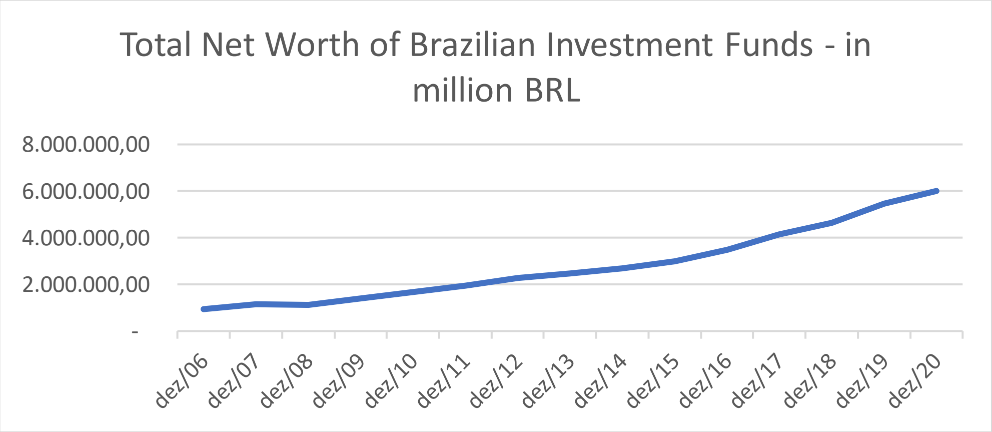
Deriving insights from Brazilian investment funds via Data Science – Final Deliverable

Rafael Y. Imai – April 2021

Description of the problem and a brief discussion about the background

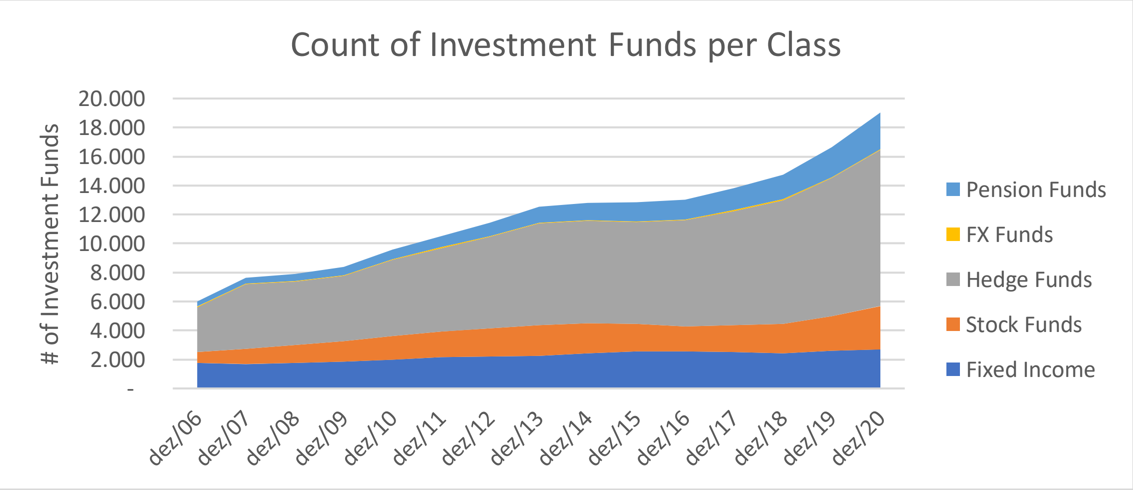
Problem

According to the December/2020 report [1] of ANBIMA (*Associação Brasileira das Entidades dos Mercados Financeiro e de Capitais*, Brazilian Financial and Capital Markets Entities Association in a free translation to English), the amount of assets managed by investment funds is steadily increasing. The graph below shows the total net worth from dec-2006 to dec-2020, in million Brazilian Reais (BRL):



Source: ANBIMA

Along with the total net worth of the Brazilian investment fund industry, the number of investment funds has risen on the same period.



Source:ANBIMA

Along with such metrics, that indicate a growth of such class of financial products, it is interesting to remark that according to a report from ANBIMA [2], Brazilians are increasing their usage of investment funds. In 2017, 4% of the investors used investment funds. In comparison, on the same year, 89% of investors used the savings account. The percentage of investment fund shareholders has risen to 5% in 2018 and 6% in 2019.

Investment funds can be analyzed both by the point of view of the asset manager and the perspective of the shareholder or the shareholder's financial planner. For instance, the asset manager may find insights about the shareholder's behavior according to given market variables useful, as they may help to project new funds or support decision-taking processes. The shareholder, on its turn, may use such insights while designing his/er portfolio, as to avoid excessive exposure on a given fund (concentration risk) or excessive variations on the portfolio's value caused by market dynamics (a concept known as market risk).

Needed Background

An investment fund, in a brief manner, consists of a pool of money belonging to several investors (named shareholders), which receive a given number of shares as they subscribe to the investment fund. Those shares represent an ideal fraction of the investment fund's assets (e.g., stocks, bonds, FX accounts, cryptocurrencies) and its liabilities (e.g., management tax, performance taxes, government taxes), and they may be redeemed if permitted to do so. Closed end funds require the shares to be sold to another willing-to-be shareholder. Open end funds, on their turn, process the redemption by themselves and are free to emit and redeem shares at will.

The fund itself is managed by a management team, consisting of diverse professionals whose functions range from brokers responsible to buy and sell assets, to operational personnel, needed to update records and calculate the share value at the end of the day. Some of the advantages of using an investment fund are the ease of access to professional managers, along with the ability of having a diversified portfolio without necessarily having a lump sum of money to buy high-value assets.

Disadvantages may include, for instance, managers acting recklessly in order to obtain short-term profits, the lack of interference of the shareholder onto the manager (if the shareholder does not agree with a change on the investment fund policy, his/er only options are to redeem their shares) and other managing-related errors.

Several classes of funds exist. Considering the Brazilian investment fund markets, some of the main classes are:

- Fixed Income Investment Fund – *(in Portuguese, Fundo de Investimento em Renda Fixa)* – such funds may invest only in fixed-income papers, such as commercial papers, bonds and deposit certificates. Bonds may range from sovereign bonds (emitted by a country, such as Treasury Bills from USA) to company emitted bonds.
- Stock Investment Fund - *(in Portuguese, Fundo de Investimento em Ações)* – Having a bit more of leeway than Fixed Income Funds, Stock Investment Funds need to have at least a minimum percentage of their portfolio in stocks.

