Data Analytics Research Project

Tarun Giduturi

This is the final documentation paper demonstrating my work on the Data Analytics Research Project under Professor Yuan An.

Yuan An

Table of Contents

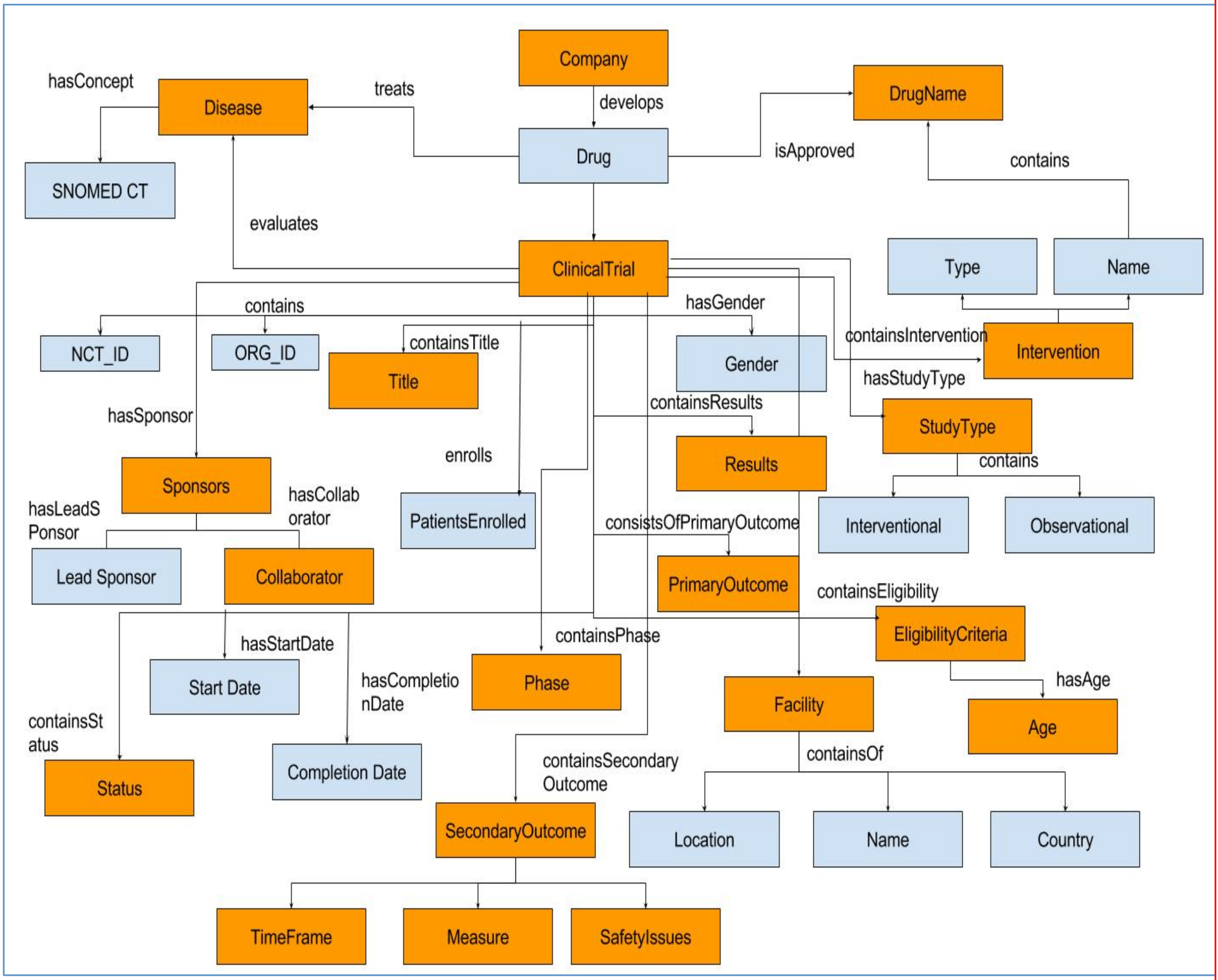
Introduction:

Data Analytics is emerging field that has many uses in the healthcare industry. In the current era, Data Analytics has display its substantial uses and the potential it has in solving complex healthcare industry problems. Despite its huge advantages, there are still a few areas within the healthcare industry, where data analytics have not been applied to its fullest potential. One of such area, where very less focus of data analytics was applied, is the Drug Discovery Area. With the rapid changes in the healthcare treatments in the new era, we need to meet the demand of creating more effective drugs that would be in market within a short time. However, on average, it takes many years for a drug to be created and sold in marketplace. Most of this time is usually taken by clinical trials, as the clinical trial procedures are very long and tidy processes. In addition, if a drug is prepared and goes out for testing it would cost an estimated cost of $2.6 billion, which also includes the failure costs [1]. Therefore, there are many chances that a pharmaceutical company might spend billions of dollars on a drug, and might not get back any money, if the drug fails the clinical trials. So, this would create a huge loss for the pharmaceutical companies, as they are spending loads of money on useless drugs. Thus, my professor Yuan An, has proposed plan of using data analytics on drug discovery data so that we can get a clear understanding of the problems existing within the drug discovery process and devise a knowledge graph that would help us to better understand the data and create an algorithm that would be able to predict accurately whether a drug would be successful at clinical trails testing or not.

In this paper, I would start with the drug discovery study ontology. Then describe how unstructured data is extracted from Clinical Trials database and how that data is refined and converted into a structured data, which could be helpful for understanding drug discovery process and also discuss the future work of this research at the very end.

Drug Discovery Study Ontology:

Ontologies usually depict the nature of the being. So, in this case, the Drug Discovery Ontology depicts all the components that Clinical Trials use to evaluate a drug, and all the fields they measure the most. Understanding the ontology would provide us with a better understanding on what areas or fields, should we focus on our next research steps. Drug Discovery Study Ontology starts with the company first producing a drug. The drug is then sent for clinical tests, which performs several tests and collects several details. The Ontology for Drug Discovery Clinical Trial Study is displayed below:



Unstructured Data Extraction:

Most of the data of clinical trials are unstructured. As the clinical trial data is collected by a various agencies all around the world, as a result the data entry for all the data will not be unique, which makes the data highly unstructured. It is very difficult to read and understand unstructured and semi-structured data. So, in order to better understand the data, we need to first extract the required fields from the unstructured data sources.

For this project, I have two methods to extract required data from unstructured data files. Those two methods are:

Method 1:

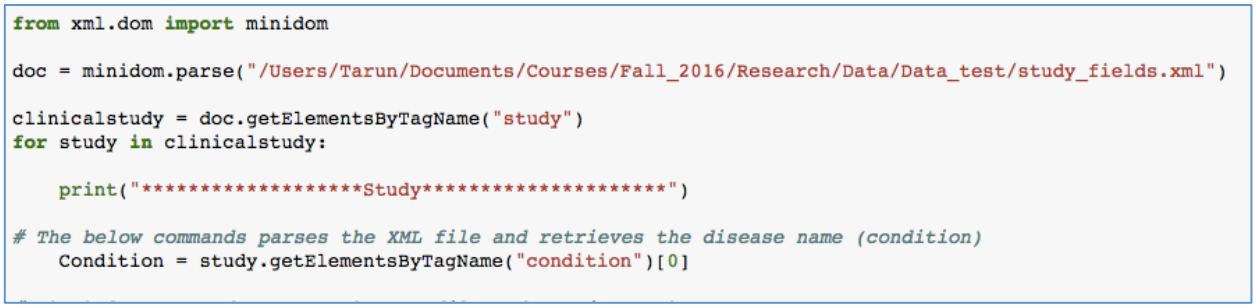
Extract Individual Clinical Trials data file separately with entire fields and then perform data analytics on each file separately.

