this way, people can know what bothers us and if the things that bother us change over time.

Here is the structure of my analysis:

First, data collection: Use Reddit API and the Python Reddit API Wrapper (PRAW) to scrape the data from Personal Finance, Wall Street Bets and Investing subreddit so we can to tokenize, normalize and vectorize the text data.

Second, counting the words and phrases: count the frequency of key words and n-grams in reddit post, do part-of-speech tagging and find the difference from personal finance reddit and other finance-related subreddit.

Third, clustering and topic modelling: Do Latent Dirichlet Allocation Topic Modelling for texts of subreddits.

Fourth, word embedding and projections to see the more information about the project.

Data

Reddit is the social news platform, web content rating and discussion website, recently including livestream functions. In reddit, there are many subreddits which are forums to a specific topic. People's online discussion is a good reflection of their real-life concerns and thinking.

I scraped top Reddit articles from different subreddits.

* Personal Finance
* Investing (Name now: lose money with friends!)
* Wall Street Bets

In Reddit world, ‘top’ means recent popular posts. Reddit API has a limitation of no more than 1000 posts scraping each time. Because I want to get the latest posts, I scraped data by myself and the size of texts from each subreddit is around 1000, and even less for Wall Street Bets because many posts are just video clips of news. For dynamic topic modeling in the third part of this paper, I download archived Reddit data from Google cloud to enlarge the size of my corpora.

Personal Finance subreddit (r/personal finance) is created on Feb 9, 2009, right after the 2008 financial depression. It has 14.4 million members and usually have 14.9k members online. It’s a very active and large subreddit compared to other relatively subreddit and well-organized. Every hour there are many new discussions about topics such as debts, loans, housing, auto, insurance, investing, retirement, taxes, budgeting and income. People post their concerns, seek for advice, or share personal experience. There is a very detailed wiki for this subreddit (Reddit, 2021) listing a summary of suggestions for different age of people. Therefore, I think it would be a good choice to scrape text data from for my personal finance concern analysis. The sample size is 977.

Wall Street Bets subreddit (r/wallstreetbets) is created at Jan 31, 2012, with 9.6 million members and 255k daily online active members. It’s smaller compared to Personal Finance subreddit in terms of members, but much more active in terms of daily online members. Participants would discuss stock and option trading strategies on it. It becomes popular and famous for its aggressive trading strategies and role in the GameStop Short Squeeze that caused losses on Wall Street hedge funds short sellers up to US$70 billion in a few days in early 2021. Perhaps due to its popularity, a lot of posts on this reddit are news videos, so the text content from this subreddit is significantly less than other two subreddits, with a sample size of only 157.

Investing subreddit (r/investing) is created at March 15, 2008, the oldest one among these three subreddit. It changed its name to ‘lose money to your friend because only after few months of its creation, the stock market crashed in 2008. But it only has 1.8 million members and 9k members online. Compared to Wall Street Bets, it’s a place to start if members don’t know anything about investing and begin to learn it. We add this subreddit into our corpora to compare the difference of content across different subreddits. The number of posts from Investing subreddit is 954.

People are anxious about money, paying debt and managing their asset and making investments. Most people who post articles on Reddit are young people, many of them are 20-30 (many people reveal their age in posts on Personal Finance discussion) and it's interesting to learn the consumption and investment patterns of these young people. They are a large group of anxious young people--we can find students who just got their first job start to consider paying back student loan, buying houses or cars with loans, starting to think of taking care of aging parents, for the first time in their life. They ask advice from others on online platform to make finance-wise decisions and many kind people offer their kind suggestions to others.

Here are some examples of the posts:

Personal Finance

**Title: “**If you can't get your emergency fund to grow because of emergencies that keep coming up, you're still doing a good job.”

**Text:** “Over the summer I made a steadfast commitment to getting my 3 month emergency fund built, which is only about 15k. I'm saving $750 a month, which is exactly 15% of my family's post-tax income. In the 3 months since I made that change, I've had $1.8k in car repairs, $600 in vet bills, and $250 to cover a friend who got towed from our guest parking (our fault). Needless to say, the needle hasn't moved as I wanted it to, and I have to keep reassuring myself that, had I not made this commitment, I'd be in real trouble covering these costs. The end goal will come eventually. EDIT: Just to clarify - this is a two person budget!.”

Wall Street Bets:

**Title:** “CLASS ACTION AGAINST ROBINHOOD. Allowing people to only sell is the definition of market manipulation. A class action must be started, Robinhood has made plenty of money off selling info about our trades to the hedge funds to be able to pay out a little for causing people to lose money now”

**Text:** “LEAVE ROBINHOOD. They don’t deserve to make money off us after the millions they caused in losses. It might take a couple of days, but send Robinhood to the ground and GME to the moon.”

Investing;

**Title:** “Robinhood and other brokers literally blocking purchase of $GME, $NOK, $BB, $AMC; allow sells”

**Text:** “See title. Can't buy these stocks on RH, but can sell. What the hell is this? How is this legal?”

Counting Words and Phrases

Word Frequency

In this part, I use Python package spaCy for a series of natural language processing method for English corpora. First, we can tokenize the texts and have a look at the top words in these three subreddits.

Table Top 20 words

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Personal Finance | | | Wall Street Bets | | | Investing | | |
| rank | word | count | rank | word | count | rank | word | count |
| 1 | $ | 2236 | 1 | $ | 295 | 1 | $ | 1677 |
| 2 | money | 1153 | 2 | gme | 285 | 2 | > | 902 |
| 3 | credit | 1006 | 3 | shares | 234 | 3 | market | 854 |
| 4 | time | 872 | 4 | 🚀 | 217 | 4 | company | 629 |
| 5 | pay | 834 | 5 | short | 191 | 5 | stock | 549 |
| 6 | edit | 811 | 6 | people | 163 | 6 | price | 536 |
| 7 | account | 809 | 7 | edit | 155 | 7 | said | 470 |
| 8 | like | 771 | 8 | market | 155 | 8 | year | 462 |
| 9 | know | 694 | 9 | like | 148 | 9 | people | 455 |
| 10 | people | 692 | 10 | buy | 147 | 10 | like | 434 |
| 11 | work | 627 | 11 | stock | 139 | 11 | time | 407 |
| 12 | years | 617 | 12 | money | 135 | 12 | = | 372 |
| 13 | year | 590 | 13 | price | 131 | 13 | money | 360 |
| 14 | job | 581 | 14 | know | 119 | 14 | billion | 358 |
| 15 | 2 | 580 | 15 | fucking | 118 | 15 | companies | 355 |
| 16 | car | 572 | 16 | time | 112 | 16 | years | 329 |
| 17 | going | 569 | 17 | going | 101 | 17 | shares | 319 |
| 18 | bank | 549 | 18 | sell | 93 | 18 | value | 318 |
| 19 | got | 541 | 19 | hedge | 90 | 19 | short | 307 |
| 20 | want | 539 | 20 | want | 88 | 20 | stocks | 295 |

We find many finance-related words, the dollar sign ($) is the most frequent word and the second is the word money, which is definitely the center of the discussion. We also have a lot of talks about credit and paying, as well as bank account and many time related words such as "year" and "month" "time". People also talk about getting a job (perhaps due to the surged unemployment rate during pandemic) and car since cars are really important in American cultures, such as car insurance and car loan. Because recent GME Short Squeeze event, the word ‘gme’ is the second most discussed word in Wall Street Bets subreddit. There are also more words related to stocks and hedge fund such as short, buy and sell. In Investing there are words like “stock” “market” and “companies”.

We can also study the relationships of ranking and counts of the words

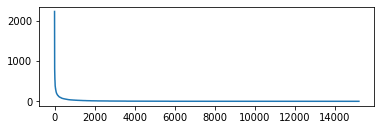


Figure Rankings and frequencies of the word

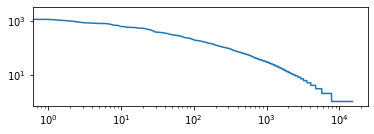


Figure log-ranking and log-frequency of the word

This shows that likelihood of a word occurring is inversely proportional to its rank. This effect is called Zipf's Law which suggests that the most frequent word will occur approximately twice as often as the second most frequent word, three times as often as the third most frequent word, etc. There is almost a linear relationship of the log(ranking) and log(frequency).

