Name:	José A Alvarado-Guzmán
Affiliation:	Faculty Practice Organization, Columbia Medical Center, Columbia University
Email:	Jaa2220@cumc.columbia.edu
Presentation type (s):	Lightning talk

Biomedical Search Engine

José A. Alvarado-Guzmán, MS, Mohammad Zaryab, MS
Data Science Institute, Columbia University, NY, NY, USA; Electronic Engineering, Columbia University, NY, NY , USA

Abstract

In this paper we propose a search engine that employs several big data concepts to make the Unified Medical Language System dataset accessible to any user. The search engine features three main components: query handling, classification, and visualization. The user can search a medical term with our system to retrieve a classification of the term determined by Mahout Naive Bayes, its definition, and a visualization of neighboring/related medical concepts using a combination of IBM System G's graph storage and a Plotly graph. The programming languages used to make this possible are PHP, HTML, Java, and Python.

Introduction

Due to the rapid advances in biomedical research and related technology, the biomedical corpus is being used at an ever-increasing rate[1]. As a consequence, without the implementation of automated algorithms, it is very difficult for researchers and medical providers to keep up with new biomedical knowledge and development as document classification and semantic role labeling are core challenges[2]. However, training models based on vectors (created from stemmed and/or stopped document word counts) have proven to be a basic and typically successful approach to solving this issue[3].

A corpus can also be provided with additional linguistic information called "annotations." This information can be of a different nature, such as semantic or historical. Natural Language Processing, which is receiving an exponential increase in interest within the biomedical field[1], enriches the biomedical corpus by adding semantic metadata. Annotated corpora constitute a very useful tool for research and for the development of computer systems and algorithms that behave as if they "understand" the language of biomedicine and health. To that end, the National Library of Medicine (NLM) produces and distributes the Unified Medical Language System Knowledge Source (UMLS).

Despite the flexibility and performance of graph databases on high connected data, most developers still prefer to model this data using the relational model due to the learning curve that graph data storage and querying requires. For this reason we design a simple and interactive application that migrates data stored in relational tables into nodes and edges (Neo4j). This migration tool (RD2GD) make it easy and straightforward for a DBA to map tables to the target of the graph database, without having to understand the many intricacies of graph databases, so that they can quickly convert the data and start utilizing the benefits of graph databases. As a test, we will convert a very large dataset in the healthcare industry from MySQL into Neo4j.

SYSTEM OVERVIEW

The dataset that our system employs is the English subset of the Metathesaurus and a subset of the Semantic

Network, both downloaded from the UMLS site. The semantic types included in our system are Body System; Body Part, Organ, or Organ Component; Finding; Laboratory or Test Result; Disease or Syndrome; Laboratory Procedure; Diagnostic Procedure; Therapeutic or Preventive Procedure; Medical Device; Pharmacologic Substance; Biomedical or Dental Material; Biologically Active Substance; Indicator, Reagent, or Diagnostic Aid; Sign or Symptom; Antibiotic; Clinical Drug.

Figure 1 depicts the six components that make up our application. These components are Input Query, Query Handling with PHP, Mahout Classifier (Java Jar File), Data Lookup in MySQL, Visualization with IBM System G & Python and Output Webpage.

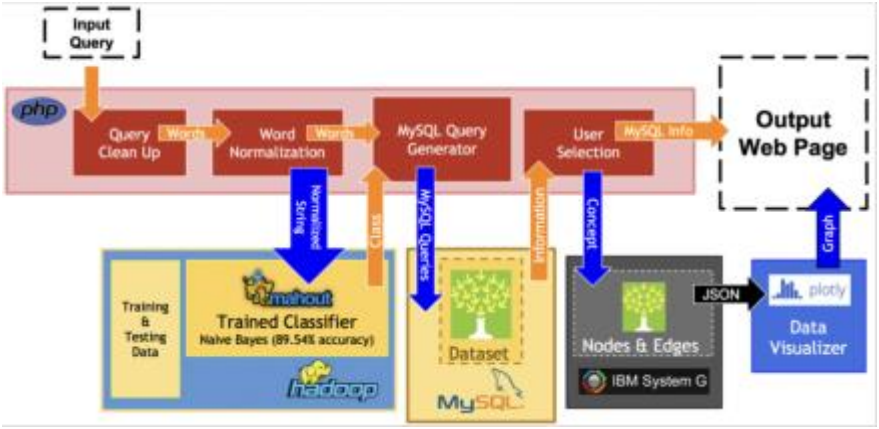

Figure 1. System Block Diagram

Conclusion

The UMLS dataset is a large and dense. Concepts and relationships are encoded; direct searches can take a long time and may present either too much information, or none at all. A search engine's quality is assessed based on its ability to retrieve relevant information for a user. When working with large datasets such as UMLS, the abundance of information is definitely a strength. However, brute force search methods can quickly strip away redeeming qualities. A term or phrase can potentially appear in thousands of unique entries, leaving the user with a data dump to manually search through.

Nevertheless, our search engine has made its data accessible in a user friendly manner. It consists of three solid components: classification, efficient data lookup, and visualization. Our classifier has a testing accuracy of 89.54% and allows queries to be slimmed to a subset of the UMLS data. Data lookup continuously yet efficiently searches for combinations of the user's query to find the closest matching concepts pertaining to the user. The first two components aim to output appropriate information even if the user's query is not found in the UMLS dataset. The last component takes the output and produces a graph depicting where the query exists within the data and its neighboring concepts.

References

1. K. Verspoor, K. Bretonnel Cohen, B. Goertzel, and I. Mani, "BioNLP'06 Linking Natural Language Processing and Biology: Towards Deeper Biological Literature Analysis", Proc. North
2. America Association for Computational Linguistic 2016 Annual Conference (NAACL HLT 06), Omnipress Inc., June 8 2016, New York City, USA.
3. J. Fan and C. Friedman, "Semantic Classification of Biomedical Concepts Using Distributional Similarity", Journal of the American Medical Informatics Association, vol. 14 num. 4 Jul. 2007, pp 467-477.
4. C. G. Chute, Y. Yang, A. Evans, "Latent Semantic Indexing of Medical Diagnosis Using UMLS Semantic Structures", Section of Medical Information Resources, Mayo Clinic,Rochest.