Method 2:

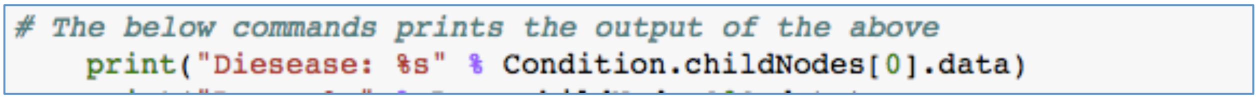
Extract only one XML file containing only limited required fields of all clinical trials and perform data analytics on that file.

In this project, I have worked on both methods. First, I have followed Method 1 and extracted each clinical trials into a separate XML with all the existing fields. Then, I have used the Pyhton DOM and Minidom Parser to write a code, which will parse the XML file and extract all the require fields and display them as output in "Individual_Unstructured_Data_files" Jupyter Notebook.

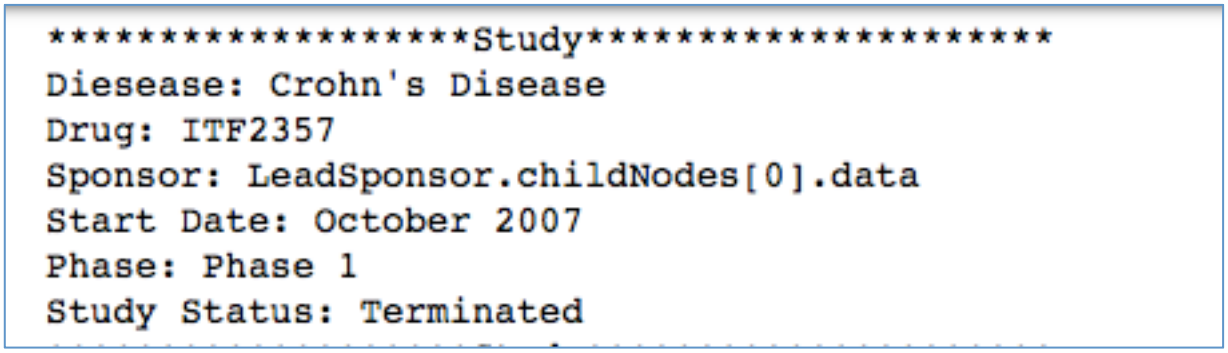
On the other hand, for method 2, I have used search criteria on Clinical Trials Database to extract only the required fields into a single XML file. Then, I wrote a different Python Code, which will extract all the fields, such as title, drug name, condition name, and all other fields in the same output file as shown below.



```
from xml.dom import minidom

doc = minidom.parse("/Users/Tarun/Documents/Courses/Fall_2016/Research/Data/Data_test/study_fields.xml")

clinicalstudy = doc.getElementsByTagName("study")
for study in clinicalstudy:

    print("*******************Study*********************")

# The below commands parses the XML file and retrieves the disease name (condition)
    Condition = study.getElementsByTagName("condition")[0]
```



```
# The below commands prints the output of the above
    print("Diesease: %s" % Condition.childNodes[0].data)
```



```
*******************Study*********************
Diesease: Crohn's Disease
Drug: ITF2357
Sponsor: LeadSponsor.childNodes[0].data
Start Date: October 2007
Phase: Phase 1
Study Status: Terminated
```

In comparison, Method 2 is very advantageous for performing data analytics, as the data from all clinical studies would be present in one file. So, while performing analytics, we can just extract the all the clinical studies data from that one file and then compare those studies internally after extraction. However, one of the disadvantages from using Method 2 is that it only includes selected fields, so the data analysis on the clinical studies will be dependent on only these fields, so we would not get a overall data results.