We can also look at when we talk about a specific word, what’s the context of the word. For example, when people use the word ‘student’ in personal finance subreddit, nearly 100% they are talking about student loan. This is due to the United States is a leader in educational loans and tuition has increased rapidly.

rding to the article employees with **student** loan debt accumulate 50 less wealth

by age 30 than their peers without **student** loan debt i think most of us with s

t loan debt i think most of us with **student** debt have at one point or another f

you would be able to make qualified **student** loan payments and have your company

ch month you made a payment on your **student** loan this does n't hurt people with

We can see this rule again in the following table showing n-gram in Personal Finance. The first bigram is ‘credit card’, followed by ‘student loan’ (‘r/personalfinance is the URL for hyperlink, it has been referred so often because people post link to their wiki for reference, so we can safely ignore it).

Table Top n-gram of personal finance

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **bigram** | **likelihood** | **bigram** | **t** | **trigram** | **t** |
| (credit, card) | 2643.3 | (credit, card) | 18.4 | (credit, card, debt) | 6.2 |
| (r, personalfinance) | 1550.0 | (student, loan) | 12.5 | (r, personalfinance, wiki) | 5.8 |
| (student, loan) | 1459.9 | ($, month) | 12.3 | (domain, core, finance) | 5.5 |
| (emergency, fund) | 867.2 | (feel, like) | 10.6 | (finance, domain, core) | 5.5 |
| (wells, fargo) | 819.7 | (r, personalfinance) | 10.4 | (economic, finance, domain) | 4.8 |
| (feel, like) | 772.1 | (year, ago) | 9.6 | (pay, credit, card) | 4.7 |
| (credit, score) | 678.8 | (edit, thank) | 9.4 | (use, credit, card) | 4.4 |
| (year, ago) | 640.5 | (credit, score) | 9.3 | ( , $) | 4.2 |
| ($, month) | 628.4 | (m, sure) | 9.3 | (long, story, short) | 4.1 |
| (debit, card) | 571.5 | (bank, account) | 9.2 | (credit, card, company) | 4.1 |

This is the lexical dispersion plot of our corpora. we find that most words appear evenly in the corpora, but “savings” and “mortgage” appear less often than others. “credit”, “bank, “loan”, “debt” and “car” are the most popular words.

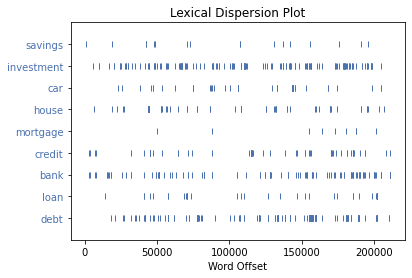


Figure Lexical Dispersion Plot

After we compare the frequency and distribution of tokenized words, we can normalize them to dig more about it. Normalization of texts means we first make all of the words lower case, drop non-word tokens, remove ‘stop words’ (we use stop words list of spaCy to do this), stem the remaining words to remove suffixes, prefixes and infixes, or lemmatize tokens by grouping variant forms of the same word.

The following plot is the conditional frequency distributions of the data by using spaCy’s conditionalFreqDist class. We use word lengths as conditions, though tags or clusters could provide more useful results. So we can find that words with largest conditional probability are different from the top words in the earlier part.

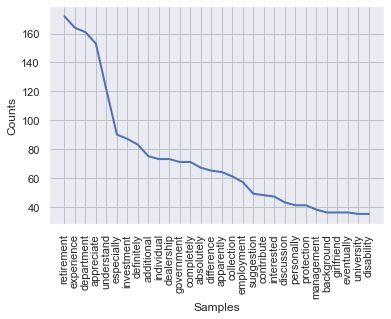
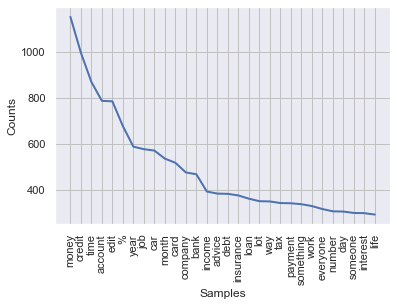


Figure Conditional Probability Distribution

Next, we can draw the Part of Speech (POS) to word conditional distribution, to inspect what are the top nouns or adjectives or verbs.



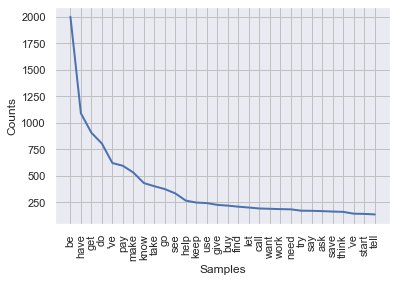
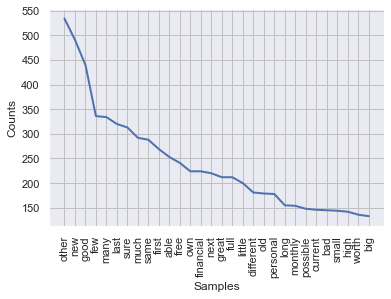


Figure Top Adjectives and verbs

From the visualization, we find that top nouns including “money”, “credit”, “time”, “account”, “year”, “job”, “car”, “month”, “card”, “company”, “income”, “advice”, “debt”, “insurance”, “bank”, “payment” and “interest”. These are the top nouns in financial word. We find that they are all about financially related.

Top adjectives used in Personal Finance are “other”, “new”, “good”, “few”, “many”, “last”, “first”, “able”, “free”, “own”, “financial” etc. They could be used to describe numbers, houses, cars, jobs, loans and debts. Top verbs are less informative, including the most frequent daily-used verb such as “be”, “have”, “get” and “make”, but we also have some uncommon words as top word in our corpora, such as “pay”, “help”, “buy”, “work” and “save”.

These are three word clouds of these three different subreddit and we can further say top words in different subreddits are different. In Personal Finance, the largest thing is “pay” for debt or loan; in Wall Street Bets, “GME” is definitely the hottest topic, and in Investing, people talk about “stock market” and “company”.

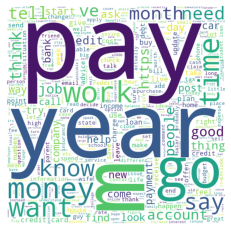


Figure Word Clouds

Distributional Distances

Distributional distance or divergence are used to compare different corpora, so we can see if the one corpus’ distribution is different from the other. First, we can draw a multi-dimensional scaling of the matrix.

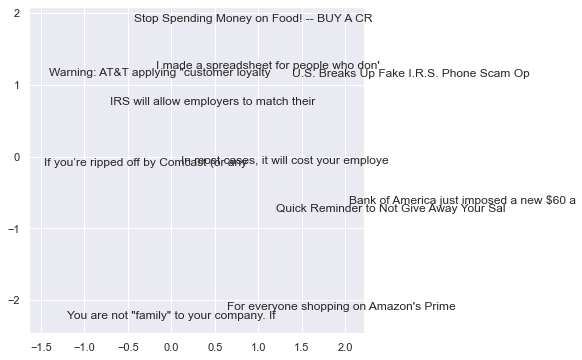


Figure A multi-dimensional scaling of the matrix

Because the title for the title for each post is usually very long, which is a feature of online posting—people always try to reveal as much information as possible in their title otherwise readers may not click in to see the whole post—I only use first 40 characters of each real time. But reading the first 40 characters can also give us an understanding of what’s the post is talking.

For example, around y = 1, there are two articles talk about IRS (Internal Revenue Service, a Federal government department for collecting taxes, especially income taxes) are near each other. In the bottom of the plot, there are two articles talking about different things. One is about “You are not "family" to your company. If you have an opportunity to better yourself, take it.” and the other is “savings from sales aren't savings if you weren't already planning on buying the item”. Despite the difference in topics, they are both giving advices to readers so they are near each other.

