PHUSE US Connect 2020

Paper ML02

Natural Language Processing for Enhanced Treatment Decision

Phani S Srinivasan Ponnapalli, Syneos Health, Raleigh, USA
Subrahmanyam Rayaprolu, Syneos Health, Raleigh, USA

ABSTRACT

Data Driven Science is opening new opportunities for assimilation and evaluation of large amounts of complex health-care data. Artificial intelligence (AI) based technologies such as Natural Language Processing (NLP) provide an opportunity for Life Science firms to extract the key facts from structured and unstructured data and transforming real world data into actionable intelligence for decision making. This paper introduces diverse data types for clinical decision support (eg, Genomics, Electronic Health Record, Physician notes, Lab data, and Clinical trial data), concepts of AI/Machine Learning and provides deep insight on Natural Language processing algorithms. Further, this paper explores Data-Driven architecture to integrate diverse Clinical data using Hadoop framework. Finally, this paper will show how to apply combination of Machine Learning and Natural Language processing on integrated data for identifying the hidden patterns and better treatment decisions using Natural Language Toolkit (NLTK), Gensim and other Python libraries.

INTRODUCTION

Advances in Digital data sciences and Bio-Informatics have resulted in the new era of precision and value based medicine.

Applied Machine Learning and NLP techniques on integrated clinical information will transform the way we look at disease and enhance treatment decisions.

NLP/AI Analytics on Clinical data would demonstrate compelling benefits, including:

- Discovery of hidden patterns and actionable insights
- Precision Medicine
- Prediction and analysis of clinical trial outcomes
- Develop new drugs faster and reduce the costs of clinical trial

This paper would explore the possibilities of NLP on “Clinical” data by discussing:

- Diverse data types for clinical decision support
- Big Data and NoSQL
- Data Science Work Flow.
- Insight of NLP Algorithms.
- Python Programming Language for NLP techniques.
- Sample Python code for applying machine NLP on Integrated Clinical data.

DIFFERENT FORMS OF CLINICAL DATA

Clinical Trials involved diverse types of data include structured, typed, unstructured data in various digital formats. Assimilation of diversified Clinical data by extracting similar kinds of information in a form that is solvable could provide deeper insights.

Genomics – A genome is an organism's complete set of DNA including all of its genes. Genomics is an interdisciplinary field of science focusing on the structure, function, evolution, mapping, and editing of genomes.

An individual's genomic information is an important determinant in influencing the state of health and disease. Examining this genetic background is of great importance for identifying mutational changes in DNA that discriminate health and disease.

Electronic Health Record - An electronic health record (EHR) is the systematized collection of patient health information and stored electronically in a digital format. EHR data may include data related to demographics, medical history, laboratory test results, vital signs, questionnaires, drug exposure and radiology data.

Sharing these Electronic Health Records across different networks and health care settings helps clinicians identify and stratify ill patients and provide better treatment. EHR data analytics can also help in improving quality care by predicting proper care and preventing hospitalizations among high-risk patients.

Physician Notes: A physician note is an entry into a medical or health record made by a Physician, nurse or lab technician or any other member of a patient's healthcare team. Physicians note may include but not limited to family history, surgical history, medications, social history, immunizations, habits, Developmental history, hospitalizations and checkups.

A complete, accurate systematic documentation of patient's medical history and diagnosis in physician notes helps in proper treatment and care.

Lab Data: Laboratory data consists of lab test results from blood chemistry, hematology and Urinalysis. Lab data may also consists of many different collections of tests such as Microbiologic laboratory tests, ECG laboratory data and other therapeutic-indication-specific clinical lab tests.

This information is fundamental in diagnosing and planning for treatment. Laboratory data is also useful to evaluate the response to a treatment and to monitor the course of disease over time.

Clinical Trial Data: Clinical trial data is the data generated from clinical trial research on a drug, biologic or device about their safety and efficacy.

