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

MACSS Jinfei Zhu (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 about every day? Do these topics 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 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 consumption and investment are individual activities that constitute our society. Therefore, studying personal finance concerns will give us a better insight into 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.

Regarding the online texts that discuss personal finance concerns, I have the following research hypothesis:

> ***Hypothesis 1***: *The most frequent personal finance concerns include saving, paying back credit cards, student loans, housing, and insurance.*

Hypothesis 2*: The topics of these concerns change over time.*

Hypothesis 3*: People talk differently in different online forums.*

Here is the structure of this paper:

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 tokenize, normalize and vectorize the text data.

Second, counting the words and phrases: count the frequency of keywords and n-grams in Reddit posts, do part-of-speech tagging, and find the difference between personal finance Reddit and other finance-related subreddit.

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

Fourth, word embedding and projections to dig 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 that 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 the 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.

The 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 has 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 advice, or share personal experiences. There is a very detailed wiki for this subreddit (Reddit, 2021) listing a summary of suggestions for different ages 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 on 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 the other two subreddits, with a sample size of only 157.

Investing subreddit (r/investing) is created on March 15, 2008, the oldest one among these three subreddits. 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 in content across different subreddits. The number of posts from Investing subreddit is 954.

People are anxious about money, paying debt and managing their assets, 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 loans, 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 platforms 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 methods for English corpora. First, we can tokenize the texts and have a look at the top words in these three subreddits.

Table 1 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 the 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 funds 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

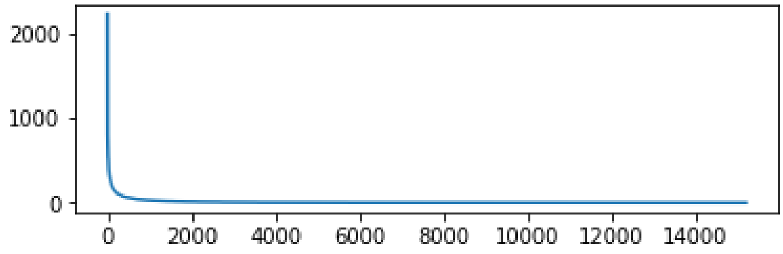


Figure 1 Rankings and frequencies of the word

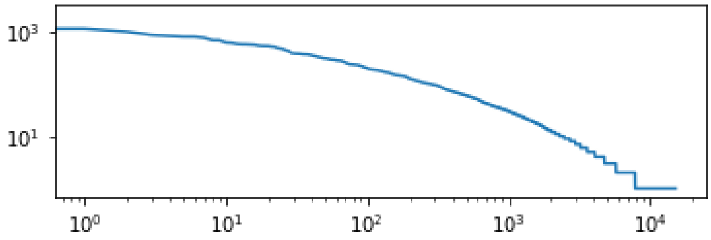


Figure 2 log-ranking and log-frequency of the word

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 between 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 the 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 many people post links to their wiki for reference, so we can safely ignore it).

Table 2 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.

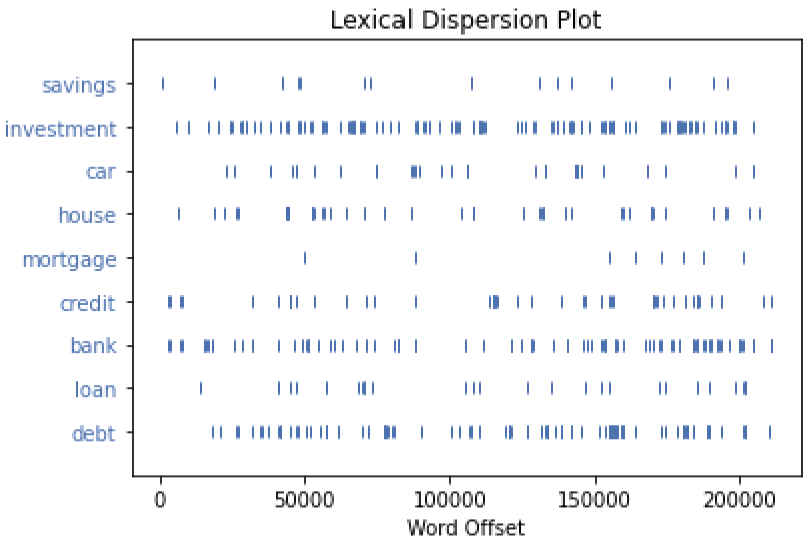


Figure 3 Lexical Dispersion Plot

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 the largest conditional probability are different from the top words in the earlier part.

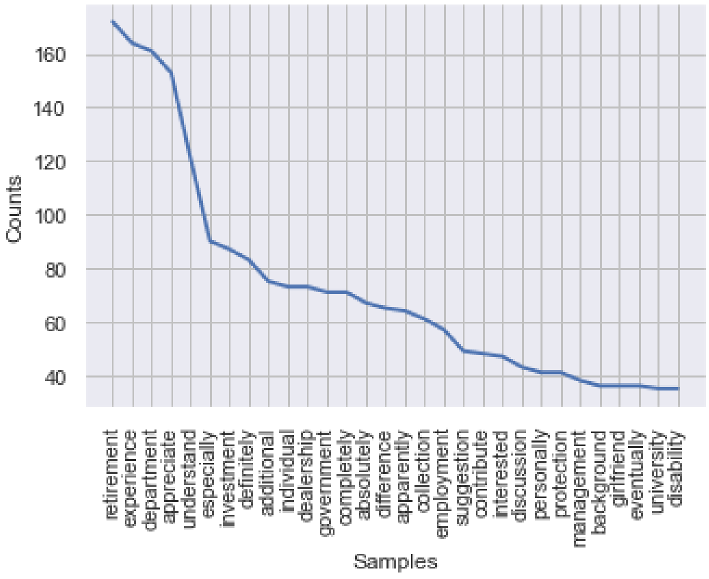
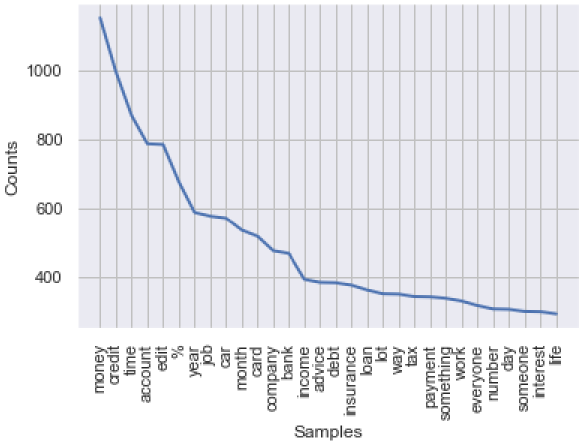


