## Research Questions

What's the most frequent concern of personal finance? Is there a heterogeneity among different groups? Does the topics of these concerns change over time?

Corpora:

\* Reddit Articles in subreddit [Personal Finance](https://www.reddit.com/r/personalfinance/)

\* Investing subreddit: [Investing](https://www.reddit.com/r/investing/)

\* Wall Street Bets subreddit: [Wall Street Bets](https://www.reddit.com/r/wallstreetbets/)

Social Game:

Consumption and investment are two import social indicators in economics. So, I would like to study people's consumption and behavior by their online postings.

Actors:

Most people who post articles on Reddit are young people, many of them are 20-30 (many of them reveal their age in posts) and it's interesting to learn the consumption and investment patterns of these young people.

World:

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/cars, 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.

What's people's biggest concerns in personal finance? Do they Change over time?

## Why my research important?

In the most widely used formula in Macroeconomics: Y = C + I + G + NX

(Total economic output = Consumption + Investment + Government spending + Net Export), consumption and investment are individual activities that constitute of our society.

A [Federal Reserve survey] (https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwiB6O2fhs3uAhXQXc0KHbeLAXUQFjAAegQIARAC&url=https%3A%2F%2Fwww.federalreserve.gov%2Fpublications%2Ffiles%2F2017-report-economic-well-being-us-households-201805.pdf&usg=AOvVaw33ULJILWvmE0JU8Dweye4R) finds almost 40% of American adults wouldn't be able to cover a $400 emergency with cash, savings or a credit-card charge that they could quickly pay off. Why do people in the United States, the most powerful country in the world, face this problem? What's the heaviest financial burden on people? Where is the money going? What are the topics that people who seek financial security talks every day? To answer these questions, we can analyze people's posting online.

\*\*The benefits people can get after they learn the results of my study\*\*

My study will report the most common financial burden on people, and the time trend of the changes most-discussed topics. So people can know what bothers us and if the things that bother us change over time.

## How

1. Collecting data and basic cleaning: Use Reddit API to get data from [Personal Finance](https://www.reddit.com/r/personalfinance/) subreddit; tokenize, normalize and vectorlize the data.

2. 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.

3. Topic Modelling: Do Latent Dirichlet Allocation topic modelling for texts of subreddits.

4. Word Embedding

## My sample

\*\*The rationale behind my proposed sample design\*\*

Collect data from online [financial discussion forums](https://www.doughroller.net/personal-finance/8-awesome-online-forums-personal-finance-investing/): Reddit-Personal Finance, myFICO Forums, YNAB Forums, Morningstar Forums, Reddit–Investing, and Bogleheads Forum. My sample will include the first four datasets.

\*\*Social Game:\*\*

People's income and financial concern.

\*\*Social Actors:\*\*

Online financial websites users: people who post their concern, seek for advice, or share personal experience.

\*\*Its virtues with respect to my research questions:\*\*

People's online discussion is a reflect of their real-life concern

\*\*Limitations:\*\*

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 boarden my dataset by scale up the time peorid to include more aticles from myFICO Forums, YNAB Forums, Morningstar Forums, Reddit–Investing, Bogleheads Forum, Fat Wallet Forums, and Bigger Pockets Forum.

# Corpora

Personal Finance

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.

'People tend to feel a sense of guilt when it comes to leaving a job like they owe them or their coworkers something. That is because America preaches this "family" culture that we are such a strong team all working together. In reality, if they need to close your entire division, they will do it without hesitation. If they can outsource something cheaper, they will do it. You do not owe them anything and if you see a better opportunity for yourself or your family, please take it and make your own financial future.'

WSB

I am proud to do my part in paying forward our good fortune with a donation of 6 Nintendo Switches and games to go with them to the Children’s Minnesota Hospital. Cant Stop. Won’t Stop. GameStop. (Still long 50 shares I WILL NOT SELL)

It runs very deep, my friends.

Crazy mannnnnnn. We can't let this slide at all

I’m so proud of you all.

WE'RE IN THE ENDGAME NOW

'LEAVE ROBINHOOD. They dont 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;

Robinhood and other brokers literally blocking purchase of $GME, $NOK, $BB, $AMC; allow sells

United Airlines stock down over 5% premarket trading

Bitcoin was nearly $20,000 a year ago today

If in 2001, you bought $399 of Apple stock instead of buying the original iPod, today that stock would be worth ~$62,000.

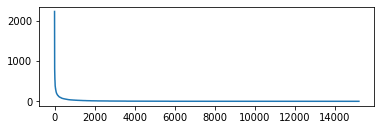
"See title. Can't buy these stocks on RH, but can sell. What the hell is this? How is this legal?"

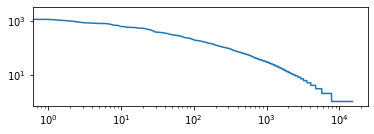
Top 20 words

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | word | count |  | word | count |  | 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 (cars are really important in American cultures, such as car insurance and car loan.

