Knowledge Discovery using PubMed to identify relations between Drugs, Chemicals and Diseases

Team Name: KDPUB
Members: Lingraj S Vannur

Problem/Motivation Statement:

Here in this project I chose to study a particular disease, **non-small cell lung cancer** (NSCLC). It has a high incidence and mortality rate: NSCLC is the most common type of lung cancer, accounting for about 85% of all cases. It is one among the leading causes of cancer-related deaths in the world. Therefore, understanding the biology and treatment of NSCLC is critical for reducing the incidence and mortality rate of lung cancer.

Using NLP we can extract information from medical records and clinical notes related to NSCLC. This information can be used to study the characteristics of the disease, treatment outcomes, and patient demographics. Extracting research articles, reviews, and clinical trial reports can be helpful to identify new targets for therapy, biomarkers, and diagnostic tools.

So in this project will take medical data from the Pubmed abstracts. Then use ML and DL algorithms to identify Named Entities Medicine and Diseases. Finally build a relation extraction model.

Dataset and analytic goals:

Data source - https://pubmed.ncbi.nlm.nih.gov/
In this project I have extracted the research abstracts from the Pubmed portal. Extracted around 500 article abstracts.

As there is a shortage of time, here I have focused on **Named Entity Recognition** using BioBert and Spacy. Analyzing the **abstracts similarity** using K Means. So that we can know how similar research is going on relating to it and who are its authors.

Overview of Data Engineering Pipeline:

Image Source: *MSDS USF Distributed Data System Course*

In our study of non-small cell lung cancer (NSCLC), we utilized various tools to analyze and extract insights from the available pubmed data. One of the primary sources of data was the PubMed database, which we accessed using a web scraper API. To store this data, we used Google Cloud Storage, which allowed us to easily store and retrieve large amounts of data.

However, the process of aggregating and analyzing this data was time-consuming and resource-intensive. To streamline this process, we utilized Apache Airflow, an open-source platform for workflow automation. With Airflow, we automated the entire data aggregation process, enabling us to easily and efficiently collect and store the data in MongoDB, a NoSQL database that is well-suited for storing large volumes of unstructured data.

Once the data was stored in MongoDB, we were able to use SparkMLib in Databricks, a cloud-based big data processing platform, to perform advanced analytics and extract insights from the data. By leveraging Airflow's automated workflow capabilities, we were

able to significantly reduce the time and resources required to process and analyze the data, allowing us to focus on developing more meaningful insights into NSCLC. Overall, our use of these advanced tools and technologies helped us to gain a deeper understanding of NSCLC and provided valuable insights into the disease and potential treatment options.

Preprocessing and Algorithms:

To effectively analyze the data extracted from PubMed for our study on non-small cell lung cancer (NSCLC), we needed to perform a series of preprocessing steps. First, we **aggregated the documents** and **identified the sentences** within them. We then **tokenize** the text, assign part-of-speech tags to each token. Next, the "attribute_ruler" component is applied to assign additional token-level attributes. The "lemmatizer" component is then used to identify the base form of each token.

POS	Tag	lemma	text
ADJ	JJ	epidermal	Epidermal
NOUN	NN	growth	growth
NOUN	NN	factor	factor
NOUN	NN	receptor	receptor
PUNCT	-LRB-	(	(
NOUN	NN	egfr	EGFR
PUNCT	-RRB-	)	)
NOUN	NNS	mutation	mutations
VERB	VBP	occur	occur
ADP	IN	in	in
ADV	RB	about	about
NUM	CD	50	50
NOUN	NN	%	%
ADP	IN	of	of
NOUN	NN	lung	lung
NOUN	NNS	adenocarcinoma	adenocarcinomas
ADP	IN	in	in
PROPN	NNP	Asia	Asia

Preprocessing pipeline outputs

Next, we utilized the **BioBert(spacy) transformer**, a deep learning algorithm designed specifically for medical named entity recognition (NER), to identify and classify the entities within the documents. This transformer is based on the pre-trained BioBert

model and has been specifically trained on medical data, allowing it to accurately identify and classify medical entities such as diseases, drugs, and chemicals.

We utilized the Spacy BioBert model to obtain a **200-dimensional vector embedding** for each abstract in our dataset. This embedding captured the semantic meaning of each abstract, making it suitable for use in **Named Entity Recognition (NER) as well as similarity analysis.**

Using the embeddings, we performed NER to identify and extract key medical entities such as diseases, treatments, and symptoms, which provided valuable information for our analysis of non-small cell lung cancer.

Time efficiency.

USed GPU cluster with 16GB RAM, 2-5 workers and 4 core driver and 8-20 cores. PySpark can perform data processing tasks such as data transformation, aggregation, and machine learning operations much faster than traditional serial computing approaches. For example, a simple PySpark operation like aggregating documents in a local PC took 5 mins, here it just took 40 seconds.