The Clinical data Interchange Standards Consortium (CDISC) Standards group have broadly categorized clinical trial data into interventions class, events class, findings class, and other special-purpose domains like demographics. Interventions are the drug administration and surgical procedures that the patient receives. Events are the unplanned clinical occurrences that the patient experiences. Findings capture the planned examinations of the patient over the course of the trial.

BIG DATA AND NOSQL:

Big Data

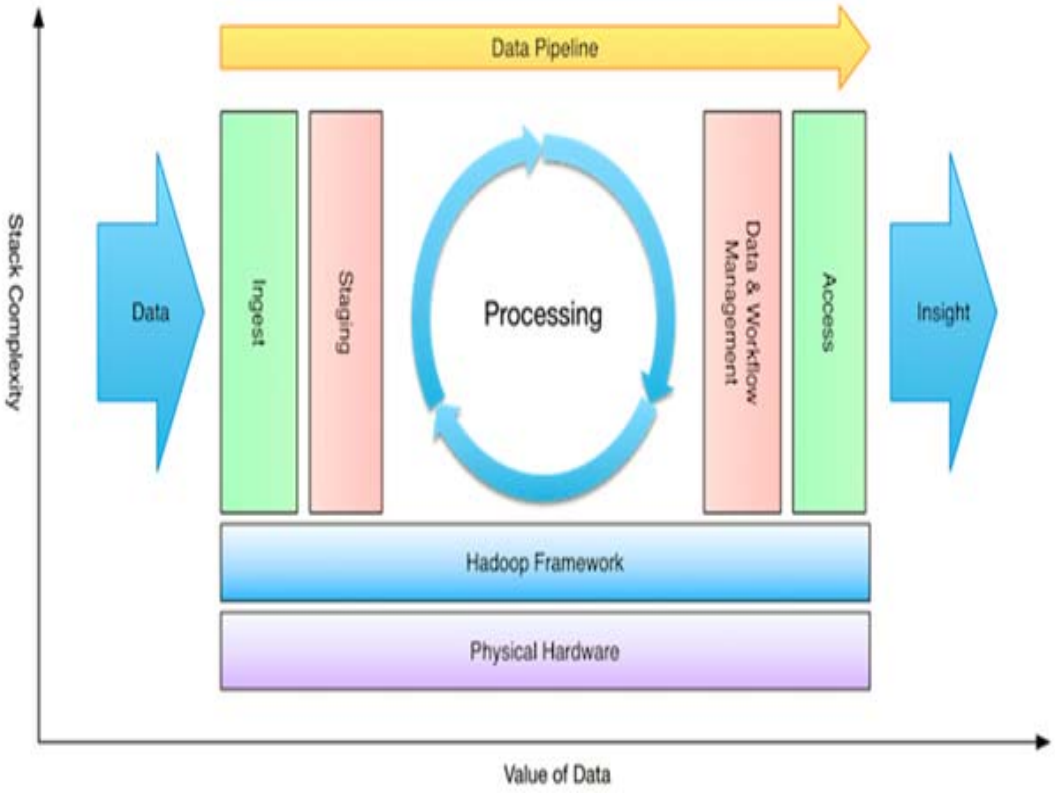
Big Data architecture is designed to handle the ingestion, processing, and analysis of data that is too large or complex for traditional database systems. Big data solutions work on a fundamentally different principle to handle device data and streaming data. Big Data architecture has the following characteristics.

- Distributed Data and parallel processing: Big Data solutions store huge data in a distributed manner in a file system. Process the data in parallel on a cluster of nodes. In simple words, Data is broken in to small pieces of information and stored in multiple blocks called HDFS. Tasks are executed in parallel by processing these blocks in parallel and results are merged back, is called map reduce.

- Fault Tolerance: Big Data solutions work on failure to tolerance by redundancy. The same information is stored in multiple places called racks. There are multiple machines available as cluster for processing. If any machine in the cluster goes down still system works due to data redundancy and multiple machines.