We can also look at the top words in Wall Street Bets and Investing subreddit.





ests/new?ticket\_form\_id=38824 banks student loans credit reports debt collector

companies to match their employees student loan repayments in the same way com

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

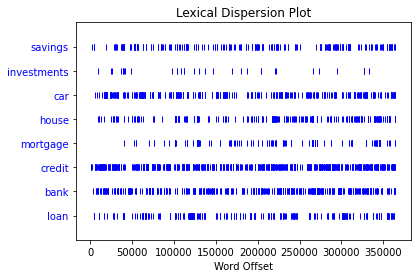
by age 30 than their peers without student loan debt i think most of us with s

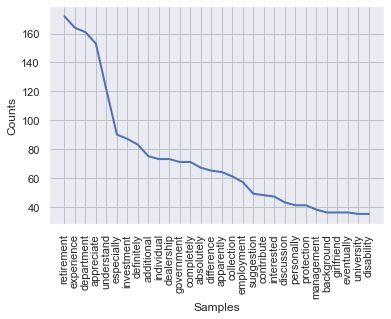
t loan debt i think most of us with student debt have at one point or another f

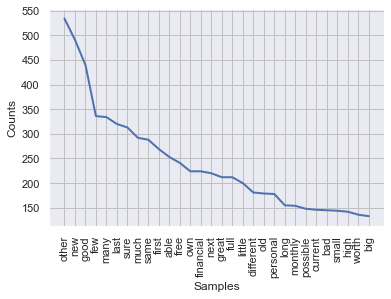
you would be able to make qualified student loan payments and have your company

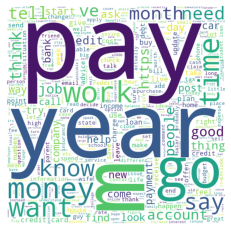
ch month you made a payment on your student loan this does n't hurt people with

g my way out of that lovely college student debt i incurred over four years and









|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | bigram | likelihood ratio | bigram | student t | trigram | student t |
| 0 | (credit, card) | 2643.262 | (credit, card) | 18.42461 | (credit, card, debt) | 6.16226 |
| 1 | (r, personalfinance) | 1549.953 | (student, loan) | 12.52463 | (r, personalfinance, wiki) | 5.830942 |
| 2 | (student, loan) | 1459.923 | ($, month) | 12.26468 | (domain, core, finance) | 5.477223 |
| 3 | (emergency, fund) | 867.1817 | (feel, like) | 10.5862 | (finance, domain, core) | 5.477223 |
| 4 | (wells, fargo) | 819.7325 | (r, personalfinance) | 10.42911 | (=, plubok8lzixw90vxgryjqwfpf4bz, tyegn) | 5.099018 |
| 5 | (feel, like) | 772.0689 | (year, ago) | 9.576361 | (economic, finance, domain) | 4.795828 |
| 6 | (credit, score) | 678.7974 | (edit, thank) | 9.389019 | (pay, credit, card) | 4.679945 |
| 7 | (year, ago) | 640.507 | (credit, score) | 9.319235 | (use, credit, card) | 4.356131 |
| 8 | ($, month) | 628.442 | (m, sure) | 9.288765 | ( , $) | 4.245525 |
| 9 | (debit, card) | 571.5363 | (bank, account) | 9.19252 | (long, story, short) | 4.123074 |
| 10 | (=, plubok8lzixw90vxgryjqwfpf4bz) | 563.0308 | (saving, account) | 8.707325 | (credit, card, company) | 4.11881 |
| 11 | (interest, rate) | 503.7708 | (emergency, fund) | 8.584085 | (file, police, report) | 3.872937 |
| 12 | (bank, america) | 503.4188 | (m, go) | 8.43796 | (thank, take, time) | 3.869329 |
| 13 | (domain, core) | 496.1812 | (year, old) | 8.221385 | (m, year, old) | 3.869165 |
| 14 | (long, term) | 495.7307 | (pay, $) | 7.954511 | (social, security, number) | 3.741621 |
| 15 | (m, sure) | 489.9854 | (debit, card) | 7.823408 | (m, gon, na) | 3.605541 |
| 16 | (gon, na) | 485.1064 | (interest, rate) | 7.750402 | ($, credit, card) | 3.58576 |
| 17 | (saving, account) | 453.3198 | (let, know) | 7.747792 | (=, =, =) | 3.46399 |
| 18 | (edit, thank) | 448.8481 | (save, money) | 7.523519 | (open, credit, card) | 3.462296 |
| 19 | (plubok8lzixw90vxgryjqwfpf4bz, tyegn) | 446.6506 | (long, term) | 7.310689 | (pay, student, loan) | 3.46094 |

## Distributional distances

If we want to compare different corpora, we need a distance or divergence that compares the two distributions.

We will use the:

+ Kullback-Leibler (KL) divergence

+ $\chi^2$ divergence

+ Kolmogorov-Smirnov (KS) distance

+ Wasserstein distance

### Kullback-Leibler and $x^2$ divergences ###

KL and $\chi^2$ divergences are members of the broader <a "href=https://en.wikipedia.org/wiki/F-divergence" target="\_blank">$f$-divergence</a> family, a function of $D\_f (P || Q)$ that calculates the difference between two probability distributions P and Q. The KL $f(t)$ is $ t \text{ log } t $, while the $\chi^2$ is $t^2-1$. KL comes from information and $\chi^2$ from measure theory. As such, the KL divergence computes the relative entropy between two distributions--how they differ in bits, while the $\chi^2$ whether the same statistical inferences can be drawn from them both.