ML Goals, Outcomes, Execute Time Efficiency:

NER was carried out using 2 different pre-trained Biobert models of spacy, named entities considered [Chemicals, Diseases, Drugs, Duration, Strength].

label	text
DISEASE	cervical cancer c...
CHEMICAL	carboplatin/cispl...
DISEASE	non-small-cell lu...
CHEMICAL	cetuximab
DISEASE	nonsquamous non-s...
CHEMICAL	pemetrexed
DISEASE	cancers
DISEASE	lung cancer
DISEASE	Non-small-cell lu...
DISEASE	anaplastic lympho...
DISEASE	tumor
DISEASE	Toxicity
DISEASE	postoperative com...
CHEMICAL	pemetrexed-platinum
DISEASE	tumors
DISEASE	NSCLCs
DISEASE	A549
DISEASE	toxicity

label	text
DURATION	for four cycles
DRUG	inhibitor
DRUG	pembrolizumab
STRENGTH	200 mg
DRUG	cetuximab
DRUG	IC50
DRUG	cisplatin
STRENGTH	10 mg/ml
FREQUENCY	once every 3 weeks
DRUG	chemotherapy
DRUG	carboplatin
DRUG	Cetuximab
DRUG	oncolytic
DRUG	platinum
DOSAGE	2:1
DRUG	Pembrolizumab
DURATION	35 cycles
DRUG	platinum-pemetrexed

NER Output

Here spaCy for medical NER, we have used a pre-trained model that is already trained on medical text data. The pre-trained model, called "en_core_med7", has been trained on a large corpus of medical text data and can identify entities such as medical conditions, treatments, procedures, drugs, and anatomy.

Outcomes:

Predicted Text Labeled with Medical Entities:

Finally By accurately labeling medical entities in text data, healthcare providers and researchers can more easily extract relevant information and gain insights into patient care, disease management, and treatment outcomes. Furthermore, the use of predictive text labeled with medical entities can help improve the accuracy and efficiency of medical coding and billing processes, which can lead to cost savings and streamlined

administrative workflows. Overall, predicted text labeled with medical entities is a powerful tool that has the potential to transform the way healthcare data is analyzed, processed, and utilized.

t was 0.700 (95% CI, 0.641-0.759) and 0.881 (95% CI, 0.842-0.920), respectively. The CT image-based radiomic feature model had good classification abi

atients with NSCLC DISEASE . Lung cancer DISEASE has become one of the leading causes of cancer DISEASE incidence and mortality worldv

NSCLC DISEASE) is the most common type among all lung cancer DISEASE cases. NSCLC DISEASE patients contained high levels of activat

ons, such as exon 19 deletion, L858R DISEASE and T790 DISEASE M. Osimertinib, a third-generation of EGFR tyrosine CHEMICAL kinase inhit

he EGFR-T790M mutation of NSCLC DISEASE patients; however, treatment of osimertinib still can induce drug resistance in lung cancer DISEASE

tolerability of DS-1205c, an oral AXL-receptor inhibitor, in combination with osimertinib DRUG in met

progression during EGFR tyrosine kinase inhibitor (TKI) treatment. An open-label, non-randomized phas

200, 400, 800, or 1200 mg STRENGTH twice daily FREQUENCY for 7 days DURATION , follow

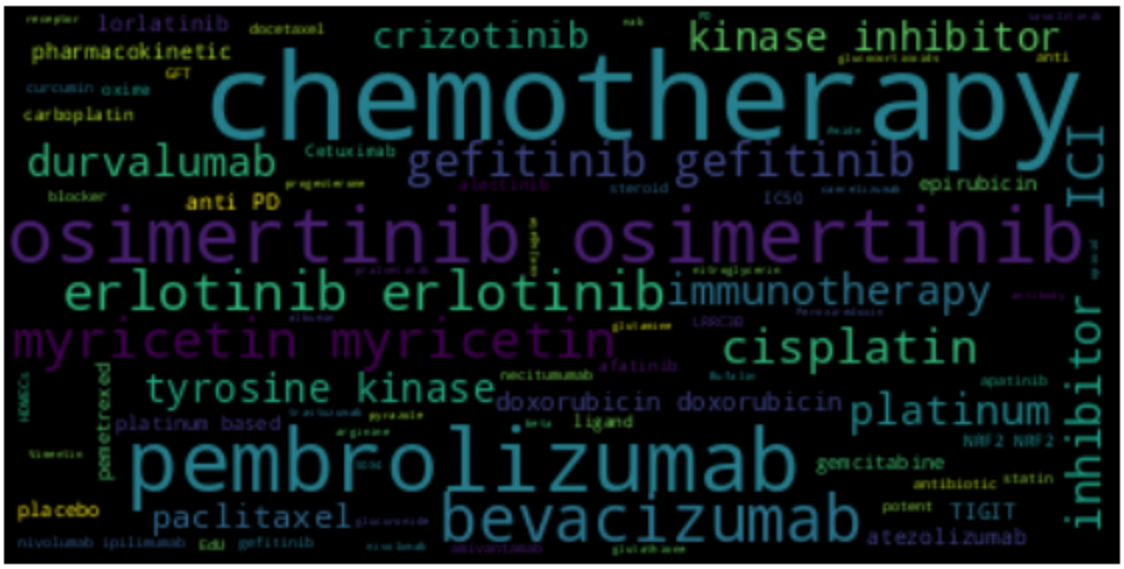
once daily FREQUENCY in 21-day DURATION cycles. Treatment continued until disease progress

