Identifying Company Similarity using NLP and Social Network Analysis

Data Science Institute
COLUMBIA UNIVERSITY

Ruxin Shen, Donghan He, Hantang Zhou, Jialei Wang, Xiaoxin Xu
Industry Mentor: Jared Peterson
Faculty Mentor: Howard Friedman

Data Science Capstone Project with Goldman Sachs

Introduction & Motivation

The project goal is to identify companies that share technological IP and that are investing in similar areas using patent data. The result of the project provides investors with more information on intersections between their investments.

Exploratory Analysis

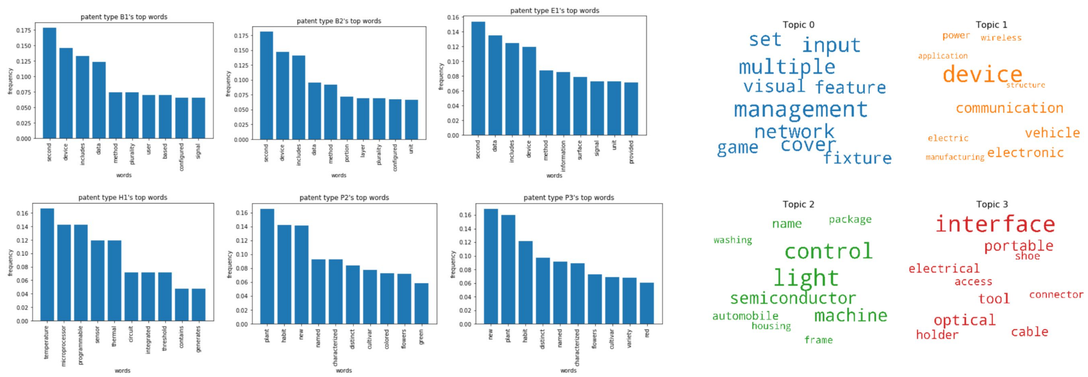

Figure 1. Most Frequent words of all Types of Patents (left); Word Clouds from Topic Modeling (right)

Methods

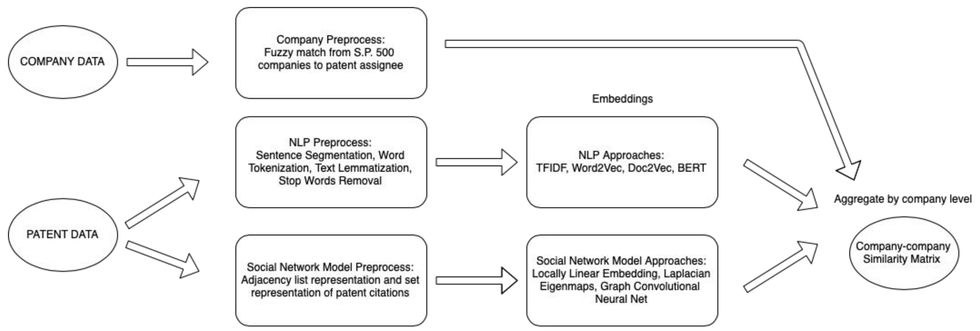

Figure 2. System Design

This project uses two approaches - NLP models and Social Network Analysis models to embed patents into vectors and then aggregates those embeddings into company embeddings. Using these embeddings, we calculated pairwise company similarity matrix through distance comparison.

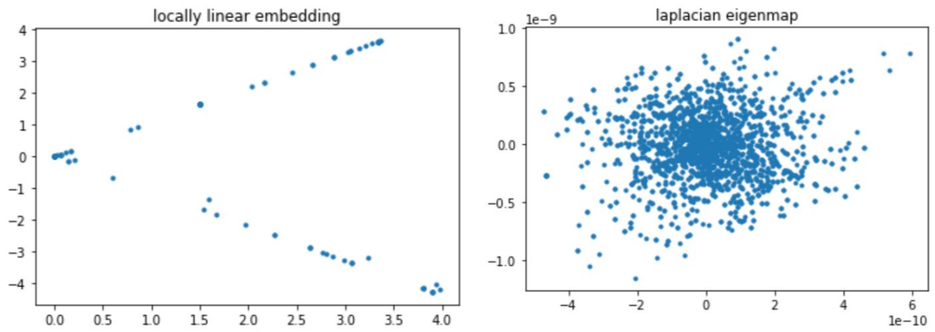

Figure 3. Locally Linear Embedding & Laplacian Eigenmap

Evaluations & Results

1. Manual Evaluation

To validate the result from this unsupervised learning problem, we manually reviewed 40 patents, scored their similarities and compared the results with the models' results. Through human score validation, we showed that our model covers more than 70% of top 30 similar patent pairs within top 100 similar patents.

2. Google Knowledge Graph

Google's API queries "people also search for" results for companies, which provides a way for validation. It also uncovers some company relationships which we did not find using patent data.

Companies Similar to **Adobe**

	Company 1	Company 2	Company 3	Company 4	Company 5
word2vec	Facebook	Microsoft	eBay	Electronic Art	Netflix
doc2vec	Facebook	Salesforce	Microsoft	Clorox	Amazon
Google	DocuSign	Salesforce	Fotolia	Microsoft	Autodesk

Figure 4. Top 5 Similar Companies of Adobe by word2vec, doc2vec, and Google

For example, searching for similar companies to Adobe, Google Search API returns other software companies in design and office. In addition to identifying Salesforce.com as a related company, our models also reveal that Adobe is similar to tech companies like Facebook and Microsoft, eBay and Amazon, which are the cross-industry similarities our project is looking for.

Conclusion & Future Directions

From the results of this project, we find cross-industry connections between companies based on their area of interests for different technologies. This result can provide more insights into company relationships and thus help investors with decision making.

For the scope of this project, we applied the models on patent data after 2018. For future directions, we recommend running those models on a larger data set with more companies and possibly find cross-year relationships.

Acknowledgments

We would like to thank our industry mentor Jared Peterson and our faculty advisor Howard Friedman who offered us this great opportunity to work on this interesting project and provided many insightful guidances on the project.

References

Belkin, Mikhail, and Partha Niyogi. "Laplacian eigenmaps for dimensionality reduction and data representation." *Neural computation* 15.6 (2003): 1373-1396.

Hamilton, Will, Zhitao Ying, and Jure Leskovec. "Inductive representation learning on large graphs." *Advances in Neural Information Processing Systems*. 2017.

Roweis, Sam T., and Lawrence K. Saul. "Nonlinear dimensionality reduction by locally linear embedding." science 290.5500 (2000): 2323-2326.

Abstract

The project’s goal is to identify companies that share technological IP and that are investing and researching in similar areas using patent data from patentsview.org. The final deliverable is a company-company similarity matrix.

This project uses two approaches - NLP models and Social Network Analysis models to generate patent embeddings and then aggregate into company embeddings. We then calculated pairwise company similarities using these embedding as our final deliverable.The patent similarity result is evaluated by human scoring and the company similarity is evaluated by Google Knowledge Graph.

From the results of this project, we find cross-industry connections between companies based on their area of interests for different technologies. This result can provide more insights into company relationships and thus help investors in quantitative investment strategies making.