Specifically, given two discrete probability distributions $P$ and $Q$, the Kullback-Leibler divergence from $Q$ to $P$ is defined as:

$D\_{\mathrm{KL}}(P\|Q) = \sum\_i P(i) \, \log\frac{P(i)}{Q(i)}$.

The [scipy.stats.entropy()](https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.entropy.html) function does the calculation for you, which takes in two arrays of probabilities and computes the KL divergence. Note that the KL divergence is in general not commutative, i.e. $D\_{\mathrm{KL}}(P\|Q) \neq D\_{\mathrm{KL}}(Q\|P)$ .

Also note that the KL divernce is the sum of elementwise divergences. Scipy provides [scipy.special.kl\_div()](https://docs.scipy.org/doc/scipy/reference/generated/scipy.special.kl\_div.html#scipy-special-kl-div) which calculates elementwise divergences for you.

The $\chi^2$ Divergence is defined as:

$D\_{\mathrm{\chi^2}}(P\|Q) = \sum\_i \left(\frac{P(i)}{Q(i)}-1\right)^2$.

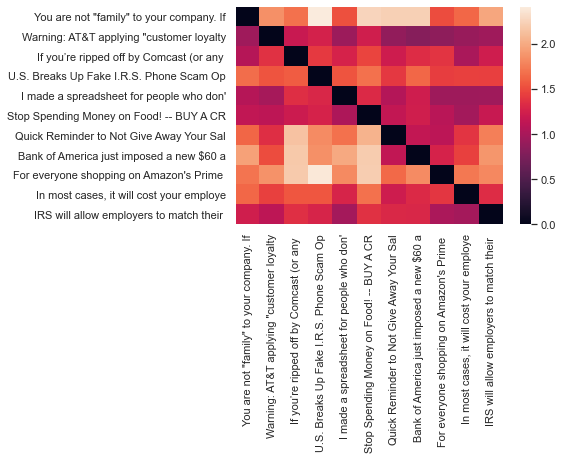
This is also noncommutative, and the code can be drawn directly from scipy.

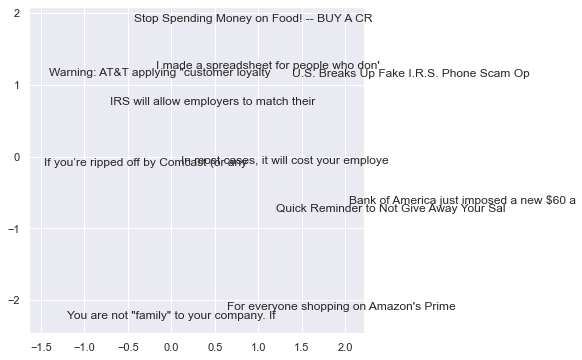
### Kolmogorov-Smirnov ###

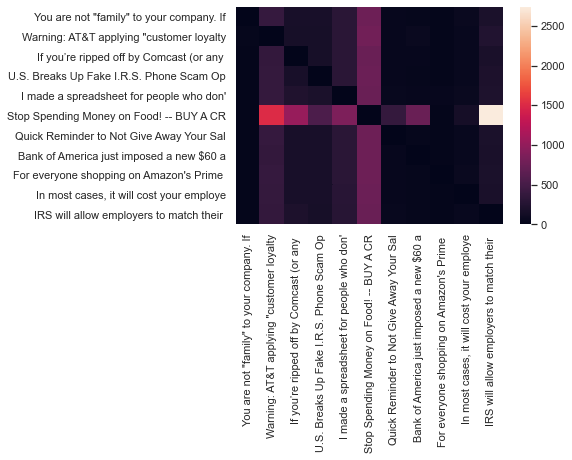
The two-sample Kolmogovorov-Smirnov test statistic calculates the distance between the cumulative distribution function of the two distributions to be compared, and, along with the $x^2$ divergence, is among the most common approaches two calculating a distance in statistics. It can be interpreted as a test of whether two distributions are drawn from the same underlying distribution. As with the others, the code is readily available in scipy.

### Wasserstein Distance ###

When this is computed on a Euclidian metric structure (e.g., numbers of words), this is also known as the earth mover’s distance, because it can be seen as the minimum amount of "work" required to transform $P$ into $Q$, where "work" is measured as the amount of distribution weight that must be moved, multiplied by the distance it has to be moved.











# Discovering Patterns, Clusters, and Topics

What do we want to do here? We want to do vectorization, i.e., converting texts into numerical features (vectors) as required by machine learning algorithms. And this is what feature\_extraction module does: to extract features from texts in a format as required by ML algorithms. feature\_extraction module has four classes: CountVectorizer, DictVectorizer, TfidfVectorizer, and FeatureHasher. Here, we use CountVectorizer, but we'll also use TfidfVectorizer as well below.