- Hedge Fund – *(in Portuguese, Fundo de Investimento Multimercado)* – Hedge Funds have a larger freedom in terms of investment abilities, being able to invest in several types of assets while still obliged to abide by the Brazilian regulations on the sector.
- Pension Fund – *(in Portuguese, Fundo de Previdência)* – Pension funds are a special class of funds that support private retirement programs. Those programs may be specific to the workers of a given company, or available for the general public.
- FX Fund – *(in Portuguese, Fundo Cambial)* – FX funds are funds whose portfolio attempts to replicate the variation of a given currency (e.g., FX Funds may attempt to replicate the variation of the United States Dollar).

Investment funds need to have three main entities:

- a management company that owns the fund, being responsible to buy and sell assets, along with managing the portfolio and its risk;
- an administrator, responsible to process trades, register the portfolio and price assets;
- a safekeeper, responsible to safeguard the assets.

Subscription and redemption of shares on closed funds involve the steps below:

- Subscription
    - A shareholder registers itself within the fund management company.
    - The shareholder transfers a given sum of money towards the fund.
    - According to the share conversion policy (that determines which day's share value will be used to compute the shareholder's position in shares), the monies are incorporated into the fund's assets and in exchange the shareholder receives a given number of shares.
- Redemption
    - The shareholder requisites the management company to redeem a given number of shares (ranging from a minimum value to the whole amount of shares detained by the shareholder).
    - The redemption request is sent to the administrator, which will process the share conversion according to the fund's regulation and, according to the payment date, transfers to a bank account specified by the shareholder the monies related to the redemption request.

Data items used on this project

All investment funds must be registered on CVM (*Comissão de Valores Mobiliários,* Securities and Exchange Commision in Portuguese), a Brazilian federal authority responsible to regulate capital markets and their participants. They are also required to send daily the value of their shares, along with the subscription and redemption volume liquidated on a given day.

A list of investment funds registered in Brazil since the late 90's [3], along with CSV files containing the daily reported share values and subscription/redemption volumes for the period between 2017 and 2020 [4] were obtained from CVM.

A historic series of IBOVESPA, the São Paulo's Stock Market Index, was obtained via the Yahoo!Finance API. More details can be found here [6]. On the context of this work, IBOVESPA is used as a proxy metric to the stock market's performance. A historic series of SELIC, the Brazilian base interest rate, was obtained by using an API from the Brazilian Central Bank [7]. The base interest rate is an interesting metric, as higher interest rates drove out investors from the stock market towards fixed income investments and vice versa.

Methodology

Before filtering the data, I picked some categories of funds out of the original set. The set in [3] contains almost all forms of investment funds, some of which are either focused on specific purposes (such as FUNCINE funds, whose objective is to support the Brazilian film industry) or belong to classes that are typically closed-end funds (such as FIIs, that are roughly equivalent to a REIT, Real Estate Investment Trust).

As the work's focus is on the dynamics of open-end funds, those funds were selected out of the [3] dataset to create a first set of funds. Other criteria used to assemble this first set were:

- Funds must have started before 2017 and must be active as of April 2021;
- Funds must not be exclusive (that is, they must be available to general public).
- Funds must be regulated only by the CVM Instruction #555 [8], as there are specific products that are also considered investment funds, while being regulated by other laws.

After this first filter, a second filtering layer was applied in order to separate different investment fund classes (for more information, please refer to the section "Needed Background"). Two main categories were chosen due to the size of their samples:

- Fixed Income Funds - *Fundo de Renda Fixa* – 473 samples
- Hedge Funds – *Fundo Multimercado* – 1537 samples

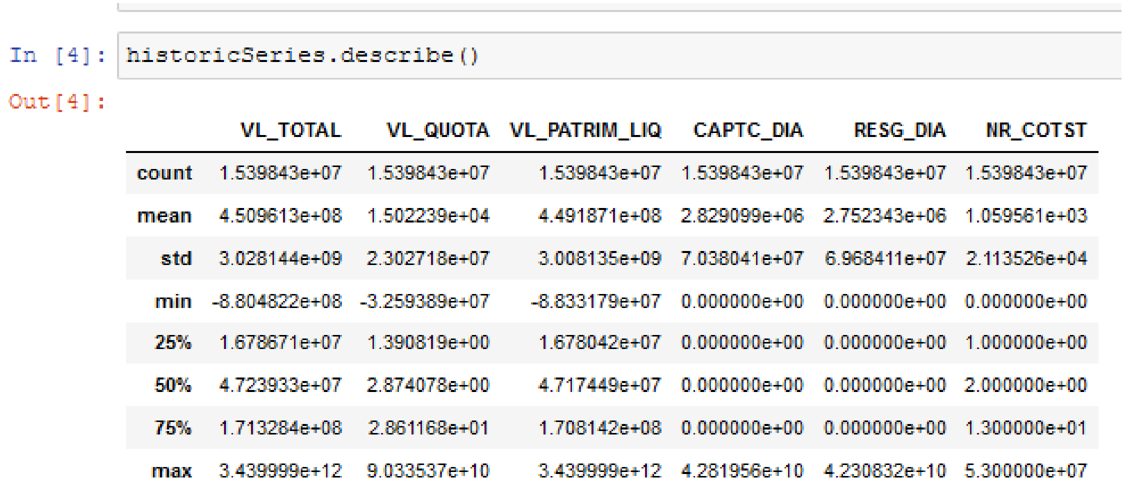
To prospect the data, a Jupyter notebook was used and the correlations between the metrics below and the number of shareholders of the fund. By analyzing such correlations and their behavior, I intend to derive insights about the behavior of shareholders. Afterwards, the k-means clustering algorithm will be used to test whether this set of correlations can be used to form clusters of funds. A list of metrics can be seen below:

- SELIC Rate;
- Stock Market Index (IBOVESPA);
- Capital Flow – defined by the difference between subscriptions and redemptions paid on a given day;
- Average Ticket – defined by the fund's net worth at a given day divided by the number of its shareholders.
- Vol – Standardized volatility, defined by the standard deviation of the share value divided by the mean of the share value. Volatility is considered according to the timespan of the analysis scenario.

In terms of capital flow, the volume of subscriptions and redemptions are based on the date that the fund has received the monies referring to a subscription or has paid the given volume of redemptions to a shareholder. According to the decisions of the management team, there might be a time interval between the redemption request and the payment of the redemption request.

This interval may vary according to the liquidity of the fund's assets – funds with highly liquid portfolios could process redemptions on the same day, while less liquid funds would demand a larger interval to avoid liquidity risk (that is, having to take a loss on an illiquid asset in order to have cash and pay the redemption or not being able to sell the given asset at all). As incorporating this information on the model would require a considerable effort to gather such data for all funds, all funds were considered as if they processed redemptions on the same day as they were requested.