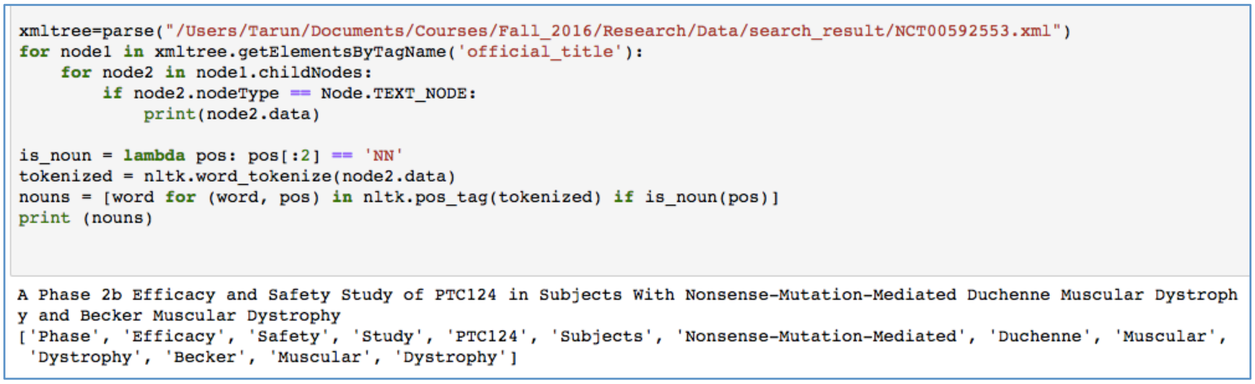
Text Mining:

After the data is extracted from the unstructured data sources, we need to mine the data, so that we can break down large complex data into smaller chunks of data. Breaking down the data into smaller chunks makes it easier for the analyst, to understand the data and be able to convert to the unstructured data into structured data. The best way to convert long unstructured data into smaller useful chunks is through text mining. Text Mining derives high quality information from text through patterns and trends of statistical pattern learning.

For the text mining part of the project, I have focused on breaking down long sentences such as title and study descriptions into nouns and noun phrases. To convert long sentences into noun and noun phrases, I have used two kinds of algorithms, which include:

Algorithm 1: Text Mining for Nouns

I have created an algorithm, which reads a sentence and only identifies and displays the nouns in that sentence. This algorithm uses lambda function, with parts of speech tagger, along with word tokenizers.

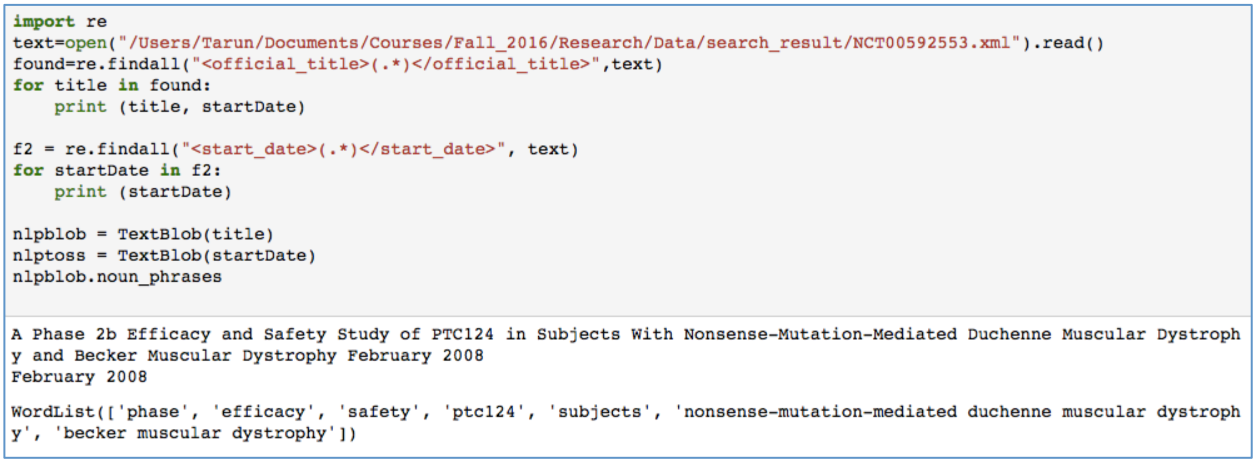


```
xmltree=parse("/Users/Tarun/Documents/Courses/Fall_2016/Research/Data/search_result/NCT00592553.xml")
for node1 in xmltree.getElementsByTagName('official_title'):
    for node2 in node1.childNodes:
        if node2.nodeType == Node.TEXT_NODE:
            print(node2.data)

is_noun = lambda pos: pos[:2] == 'NN'
tokenized = nltk.word_tokenize(node2.data)
nouns = [word for (word, pos) in nltk.pos_tag(tokenized) if is_noun(pos)]
print (nouns)
```

```
A Phase 2b Efficacy and Safety Study of PTC124 in Subjects With Nonsense-Mutation-Mediated Duchenne Muscular Dystroph
y and Becker Muscular Dystrophy
['Phase', 'Efficacy', 'Safety', 'Study', 'PTC124', 'Subjects', 'Nonsense-Mutation-Mediated', 'Duchenne', 'Muscular',
 'Dystrophy', 'Becker', 'Muscular', 'Dystrophy']
```