There are various strategies by which we extract features. Here, we use CountVectorizer, and, in particular, we use 'Bag of Words' representation. In other words, the features we hope to extract from the texts are each individual token occurrence frequency. We simply count the the occurrence of each token in each document. So, here, we get a document-term-matrix, in which documents are characterized by the occurrences of tokens. Other forms of features, such as the relative position information of words, are ignored. We'll see other types of representations and strategies as well soon, such as N-gram (by the way, we can do n-gram with CountVectorizer. CountVectorizer class takes a set of parameters, such as analyzer, which you can specify the n-gram).

tf-idf data

|  |  |  |
| --- | --- | --- |
|  | **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

介绍一下 k-means和homogeneity completeness这些

Homogeneity: 0.504

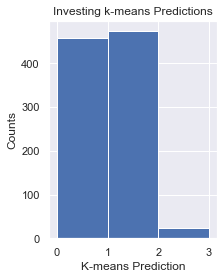
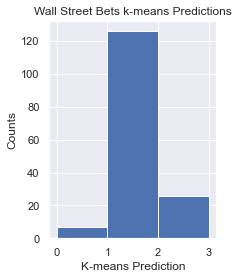
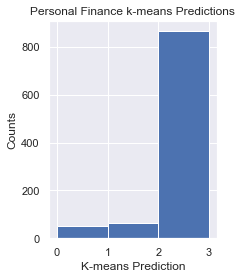
Completeness: 0.428

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 exacterly one class.

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



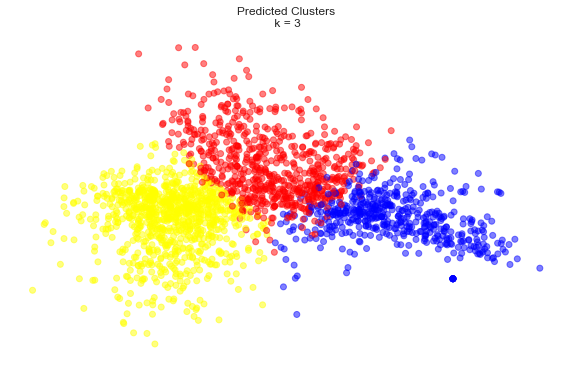
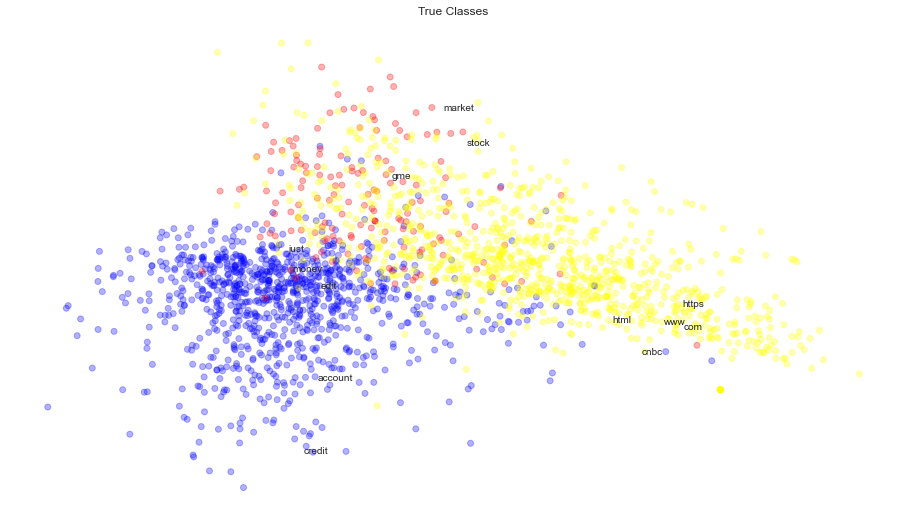
We found that, almost all texts in personal finance category are in cluster 2, which means they are really different (far) from others.

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 some times they could be devided in cluster 2 (which means they have personal finance problem such as tax, debt or retirement), and they haven't realize subreddit 'personal finance' is the best choice

However, for category 'investing', the k-means cluster label is really unstable and inconsistent, in 10 posts, there are 4 in cluster 0, 5 in cluster 1,

|  |  |  |
| --- | --- | --- |
| 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 |
| 2018 | company | pay |
| news | year | job |
| http | people | card |

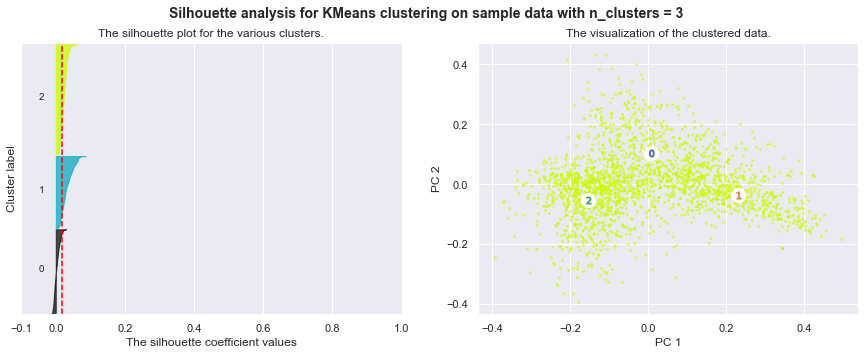
### Plot clusters & features after reducing with PCA