Figure 4 Conditional Probability Distribution

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



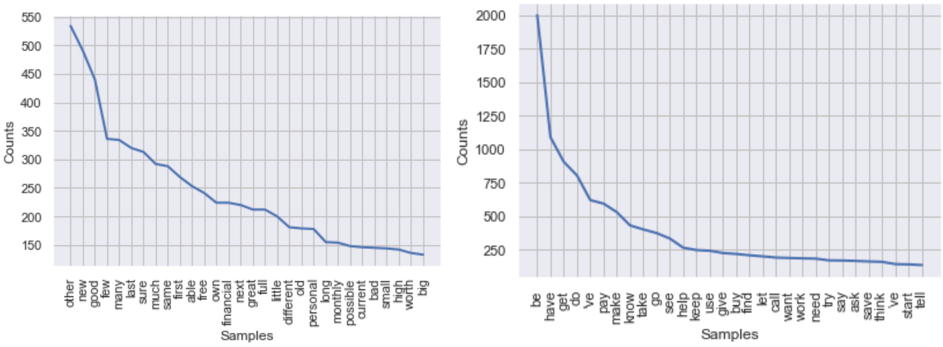


Figure 5 Top Adjectives and verbs

From the visualization, we find those top nouns including “money”, “credit”, “time”, “account”, “year”, “job”, “car”, “month”, “card”, “company”, “income”, “advice”, “debt”, “insurance”, “bank”, “payment” and “interest”. We find that they are all 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 words in our corpora, such as “pay”, “help”, “buy”, “work” and “save”.

These are three word clouds of these three different subreddits 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”.

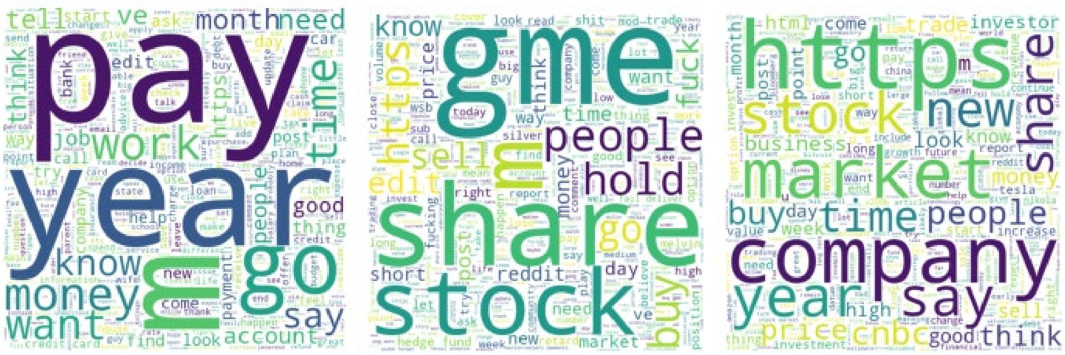


Figure 6 Word Clouds

Distributional Distances

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

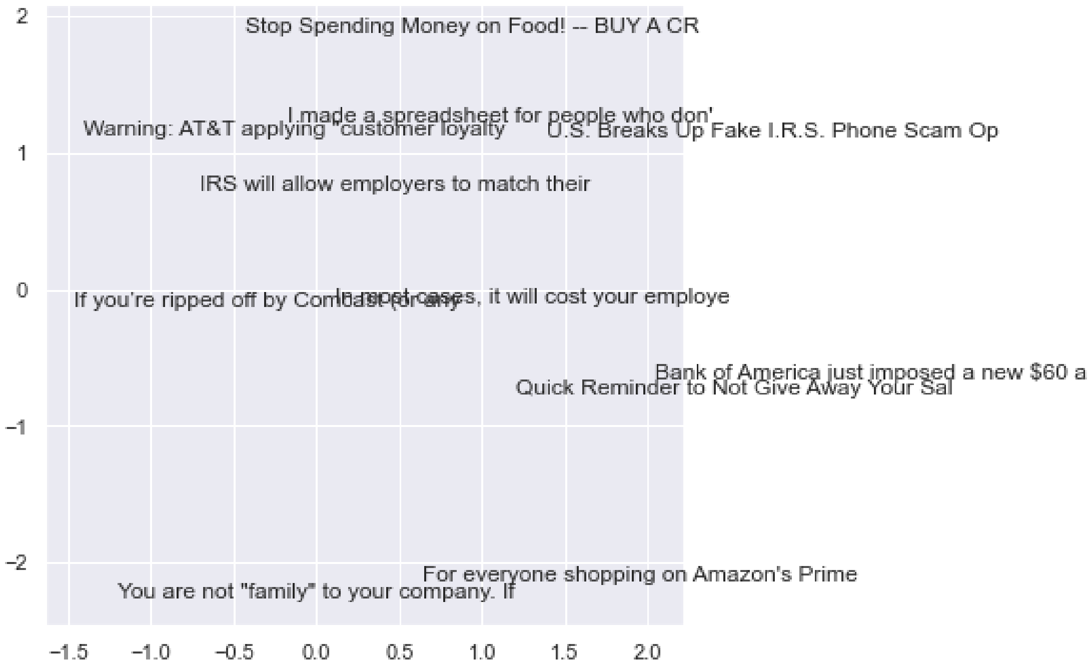


Figure 7 A multi-dimensional scaling of the matrix

Because 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 the first 40 characters of each article. But reading the first 40 characters can also give us an understanding of what's the post is talking about.