Before processing the correlations, a checkup was performed on the data from CVM, by using the method describe:



```
In [4]: historicSeries.describe()
```

Out[4]:

	VL_TOTAL	VL_QUOTA	VL_PATRIM_LIQ	CAPTC_DIA	RESG_DIA	NR_COTST
count	1.539843e+07	1.539843e+07	1.539843e+07	1.539843e+07	1.539843e+07	1.539843e+07
mean	4.509613e+08	1.502239e+04	4.491871e+08	2.829099e+06	2.752343e+06	1.059561e+03
std	3.028144e+09	2.302718e+07	3.008135e+09	7.038041e+07	6.968411e+07	2.113526e+04
min	-8.804822e+08	-3.259389e+07	-8.833179e+07	0.000000e+00	0.000000e+00	0.000000e+00
25%	1.678671e+07	1.390819e+00	1.678042e+07	0.000000e+00	0.000000e+00	1.000000e+00
50%	4.723933e+07	2.874078e+00	4.717449e+07	0.000000e+00	0.000000e+00	2.000000e+00
75%	1.713284e+08	2.861168e+01	1.708142e+08	0.000000e+00	0.000000e+00	1.300000e+01
max	3.439999e+12	9.033537e+10	3.439999e+12	4.281956e+10	4.230832e+10	5.300000e+07

The net worth of a fund (VL_PATRIM_LIQ) consist of the amount of outstanding shares of the fund times the share value (VL_QUOTA). Under specific scenarios, it is possible for a fund to have a negative net worth. In such cases, shareholders would need to transfer monies to the fund in order to liquidate it, indicating therefore either some sort of distressed fund or a possible mistake on the data. Those cases were filtered out before analyzing the correlations.

Another issue observed with the data was that some of the funds lacked data points. Between 2017 and 2020, for instance, there should be 1003 entries for each fund. Some of the funds had as low as 300 entries, indicating a possible failure to insert the fund/share values for the given date. As the correlations of each fund towards a given metric are considered, only the matching entries of the metric will be considered and therefore we should not expect a relevant impact of this failure. In other contexts, however, it would be interesting to filter out funds that are below a given threshold in terms of data entries.

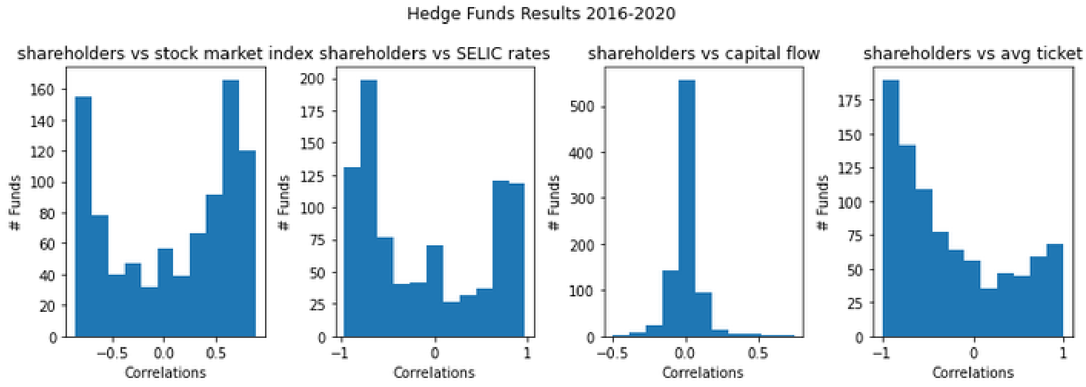
A more general analysis has been done using the time interval between 2017 and 2020, creating a distribution of correlations for each metric. Afterwards, yearly analysis using each year were taken to observe the yearly variation of the distribution of the correlations. Afterwards, the pair plots (via Seaborn) were taken for both types of funds, in order to seek for possible clusters. At the end, k-means algorithm is used to check if its use to find subtypes of hedge funds and fixed income funds return a meaningful difference.

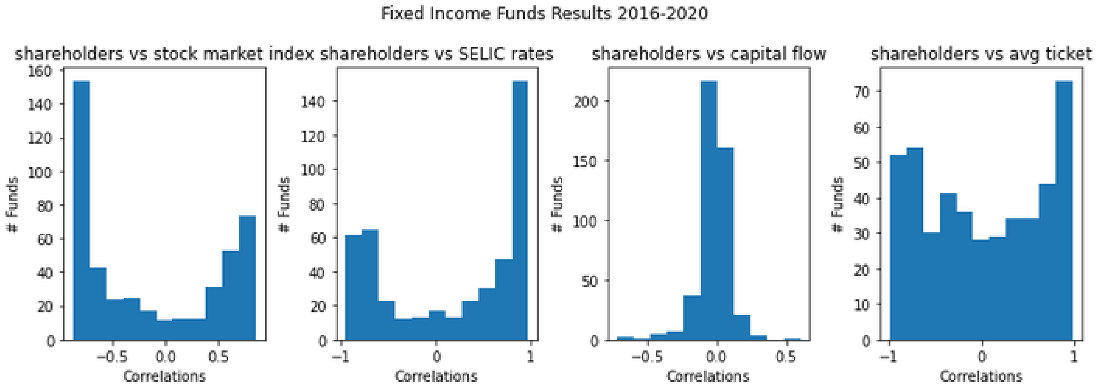
On the latter case, I will run the algorithm within each set of funds, considering the correlations observed over the period 2017-2020. Using the clusters and the labels assigned to each fund, I will check whether subtypes of funds within a fund set exhibit similar behavior (in terms of the considered variables) or not.

Results

Correlation Analysis and Discussion

The figures below show the resulting correlation distributions between 2017 and 2020:





Although simplistic, the idea that “an increase of the base rate will drive shareholders towards the fixed income fund market” (which, in statistical terms, means that the shareholders on hedge funds and the base interest rate should be negatively correlated) is challenged by the data above.