Full titles for article in the scatter plot and heatmap are followed.

['You are not "family" to your company. If you have an opportunity to better yourself, take it. They will do the same when it comes to cutting ties with you.',

'Warning: AT&T applying "customer loyalty speed upgrades" without customer consent',

'If you’re ripped off by Comcast (or any internet company), Wells Fargo (or any bank/student lender), or Aetna (or any health insurance company), here’s how to get your money back.',

'U.S. Breaks Up Fake I.R.S. Phone Scam Operation -- 21 people sentenced for up to 20 yrs, 32 in India indicted',

"I made a spreadsheet for people who don't know how to budget!",

'Stop Spending Money on Food! -- BUY A CROCKPOT',

'Quick Reminder to Not Give Away Your Salary Requirement in a Job Interview',

'Bank of America just imposed a new $60 annual fee on their previously free personal savings account.',

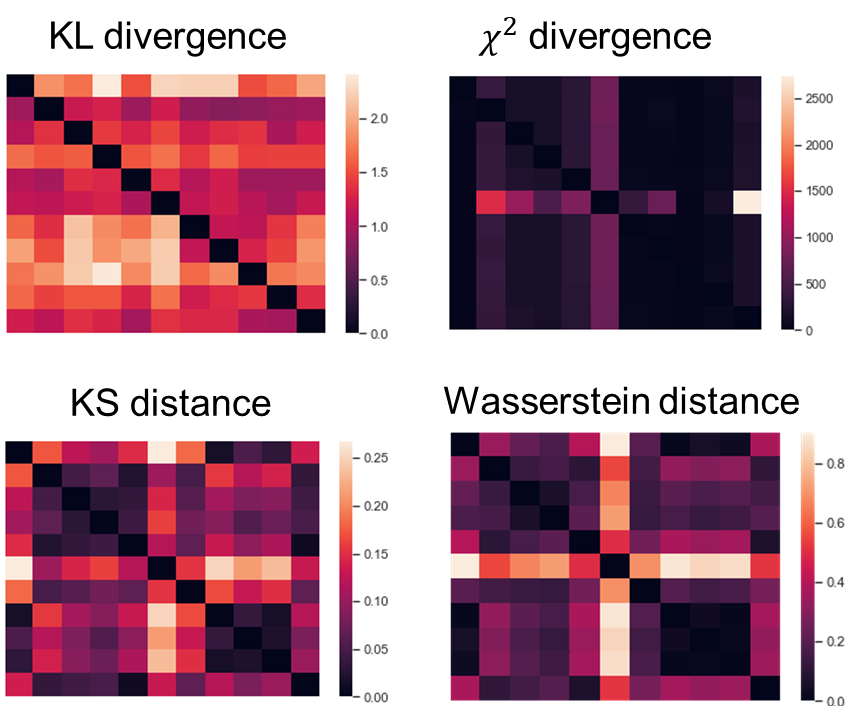
'For everyone shopping on Amazon\'s Prime Day: "savings" from sales aren\'t savings if you weren\'t already planning on buying the item.',

'In most cases, it will cost your employer far more to replace you than it would to give you a raise. So ask firmly.',

"IRS will allow employers to match their employees' student loan repayments"]

We can also calculate the distance or divergence that compares the two distributions. Here are some mostly used divergence measures and their heatmaps are followed. They are the same ten articles mentioned above and their titles are left out again due to their exceptional length.

* Kullback-Leibler (KL) divergence
* divergence
* Kolmogorov-Smirnov (KS) distance
* Wasserstein distance



From the heatmaps of articles with different measures distance, we can find that although the exact value varies across different measures,

Discovering Patterns, Clusters, and Topics

In this part, I begin to dig deeper to discover patterns in my corpora by clustering and topic modeling. For clustering, I will apply both a flat clustering algorithm, k-means, as well as a hierarchical clustering method, Ward’s minimum clustering method. I use the silhouette analysis to compare the shape of silhouette of clusters of cluster number, as well as the silhouette score to evaluate the quality of unsupervised clusters, to determine the best number of clusters. I also did a topic modeling, a two-dimensional content clustering method, which can find words cluster in topics and topic cluster in documents. Finally, I also did a dynamic topic modeling to find if different topics change from 2015 to 2021.

In this step, we do vectorization, converting texts into numerical vectors by machine learning algorithm by counting vectorizer (Scikit-learn, 2021).

We can also calculate the TF-IDF (term-frequency times inverse document-frequency) for our data to calculate the words weights for future clustering method:

Table TF-IDF of the corpora

|  |  |  |
| --- | --- | --- |
|  | word | tf-idf |
| 0 | people | 0.113324 |
| 1 | tend | 0.183513 |
| 2 | to | 0.096107 |
| 3 | feel | 0.0947 |
| 4 | sense | 0.089033 |
| 5 | of | 0.107934 |
| 6 | guilt | 0.060562 |
| 7 | when | 0.060865 |
| 8 | it | 0.111935 |
| 9 | comes | 0.097494 |

Flat clustering with K-means

K-means defines clusters by the centroid, barycenter or center of mass of its members. It randomly allocates text objects to clusters and constantly evaluate the criterion functions. If the new allocation earns better score in the criterion function, then the text objects shift, if not, they stay. It repeats this process many times until stability is achieved.

Note that the true classes for our corpora are the subreddits the texts are from, which are “Personal Finance”, “Wall Street Bets” and “Investing”. But when we use the k-means model, we don’t use the true classes and let the model tell us which cluster each post belongs. Therefore, we can evaluate the performance of clustering algorithm because we know the true class and also compare the prediction and true class.

There are several indicators for the k-means. First, we need to define the conditional entropy, which is the likelihood that given object shows up in the given cluster:

Where denotes the probability density function.

**Homogeneity** is defined as:

Where C denotes the true class labels and K denotes clusters it belongs. Homogeneity measures the degree that all objects in one cluster belongs to the same category.

**Completeness** is defined as:

Completeness measures if objects in one cluster contain all objects in the same true category.

**V-measure** is a harmonic mean of homogeneity and completeness, which is defined as:

is a tuning parameter. The greater , the greater value plays homogeneity.

**Adjusted Rand Index** measures if the cluster is a random guess or perfect alignment. It belongs to [-1, 1], where 0 means a random alignment, -1 means much worse than random alignment and 1 means perfect alignment.

After introducing these necessary concepts, we can use these four measures to evaluate the performance of k-means in our corpora. Here are the value for these four measures:

Homogeneity: 0.504

Completeness: 0.428

V-measure: 0.463

Adjusted Rand Score: 0.513

We can find that our data has homogeneity and completeness all nearly 0.50, which means that there is no perfect alignment. No cluster contains all texts from one class, and no cluster is exactly one class.

According to the Adjusted Rand Index, our clusters are much better than a random assignment, of which ARI is 0 (so compared to 0, 0.513 is a satisfactory value).

We can also have a close look at k-means clustering results:

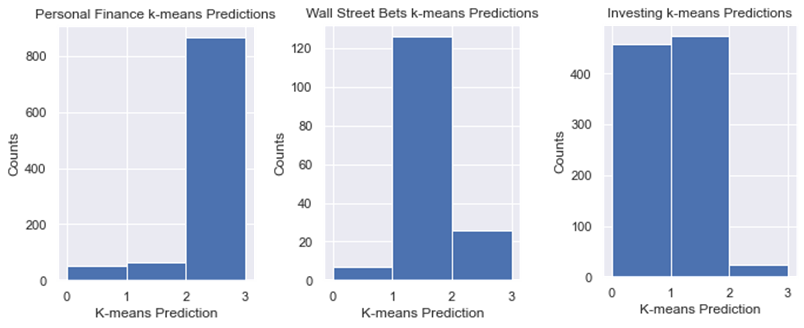


Figure k-means clustering results

We found that, almost all texts in personal finance category are in cluster 2, which means they are really different (far) from others.

