Uncovering Funding Pathways by Leveraging Academic Data

Vishal K. Panda, MSDS; Nithya Mylakumar, MSDS; Parth Maheshwari MSDS; Kuldeep Singh, MSDS
Michigan State University

Abstract

We conducted a thorough analysis of **Refinitive Workspace** financial dataset and the **S2ORC** academic publications dataset. Utilizing advanced NLP techniques (**BertTopic** and **Large Language Models**), we performed precise **topic modeling**. Additionally, we developed a cross-domain ontology using semantic representation of topics using **GPT-4**.
Leveraging topic modelling, concept identification and ontology, we connected financial and academic datasets, uncovering patterns linking academic research to financial institution funding. This research sheds light on the synergy between academia and finance, demonstrating the power of **NLP techniques and cross-domain ontologies to reveal hidden connections**. Our findings inform future investments and research initiatives in the financial domain.

Introduction

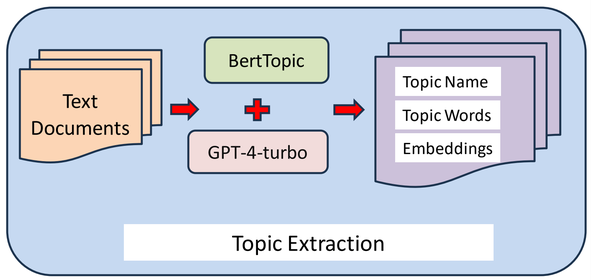
In an era marked by the rapid evolution of technology and finance, the interplay between academic research and financial trends has become increasingly pivotal. This poster presentation delves into a data-driven approach aimed at uncovering the intricate relationships between these two domains. To tackle this multifaceted challenge, we adopted a carefully curated methodology designed to extract meaningful insights from vast and diverse datasets.

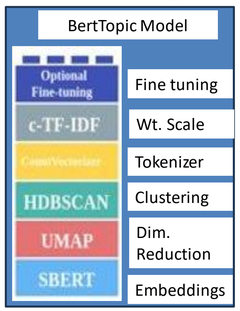
Methodology

We restricted the academic data to **computer science** publications and the financial data to **technology companies**. Next, standard text processing techniques were applied to both datasets, including cleansing and formatting company descriptions, publication titles, and abstracts. This crucial preprocessing allowed us to do subsequent analyses.

For **topic extraction** and categorization within each domain, we deployed **BertTopic** which was used for topic extraction. **GPT-4-turbo** was employed for labelling topics , adding interpretability to extracted topics. Additionally, We utilized OpenAI's **text-embedding-ada-002** model to construct semantic representations of these topics

Using semantic topic representations, we established a connection between academic research and financial trends, using cosine similarity and identified heuristics. This allowed us to examine and compare trends within the financial domain relative to their alignment with academic research.





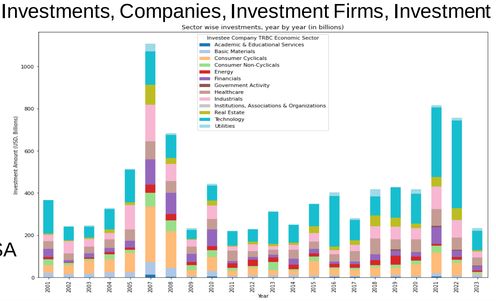
Cross Domain Ontology

To create a mapping between different domains using embeddings, we first represent topics or concepts within each domain as embeddings. These embeddings encapsulate the semantic meaning of each topic. Next, we calculate similarity scores between these embeddings, enabling us to measure the relatedness of topics across domains. By identifying high similarity pairs, we establish valuable cross-domain connections, facilitating the exploration of relationships and insights between disparate knowledge areas.

Finance Data Analysis

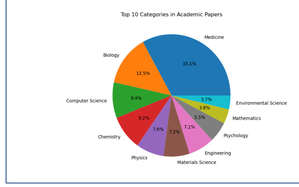
We acquired the financial dataset from the Refinitiv Workspace. The dataset was divided into 4 tables, namely – Investments, Companies, Investment Firms, Investment Funds.