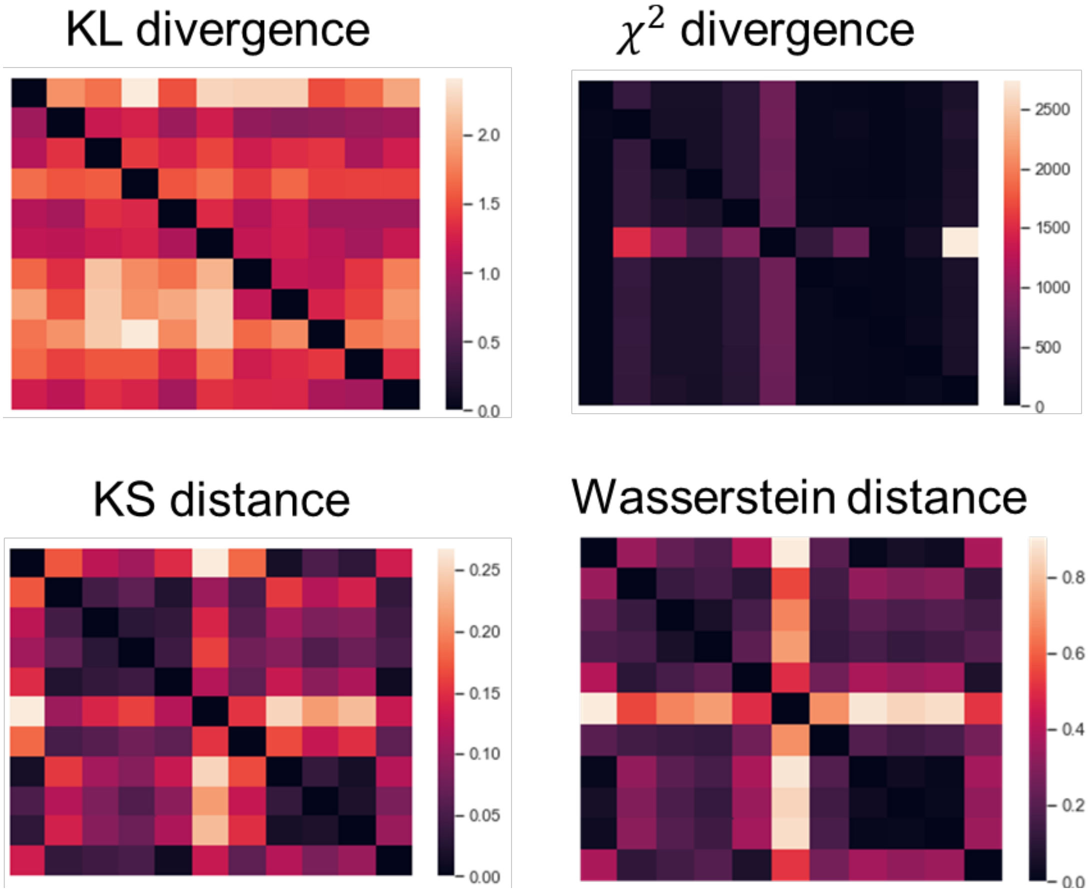
For example, around y = 1, there are two articles that talk about IRS (Internal Revenue Service, a Federal government department for collecting taxes, especially income taxes) are near each other. At the bottom of the plot, there are two articles talking about different things. One is about “You are 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 advice to readers so they are near each other.

Full titles for articles 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
- χ^2divergence
- Kolmogorov-Smirnov (KS) distance
- Wasserstein distance



KL divergence

χ^2 divergence

KS distance

Wasserstein distance

Figure 8 Heatmaps with different distance measures

From the heatmaps of articles with different measures distance, we can find that although the exact value varies across different measures, the shapes are similar across four distances.

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 the 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 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 weights of the words for future clustering method:

Table 3 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 evaluates the criterion functions. If the new allocation earns a 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 to. Therefore, we can evaluate the performance of the 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 a given object shows up in the given cluster:

$$H(Y|X) = \sum_{x\in X,\ y\in Y} p(x,y)\log\frac{p(x)}{p(x,y)}$$

Where $p(\cdot)$ denotes the probability density function.

Homogeneity is defined as:

$$h = \begin{cases} 1 & if\ H(C,K) = 0 \\ 1 - \frac{H(C|K)}{H(C)} & else \end{cases}$$

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

Completeness is defined as:

$$c = \begin{cases} 1 & if\ H(K,C) = 0 \\ 1 - \dfrac{H(K|C)}{H(K)} & else \end{cases}$$

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:

$$V_\beta = \frac{(1+\beta) * h * c}{(\beta * h) + c}$$

β is a tuning parameter. The greater β, 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 values 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:

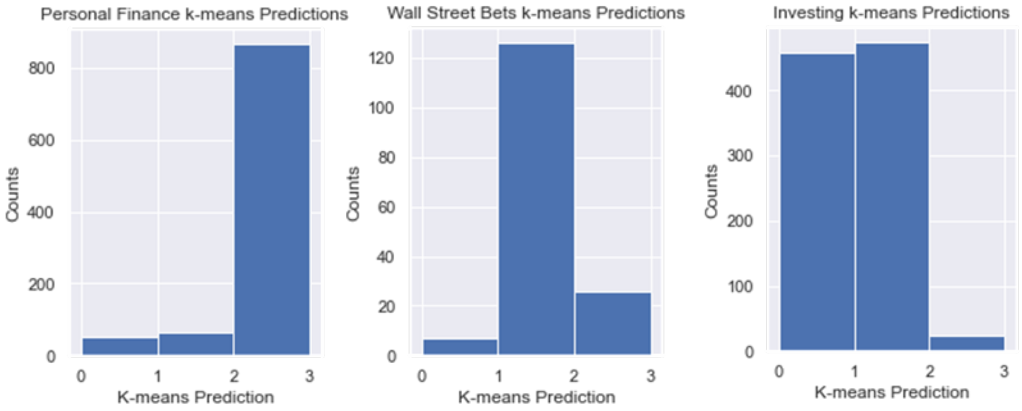


Figure 9 k-means clustering results

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

Almost 80% of 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 concerns such as debts or loans).

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 4 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 the year and URL components, which are not very useful. Cluster 1 is more about the 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 of as a place for beginners 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.