Algorithm 2: Text Mining for Noun Phrases

For this algorithm, I have used an in-built python toolkit known as Natural Language Toolkit (NLTK). In the Natural Language Toolkit, there is an inbuilt classifier called as TextBlob. I have used textblob classifier using NLTK to extract the sentences in the form of Noun Phrases.



```
import re
text=open("/Users/Tarun/Documents/Courses/Fall_2016/Research/Data/search_result/NCT00592553.xml").read()
found=re.findall("<official_title>(.*)</official_title>",text)
for title in found:
    print (title, startDate)

f2 = re.findall("<start_date>(.*)</start_date>", text)
for startDate in f2:
    print (startDate)

nlpblob = TextBlob(title)
nlptoss = TextBlob(startDate)
nlpblob.noun_phrases
```

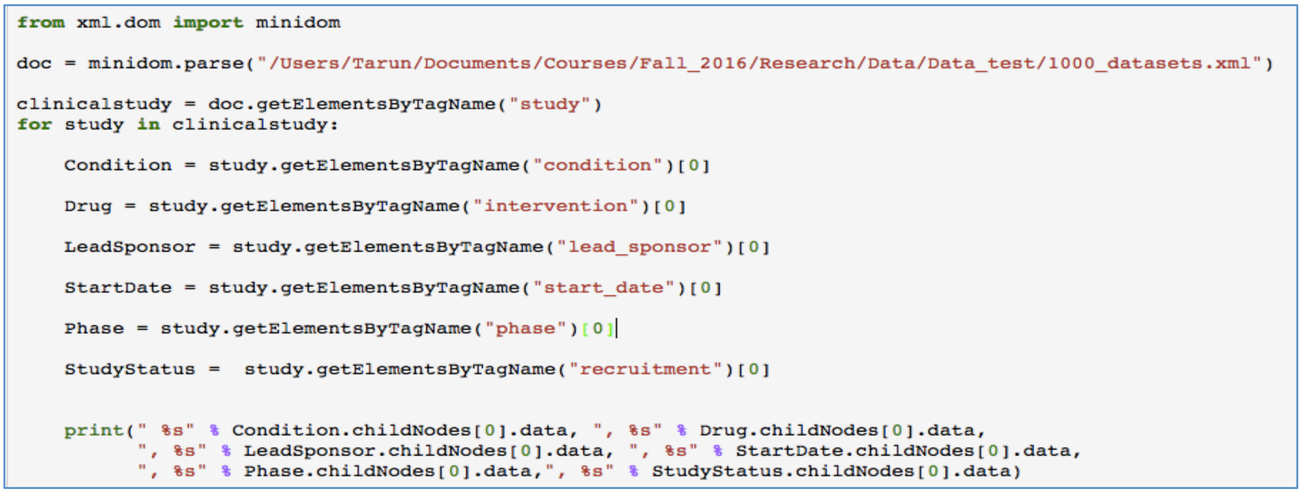
```
A Phase 2b Efficacy and Safety Study of PTC124 in Subjects With Nonsense-Mutation-Mediated Duchenne Muscular Dystroph
y and Becker Muscular Dystrophy February 2008
February 2008

WordList(['phase', 'efficacy', 'safety', 'ptc124', 'subjects', 'nonsense-mutation-mediated duchenne muscular dystroph
y', 'becker muscular dystrophy'])
```