- Scalability: Big Data systems are very flexible in scaling storage space and computing power on fly whenever required.

- Cost effectiveness: Big Data systems like Hadoop are open source and uses commodity hardware. They do not require a very high-end server with large memory and processing power. This makes the system very cost effective.



NoSQL

NoSQL is an approach to databases that represents a shift away from traditional relational database management systems (RDBMS).NoSQL databases do not rely on predefined schema structures and use more flexible data models. NoSQL can mean “not SQL” or “not only SQL.” NoSQL is particularly useful for storing unstructured/semi-structured data.

Features of NoSQL:

- Minimal data modeling
- Minimal/no ETL
- No pre-defined schema necessary

Types of NoSQL Databases

Several different varieties of NoSQL databases have been created to support specific needs and use cases. These fall into four main categories:

Key-value data stores: Key-value NoSQL databases emphasize simplicity and are very useful in accelerating an application to support high-speed read and write processing of non-transactional data. Stored values can be any type of binary object (text, video, JSON document, etc.) and are accessed via a key. The application has complete control over what is stored in the value, making this the most flexible NoSQL model. Data is partitioned and replicated across a cluster to get scalability and availability. For this reason, key value stores often do not support transactions. However, they are highly effective at scaling applications that deal with high-velocity, non-transactional data.

Document stores: Document databases typically store self-describing JSON, XML, and BSON documents. They are similar to key-value stores, but in this case, a value is a single document that stores all data related to a specific key. Popular fields in the document can be indexed to provide fast retrieval without knowing the key. Each document can have the same or a different structure.

Wide-column stores: Wide-column NoSQL databases store data in tables with rows and columns similar to RDBMS, but names and formats of columns can vary from row to row across the table. Wide-column databases group columns of related data together. A query can retrieve related data in a single operation because only the columns associated with the query are retrieved. In an RDBMS, the data would be in different rows stored in different places on disk, requiring multiple disk operations for retrieval.

Graph stores: A graph database uses graph structures to store, map, and query relationships. They provide index-free adjacency, so that adjacent elements are linked together without using an index.

Multi-modal databases leverage some combination of the four types described above and therefore can support a wider range of applications.

Relational Table Vs NoSQL

Relational Table

Product_id	Product_name	Product_price
2	Product 2	3000

NO SQL

```
<Table >
<product >
<product_id>2</product_id>
<product_name>product 2</product_name>
<product_price>3000</product_price>
</product>
```

NATURAL LANGUAGE PROCESSING (NLP)

NLP is a way for computers to analyze, understand, and derive meaning from human language in a smart and useful way. By utilizing NLP, developers can organize and structure knowledge to perform tasks such as automatic summarization, translation, named entity recognition, relationship extraction, sentiment analysis, speech recognition, and topic segmentation.

INSIGHT ON NLP ALGORITHMS

- **Sentiment analysis**

Sentiment analysis is a machine learning technique that automatically analyze the data and detects the Positive or Negative polarity within a text whether it is a whole document, paragraph, sentence or clause.

Much of the information of Clinical data i.e. Genomics, Health care records and Physician notes lies buried in unstructured text and it is very hard to analyze this data. Sentiment analysis helps in making sense of the unstructured Clinical data by automatically tagging and detecting the polarity.

Example:

Physician Notes	**Sentiment**
Adverse Event reported. Patient admitted for serious condition.	Negative
Tumor size decreased by 10%.	Positive

- **Topic segmentation**

Topic segmentation is the process of dividing written text into meaning blocks such as words, sentences or topics. Unstructured Clinical data could be segmented into meaningful blocks such as demographics, medical history, laboratory test results, hospitalizations and social history for further analysis.

- **Text Generation**

Text generation is the task of generating text with the goal of appearing indistinguishable or missing text. A trained language model predicts the likelihood of occurrence of a word based on the previous sequence of words used in the text. Language models can be operated at character level, sentence level or even paragraph level.