Almost 80% posts from wall street bets are in cluster 1, 16.7% of its posts are in cluster 2, so we have the conclusion that usually posts from wall street bets are different from others, but sometimes they could be divided in cluster 2 (which means they have personal finance problem such as tax, debt or retirement).

However, for category 'investing', the k-means cluster label is really unstable and inconsistent, there are almost half posts from ‘investing’ subreddits are in cluster 0, and half in cluster 1.

We can also look at top words in each cluster:

Table Top words in each cluster

|  |  |  |
| --- | --- | --- |
| Cluster 0 | Cluster 1 | Cluster 2 |
| com | market | just |
| https | stock | money |
| www | https | credit |
| cnbc | com | account |
| html | gme | edit |
| 2020 | stocks | ve |
| 2019 | price | don't |
| news | year | job |
| http | people | card |

By looking into the top words in each cluster, we found that cluster 0 contains some common words about year, URL components, which are not very useful. Cluster 1 is more about stock market, investment strategies. Cluster 2 is about many financial concerns such as credit card, bank account, money, etc. So now we understand why posts from Personal Finance subreddit are predicted to be in cluster 2, posts from Wall Street Bets subreddit are predicted from cluster 1, and posts from Investing are predicted from cluster 0 and 1 (it has been thought as a starter place so people may discuss a lot about basic stuff and share web links, so it will have a lot of words that are not talking about investing strategies).

Plot clusters & features after reducing with PCA

We can use Principal Component Analysis (PCA) to reduce the dimension, and then draw the distribution plot for different clusters predicted by k-means.

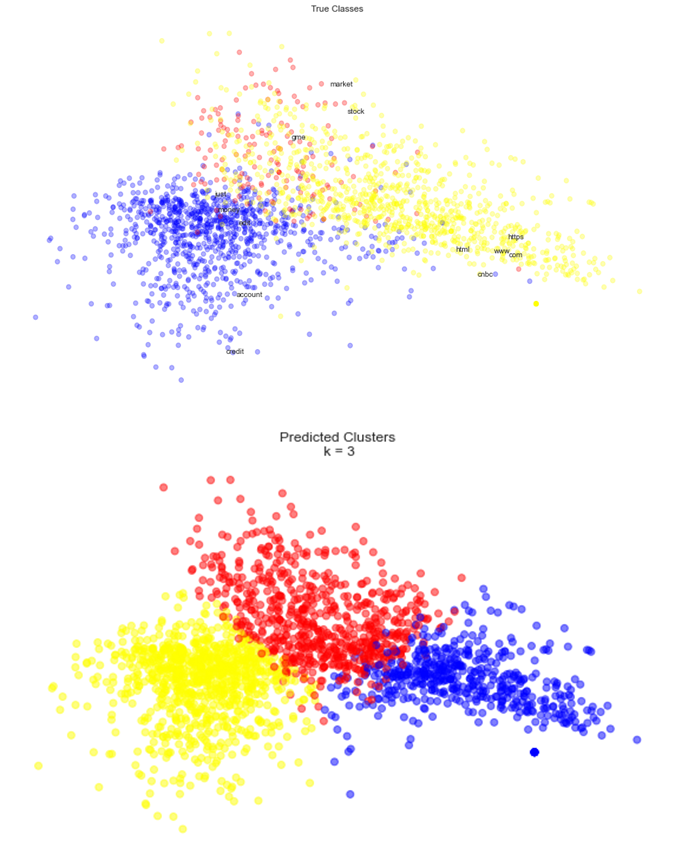


Figure True classes and predicted clusters for the data

The result shows that that, similar colors in predicted clusters are nearer to each other. Maybe because the size of data frames is not equal, (there are so many posts in Wall Street Bets subreddit lack content--they only use photo, video, gif or emoji).

In these three subreddits, many posts are share common topics such as planning, budgeting, and investing strategies, so there are a lot of overlaps for real classes. But in machine learning model, the boundaries are in fact clearer and the posts are more separated.

Identify the optimal cluster number with Silhouette analysis

Due to the poor performance of k-means when clustering the posts from Investing subreddit, I begin to think if 3 is not the best number for clusters. To compare the performance of different clusters, we can conduct silhouette analysis by eyeballing if the shape of the silhouette plot for different clusters is even, and by comparing the silhouette score.

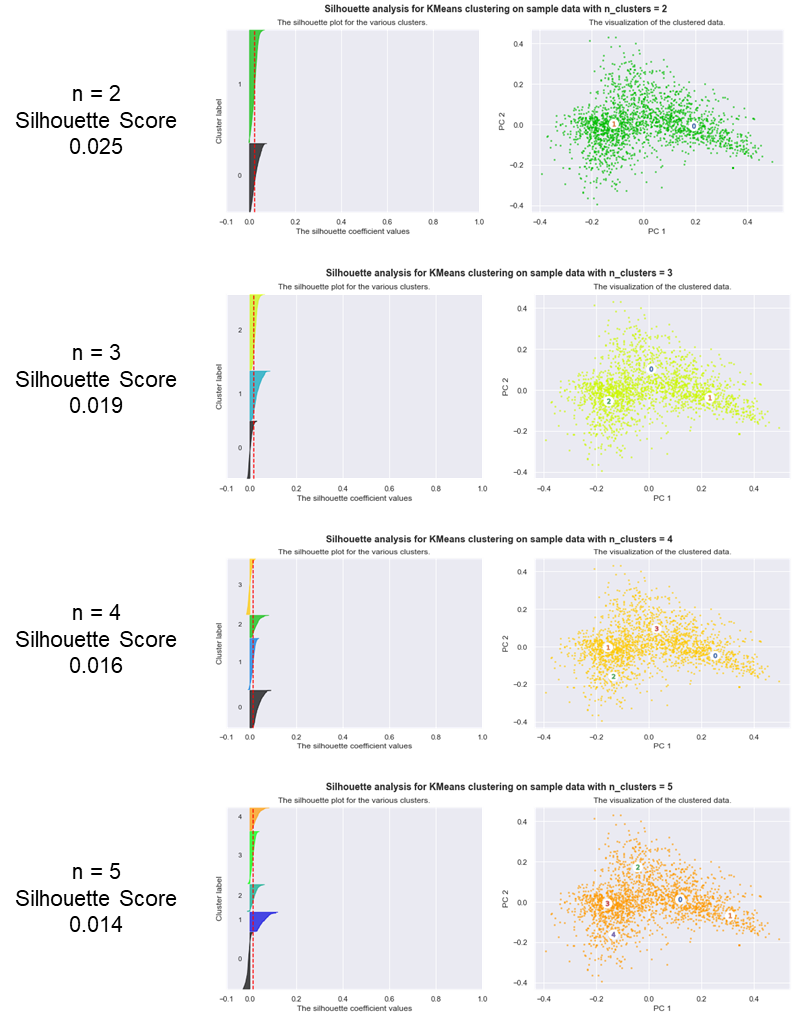


Figure Silhouette score for different number of clusters

After comparing the silhouette score, we find that instead of 3, the number of 2 is actually a better number for clusters.

Hierarchical clustering: Ward’s method

Instead of requiring us to pre-specify the number of clusters K in K-means clustering, Hierarchical clustering does not require that we commit to a particular choice of K, and it can result in an attractive tree-based representation of the observations, called dendrogram(Gareth James, 2013).

The dendrogram is built starting from the leaves and combining clusters up to the trunk. It can be cut at certain level of height and result in distinct clusters. Observations that fuse at the very bottom of the tree are quite similar to each other, whereas observations that fuse close to the top of the tree will tend to be quite different.

Using different distance measures sometimes can result in very different clustering result. Following are four dendrogram with four different distance measures.

