Finding peer stocks using business description of a company

The article demonstrates a method for finding stock peers by natural language processing (NLP) on the business descriptions of a company.

AUTHOR

Rahul

AFFILIATION

Quantitative Analyst

PUBLISHED

June 30, 2022

CITATION

Rahul, 2022

Contents

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.

Methods

Since this is a methodology paper to keep the data size and computation time manageable I used EuroStoxx 50 ("EURO STOXX 50" 2022) as a universe. Furthermore, the data for this universe is available for

all stocks on the Yahoo finance ("Yahoo Finance" 2022). So, I pulled the data from Yahoo finance using yfinance library (Aroussi 2022). 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) ("Term Frequency–Inverse Document Frequency" 2022) 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.

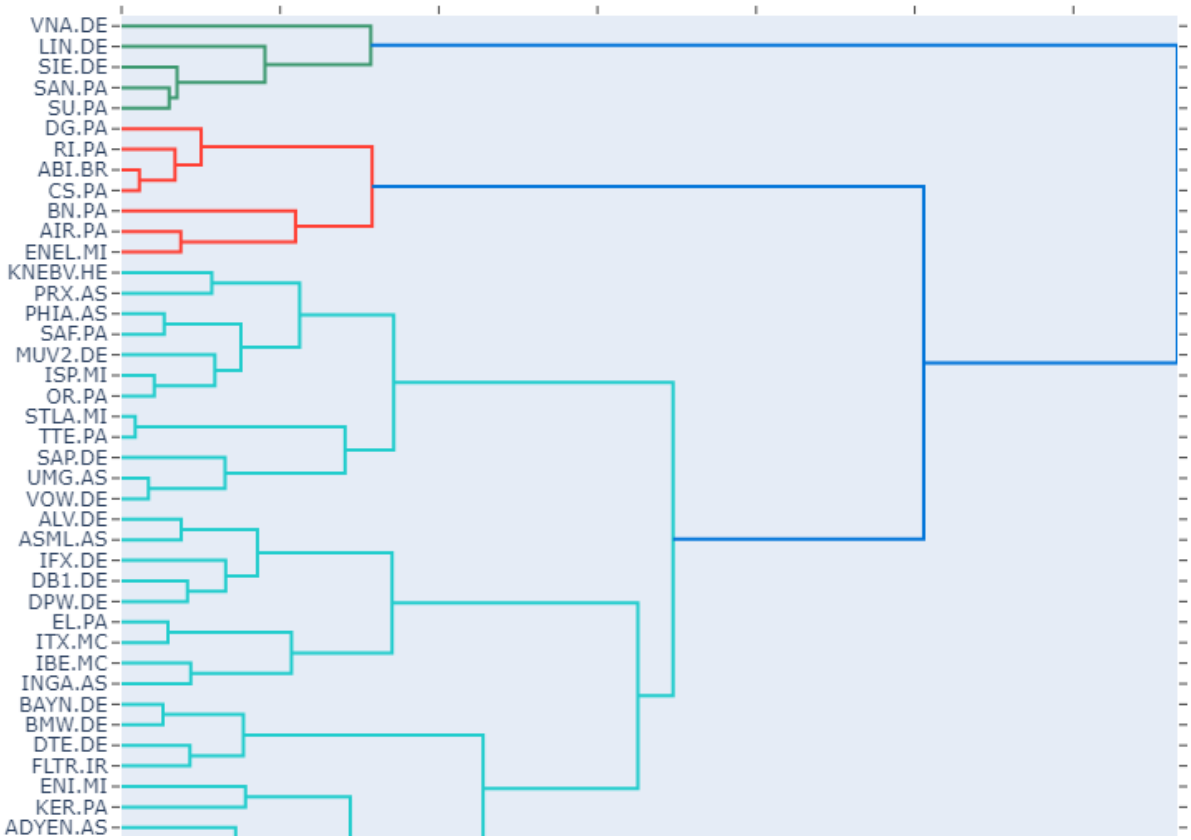
Apart from removing standard stop-words I also removed stop words like *co, ltd, llc*. Next, I demonstrate the effect of customized data cleaning on the business descriptions through word cloud. The panel A in figure 1 shows the word frequency of business description without customized data cleaning.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

Figure 1: Wordcloud comparison showing effect of customized data cleaning

For the technology I used python data science stack "Natural Language Toolkit" (2022). Besides, I used MLflow ("An Open Source Platform for the Machine Learning Lifecycle" 2022) and Snakemake (Molder et al. 2021) 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. 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.