INTEGRATION OF ALLFORMS OF CLINICAL DATA

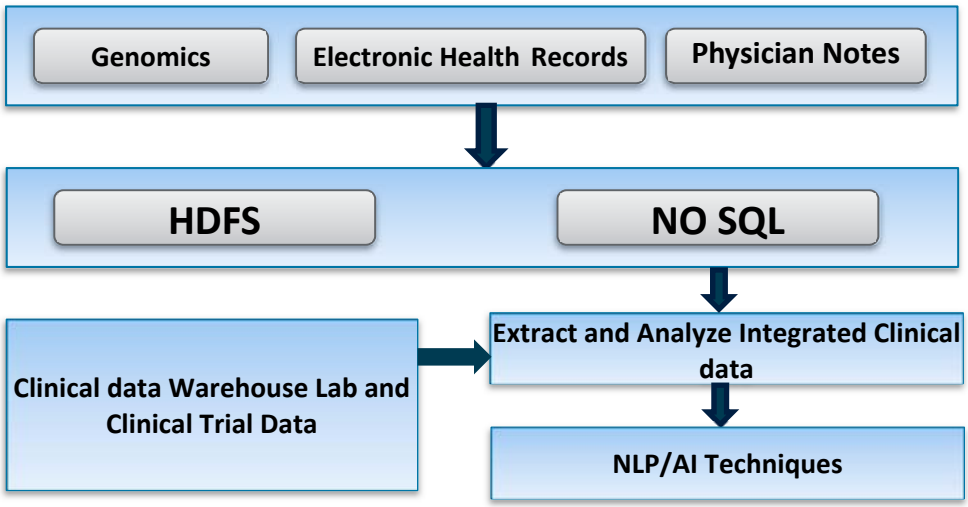
Latest technological advancements in Pharma has resulted in massive amounts of life-sciences data consisting of characteristics like Variety, Velocity, Veracity, Volume and Potential Value. In clinical domain, data of many patients is collected for intelligent decision-making. This data needs to be stored and analyzed in an efficient way that would be helpful for making best decisions. Different data mining and machine learning algorithms require input data in specific types and formats.

Lifesciences data is available for processing in various formats such as

1. Hospital Notes
2. Pathology Reports
3. *EMR data*
4. Clinical database

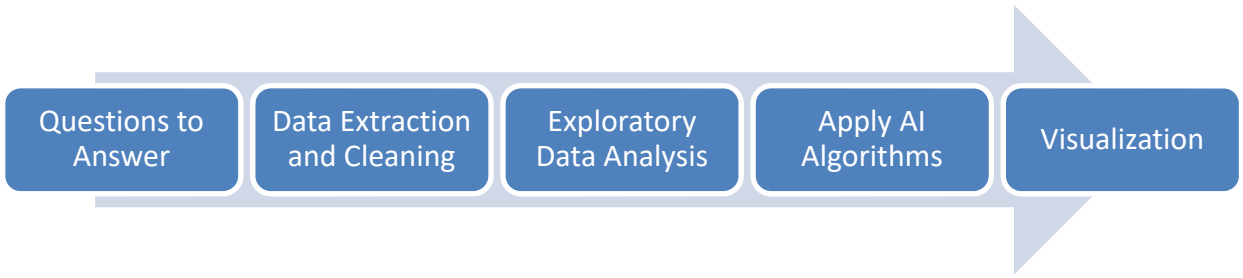
Big Data and NOSQL framework is the most suitable architecture for loading this unstructured Data.
Big data is to parsed and integrated with clinical trial database to obtain value out of it. Applying NLP and AI algorithms on the integrated information will open the door of hidden insights and better treatment decisions

Architecture



DATA SCIENCE WORK FLOW WITH PYTHON LIBRARIES

Although data science projects can range widely in terms of their aims, scale, and technologies used, at a certain level of abstraction most of them implemented as the following workflow



STEP 1: IDENTIFY THE QUESTIONS TO BE ANSWERED

Analytics is about asking—and answering—smarter questions.