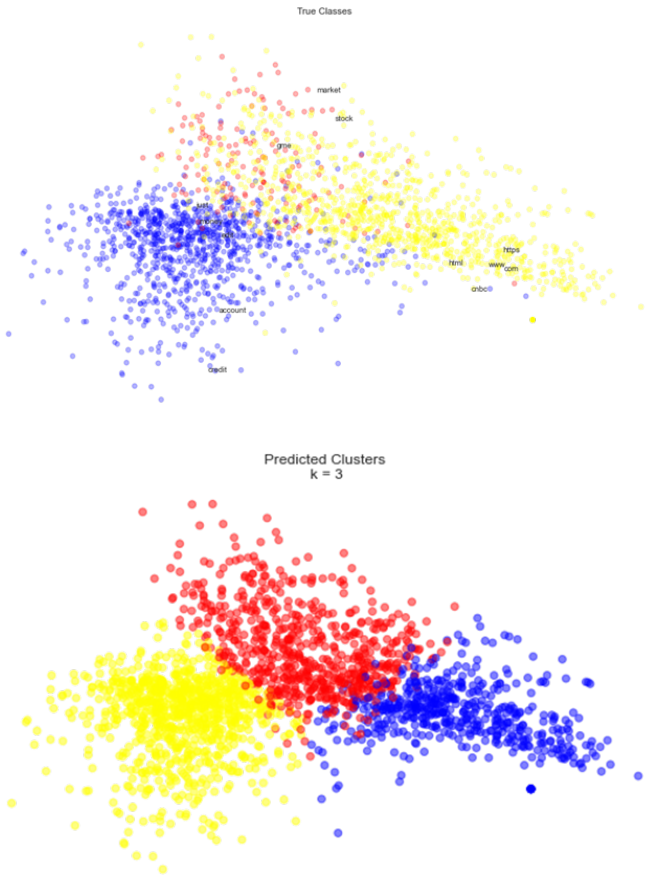


Figure 10 True classes and predicted clusters for the data

The result shows 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 photos, videos, GIFs, or emoji).

In these three subreddits, many posts share common topics such as planning, budgeting, and investing strategies, so there are a lot of overlaps for real classes. But in the 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.

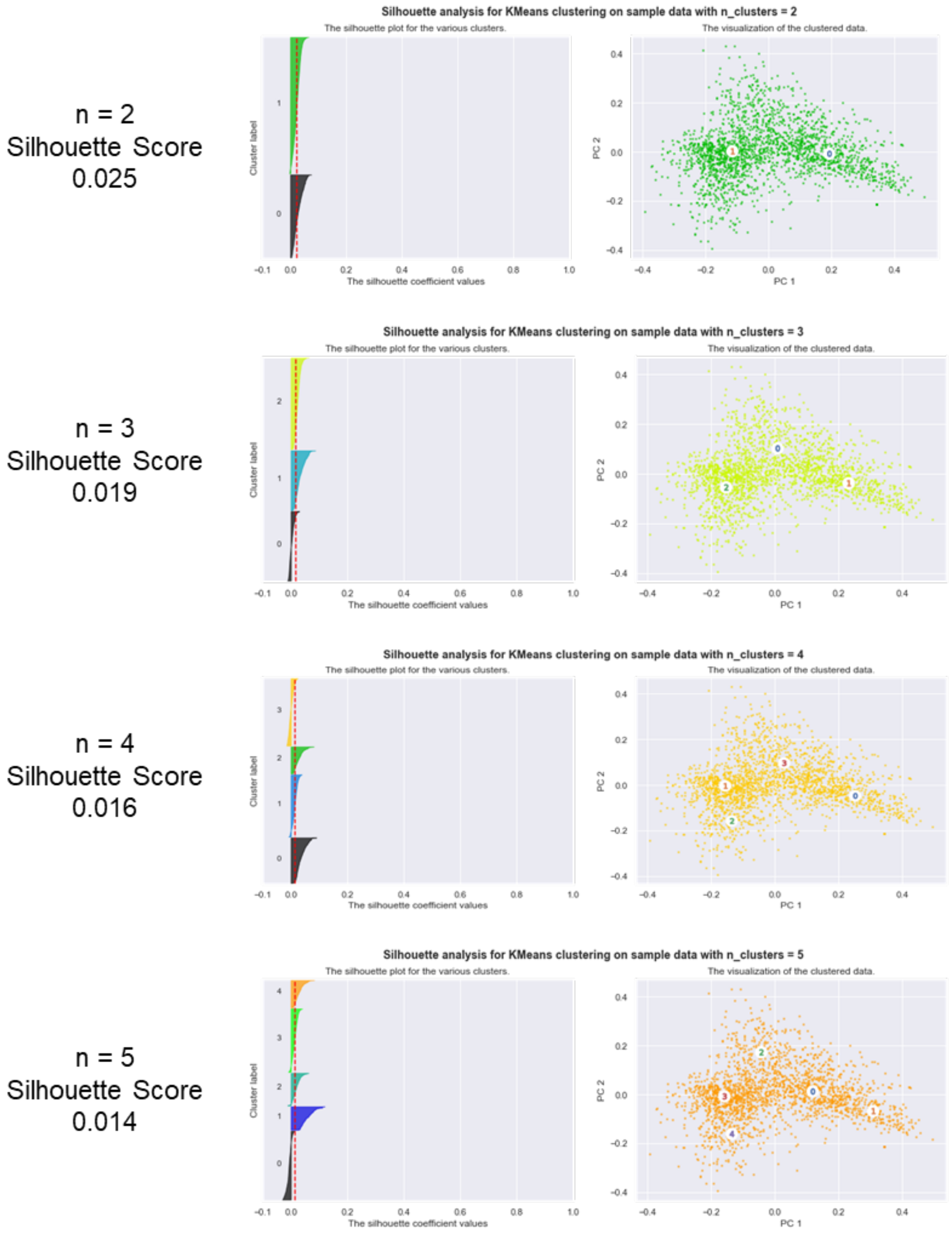


Figure 11 Silhouette score for different number of clusters

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 a 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 results. Following are four dendrograms with four different distance measures.

