Tracking Companies through Time: Wikipedia-Based Data Harvesting for Financial Analysis

Keywords: Web Crawling, Wikipedia, Historical Company Data, Natural Language Processing

Project description

Wikipedia serves as a rich repository of information on companies, encapsulating both current and historical data. Given that these pages are continually updated, Wikipedia offers an invaluable source for temporal insights into companies.

In this project, the student should implement a web crawler to collect company information from Wikipedia at the beginning of each year. The crawler will not only gather current data but also locate URLs for historical versions of Wikipedia pages to collect retrospective company information.

To do so, the student will need to map Wikipedia URLs to the Refinitiv stock market database, implementing a crawler that locates URLs directing to the historical versions of a Wikipedia page and construct another crawler that downloads the historical company information at the start of each calendar year.

The ultimate goal is to leverage this dataset to create a corpus of short historical business descriptions. Summarization or question-answering tools will be employed to extract the most salient information from these Wikipedia pages. This summary data will then be used to determine the similarity of companies over time and correlate these findings with existing company similarity networks (e.g. Breitung and Müller, 2023).

To accomplish these objectives, the student should have a robust background in programming, particularly in Python. Knowledge in web scraping techniques is essential, as is a foundational understanding of natural language processing tools for summarization and question-answering functionalities.

Overall, this project aims to contribute significantly to the field of finance and business analysis by offering a novel method to capture and analyze temporal changes in company descriptions and their correlations with financial markets.
What we are looking for

- Strong analytical and project management skills
- Determination and passion for your areas of expertise
- Good Python programming skills
- Interest to work at the intersection of finance and IT
- 1 or 2 persons

What we offer

- Knowledge in quantitative finance, corporate finance, machine learning and textual analysis
- Kick-off session including introduction to relevant finance and/or business topics
- Experience with IDPs
- Open dialogue and support
- Access to prime capital markets databases (Bloomberg, Datastream, Thomson Reuters, etc)
- Potential for publication and/or evaluation of future use cases
- Both single and group projects are possible

Interested?

Please send an e-mail with CV, academic transcript and your preference for this project to christian.breitung@tum.de

Questions?

In case of any (e.g. topic related) questions, please contact Christian Breitung (christian.breitung@tum.de)

References