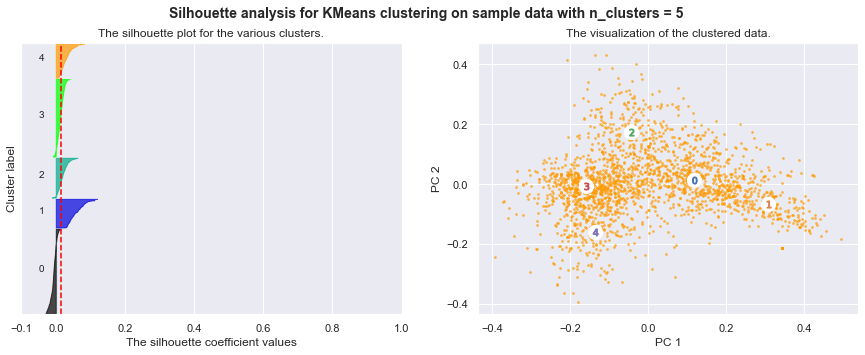
We can find that, similar colors in predicted clusters are nearer to each other. Maybe bacause the size of dataframes are not equal, (there are so many posts in wall street bets subreddit lack content--they only use photo, video, gif or emoji), many topics are related such as planning and budgeting. So in manual labeling they are a lot of overlaps but in machine learning, they could be seperated better.

### Identify the optimal cluster number with Silhouette analysis

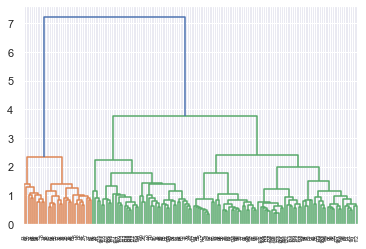




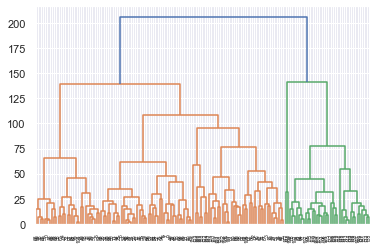




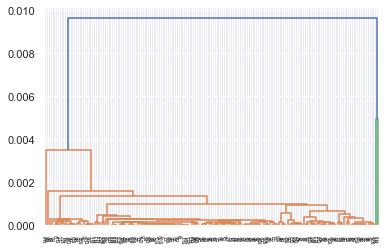
*method='single'*, *metric='euclidean'*, distance



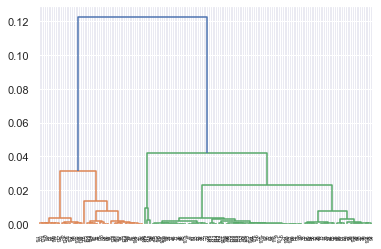
Average distance



Correlation distance



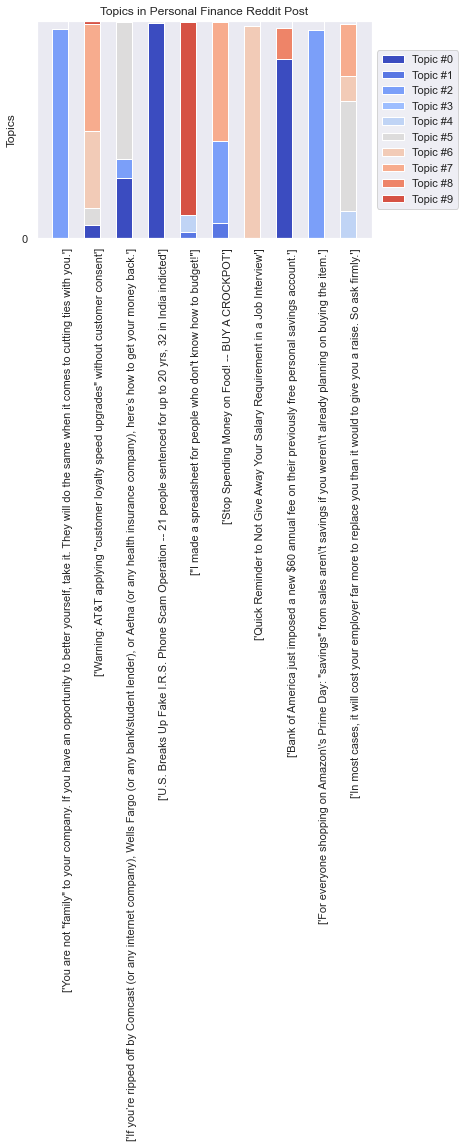
Cosine distance

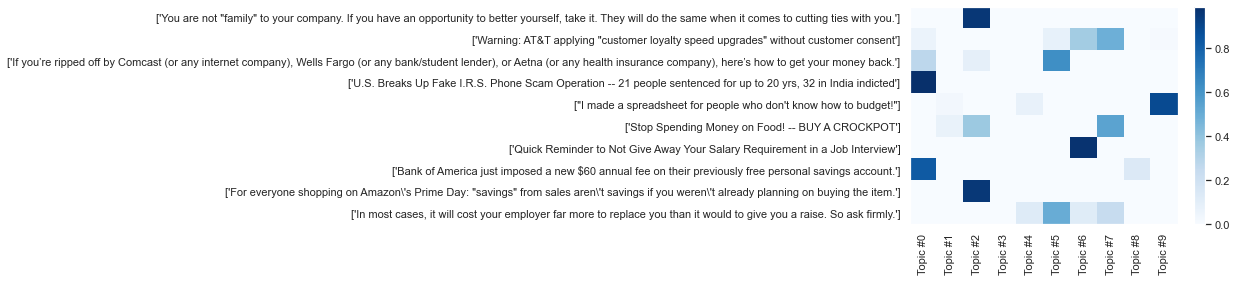


|  |  |  |
| --- | --- | --- |
|  | k-means | Ward |
| Homogeneity | 0.504 | 0.304 |
| Completeness | 0.428 | 0.268 |
| V-measure | 0.463 | 0.285 |
| Adjusted Rand Score | 0.513 | 0.367 |