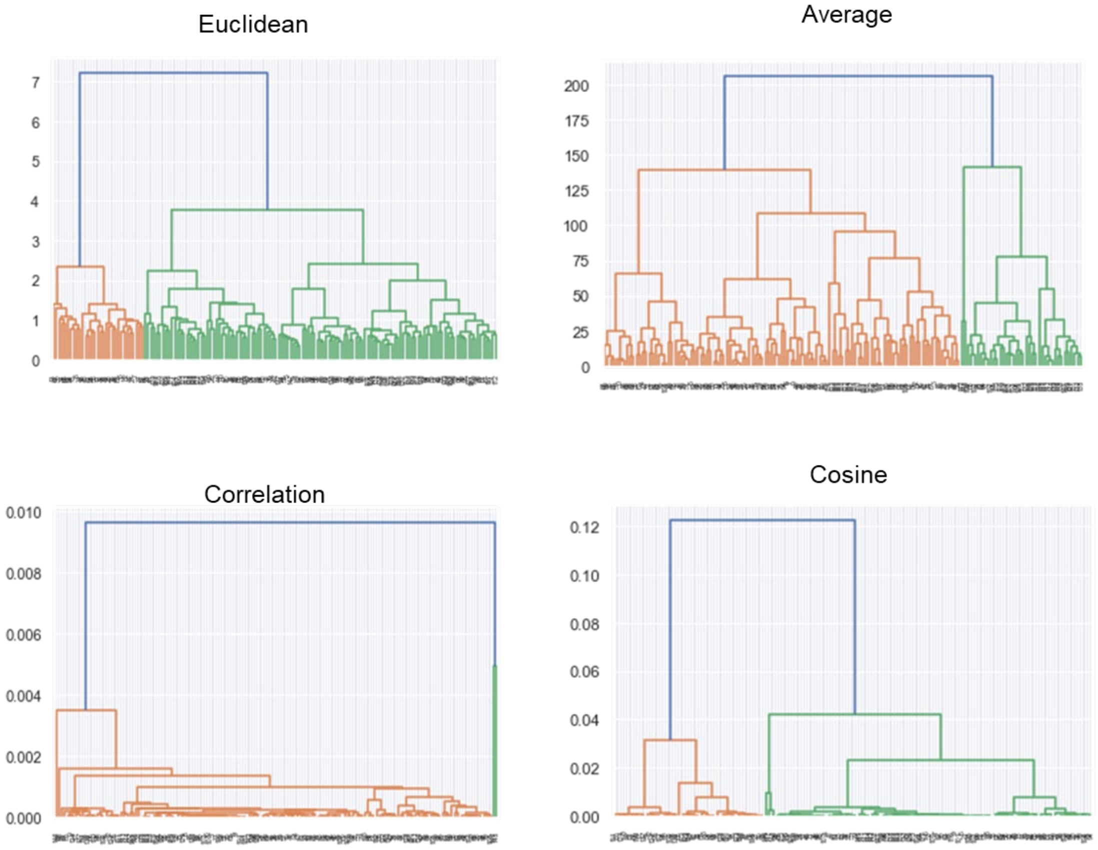


Figure 12 Hierarchical clustering result with different distance measures

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

Table 5 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 than 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 the first 10 posts in Personal Finance subreddit. Here I use Latent Dirichlet Allocation (LDA) (Blei et al., 2003)

I choose 10 as the number of topics here because we may have some sub-topics such as tax, car, insurance, housing, loan, etc.

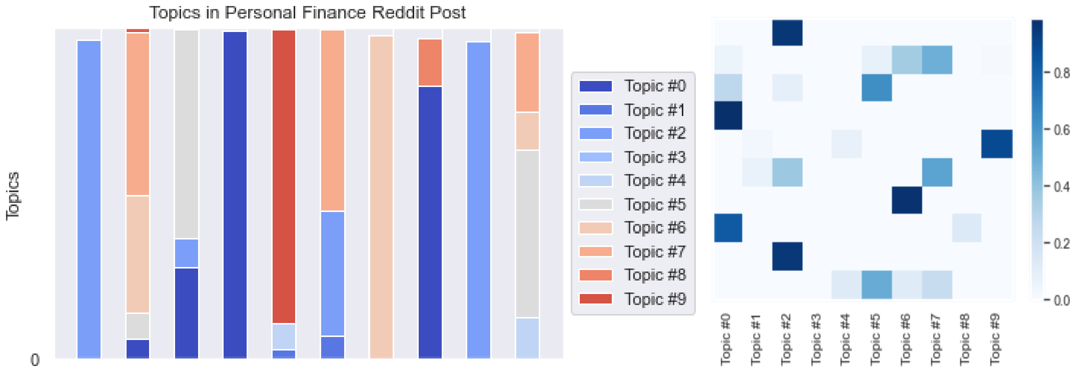


Figure 13 Topics distribution in first 10 posts from Personal Finance Subreddit

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 6 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 the same words that appear on many topics. For example, 'pay' and 'year' appears on almost every topic! But there are also some distinctions between each one, for example, I think topic 1 and topic 7 have referred to time, containing words such as ‘year’, ‘month’, ‘day’, and ‘time’. Topic 4 and topic 5 are about cars. But overall, I think the distinction between each topic is not very significant, perhaps because we choose the 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.