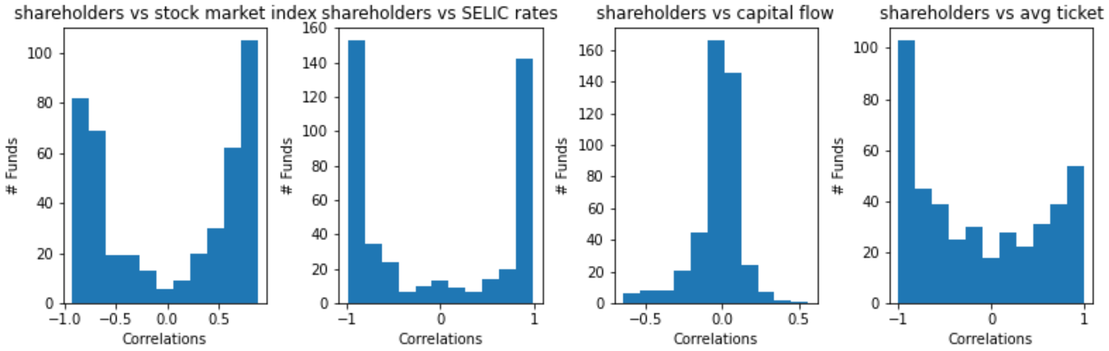
Had the data corroborated the idea, little to no bins with positive correlations both on the “shareholders vs. SELIC rates” graphs on hedge funds, and a similar pattern should be observed on the same graph on fixed income funds, although inverted. In both cases, however, there seem to be a set of “contrarian” funds – that is, funds whose shareholder trends go against the theoretical trend of their classes.

As it is possible to perceive on the graphs mentioned above, the trend of each product can still be seen (that is, fixed income funds having a positive correlation with the base rate and hedge funds having a negative correlation), but a contrarian trend does not seem weak enough to be neglected.

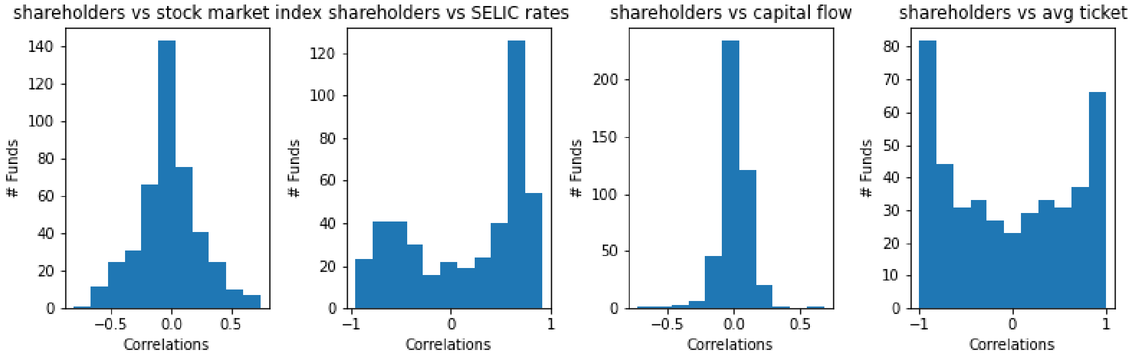
A bit of leeway should be considered while analyzing such correlations for hedge funds, as the enhanced freedom experienced by their managers allow the construction of various types of portfolios. In this case, the correlations could be used as an evidence of the existence of groups with noticeable differences on their strategies.

The sequence of pictures on the next page depict the yearly analysis for fixed income funds.