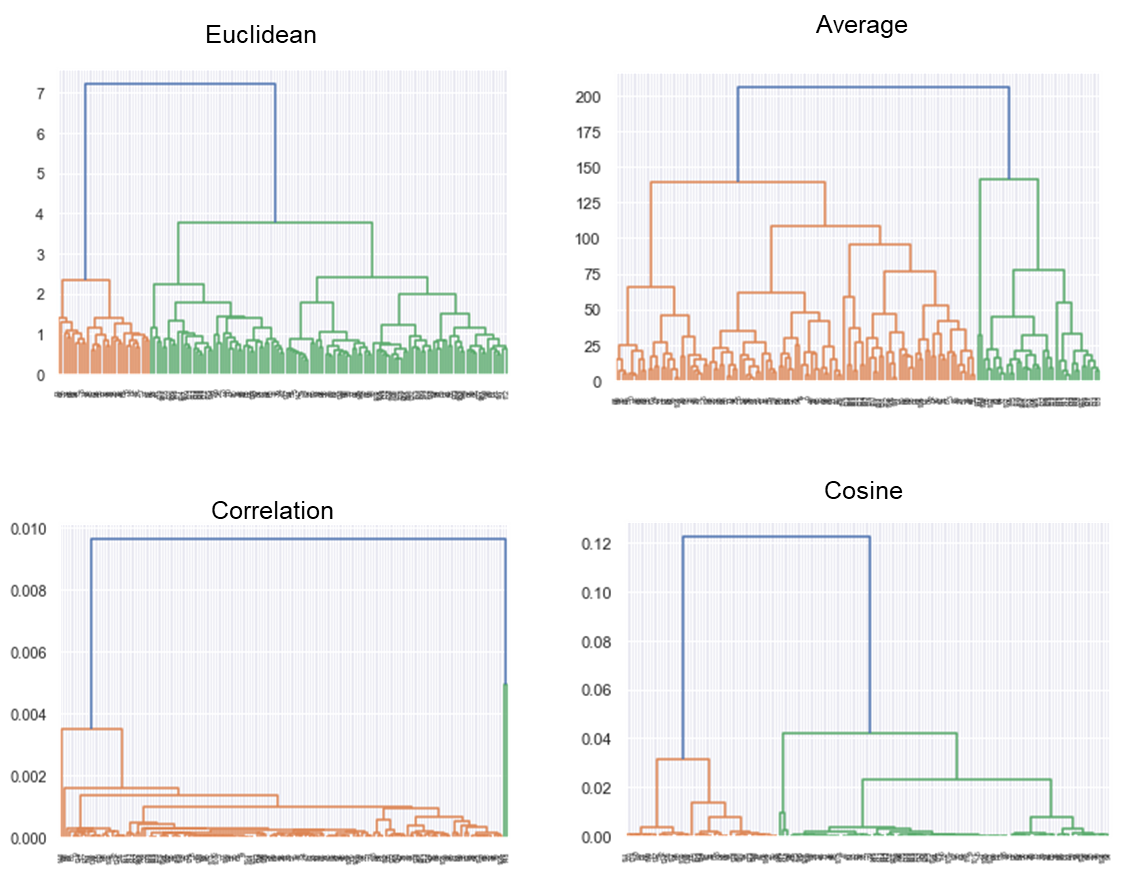


Figure Hierarchical clustering result with different distance measures

We find that the scale of distance results in different clustering result. We can also compare the performance of k-means to Ward’s Hierarchical clustering.

Table The comparison of the performance between K-means and Ward

|  |  |  |
| --- | --- | --- |
|  | k-means | Ward |
| Homogeneity | 0.504 | 0.304 |
| Completeness | 0.428 | 0.268 |
| V-measure | 0.463 | 0.285 |
| Adjusted Rand Score | 0.513 | 0.367 |

The result shows that k-means does a better job overall then Ward. Maybe there are too many words for Ward to build the hierarchy clusters or we shouldn't use TF-IDF since TF-IDF compresses the space.

Topic Modeling

Topic modeling is a two-dimensional content clustering method, which can find words cluster in topics and topic cluster in documents. Because there are so many posts and it’s hard to present the results for all of them, I analyze the topics for first 10 posts in Personal Finance subreddit. We choose 10 as the number of topics here because we may have some sub-topics such as tax, car, insurance, housing, loan, etc.

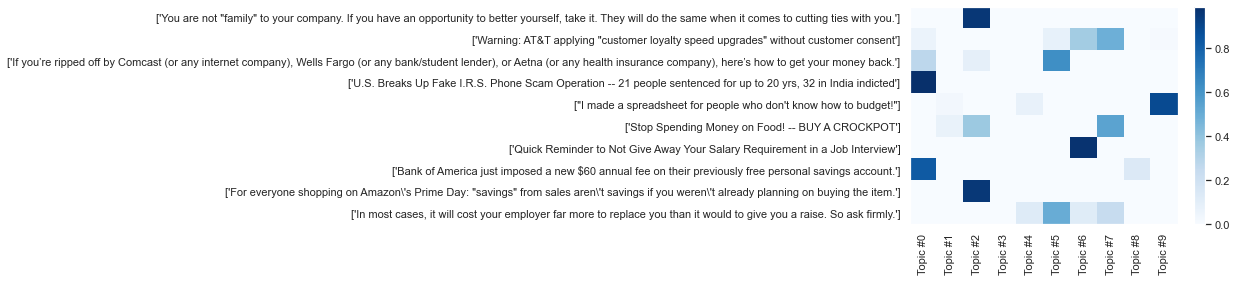
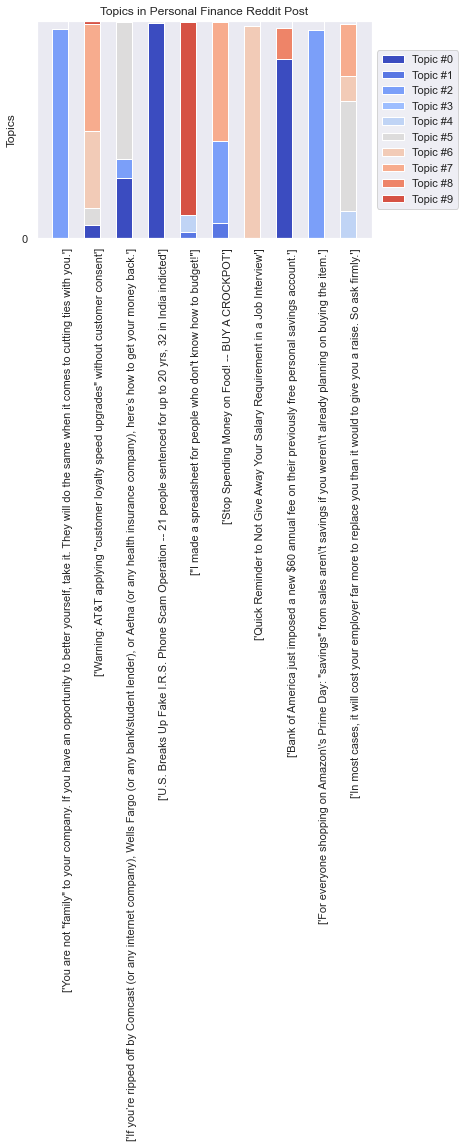


Figure Topics distribution in first 10 posts from Personal Finance Subreddit

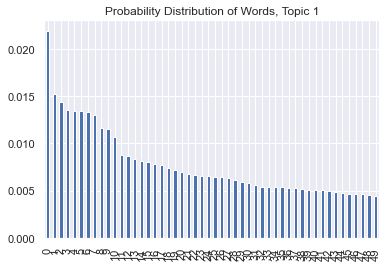
By doing topic modeling, we find that usually one post only has a significant topic, or at most two. There are top words for each topic.