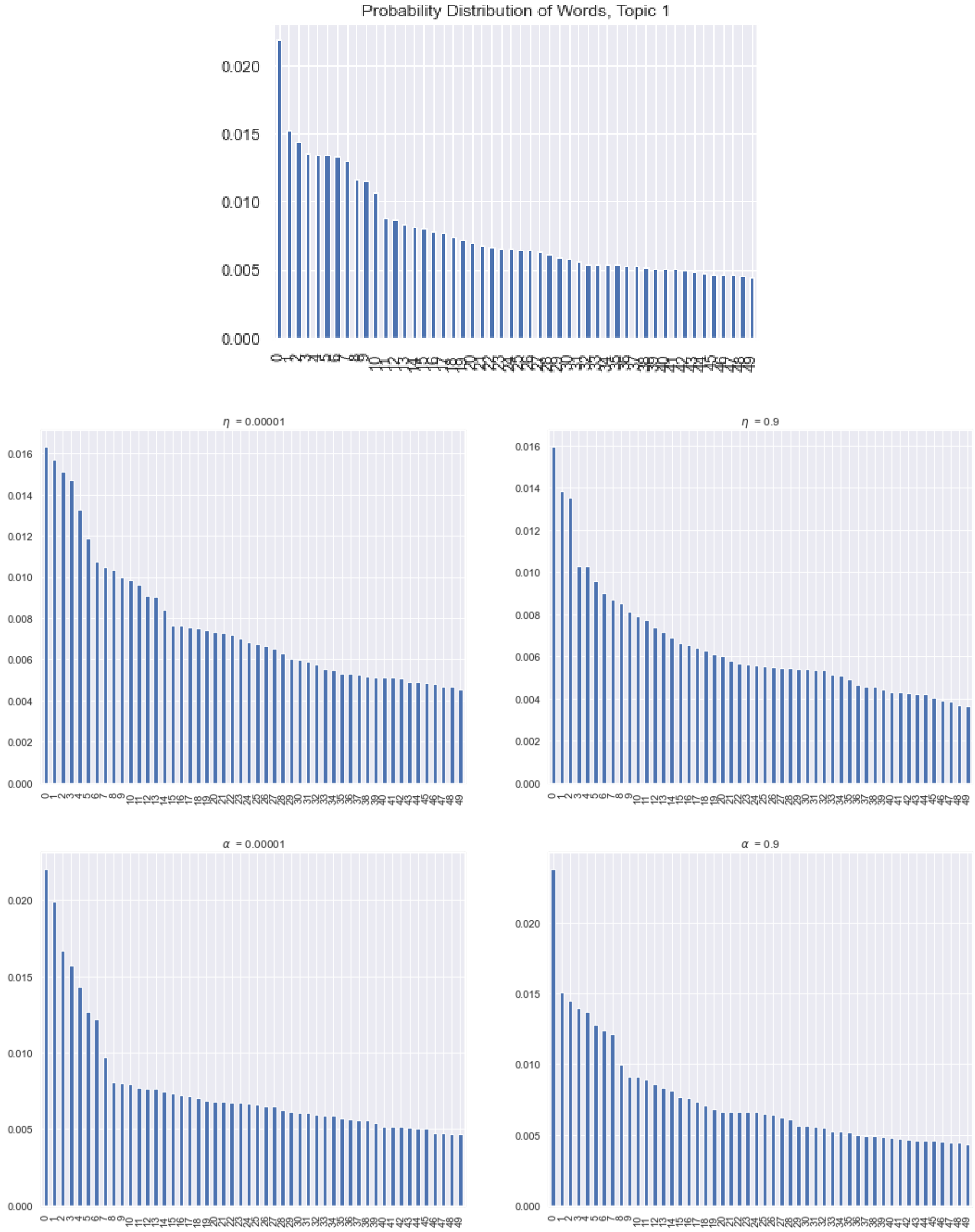


Figure 14 Probabilities of words and parameters tuning for topic modeling

α controls the sparsity of document-topic loadings, which means if one document is made of one topic or more. η 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 η changes the topic a lot, while α 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 η, the probability of different words becomes similar, but when we increase α, 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. Because Reddit API has the restriction of at most 1000 posts for each subreddit, I turn to the Google Cloud Big Query platform to download the archived data. The earliest year for the database is 2015. However, the database for Archived Reddit is not very up-to-date and its manager stopped maintaining it in 2019. So I only have data in 2015, 2016, 2017, 2018, 2019, and 2021. I also limit my corpora to only the Personal Finance subreddit because it contains more on people's overall concerns instead of just investment.

Following are the tables showed the results of words evolution for 5 topics across the six years.

Table 7 Evolution of top words in topic 1

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

Table 8 Evolution of top words in topic 2

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

Table 9 Evolution of top words in topic 3

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

Table 10 Evolution of top words in topic 4

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

Table 11 Evolution of top words in topic 5

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

From the top words in each topic, we find that the first topic is more about job and company; the second topic is about car and paying back car loans; the third topic is about fund, stock, market, vanguard account, etc., the fourth topic is about time and retirement, and the fifth topic is about credit card, credit score, bank account and check, etc. We found the topic here is more significantly separated and featured than the last part, maybe because we use a large size of data.

However, despite the clear meaning for each topic, the change in language is not very significant. There are some minor changes, for example, in topic 1, the word *job*'s rank increases from second place to first place after 2016, and in topic 2, the word car's rank increases from sixth place in 2015 to third place in 2021. But overall, it's a very slight change. I think the reason could be 6 years is relatively a short time compared to the process of language evolution.

Vector Space and Word Embeddings

In this part, we train a model with two different Google's word2vec algorithms, CBOW (The Continuous Bag of Words) and Skip-gram to embed our corpus words in their local linguistic contexts, and explore their locations in the resulting space to learn about the discursive culture that produces them. (Evans et al., 2021) We also use doc2vec to find the relationship of documents

Next, instead of just looking at how words are embedded in the vector space, we can also embed documents within the space by using the Doc2vec model. Finally, we can create two dimensions to make projections to them to further understand the structure of our corpora.

Word2Vec

CBOW is trained to predict a single word from a fixed window size of context words, whereas Skip-gram does the opposite. We can see their difference in the following plot.

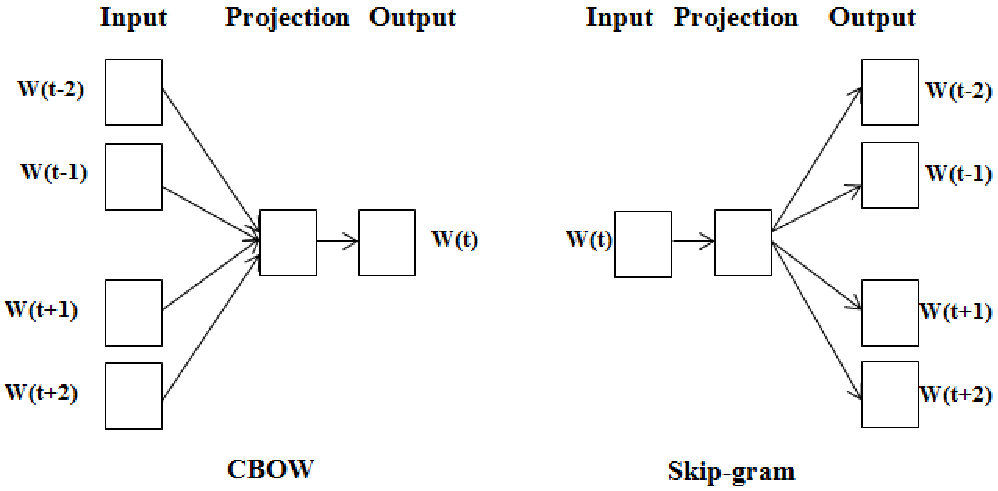


Figure 15 The difference between CBOW and Skip-gram

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. (Evans et al., 2021)

After training with word2vec, we can use PCA to reduce the dimension and T-SNE to project them down to two in order to visualize the words' locations in the semantic spaces.