VNA.DE
LIN.DE
SIE.DE
SAN.PA
SU.PA
DG.PA
RI.PA
ABI.BR
CS.PA
BN.PA
AIR.PA
ENEL.MI
KNEBV.HE
PRX.AS
PHIA.AS
SAF.PA
MUV2.DE
ISP.MI
OR.PA
STLA.MI
TTE.PA
SAP.DE
UMG.AS
VOW.DE
ALV.DE
ASML.AS
IFX.DE
DB1.DE
DPW.DE
EL.PA
ITX.MC
IBE.MC
INGA.AS
BAYN.DE
BMW.DE
DTE.DE
FLTR.IR
ENI.MI
KER.PA
ADYEN.AS

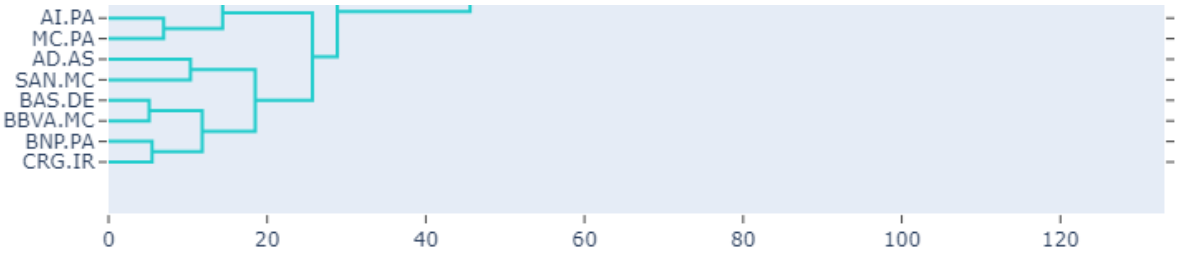


Figure 2: The figure shows the final cluster groups based on the business description

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 3 compares the word cloud of the cluster group. Therefore, demonstrating that it is feasible to find stock peers through natural language processing.

Figure 3: The wordcloud showing comparison of words in cluster in red

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 ("Yahoo Finance" 2022) 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

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.

References

"An Open Source Platform for the Machine Learning Lifecycle." 2022. https://www.mlflow.org/.

Aroussi, Ran. 2022. "Download Market Data from Yahoo! Finance's API." https://aroussi.com/post/python-yahoo-finance.

"Clustering Text Documents Using k-Means." 2022. https://scikit-learn.org/stable/auto_examples/text/plot_document_clustering.html.

"EURO STOXX 50." 2022. https://en.wikipedia.org/wiki/EURO_STOXX_50.

Molder, F, KP Jablonski, B Letcher, MB Hall, CH Tomkins-Tinch, V Sochat, J Forster, et al. 2021. "Sustainable Data Analysis with Snakemake [Version 1; Peer Review: 1 Approved, 1 Approved with Reservations]." *F1000Research* 10 (33). https://doi.org/10.12688/f1000research.29032.1.

"Natural Language Toolkit." 2022. https://www.nltk.org/.

"Term Frequency–Inverse Document Frequency." 2022. https://en.wikipedia.org/wiki/Tf%E2%80%93idf.

"TextBlob: Simplified Text Processing." 2022. https://textblob.readthedocs.io/en/dev/.

"Yahoo Finance." 2022. https://finance.yahoo.com/.

Corrections

If you see mistakes or want to suggest changes, please create an issue on the source repository.

Reuse

Citation

For attribution, please cite this work as

```
Rahul (2022, June 30). Finding peer stocks using business description of a company. Retrieved
from https://rpubs.com/r2rahul/874302
```

BibTeX citation

```
@misc{rahul2022,
  author = {Rahul, },
```

```
  title = {Finding peer stocks using business description of a company},
  url = {https://rpubs.com/r2rahul/874302},
  year = {2022}
}
```