What questions should you are asking to drive more value in your organization? What information needed to answer these questions? Which tools can efficiently deliver the answers? Clinical data is complex and diverse, Pharma Organizations have to identify and prioritize the questions answered.

For example:

- In a typical clinical trial, each patient responds differently to the same drug. Pharma companies challenge is to find a pattern associated with individual characteristics.
- Patient enrolment for a rare disease.
- Plan for monitoring trials in real time to identify signals requiring action to avoid significant and potentially costly issues.

STEP 2: DATA EXTRACTION AND CLEANING

At this step, scope of project is defined using technical and domain expertise.

Data Acquisition:

Data can be acquired from a variety of sources.

- Data files can be downloaded from internal/external online repositories.
- Data can be automatically generated by physical apparatus
- Physician or Hospital can manually enter data into application notes.

Bigdata frame works like apache Hadoop and Ingestions tools are used to acquire the data and load in to enterprise Data Lake. Big Data Ecosystems contain tons of data in raw format. We need to extract the relevant information with which we will be able to answer the questions, which are not available in Clinical database.

Data extraction and cleaning processing step is to turn the raw source data into a system analyzable format. Different types of analysis require type of data formats.

- Make text all lower case
- Remove punctuation
- Remove numerical values
- Tokenize text
- Remove stop words

Python Programing Paradigm:

Input: Raw Data

Python Libraries:

- pandas – is a data analysis library that includes dataframes
- pickle - The pickle module implements binary protocols for serializing and de-serializing a Python object structure.
- NLTK (Natural Language Toolkit) is used for such tasks as tokenization, lemmatization, stemming, parsing, POS tagging, etc. This library has tools for almost all NLP tasks.

Output:

- Pickle file – serialized binary file
- Corpus - a collection of text
- Document-Term Matrix - word counts in matrix format

STEP 3: EXPLORATORY ANALYSIS

Exploratory Data Analysis refers to the critical process of performing initial investigations on data to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations

At this step, we decide if data is making sense and see if any additional cleaning or processing is required on data.

Python Programing Paradigm:

Input: Document Matrix, Pickled file

Python Libraries:

- numpy - is used for its N-dimensional array objects
- matplotlib – is 2D plotting library for creating graphs and plots

Output:

- Distribution plots
- Simple Graphs and Charts

STEP 4: APPLY AI ALGORITHMS

At this point data is integrated and converted to format required by particular NLP/AI techniques.

Common Tasks solved by NLP algorithm include:

- Text classification – definitions of text semantics for further processing
- Named-entity recognition (NER) – definition and selection of entities with a predefined meaning (used to filter text information and understand general semantics);
- Summarization – the text generalization to a simplified version form (re-interpretation the content of the texts);
- Text generation – one of the tasks that are used to build AI-systems;
- Topic modeling – technique for extracting hidden topics from large text volumes.

Based on the questions we are trying to answer we apply Specific NLP algorithm to gain deeper insight and decision-making.

Python Programing Paradigm:

Input: Corpus, Document Matrix, Pickled file

Python Libraries:

- Gensim is the package for topic and vector space modeling, document similarity.
- TextBlob– is a package for part-of-speech tagging, noun phrase extraction, sentiment analysis, classification

STEP 5: VISUALIZATION

Final step in data science workflow include interpretation and presentation of the results in form of graphs, reports. It provides easy-to-understand format and helps in gaining insights of hidden patterns in data.

Python Programing Paradigm:

Input: Corpus, Document Matrix, Pickled file

Python Libraries:

- seaborn – a data visualization library based on matplotlib
- scikit-learn - the algorithms used for data analysis and data mining tasks
- seaborn – a data visualization library based on matplotlib

Output:

- Graphs and Charts
- Visualizations.