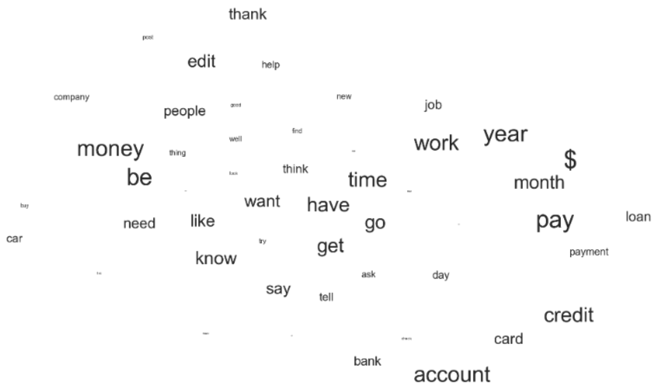


Figure 16 Visualization of CBOW

I repeat the visualization several times because the plot is kind of nondeterministic. In this visualization, we find that "bank", "account", "card", and "credit" are near each other. "year", "month", "pay", "payment", and "loan" are also located together. "car" is very far away from others.

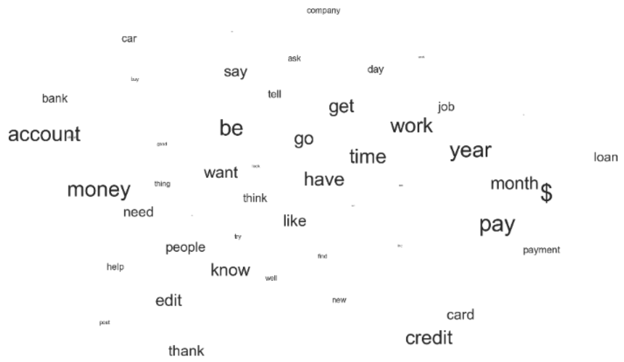


Figure 17 Visualization of skip-gram

The result of skip-gram is very similar to CBOW. Time and payment related words such as “year”, “month”, “pay”, “payment”, and “loan” are also located together. “credit” and “card” are in a group. “get”, “work” and “job” are near each other.

We can also find other patterns that reveal the structure of the semantic organization of words in the Personal Finance corpora. For example, in a word list of ['student', 'loan', 'debt', 'payment', 'account', 'investment', 'tax'], the most different word in is 'investment'. Probably it's because the investment is an action for people who 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 or subtraction relationships, we got following relationships:

pay + debt = loan + payment

student + loan = pay + $

credit + score = account + start

mortgage + house = lose + refinance

We can find the most similar words to a topic or topic by CBOW or skip-gram. The following table shows the result. The number followed by each word is the cosine similarity between the word and the topic. For example, the cosine distance between ‘finance’ and ‘economics’ is 0.9970, which is a quite high number and means they are very similar.

Table 12 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 13 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 14 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)

From the table, we can find that the most similar word found by CBOW and Skip-gram are very different! There are almost zero overlap for the top 10 words. For example, for the word ‘loan’, CBOW predicts the most similar word is ‘pay’ but Skip-gram’s prediction is ‘student’ (another evidence that the student loan plays such an important role in contemporary financial concerns!) Besides, for the word ‘house’, top words in

CBOW include ‘buy house’, ‘house expense’, but for Skip-gram, ‘cheap house’, ‘own house’, ‘sell house’ and ‘rent house’ rank higher.

Doc2Vec

With the doc2vec model, instead of only get vectors for words, we can also get them from documents and tags. Documents are the centroid of their words.

Since the documents now have vectors, we can compute the distance between words and words, documents and documents, and also words and documents. First, we can calculate the distance between words:

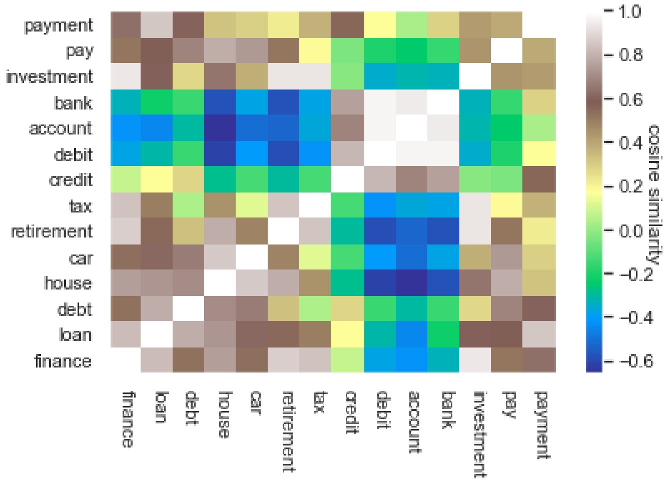


Figure 18 Doc2vec: heatmap for the distance between words

Then, we can look at the distance between words and documents:

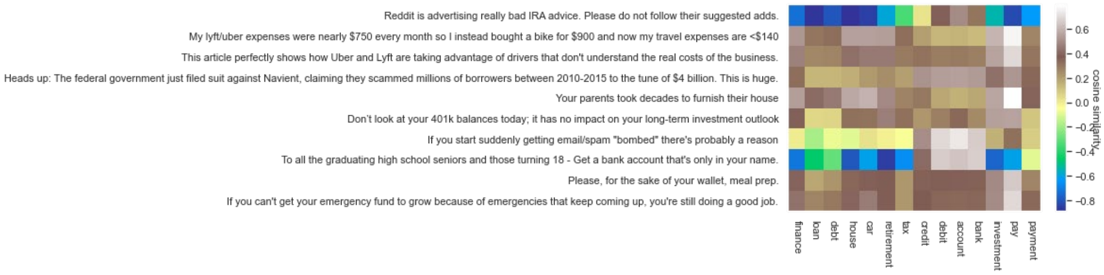


Figure 19 Doc2vec: heatmap for the distance between word and distance

We have some positive as well as negative correlations. For example, the word ‘payment’ has positive cosine similarity to words ‘credit’, ‘tax’, and ‘debt’.