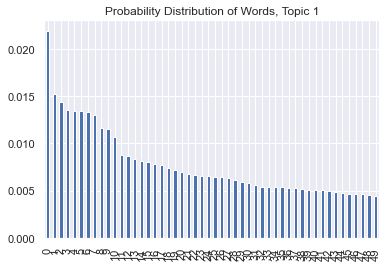
## Topic Modeling

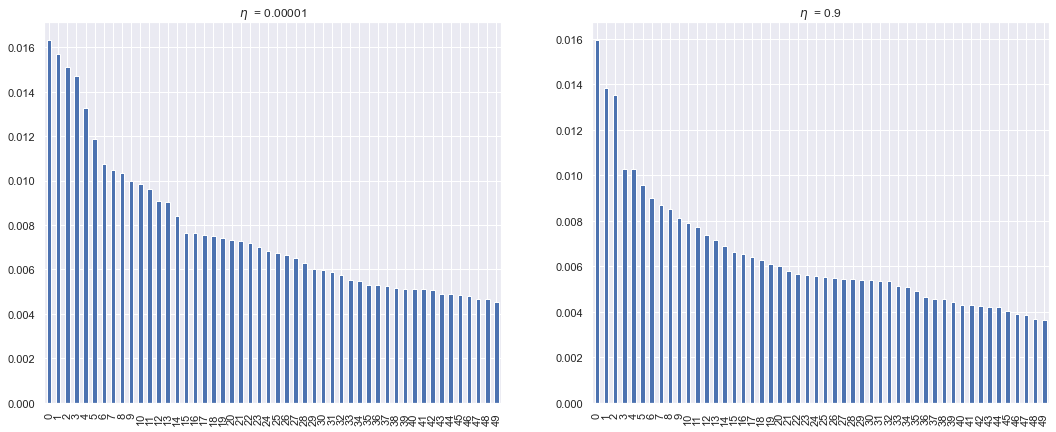
|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| topic\_0 | topic\_1 | topic\_2 | topic\_3 | topic\_4 | topic\_5 | topic\_6 | topic\_7 | topic\_8 | topic\_9 |
| 0.06 | 0.00 | 0.00 | 0.00 | 0.00 | 0.08 | 0.36 | 0.49 | 0.00 | 0.01 |
| 0.00 | 0.21 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.09 | 0.69 |
| 0.00 | 0.00 | 0.36 | 0.00 | 0.00 | 0.00 | 0.00 | 0.63 | 0.00 | 0.00 |
| 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.79 | 0.00 | 0.19 | 0.00 |
| 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.14 | 0.86 |
| 0.30 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.53 | 0.08 | 0.00 | 0.08 |
| 0.14 | 0.00 | 0.00 | 0.00 | 0.00 | 0.47 | 0.00 | 0.39 | 0.00 | 0.00 |
| 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.98 | 0.00 | 0.00 |
| 0.87 | 0.02 | 0.01 | 0.01 | 0.01 | 0.02 | 0.02 | 0.02 | 0.01 | 0.01 |
| 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.99 | 0.00 | 0.00 | 0.00 | 0.00 |

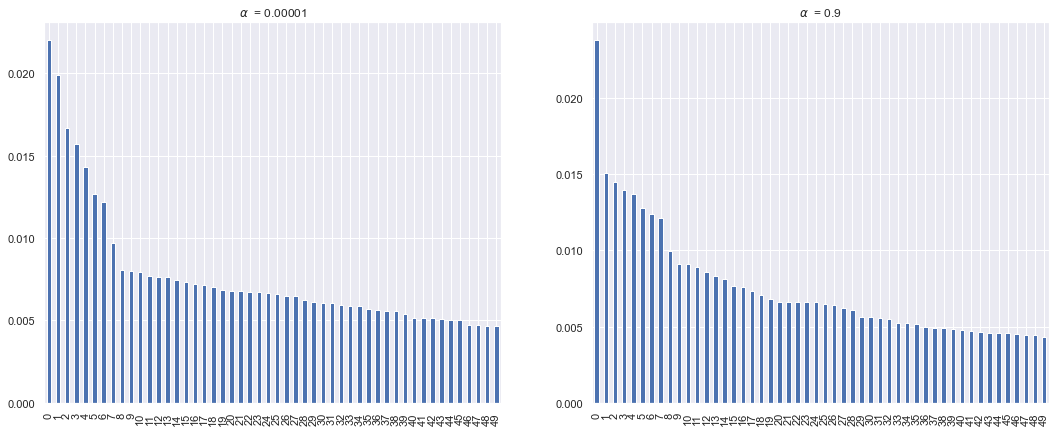




|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Topic\_0 | Topic\_1 | Topic\_2 | Topic\_3 | Topic\_4 | Topic\_5 | Topic\_6 | Topic\_7 | Topic\_8 | Topic\_9 |
| credit | year | money | pay | pay | pay | work | pay | account | credit |
| pay | loan | pay | 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 |