Though it is easier to use Nouns for data analysis, Noun Phrases provide a better result. Our project was focused on analyzing the medical drugs, so we need the names of the drugs, but using algorithm 1 would breakdown the drug name, or disease name into 2 nouns as it aims breaks the sentences into as many nouns as possible. However, using algorithm 2, we would be able to retrieve the drug name, or disease name as it is, without any breakages.

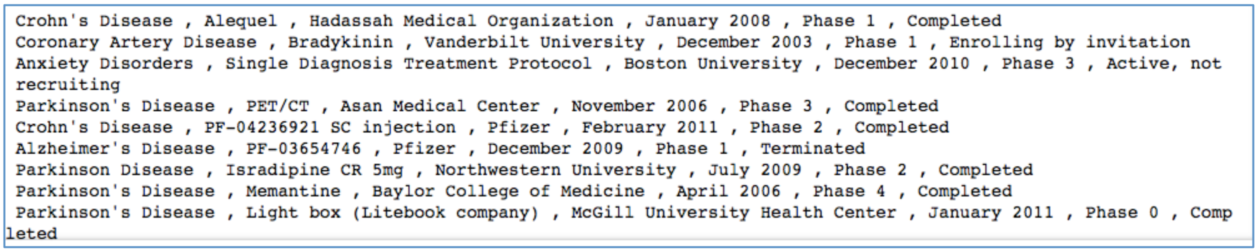
Conversion of Unstructured Data into Structured Data:

After the unstructured data is data is extracted and refined through text mining, we will focus on converting the refined unstructured data into structured data. By converting unstructured data into structured data, we will get quality data, which can be transformed into information. With structured data, it will be easier for data analyst to understand the landscape of drug discovery and be able to perform data analysis on the clinical studies. In this stage, the main focus is on knowledge extraction, so I have focused on extracting the most important fields that would be help data analysts in understanding some questions such as what diseases are the drugs helpful in curing, which company is successful in sponsoring most drugs, and what phases are most of the drug testing's in?

