Finding peer stocks using business description of a company

Rahul, PhD.

June 30, 2022

Introduction

The report outlines an alternative method to find stock peers through natural language processing of business descriptions of a company. The company value chain is highly connected in the current globalized world. The specialized patented products, commodity availability, cost, and similar factors contributing to value-chain connectivity. Therefore, looking into same sector peers won't give the whole picture. The report describes the novel method to find peers by applying text clustering on the business description. The method can be used as a complementary method to identify stock peers.
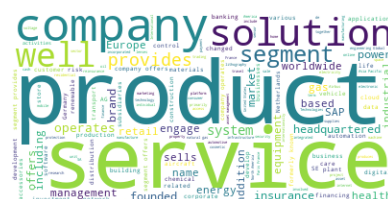
The report is organized in three sections. The methods section describe the data source, algorithm, and technology used for the analysis. The results section outline the results of the analysis. Finally, conclusion provides the discussion and future work.

Method

Since this is a methodology paper to keep the data size and computation time manageable I used EuroStoxx 50 [1] as a universe. Furthermore, the data for this universe is available for all stocks on the Yahoo finance [2]. So, I pulled the data from Yahoo finance using yfinance library [3]. The fields symbol, sector, industry, and longBusinessSummary are used in the analysis. The *longBusinessSummary* is used as input text.

Next, term frequency-inverse document frequency (TF-IDF) [4] transformation is used to create features. I decided to choose Hierarchical cluster as a preferred method for the clustering, since the algorithm does not require initial number of cluster as an input parameter. Also, I use l1 distance measure, since data is very sparse.

Apart from removing standard stop-words I also removed stop words like *co, ltd, llc*. Next, I demonstrate the effect of customized data cleaning on the business descriptions through word cloud. The figure 1 shows the word frequency of business description without customized data cleaning. The figure 2 shows the effect of customized stop-words removal.

Figure 1: Wordcloud comparison showing effect of customized data cleaning

Figure 2: Wordcloud comparison showing effect of customized data cleaning

[1] Euro stoxx 50. https://en.wikipedia.org/wiki/EURO_STOXX_50, 2022. Accessed: 2022-03-04

[2] Yahoo finance. https://finance.yahoo.com/, 2022. Accessed: 2022-03-04

[3] Ran Aroussi. Download market data from yahoo! finance's api. https://aroussi.com/post/python-yahoo-finance, 2022. Accessed: 2022-03-04

[4]

For the technology I used python data science stack [5]. Besides, I used [6] and Snakemake [7] for workflow management. The source code is available on Github `https://github.com/r2rahul/thematic`.

Results

In the section I present two results from the analysis. First, I show the hierarchical cluster dendrogram 3. Given, the text is very sparse, still the cluster is able to capture sector level homogeneity. If I cut the tree around 30, I get 10 clusters, which is close to 11 sector classification available in the universe. In addition, the distribution of sector is also skewed in the universe with Financial Services and consumer cyclical are most represented, while Real Estate is least represented.

Second, the model is quite homogeneous at the terminal nodes with one or two cross-sector symbols in the group. For example, in red cluster ABI.BR and RI.PA belong to same sector, whereas CS.PA and RI.PA belong to different sectors. Figure 4 compares the word cloud of the cluster group. Therefore, demonstrating that it is feasible to find stock peers through natural language processing.

Conclusion

In this report I described the methodology of finding peers through business description and natural language processing. I demonstrated the methodology by mining the public information through Yahoo finance [8] and using hierarchical cluster to identify clusters based on business descriptions. The method is able to find homogeneous cluster with non-intuitive peers based on the business words.

The method provides two important knobs to tweak, when working on actual data provided by commercial data vendors and texts from annual reports. First, input data the description can be more detailed as mentioned in the annual reports. The large corpus will need more careful feature engineering, like log transformation to represent all words meaningfully. Second, hyper-parameter tuning of the cluster algorithm. Since, the algorithm is very sensitive to choice of distance and linkage method, a more detailed training will be required. The report outlined a complimentary method for finding non-intuitive peer stocks using NLP and method will be helpful in the portfolio management.

[5] Natural language toolkit. `https://www.nltk.org/`, 2022. Accessed: 2022-03-04

[6] An open source platform for the machine learning lifecycle. `https://www.mlflow.org/`, 2022. Accessed: 2022-03-04

[7] F Molder, KP Jablonski, B Letcher, MB Hall, CH Tomkins-Tinch, V Sochat, J Forster, S Lee, SO Twardziok, A Kanitz, A Wilm, M Holtgrewe, S Rahmann, S Nahnsen, and J Koster. Sustainable data analysis with snakemake [version 1; peer review: 1 approved, 1 approved with reservations]. *F1000Research*, 10(33), 2021. DOI: 10.12688/f1000research.29032.1

Figure 3: The figure shows the final cluster groups based on the business description

Figure 4: The wordcloud showing comparison of words in cluster in red

[8] Yahoo finance. `https://finance.yahoo.com/`, 2022. Accessed: 2022-03-04

References

An open source platform for the machine learning lifecycle. `https://www.mlflow.org/`, 2022. Accessed: 2022-03-04.

Natural language toolkit. `https://www.nltk.org/`, 2022. Accessed: 2022-03-04.

Euro stoxx 50. `https://en.wikipedia.org/wiki/EURO_STOXX_50`, 2022. Accessed: 2022-03-04.

Term frequency–inverse document frequency. `https://en.wikipedia.org/wiki/Tf%E2%80%93idf`, 2022. Accessed: 2022-03-04.

Yahoo finance. `https://finance.yahoo.com/`, 2022. Accessed: 2022-03-04.

Ran Aroussi. Download market data from yahoo! finance's api. `https://aroussi.com/post/python-yahoo-finance`, 2022. Accessed: 2022-03-04.

F Molder, KP Jablonski, B Letcher, MB Hall, CH Tomkins-Tinch, V Sochat, J Forster, S Lee, SO Twardziok, A Kanitz, A Wilm, M Holtgrewe, S Rahmann, S Nahnsen, and J Koster. Sustainable data analysis with snakemake [version 1; peer review: 1 approved, 1 approved with reservations]. *F1000Research*, 10(33), 2021. DOI: 10.12688/f1000research.29032.1.