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

## Dynamic Topic Modeling

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 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

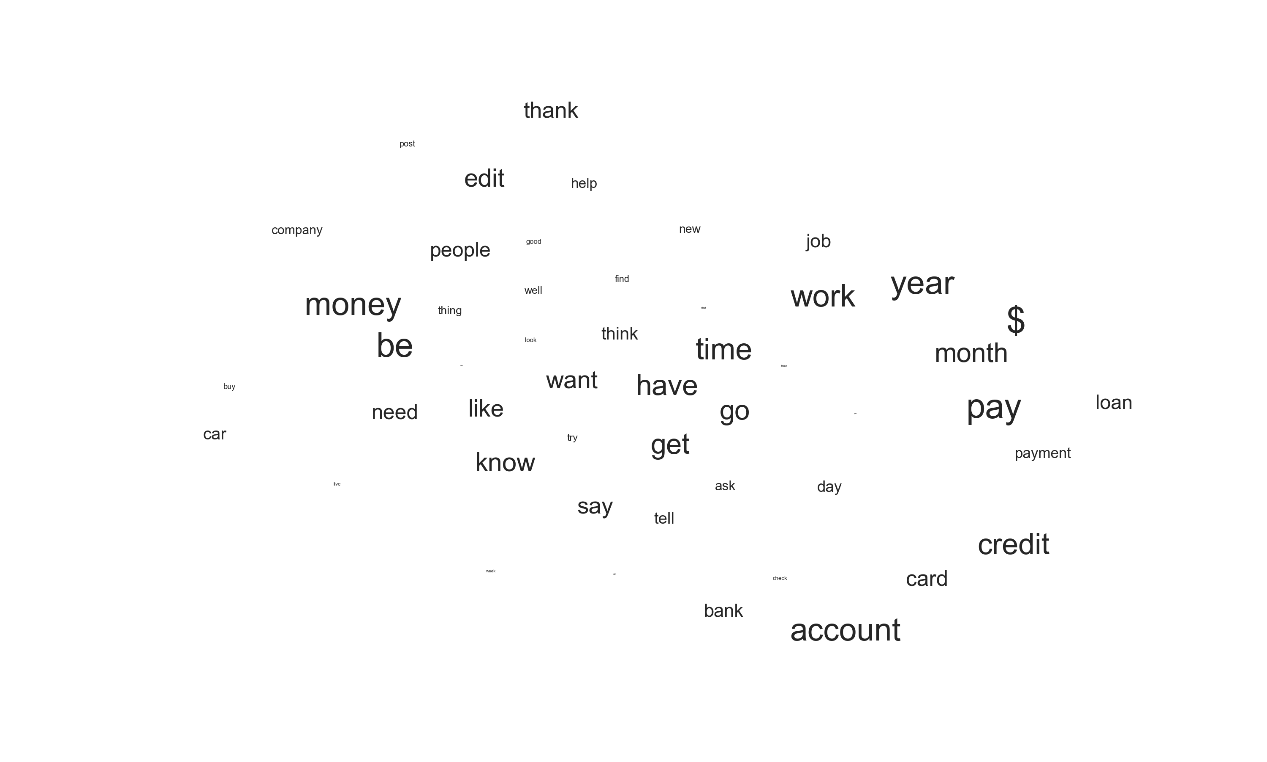
This part, we build on last part's topic modeling techniques by taking a text corpus we have developed, specifying an underlying number of dimensions, and training a model with a neural network auto-encoder (one of Google's word2vec algorithms) that best describes corpus words in their local linguistic 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.