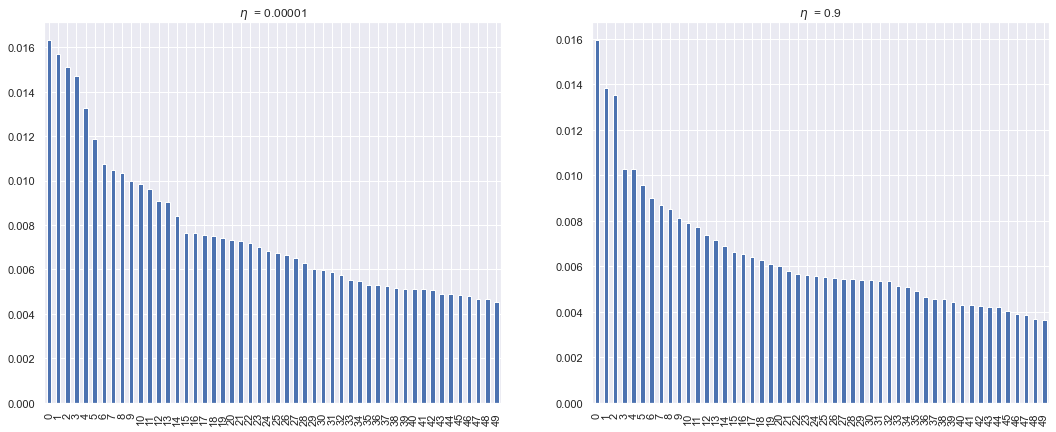
Table Top words for each topic

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Topic0 | Topic1 | Topic2 | Topic3 | Topic4 | Topic5 | Topic6 | Topic7 | Topic8 | Topic9 |
| credit | year | money | pay | pay | pay | work | pay | account | credit |
| pay | loan | pay | car | year | work | account | month | card | pay |
| account | know | need | tell | car | year | year | account | money | card |
| edit | work | edit | ask | money | time | pay | day | credit | year |
| know | time | year | work | work | job | time | know | pay | like |
| year | pay | bank | say | cost | company | card | time | bank | money |
| loan | want | car | time | say | think | tell | people | finance | account |
| bank | month | like | money | need | money | try | edit | year | want |
| job | money | work | year | payment | say | thank | money | spend | time |
| money | payment | say | month | income | thank | say | loan | time | month |

From the table, we find that there are a lot of same words that appear in many topics. For example, ‘pay’ and ‘year’ appears in almost every topic! But there are also some distinctions between each one, for example, I think topic 1 and topic 7 have refers to time, containing words such as ‘year’, ‘month’, ‘day’, and ‘time’. Topic 4 and topic 5 are about car. But overall, I think the distinction between each topic is not very significant, perhaps because we choose a wrong number or choosing the first 10 is not enough.

We can also look into the probability distribution of words in different topics and tune some parameters.





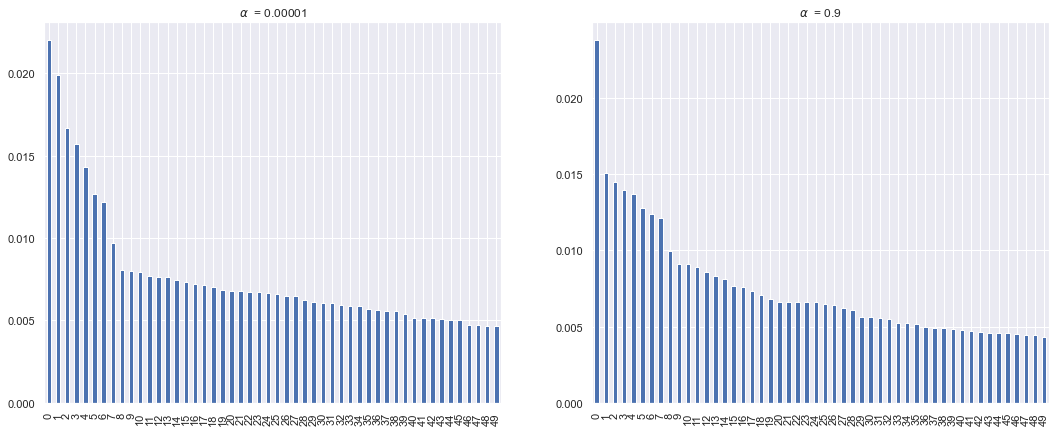


Figure Probabilities of words and parameters tuning for topic modeling

controls the sparsity of document-topic loadings, which means if one document is made of one topic or more. controls the sparsity of topic-word loadings, which measures if one topic is represented by a small number of words or a variety of words. We can find that changes the topic a lot, while doesn't change the graph much. The reason could be that my topics have some cross-over and some of them have similar contents. Therefore, when we increase , the probability of different words becomes similar, but when we increase , it doesn't change much.

Dynamic Topic Modeling

Dynamic Topic Modelling is a time-based topic model method introduced by David Blei and John Lafferty (Blei & Lafferty, 2006). It allows one to see topics evolve over a time annotated corpus.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 2015 | 2016 | 2017 | 2018 | 2019 | 2021 |
| = | = | job | job | job | job |
| job | job | = | = | = | = |
| people | people | people | people | people | people |
| work | work | work | time | time | time |
| time | time | time | work | work | work |
| ask | ask | ask | ask | ask | ask |
| company | company | company | like | offer | offer |
| like | like | like | company | like | company |
| offer | offer | offer | offer | company | like |
| know | know | know | know | know | know |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 2015 | 2016 | 2017 | 2018 | 2019 | 2021 |
| $ | $ | $ | $ | $ | $ |
| pay | pay | pay | pay | pay | pay |
| loan | loan | loan | loan | loan | car |
| year | year | year | year | car | year |
| month | month | month | month | year | loan |
| car | car | car | car | month | month |
| debt | debt | debt | debt | debt | debt |
| payment | payment | payment | payment | work | work |
| work | work | work | work | payment | payment |
| house | live | live | live | get | get |

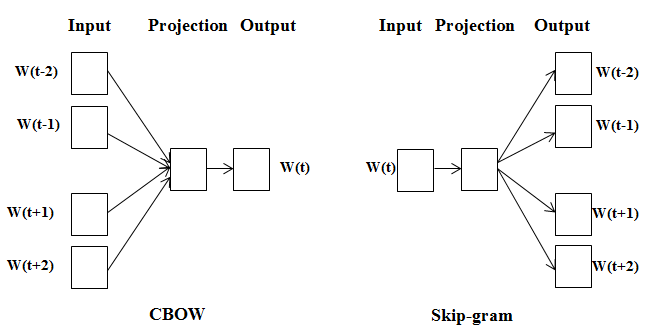
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 2015 | 2016 | 2017 | 2018 | 2019 | 2021 |
| delete | delete | delete | delete | delete | delete |
| fund | fund | fund | remove | remove | remove |
| remove | remove | remove | fund | fund | fund |
| stock | stock | stock | stock | stock | stock |
| sell | sell | sell | sell | sell | sell |
| buy | buy | buy | buy | buy | buy |
| market | market | market | market | market | market |
| share | share | share | share | vanguard | vanguard |
| vanguard | vanguard | vanguard | vanguard | share | share |
| portfolio | portfolio | portfolio | portfolio | portfolio | portfol |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 2015 | 2016 | 2017 | 2018 | 2019 | 2021 |
| $ | $ | $ | $ | $ | $ |
| year | year | year | year | year | year |
| money | money | account | account | account | account |
| account | account | money | money | money | money |
| tax | tax | tax | tax | tax | tax |
| saving | saving | saving | ira | ira | saving |
| 401k | ira | ira | 401k | 401k | 401k |
| income | 401k | 401k | saving | saving | ira |
| ira | income | income | invest | invest | invest |
| invest | invest | invest | roth | roth | roth |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 2015 | 2016 | 2017 | 2018 | 2019 | 2021 |
| credit | credit | credit | credit | credit | credit |
| card | card | card | card | card | card |
| account | account | account | account | account | account |
| bank | bank | bank | bank | bank | bank |
| pay | pay | pay | check | check | check |
| get | check | check | pay | $ | $ |
| check | get | get | get | get | get |
| say | say | say | $ | pay | pay |
| $ | score | $ | say | score | score |
| score | $ | score | score | say | say |

 Vector Space and Word Embeddings

This part, we build on last part's topic modeling techniques by taking a text corpus we have developed, specifying an underlying number of dimensions, and training a model with a neural network auto-encoder (one of Google's word2vec algorithms) that best describes corpus words in their local linguistic contexts, and exploring their locations in the resulting space to learn about the discursive culture that produced them. Documents here are represented as densely indexed locations in dimensions, rather than sparse mixtures of topics (as in LDA topic modeling), so that distances between those documents (and words) are consistently superior, though they require the full vector of dimension loadings (rather than just a few selected topic loadings) to describe. We will explore these spaces to understand complex, semantic relationships between words, index documents with descriptive words, identify the likelihood that a given document would have been produced by a given vector model, and explore how semantic categories can help us understand the cultures that produced them.