Document Term Matrix

	Terms			
Document	adverse event	visit	admit	result
Note1	20	6	45	67
Note 2	4	4	5	50
Note 3	6	8	6	66
Note 4	0	7	0	88

Corpus

Patient ID	Physician Notes
CAB101	Patient admitted with adverse reaction to Drug and Oxygen support is Given
CAB102	Patient was called on phone to find out adherence to Drug Schedule

PYTHON NLP SAMPLE CODE

Python is a popular open source programming language and it is one of the most-used languages in artificial intelligence and Machine Learning fields.

Anatomy of Basic Python Program

```
# Import Statements of Libraries
# Variable Definitions
# Classes Definitions
# Function Definitions
# Processing
```

Patient Response trend:
Sentiment Analysis and Segmentation on clinical trial site notes and Investigator notes, which are sometimes available in comments and not loaded to Clinical database could be used to gain more insight and decision making during a trail.

- Pattern of Patients response positive or Negative to a drug across the visit?
- Any additional reasons of patient's response like an event in family or individual Genome?

Sentiment Analysis:

Each word in a corpus is labeled in terms of polarity and subjectivity. A corpus sentiment is the average of these.

- Polarity: How positive or negative a word is. -1 is very negative. +1 is very positive.
- Subjectivity: How subjective or opinionated a word is. 0 is fact. +1 is very much an opinion.

Code Snippet

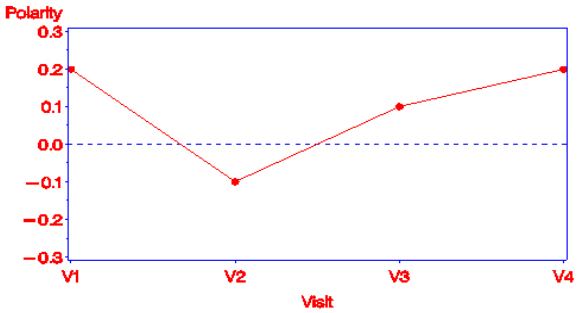
```
# Calculate the average polarity of each patient

polarity_patient = []
for lp in notes_visit:
    polarity_visit = []
    for p in lp:
        polarity_visit.append(TextBlob(p).sentiment.polarity)
    polarity_patient.append(polarity_visit)

polarity_patient
```

```
# Show the plot for one patient
plt.plot(polarity_Patient[0])
plt.title(data['Patient_name'].index[0])
plt.show()
```

Sample Output



CONCLUSION

Machine learning offers tremendous opportunities to more efficiently access and understand vast amounts of data produced by Life-sciences Industry. NLP can provide access to most of the data otherwise trapped in unstructured text, turning it at large scale into features that can drive predictive models. The combination of NLP and ML provides a powerful paradigm for developing predictive models in healthcare and life science applications.

REFERENCES

1. https://github.com/adashofdata/nlp-in-python-tutorial/
2. http://alttox.org/mapp/emerging-technologies/omics-bioinformatics-computational-biology/
3. https://www.sciencedirect.com/science/article/pii/S2001037014000464
4. http://blog.cloudera.com/blog/2014/09/getting-started-with-big-data-architecture/
5. https://monkeylearn.com/sentiment-analysis/
6. https://en.wikipedia.org/wiki/Electronic_health_record
7. https://practicefusion.com/medical-notes/
8. https://www.analyticsvidhya.com/blog/2013/07/big-data/
9. https://emerj.com/ai-sector-overviews/natural-language-processing-in-pharma-current-applications/
10. https://www.sciencedirect.com/science/article/pii/S1532046418302016

ACKNOWLEDGMENTS

The authors would like to thank Carol Frazee, Associate Director, Syneos Health, for her valuable comments and feedback.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Name: Phani Ponnapalli
Enterprise: Syneos Health
Work Phone: +1 732.666.8572
E-mail: phani.ponnapalli@syneoshealth.com

Name: Subrahmanyam Rayaprolu
Enterprise: Syneos Health
Work Phone: +1 972.370.3552
E-mail: subrahmanyam.rayaprolu@syneoshealth.com