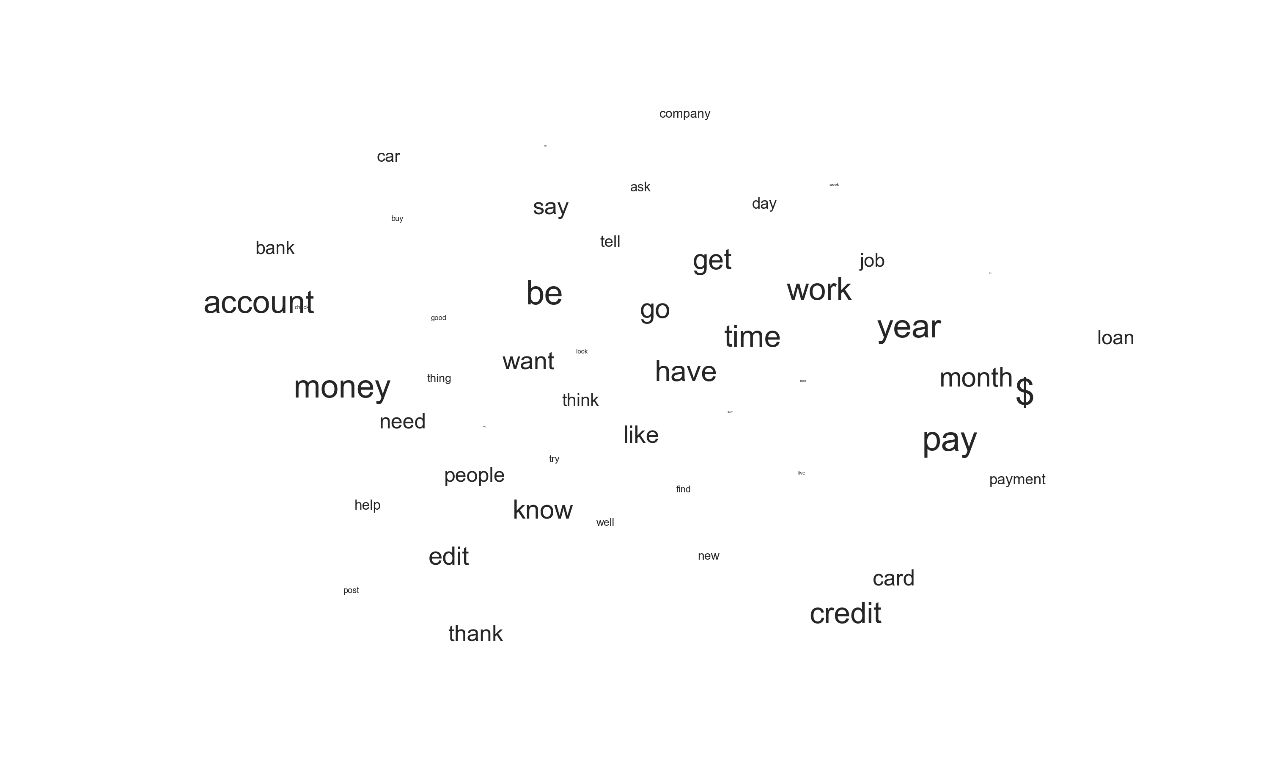
## 5.1  CBOW(The Continuous Bag of Words) Model

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 1 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 2 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 3 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) |

Attentions: We have different columns:

'tokenized\_text',

'normalized\_tokens',

'reduced\_tokens',

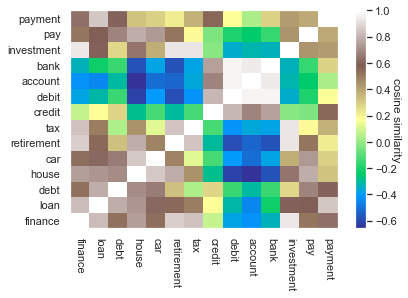
'tokenized\_sents',

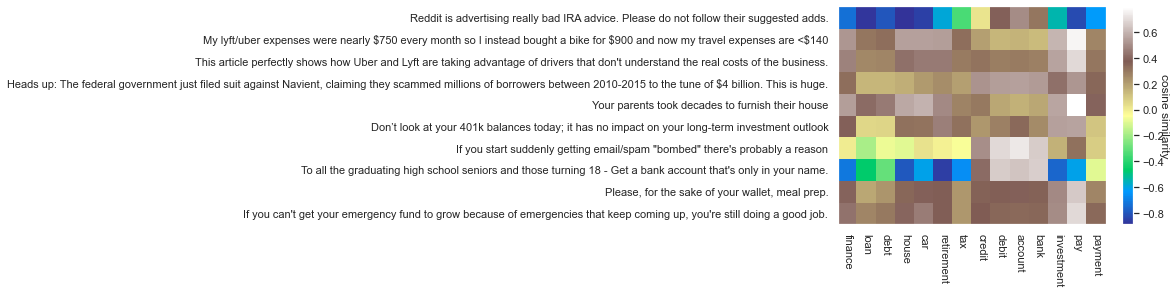
'normalized\_sents',

'tokenized\_words',

'normalized\_words'

# D2V





We have some positive as well as negative correlations.

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

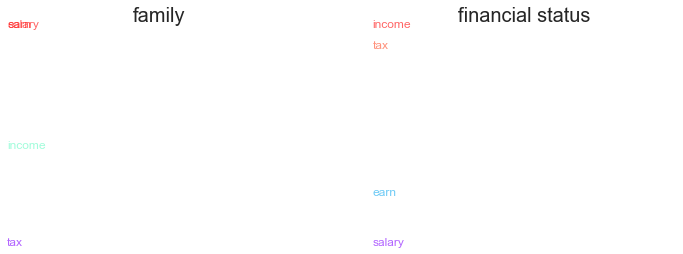
## 5.4 Projections

Create 2 dimensions

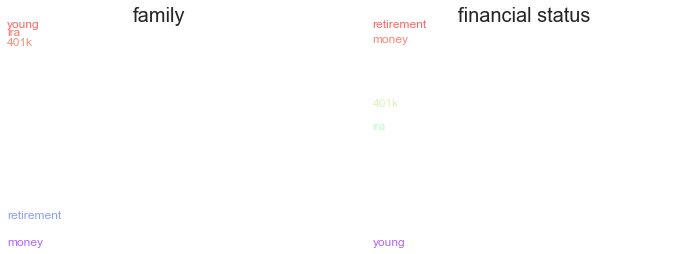
|  |  |  |
| --- | --- | --- |
|  | positive | negative |
| family | ['mom','parent','father','dad','mother'] | ['son','daughter','child'] |
| financial status | ['loan','debt','poor','unemployment'] | ['saving','investment','rich','job'] |

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.

# Reference

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