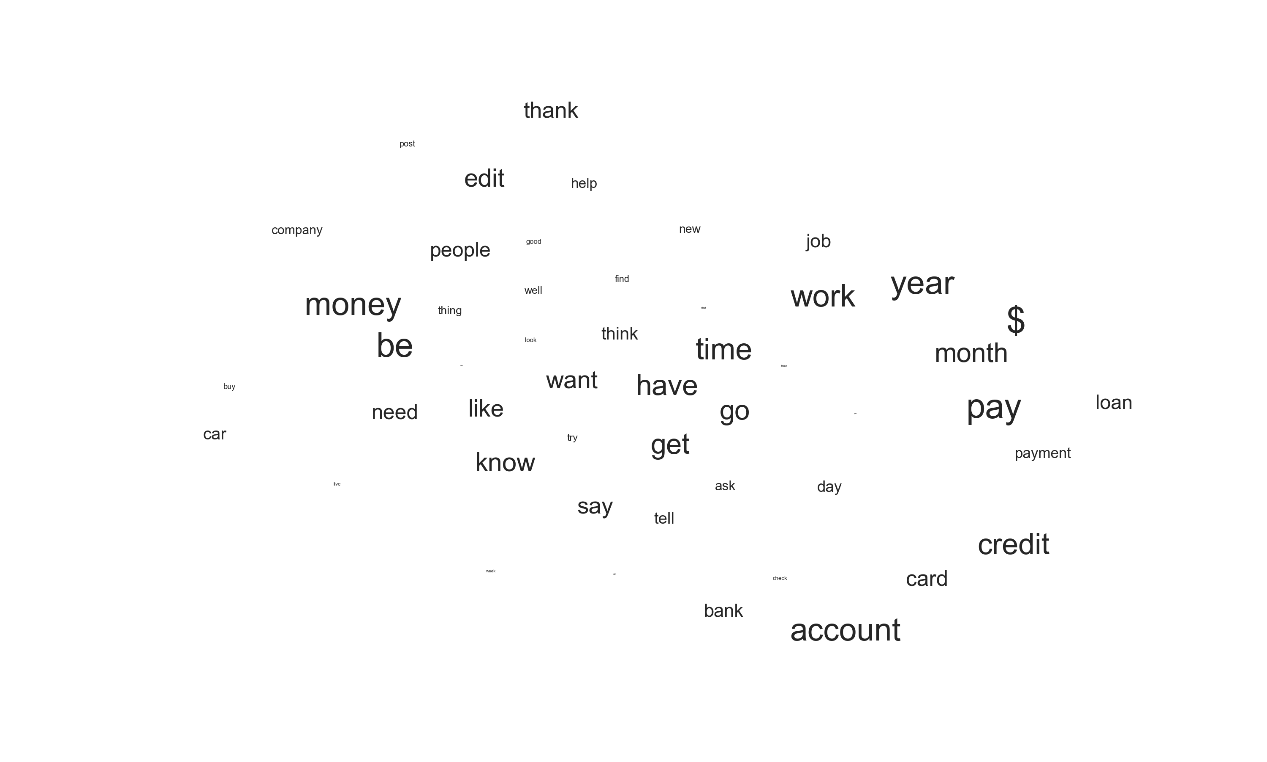


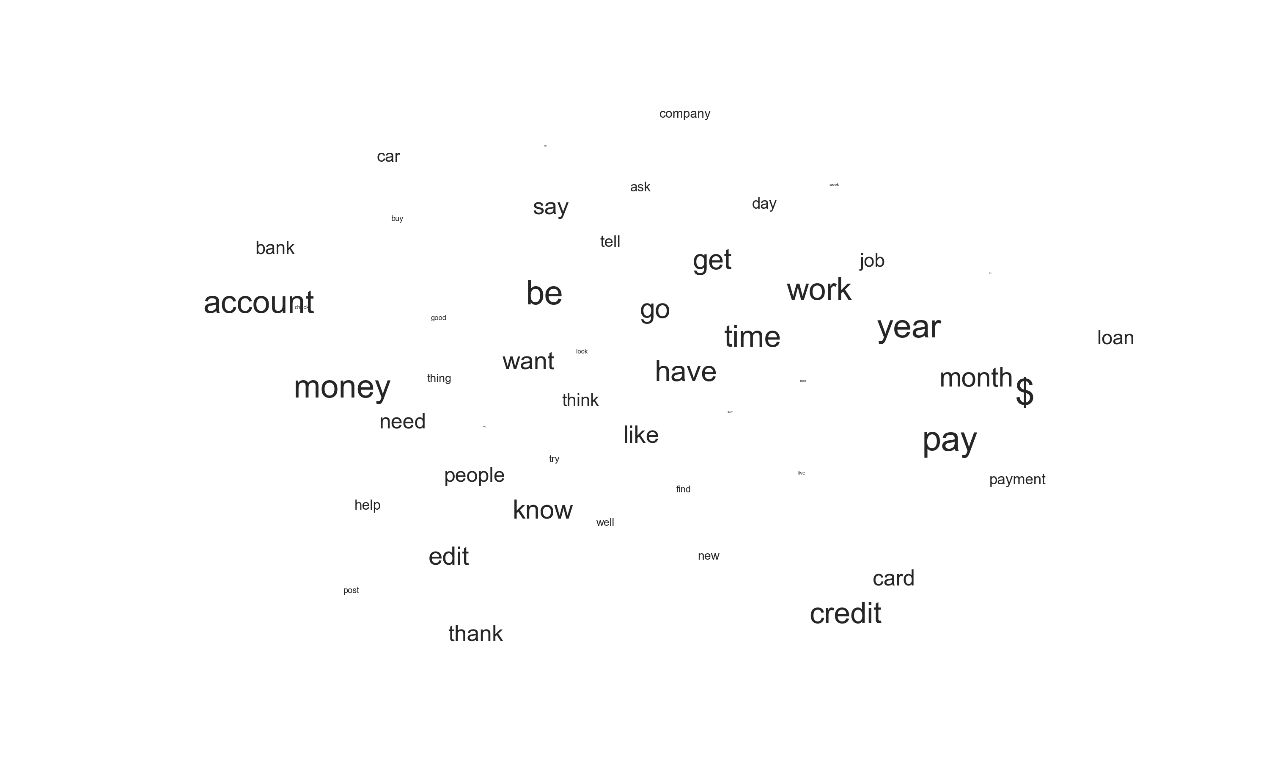
CBOW (The Continuous Bag of Words) Model

Word2Vec needs to retain the sentence structure so as to capture a "continuous bag of words (CBOW)" and all of the skip-grams within a word window. The algorithm tries to preserve the distances induced by one of these two local structures. This is very different from clustering and LDA topic modeling which extract unordered words alone.

When we normalize here, we don't use the lematized form of the word because we might lose information. Note the paramter in the normalize tokens function.

CBOW





Skip gram

\*\*What does this pattern reveal about the semantic organization of words in your corpora?\*\*

We can find that the most different word in ['student', 'loan', 'debt', 'payment', 'account','investment', 'tax'] is 'investment'. Probabaly it's because investment is an action for people have extra money and relatively more affluent, but other words such as 'loan' 'debt' 'payment' are more for people who have more financial limitations.

