Pharmaceutical R&D Similarity using FDA Clinical Data

Data Science Institute
COLUMBIA UNIVERSITY

Authors:
Aishwarya Srinivasan(as5431), Brian Allen(ba2542),
Harsheel Singh Soin (hss2148), Xiangzi Meng(xmm2103), Yiwen Zhang(yz3310)

Data Science
Capstone Project

Goldman Sachs

Industry Mentors: Jared Peterson
Faculty Mentors: Zoran Kostic, Smaranda Muresan

Introduction

Given the importance of the research pipeline to a pharmaceutical firm, it is of interest to us to seek to determine – based on clinical trials being conducted – which firms are engaged in similar R&D pipelines. Our project is to develop appropriate evaluation criteria that will allow for the similarity comparison of clinical trials at company level.

Background & Models

We use AACT database, which is an open source relational database that contains details about clinical trials ranging from descriptions to study metadata. We include three methods of building similarity matrices:

- Keyword model: MeSH keyword vectorization
- Text model: NLP and text vectorization
- Graph model: Graphical representations of clinical metadata

Results & Conclusion

After regressing against manually validated scores and learning the weights for ensembling clinical trial similarity matrices generated from different approaches, a composite trial-level similarity matrix per year is generated. These matrices are convoluted to form a company-level matrix which takes into account similarities across all trials conducted for all pairwise combinations of trial sponsors (companies)

The final deliverables are the trial-wise and company-wise similarity matrices. With this, Goldman Sachs can apply investment techniques to capitalize on stock performance of companies that exhibit high similarities. With our additional time, we hope to build a tool beyond their requests to help visualize the stock performance for select companies to aid their analysis.

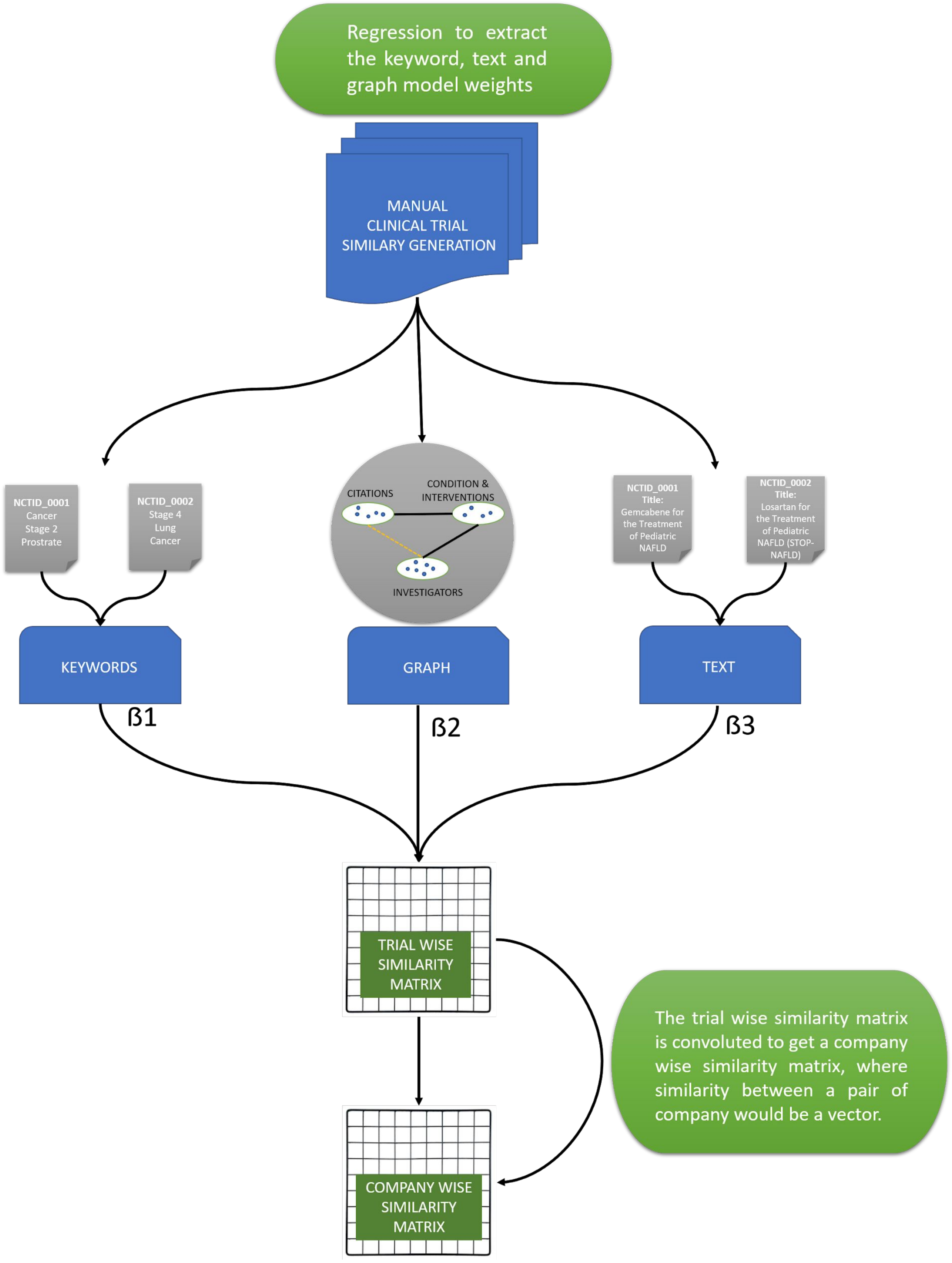
Acknowledgments

Given the unstructured and unsupervised nature of the problem statement, our team had a great learning experience working through diverse methodologies and combining them in a meaningful way towards achieving the end goals. We would also like to thank Goldman Sachs and Columbia University for the opportunity to work on such a challenging objective driven by real-world outcomes.

References

D. H. Wei and T. Campbell, "A similarity measurement of clinical trials using SNOMED — A preliminary study," 2014 International Conference on Collaboration Technologies and Systems (CTS), Minneapolis, MN, 2014, pp. 457-460.

Hao TY, Rusanov A, Boland MR, Weng CH. Clustering clinical trials with similar eligibility criteria features. J Biomed Inform. 2014;52:112–120.



Regression to extract the keyword, text and graph model weights

MANUAL CLINICAL TRIAL SIMILARY GENERATION

NCTID_0001
Cancer
Stage 2
Prostrate
NCTID_0002
Stage 4
Lung
Cancer
CITATIONS
CONDITION & INTERVENTIONS
INVESTIGATORS
NCTID_0001
Title:
Gemcabene for the Treatment of Pediatric NAFLD
NCTID_0002
Title:
Losartan for the Treatment of Pediatric NAFLD (STOP-NAFLD)
KEYWORDS
GRAPH
TEXT
ß1
ß2
ß3
TRIAL WISE SIMILARITY MATRIX
COMPANY WISE SIMILARITY MATRIX

The trial wise similarity matrix is convoluted to get a company wise similarity matrix, where similarity between a pair of company would be a vector.

Figure 1. Methodology - Clinical data to company level similarity matrix