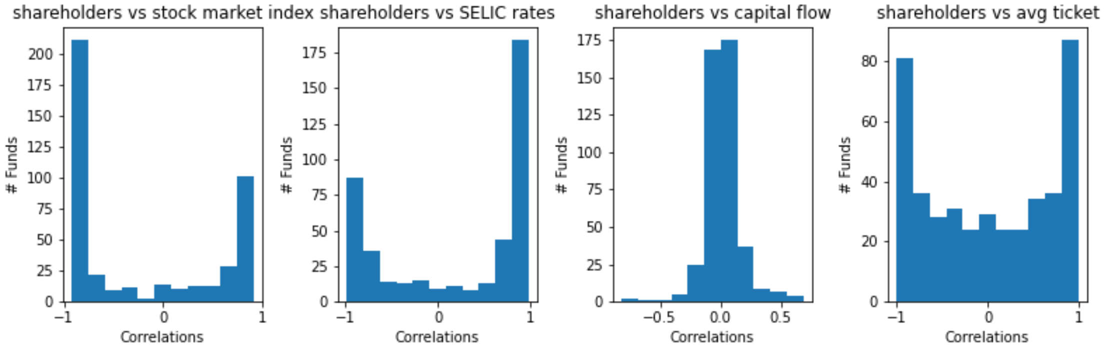
Fixed Income Funds Results 2017-2017



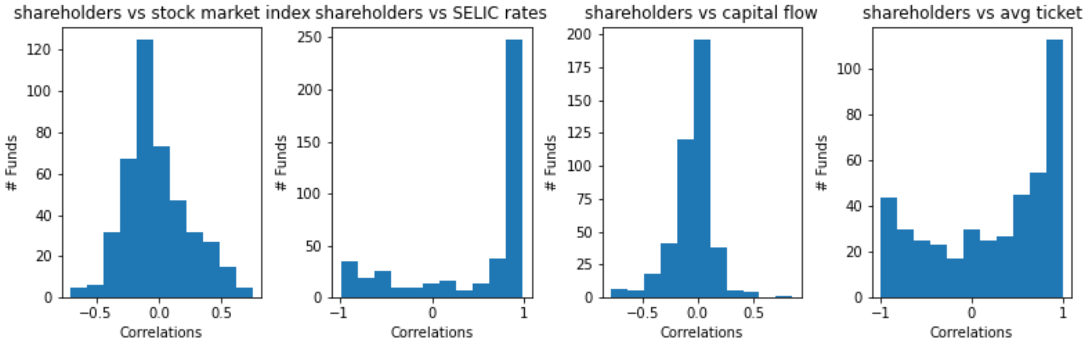
Fixed Income Funds Results 2018-2018



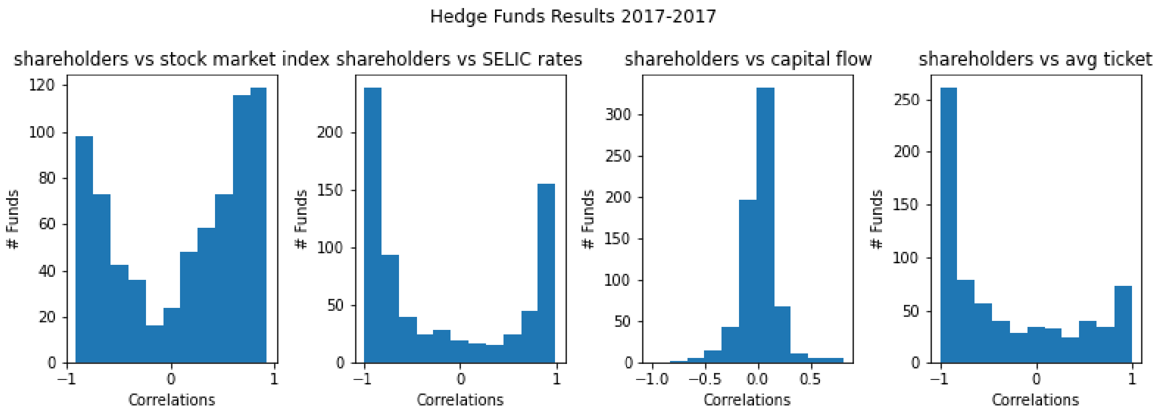
Fixed Income Funds Results 2019-2019

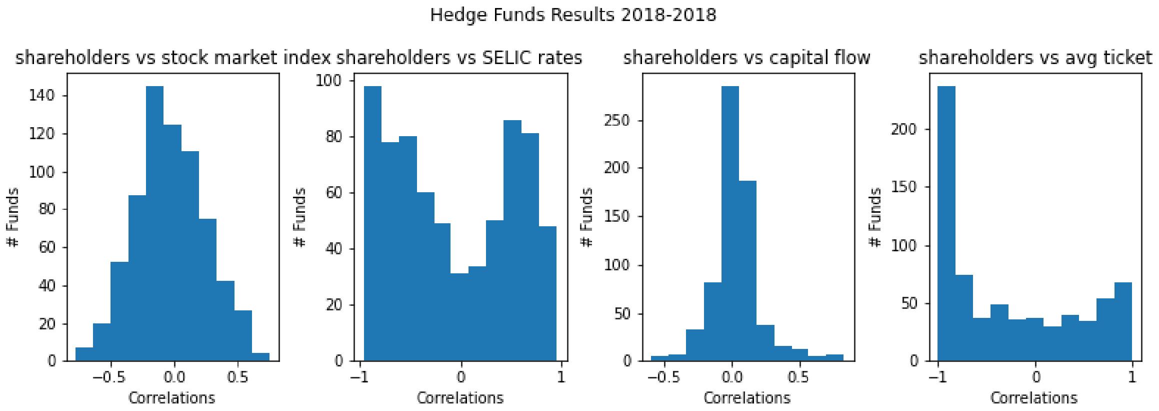


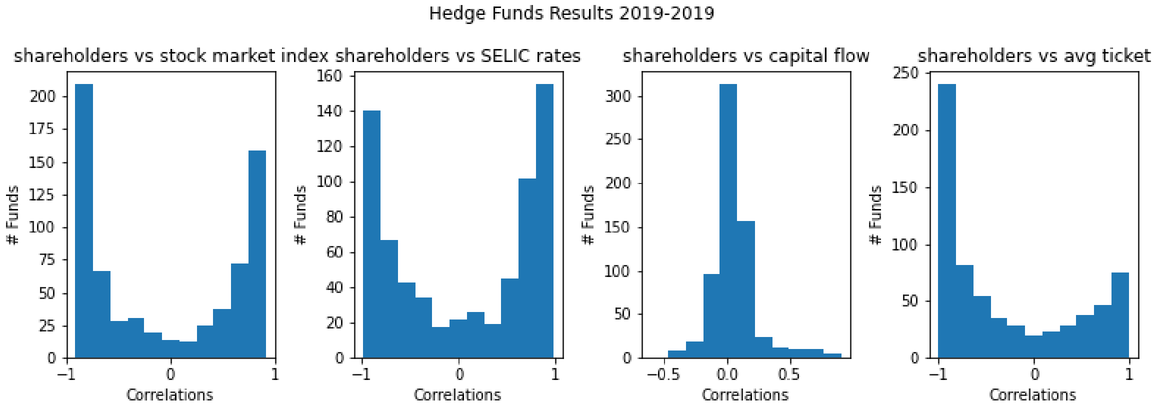
Fixed Income Funds Results 2020-2020

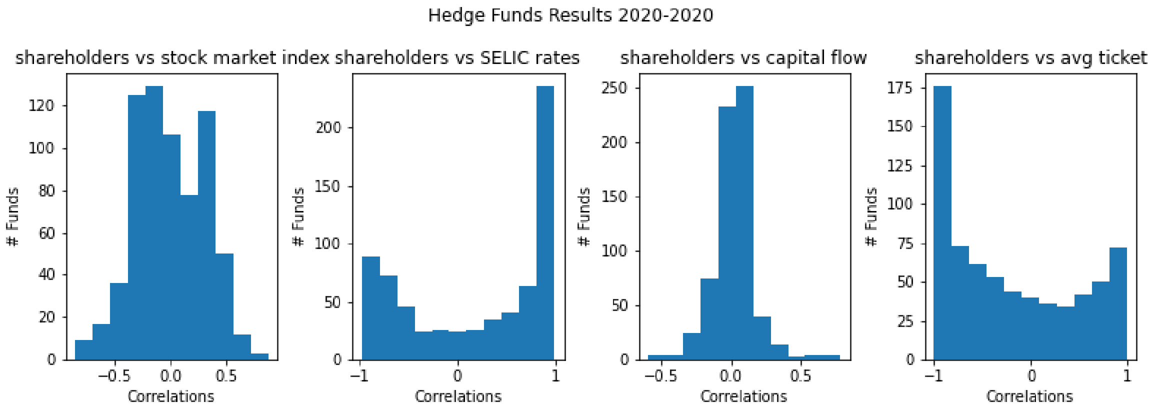


On their turn, the yearly analysis for hedge funds can be seen below:





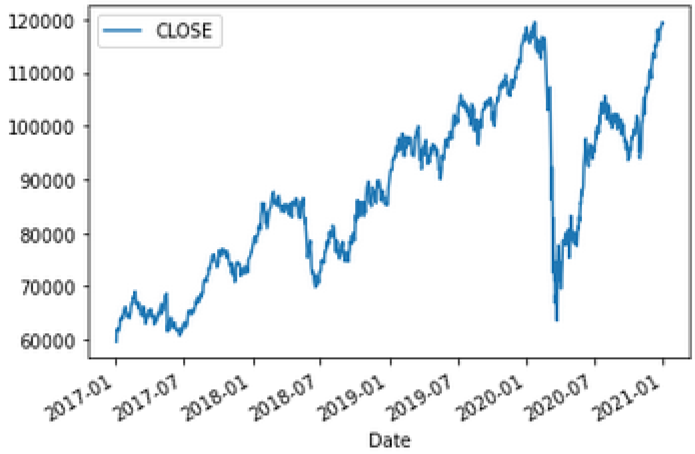




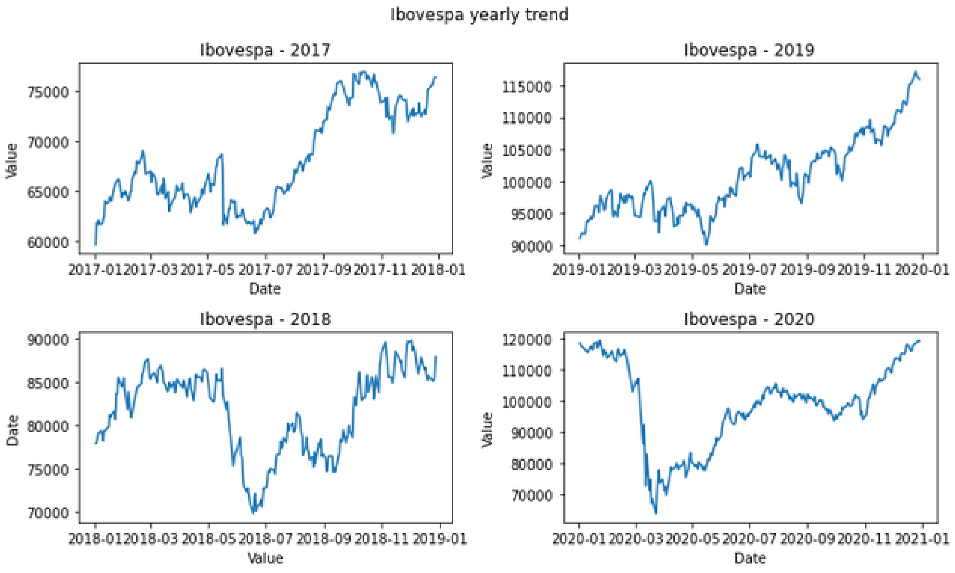
One of the interesting observations that can be done about hedge funds on the context of a yearly analysis is that, unlike fixed income funds, there is a consistent trend of negative correlations between shareholders and the average ticket on a fund. As this indicates, for instance, that individual investors may be especially prone to invest on hedge funds, as their tickets tend to be smaller than institutional investors (that is, other funds, corporations and the like).

Another phenomenon observed in all funds is that the histogram of the correlations between the number of shareholders and the stock market index seems to follow a cyclical pattern. In 2018 and 2020, correlations tend to stay close to zero (indicating a irrelevant correlation between the two variables on the period), while on 2017 and 2019 correlations are displaced to both extremes, showing sets of funds that have either a positive or negative correlation.

A possible explanation could involve the observation of the stock market behavior as a proxy of the investors' expectations at a given time:



By dividing the graph above in periods corresponding to each year, correlations between the number of shareholders and the stock market index tend to vary with the IBOVESPA yearly performance. The graph below shows the variation of IBOVESPA within each selected year.

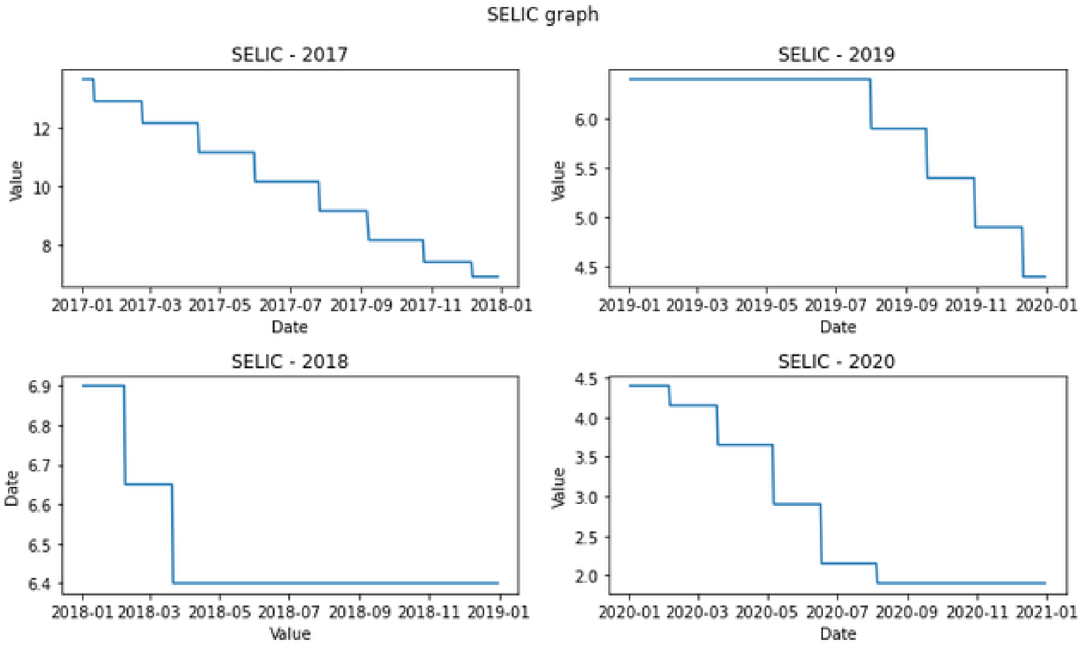


Major events occurred on the stock market (in the sense of sharp, short-spanned movements) in each one of the years, with 2019 being an exception. On May 2017, a recording that composed the testimony of Joesley Batista (owner of JBS, a company investigated on corruption schemes) to “Operação Lava Jato”, was made public. As Brazil had impeached a president on the previous year and the current president was involved on Joesley’s testimony, uncertainty regarding Brazil’s political stability struck the stock market, driving it down by 12% over an trading hour [9].

On May 2018, a nationwide strike of truck drivers [10] sparked uncertainty on the country’s economic perspectives. This strike led to a temporary spike on inflation and a short-term loss on the IBOVESPA Index. Finally, on March 2020, a worldwide crisis caused by the impacts of Covid-19 on the global economy plummeted stock markets around the world [11], and IBOVESPA fell by almost 40% over the course of a month.

If the presence of such events had a consistent interference on the correlation between the number of shareholders and the stock market index, the similarity of the histograms from 2017 and 2019 should not be observed. One could argue, however, that as most of those events were strongly linked to the political scenario at the time, an analysis focusing on the interference of the political scenario and the variation of shareholders on investment funds could yield more relevant results.