For addition/subtraction, we got following relationships:

'pay' + 'debt' = 'loan' + 'payment'

'student' + 'loan' = 'pay' + '$'

'credit' + 'score' = 'account' + 'start'

'mortgage' + 'house' = 'lose' + 'refinance'

\*\*Which estimation and visualization specification generate the most insight and appear the most robustly supported and why?\*\*

I repeat the visualization several times (because the plot is kind of nondeterministic). Among them, 'payment', 'pay', 'credit', 'loan' are usually together. 'house' and 'mortgage' are side by side, as well as 'tax' and 'income'

Table 7 Most Similar Words to ‘finance’

|  |  |
| --- | --- |
| CBOW | SG |
| (domain, 0.9975020885467529) | (domain, 0.9038022756576538) |
| (core, 0.9973710775375366) | (core, 0.8971995711326599) |
| (economics, 0.9970412850379944) | (economics, 0.8908737897872925) |
| (introduction, 0.9942911863327026) | (tutorial, 0.8825550079345703) |
| (tutorial, 0.9941723942756653) | (personal, 0.8703951835632324) |
| (inflation, 0.9931447505950928) | (v, 0.8317123651504517) |
| (investment, 0.9930504560470581) | (introduction, 0.8269104957580566) |
| (v, 0.9929893016815186) | (investment, 0.8173561096191406) |
| (personal, 0.9928480386734009) | (vehicle, 0.8171368837356567) |
| (gain, 0.9927452206611633) | (inflation, 0.8123129606246948) |

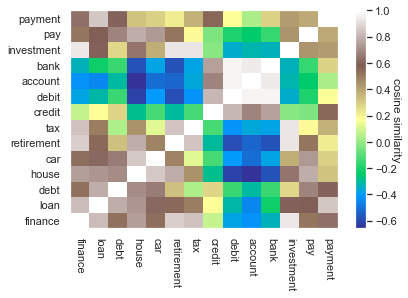
Table 8 Most Similar Word to ‘loan’

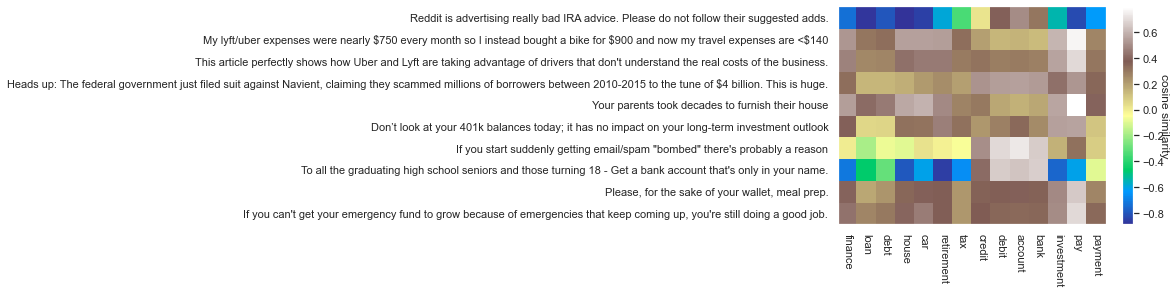
|  |  |
| --- | --- |
| CBOW | SG |
| (pay, 0.9886571168899536) | (student, 0.947920560836792) |
| (interest, 0.986750602722168) | (forgiveness, 0.9250237345695496) |
| (payment, 0.9842690229415894) | (program, 0.9044044613838196) |
| (student, 0.982388973236084) | (borrower, 0.8913569450378418) |
| (month, 0.9769452810287476) | (qualify, 0.8813320398330688) |
| ($, 0.9758726358413696) | (graduate, 0.8758156299591064) |
| (debt, 0.975852370262146) | (forgive, 0.8746172189712524) |
| (rate, 0.9507907629013062) | (consolidate, 0.8732489347457886) |
| (year, 0.9503719806671143) | (repay, 0.8646396398544312) |
| (car, 0.9470028281211853) | (discharge, 0.8627040982246399) |

Table 9 Most Similar Word to ‘house’

|  |  |
| --- | --- |
| CBOW | SG |
| (home, 0.9981556534767151) | (cheap, 0.9140413999557495) |
| (make, 0.9973364472389221) | (home, 0.90831458568573) |
| (buy, 0.9973288774490356) | (own, 0.8955202102661133) |
| (expense, 0.9969913959503174) | (sell, 0.893168032169342) |
| (college, 0.9969626665115356) | (rent, 0.8855429887771606) |
| (old, 0.9969006180763245) | (buy, 0.8843915462493896) |
| (plan, 0.996694803237915) | (nice, 0.8827567100524902) |
| (school, 0.9964942336082458) | (clothe, 0.879166841506958) |
| (price, 0.9963560700416565) | (car, 0.8713674545288086) |
| (cost, 0.9963399171829224) | (reliable, 0.8700470924377441) |

D2V





We have some positive as well as negative correlations.

What does this pattern reveal about the documentary organization of your semantic space?

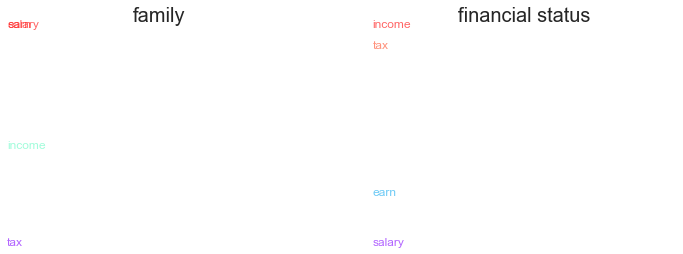
## 5.4 Projections

Create 2 dimensions

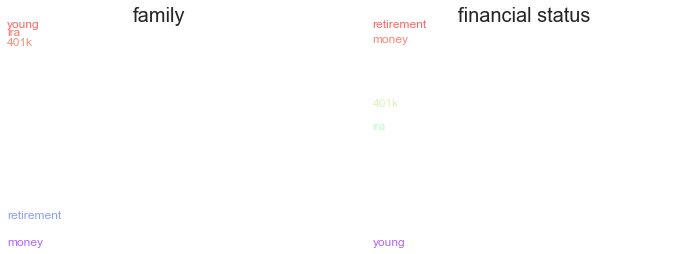
|  |  |  |
| --- | --- | --- |
|  | positive | negative |
| family | ['mom','parent','father','dad','mother'] | ['son','daughter','child'] |
| financial status | ['loan','debt','poor','unemployment'] | ['saving','investment','rich','job'] |

Words to project:

tax: 'tax', 'income', 'salary', 'earn',



retirement: 'retirement', '401k', 'ira', 'young', 'money'



Interpretation:

Which of the dimensions you analyze explain the most variation in the projection of your words and why?

Family dimension explains the most variation in the 'retirement' word list, because the concept of family is more related to retirement (for example, we may care our parents' retirement.) Financial status explain the most variation in 'tax' dimension, maybe because tax is more related to finance compared to family.

Discussions

Generalization Bias--most users of online platforms are young people who are used to the internet. Middle-age people may not be willing to disclose their financial concerns online.

\*\*Alternatives:\*\*

Other discussion websites.

Methods to scale up my sample

I can broaden my dataset by scale up the time period to include more articles from myFICO Forums, YNAB Forums, Morningstar Forums, Reddit–Investing, Bogleheads Forum, Fat Wallet Forums, and Bigger Pockets Forum.

# Reference

scipy.cluster.hierarchy.linkage https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.linkage.html#scipy.cluster.hierarchy.linkage

**References:**

Blei, D., & Lafferty, J. (2006). *Dynamic Topic Models* (Vol. 2006). https://doi.org/10.1145/1143844.1143859

Gareth James, D. W. T. H. (2013). *An introduction to statistical learning : with applications in R*. New York : Springer, [2013] ©2013.

Reddit. (2021). *Personal Finance Wiki* https://www.reddit.com/r/personalfinance/wiki/index

Researve, F. (2018). Report on the Economic Well-Being of U.S. Households in 2017. (Reprinted.

Scikit-learn. (2021). *CountVectorizer* https://scikit-learn.org/stable/modules/generated/sklearn.feature\_extraction.text.CountVectorizer.html