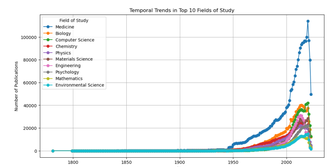
2000-2023 data statistics -
Countries: 174
Companies: 70027 in USA
181733 total
Investments: 408558 in USA
786905 total
Total Investment: 18T in USA
33.3T total
Tech companies – 14282 in USA
30657 total



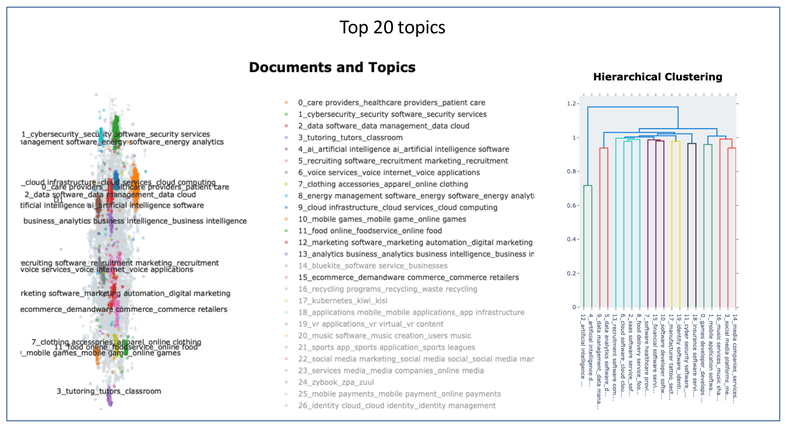
Academic Data Analysis

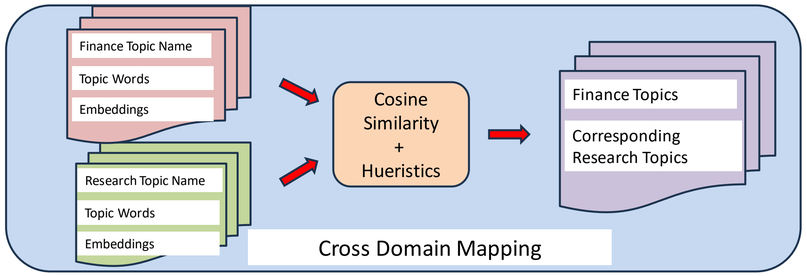
Data has been extracted from the Semantic Scholar API, covering a diverse range of academic materials including papers, abstracts, author details, and TLDRs (short summaries), each segmented into separate datasets for comprehensive analysis.



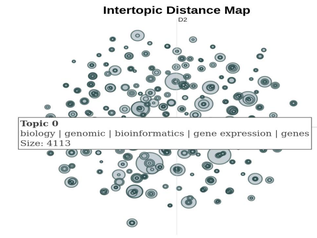


Financial Topic Modelling

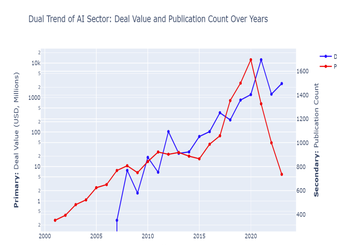


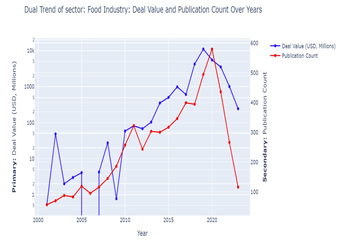


Academic Topic Modelling



Results





The first chart depicts a two-decade trend analysis of the Artificial Intelligence sector, highlighting the interplay between financial influx, measured in deal values, and academic vigor, represented by publication counts. The concurrent rise in both metrics suggests a robust growth pattern and a synergy between economic investment and research development in AI. Similarly, the second chart illustrates a comparable pattern within the Food Industry sector. Here, we observe a fluctuating trend in finance deals and a relatively steady, positive trend in research publication counts.

Conclusion

We were able to successfully map two very distinct datasets (Financial and Academic) using Topic Modelling and Similarity based methods. This opens up a wide range of applications and research areas such as -

1. Predicting Private Equity Fund Flows using Research Activity
2. Identifying underfunded research areas and countries
3. Portfolio Optimization & Risk Management (Overvaluation, Research saturation)
4. Stock Market prediction (on integration with stock market dataset)

We are currently working on a dashboard that could automatically integrate different data sources, generate topics, and provides user with the aforementioned functionalities. Moreover, we are working on integrating more well-defined ontologies such as OpenAlex Concept Tagger, such that we can perform many-to-many joins.

References

1. BERTopic: Neural topic modeling with a class-based TF-IDF procedure (https://arxiv.org/abs/2203.05794)
2. S2ORC: The Semantic Scholar Open Research Corpus (https://aclanthology.org/2020.acl-main.447/)
3. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts (https://arxiv.org/abs/2205.01833)
4. GPT-4 Technical Report (https://arxiv.org/abs/2303.08774)
5. A Web-scale system for scientific knowledge exploration (https://arxiv.org/pdf/1805.12216.pdf)
6. Building a Concept Hierarchy from a Distance Matrix (https://link.springer.com/chapter/10.1007/3-540-32392-9_10)