In terms of SELIC rates, a similar graph can be plotted:



A possible limitation of this study is that, during the studied time span, the base rate has consistently fallen (as of today, the base rate has already been increased and it is expected that it will raise more on the next Brazilian Central Bank meetings). On the context of hedge funds, one could affirm that during the fall of the base rate, some funds gained shareholders and some others lost shareholders, but per se this does not indicate that such a correlation pattern would be kept if the base rate increased steadily over time.

A noteworthy fact is that, along all intervals and funds, there does not seem to be a relevant correlation between the number of shareholders and the capital flow on the funds. Such effect may be caused by the diversity of clients (in terms of the subscription amount) and the consequent variation on the average ticket.

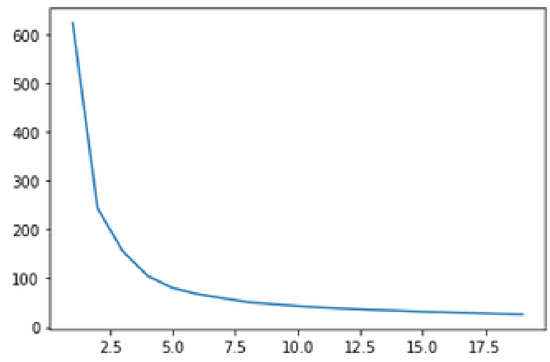
K-means algorithm and wordclouds

As an attempt to use the k-means algorithm to derive similarities between the behavior of given funds, I will test if it is able to clusterize the set of funds (based on their correlations between the shareholder count and the chosen metrics) by checking if there are predominant terms in wordclouds made of the fund's names. Depending on the assets that they hold and, on the strategy adopted by the fund manager, funds are either required to adopt certain words on their names, or they adopt the words as a way to denote their strategies. Some common examples can be seen below (for each kind of fund):

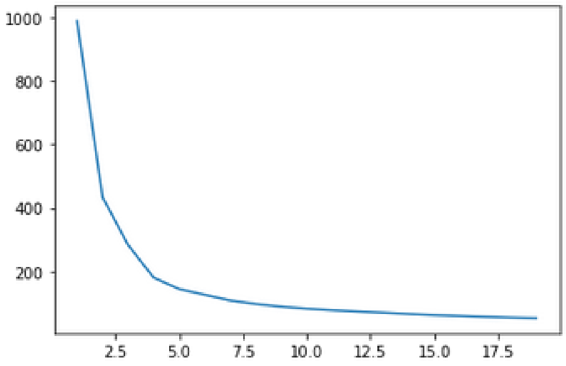
- Hedge Funds
    - Long Bias/Long Only – denote funds whose stock portfolio is either biased towards long positions or accepts only long positions.
    - Macro – denote funds whose investment strategy focus on macroeconomic theories, using sovereign bonds and currencies to form their portfolios.
    - IE (*Investimento no Exterior*/Overseas Investment in English) – used by funds that have exposition, to some degree, to assets located outside Brazil.
- Fixed Income Funds
    - Referenciado/Referenciado DI – "Referenciado" stands for referenced in portuguese. This means that the investment fund's benchmark follows a given index. On the case of Referenciado DI, the CDI (*Certificado de Depósito Interbancário)* Rate is used as a benchmark. This rate is the interest rate adopted between banks, usually staying close to the SELIC rate.
    - Crédito Privado – Crédito privado (private credit, in a free translation) encompasses fixed income assets that are issued by private entities, such as commercial papers and debentures. Funds that contain either "crédito privado" or "CP" (its abbreviation) can have a fraction of their net worths on this kind of asset.
    - Longo prazo/Curto prazo – Stands for "long-term/short-term" in English. In the context of Brazilian investment funds, being a long term or a short-term fund refers to the average time to expiration of its portfolio as a definition of the fund management policy. If the average time to expiration of its assets must be less than one year, then it is classified as a short-term fund. Otherwise, it is considered a long-term fund.

In order to determine an optimal number of clusters for each type of fund, a WCSS (Within-Cluster Sum of Squares – the sum of the squares of the distances between each point and its cluster centroid) graph was plotted for each category, considering one to twenty clusters. An approximation of the "elbow" point was considered the optimal number of clusters.

The graph below shows the WCSS values observed as a function of the respective cluster count for fixed income funds. After all, a total of four clusters were used to this category.



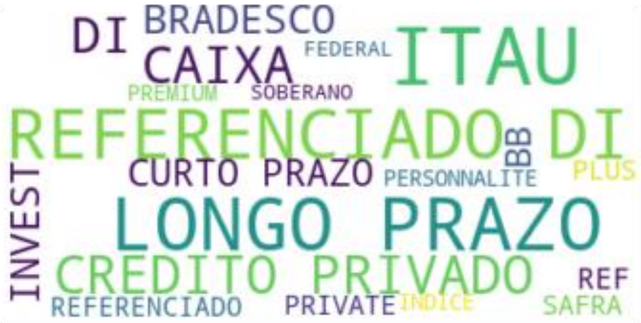
On its turn, hedge funds were assessed in the same way, and a total of three clusters was used:



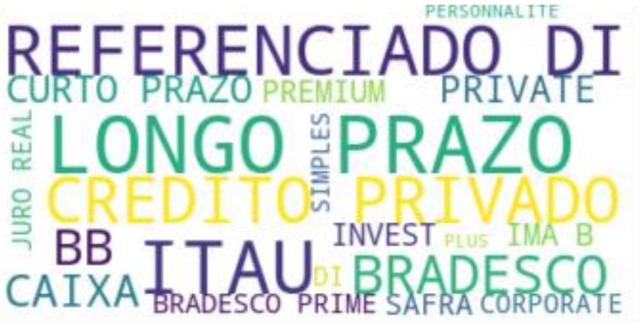
A plot of each wordcloud set can be seen on the following pages, followed by a brief discussion about the observed results.