In the heatmap for the distance between words and documents, we can find that the word ‘retirement’ is negatively correlated to an article named “To all the graduating high school seniors and those turning 18”. Of course, they should not be very similar.

Projections

In the last part, I create 2 dimensions, namely ‘family’ which ranges from the older generation in a family to sons and daughters, and ‘financial status’, which ranges from poorer status with loans, debts, and poor to more affluent status with savings and investments.

Table 15 Dimensions for projection

	Dimensions	
family	[‘mom’, 'parent’, 'father’, 'dad’, 'mother']	['son’, 'daughter’, 'child']
financial status	['loan’, 'debt’, 'poor’, 'unemployed']	['saving’, 'investment’, 'rich’, 'job']

Two lists of words are projected, they are about tax and retirement

The first list is about **tax**, including: *'tax', 'income', 'salary', 'earn'*

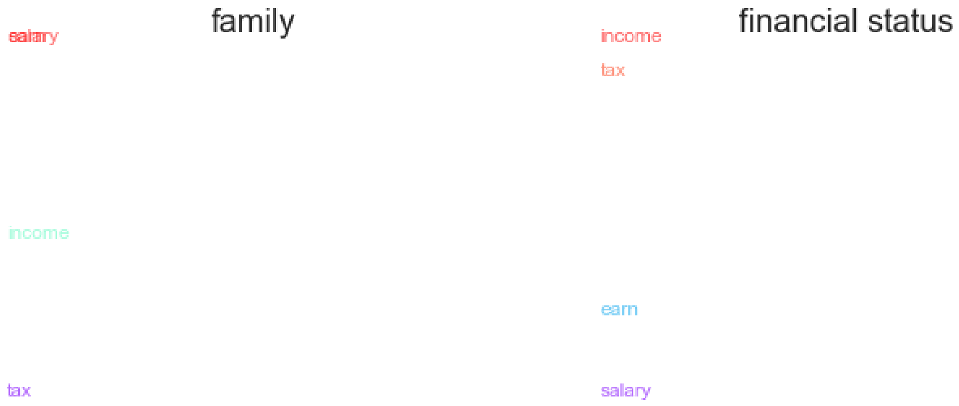


Figure 20 Projection results for ‘tax’ related words

Financial status explains the most variation in the 'tax' dimension, maybe because the tax is more related to finance compared to family.

There is a very strange result for me. People in poor financial status use the word ‘income’ more and those in better financial status use the word ‘salary’ more. I don’t quite get the subtle difference between these two words.

The second list is about **retirement**, including: *'retirement', '401k', 'ira', 'young', ‘old’, ’plan’*

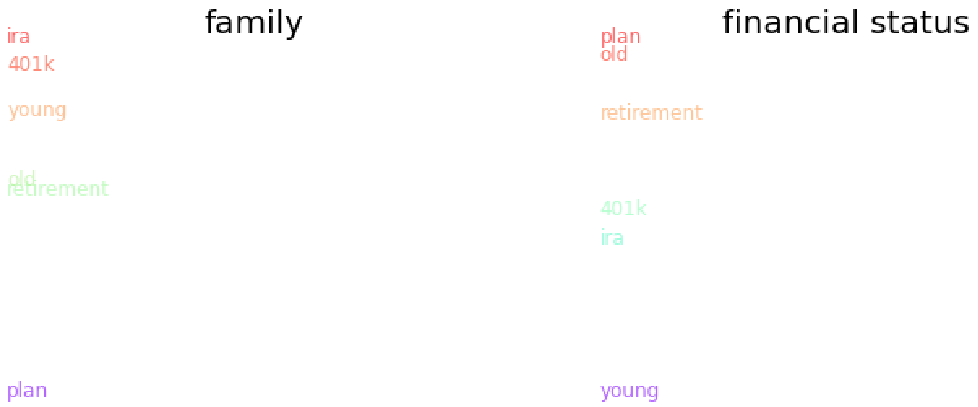


Figure 21 Projection results for ‘tax’ related words

Both dimensions explain the variation in the 'retirement' word list well because the concept of retirement is important for both family and financial conditions. But I find it a little surprising that words ‘plan’, ‘old’ and ‘retirement’ are nearer to poor financial conditions. The reason could be a sample bias, that only people in poor financial conditions worry about their retirement and discuss it.

Conclusion

In conclusion, for the three hypotheses listed in the introduction part:

> ***Hypothesis 1***: *The most frequent personal finance concerns include saving, paying back credit cards, student loans, housing, and insurance.*
>
> ***Hypothesis 2***: *The topics of these concerns change over time.*
>
> ***Hypothesis 3***: *People talk differently in different online forums.*

We find that the first Hypothesis is correct, top financial concerns are about saving, paying back credit card, bank account, debts, student loans, cars, housing, job, insurance (health, car, and house insurance), tax, and retirement.

There is no evidence in our corpora from 2015 to 2021 that supports the change of topics, maybe because 6 years is too short for topics to change. If we can get corpora 10 years or 20 years earlier, we might be able to discover some patterns.

The themes and posts are very different across Personal Finance and Wall Street Bets / Investing. The former is more about problem-solving, the latter is more about trading and investing strategies. Wall Street Bets is not that different from Investing, because they both talk about the stock market a lot.

Discussion

I think my analysis mainly has the following weakness:

Generalization Bias: most users of online platforms are young people who are used to the internet. Middle-aged people may not be willing to disclose their financial concerns online. But the good news is, a lot of young people are describing their parents' financial problems and ask advice to help them. So, although there may not be enough direct information about the financial concerns for middle-aged or old people, there are many posts that indirectly mention them.

Sample Size: Although I include three different forums and texts over 6 years in my corpora, my data size is still not big enough or diverse enough since all articles are from Reddit. Maybe in the future, I can dig more discussion about personal finance from digitalized books so I can get people's discussion even before 2008, the year Reddit becomes popular and people start to talk about it.

The biggest strength of this paper is that we take advantage of online discussion and natural language processing models to find out people's financial concerns, which is a convenient and low-cost way compared to the traditional survey method.

References:

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, 3, 993-1022.

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

Evans, J., Desikan, B. S., & Kwon, H. (2021). Computational Contetn Analysis Notebooks and Slides. (Reprinted.

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