The fields I have extracted for solving the above questions are Disease, Drug Name, Sponsor, Start Date, Phase, and Status of the clinical trials. While extracting these fields, I delimited the fields using a comma (",") in Jupyter Notebook so that I can extract these data and csv file. Then, I have crated a database with those fields and then imported the file into the database table using MySQL for future research.



```
from xml.dom import minidom

doc = minidom.parse("/Users/Tarun/Documents/Courses/Fall_2016/Research/Data/Data_test/1000_datasets.xml")

clinicalstudy = doc.getElementsByTagName("study")
for study in clinicalstudy:

    Condition = study.getElementsByTagName("condition")[0]

    Drug = study.getElementsByTagName("intervention")[0]

    LeadSponsor = study.getElementsByTagName("lead_sponsor")[0]

    StartDate = study.getElementsByTagName("start_date")[0]

    Phase = study.getElementsByTagName("phase")[0]

    StudyStatus =  study.getElementsByTagName("recruitment")[0]


    print(" %s" % Condition.childNodes[0].data, ", %s" % Drug.childNodes[0].data,
          ", %s" % LeadSponsor.childNodes[0].data, ", %s" % StartDate.childNodes[0].data,
          ", %s" % Phase.childNodes[0].data,", %s" % StudyStatus.childNodes[0].data)
```



```
 Crohn's Disease , Alequel , Hadassah Medical Organization , January 2008 , Phase 1 , Completed
 Coronary Artery Disease , Bradykinin , Vanderbilt University , December 2003 , Phase 1 , Enrolling by invitation
 Anxiety Disorders , Single Diagnosis Treatment Protocol , Boston University , December 2010 , Phase 3 , Active, not
 recruiting
 Parkinson's Disease , PET/CT , Asan Medical Center , November 2006 , Phase 3 , Completed
 Crohn's Disease , PF-04236921 SC injection , Pfizer , February 2011 , Phase 2 , Completed
 Alzheimer's Disease , PF-03654746 , Pfizer , December 2009 , Phase 1 , Terminated
 Parkinson Disease , Isradipine CR 5mg , Northwestern University , July 2009 , Phase 2 , Completed
 Parkinson's Disease , Memantine , Baylor College of Medicine , April 2006 , Phase 4 , Completed
 Parkinson's Disease , Light box (Litebook company) , McGill University Health Center , January 2011 , Phase 0 , Comp
leted
```

The SQL code I have used for creating my database table is displayed below:

```
Create Table Drug_Discovery
 (
 ID              INT                      NOT NULL    AUTO_INCREMENT,
 DISEASE         VARCHAR2(1055),
 DRUG_NAME                 VARCHAR2(1055),
 SPONSOR         VARCHAR2(1055),
 START_DATE  DATE,
 PHASE           VARCHAR2(50),
 STUDY_STATUS  VARCHAR2(200),
 RECRUITMENT_STATUS  VARCHAR2(100)
 );
```

This part of the research is a crucial part, as we convert the entire drug discoveries related clinical studies into our local database. So far, due to my computer limitations, I was able to convert 1000 data records of the above-mentioned fields into the database. We can query those converted structured data and draw concepts and relations between the data discoveries, which would help the data analyst better understand the Drug Discoveries related data and make clear judgments on the relationships between the data.

Future Research:

Currently, I have worked on extraction and parsing of unstructured data, text mining of that data and then conversion of refined unstructured data into structured data. As my computer was only able to compute 1000 data values, for the next steps of my project, I am going to switch to a computer with high processing power and use it to compute my code on the entire drug discovery data sets.

Now, the next steps of this research includes querying on the structured database to draw patterns between the drugs and the diseases and also compare how long it is taking for the drugs to be completed and accepted by the clinical trials. Using that information, we will use NEO4j tool to create a knowledge graph, which would represent the data in graphical formation. This graphical formation of data clearly demonstrates the relationships between different drugs and their diseases, which could be helpful in identifying new knowledge relationships between different clinical trail studies.

Conclusion:

Overall, from this project, I was able to learn a lot about the python, and dealt with solving some real world data problems. In this project, first we have designed the ontology of the drug discovery clinical trial model, so that we can understand the key fields that we need to focus on and remove unnecessary fields. Then, we have extracted all the clinical trials studies on drug discoveries into a single XML file and then used DOM Minidom parser to retrieve the key fields such as Disease, Drug Name, Sponsor, Start Date, Phase, and Study Status. After these fields are extracted, we perform text mining on these fields, so that we can break down any unstructured data into smaller chunks of useful data. After the text mining is performed, we separate the data fields by a comma (",") and extract the data as a csv file. Once the data is extracted, we can create a structured database with the Drug discovery table and import the csv data file using MySQL import wizard. Once, the data is entered into the structured database and converted into graphical representation, we can understand the data and design new algorithms that could provide with predictive analytics on the drug discovery data.

References:

[1] Dr. Yuan An, (n.d.) Harnessing Open Data Resplan Proposal.