Fixed Income funds

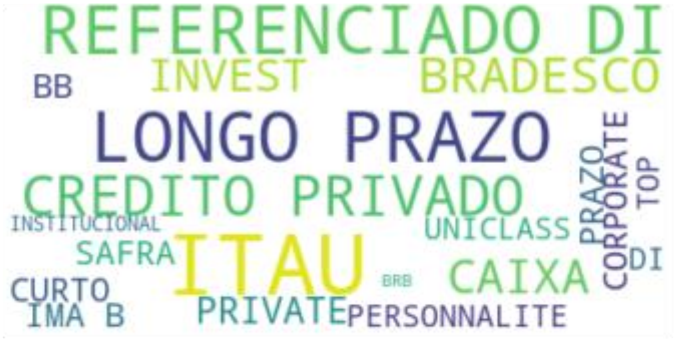
Fixed Income wordcloud - Cluster #1



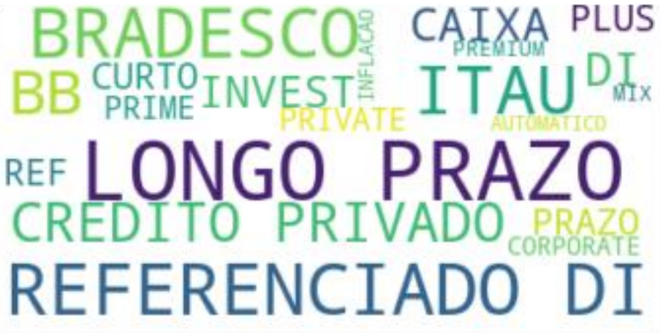
Fixed Income wordcloud - Cluster #2



Fixed Income wordcloud - Cluster #3



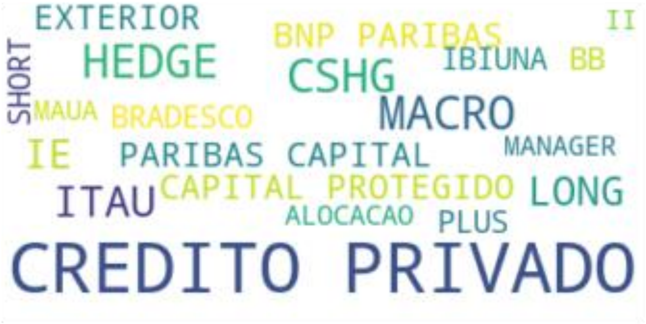
Fixed Income wordcloud - Cluster #4



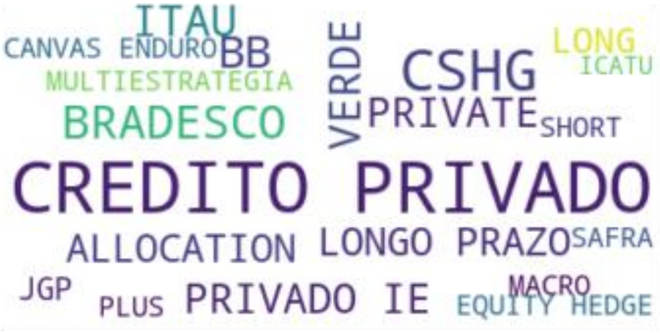
At a first glance, the clusters seem similar. The words “REFERENCIADO DI”, “LONGO PRAZO” and “CREDITO PRIVADO” are present in all of four clusters. Itau, Caixa, Safra, Bradesco and BB are the names of major financial institutions in Brazil. As they are also present among all four clusters, the most likely conclusion is that this method is not able to provide meaningful insights about the funds contained on each cluster.

A similar phenomenon can be observed on the clusters below, derived from the hedge fund dataset. The word “Macro” appears on cluster number one, along with “IE” and “Exterior”, indicating that “macro” funds behave like funds that invest overseas. This relation makes sense to some extent, as “macro” funds may invest in foreign currency as a strategy, showing therefore some influence from the FX market. All the four clusters show the words “Crédito Privado”, indicating that several hedge funds may use debt assets issued by private entities on their strategies (mixing debt assets with stocks, for instance). Understanding this phenomenon would require a deeper analysis.

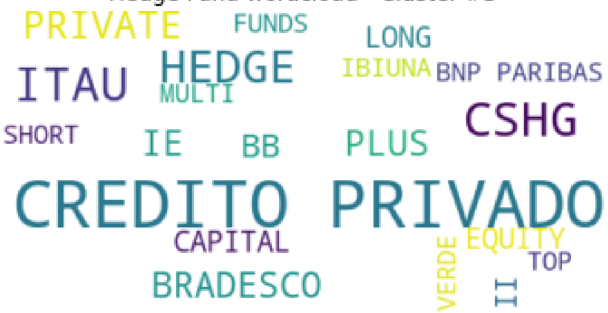
Hedge Fund wordcloud - Cluster #1



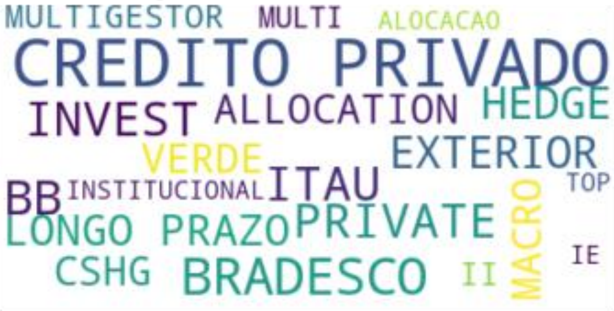
Hedge Fund wordcloud - Cluster #2



Hedge Fund wordcloud - Cluster #3



Hedge Fund wordcloud - Cluster #4



Conclusion

Some insights could be obtained with the analysis of correlations, both on the whole period and on a yearly basis. However, clusterization did not return results that seem strong enough to be considered as meaningful, given that certain words are present along all observed clusters.

References

[1] - https://www.anbima.com.br/pt_br/informar/relatorios/fundos-de-investimento/boletim-de-fundos-de-investimentos/classe-multimercados-registra-recorde-em-entrada-liquida-de-recursos-no-ano.htm

[2] - https://www.anbima.com.br/pt_br/especial/raio-x-do-investidor-2020.htm#Onde

[3] - http://dados.cvm.gov.br/dataset/fi-cad/resource/1baccbb6-cd82-49f6-b70f-5a7d5ad7d616

[4] - http://dados.cvm.gov.br/dataset/fi-doc-inf_diario

[5] - https://www.anbima.com.br/data/files/04/35/26/3E/E9E057106A070057882BA2A8/IHFA%20-%204T20.xlsx

[6] - https://pypi.org/project/yfinance/

[7] - http://api.bcb.gov.br/dados/serie/bcdata.sgs.4189/dados?formato=json&dataInicial=01/01/2017&dataFinal=31/12/2020

[8] - http://dados.cvm.gov.br/dataset/fi-doc-cda

[9] - https://cointimes.com.br/joesley-day-um-dos-dias-mais-loucos-da-historia-da-bovespa/

[10] - https://g1.globo.com/economia/noticia/bovespa-28-05-2018.ghtml

[11] - https://www.theguardian.com/business/live/2020/mar/12/stock-markets-tumble-trump-europe-travel-ban-ecb-christine-lagarde-business-live