Merck: Natural Language Processing

The Data Mine Corporate Partners Symposium 2023

Introduction

- Bioanalytical “BP” documents submitted by scientists & uploaded to app
- Can be both read and written

Result: Helps forecast future drug development procedures

- Terms extracted and stored into database
- Searches yield BP documents with attribute

Methods

**Researcher provides BP document with chemical information**

Researcher uploads document into the app

The app extracts the key terms identified in ‘blue’ using RegEx

Extracted terms are stored in the Graph Database

**BP-xxxxx (Draft)**

This analytical method is based on an automated 96-well format extraction method of drug from species matrix. MK-xxxxx and stable isotope labeled internal standard (xxx) are chromatographed using chromatography and detected with tandem mass spectrometric detection employing a turbo ionspray (TIS) interface in the polarity ion mode.

**BP-xxxxx (Draft)**

This analytical method is based on an automated 96-well format extraction method of drug from species matrix. MK-xxxxx and stable isotope labeled internal standard (xxx) are chromatographed using chromatography and detected with tandem mass spectrometric detection employing a turbo ionspray (TIS) interface in the polarity ion mode.

Frontend

Fuzzy search allows for accurately returned results regardless of variations

- Original
- Variations
- Misspelling
- Casing
- Other

- BP documents ranked with relevancy
- Checklist-style downloading

Database

- Python Scripts
- CSV Connector
- Neo4j

As the graph database grows, more relationships are established between existing chemicals, which will facilitate future drug development.

**Node Class:**
- Formats attributes as nodes
- Follows graph database relationship structure
- GraphPopulator Class:
- Uses nodes generated from node class to populate database
- Omits “repeated” nodes, creating relationships between BPs

**GraphPopulator Class:**

- Uses nodes generated from node class to populate database
- Omits “repeated” nodes, creating relationships between BPs

Conclusions

We created an app that...

- Extracts data of interest from provided BP documents using RegEx, Bert, and ChatGPT
- Maps extracted data to Neo4j-hosted graph database
- Provides a front-end application through React that has upload, search, and download functionalities

Next Steps

- Incorporate dynamic tables into search results page (arrangeable and filterable)
- Modeling improvements (bettering parsing scripts & Bert model)

Acknowledgements

**Merck:** Xiang Yu, Terri Bui, Kenneth Wilson, Rajesh Desai

**Purdue Data Mine:** Dr. Mark Ward, Maggie Betz, Nicholas Rosenorn, Kali Lacy, Mihir Ahlawat

**Fall Semester Members:** Dante Moon, Yi Zhang, Shah Mahir

Sarah Firestone, Brian Tam, Adhvitha Sivaganes, Arsh Batth, Jagannath Prabhakaran, Cheyenne Ward, Pradyun Kamaraju, Lingyu Li, Varun Manish Karlekar, Jayden Cheung