reported in all 13 patients treated with DS-1205c plus osimertinib DRUG ; with ≥ 1 grade 3 TEAE in 6

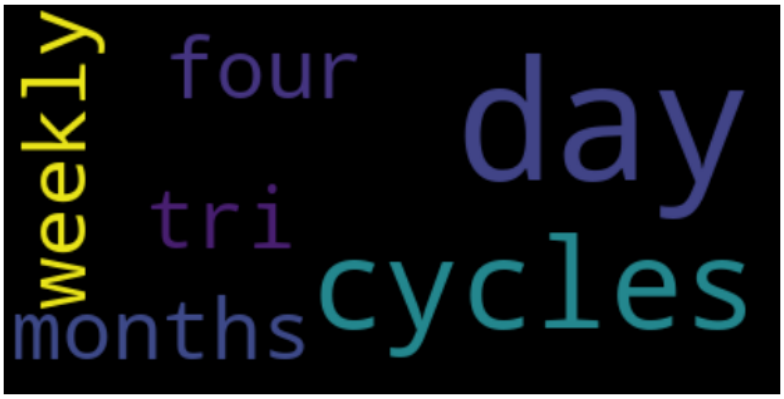
Word Cloud:

The word cloud here displays the most commonly occurring medical terms, procedures, drugs, and conditions visually. This type of visualization can provide insights into the most relevant and significant medical entities in a particular field or specialty.

Wordcloud: Chemicals



Wordcloud: Drug Tag

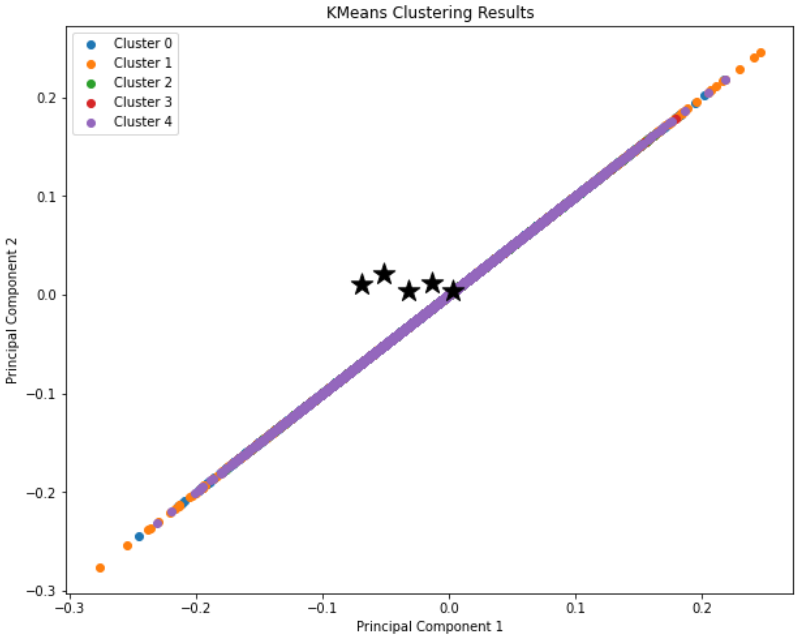


Wordcloud: Duration Tag

Execution Time Efficiency:

PySpark can perform machine learning operations much faster than traditional serial computing approaches. For example, KMeans took less time to run in databricks than my traditional CPU machine. SInce the data was bit small I could not see significant difference.

K-Means Clustering for Medical Abstracts and author grouping



Further the embeddings were used to build the cluster to identify how unique research topics are being carried out. It will also be helpful for us to segregate the authors, research areas and medical treatments and finding going on. We can see that the optimal 5 clusters were defined by using `Silhouette Score.` But the plot is very bad, showing that more research is highly similar.

Lessons Learned:

Lessons learned from building an end-to-end automated pipeline on Big Data application using Airflow, web scraping data, storing in GCP, saving scraped data in MongoDB, using Databricks to run the models, handling NLP data using spaCy for NER, and K-means clustering for grouping medical research and treatments include the importance of having a well-defined data pipeline, leveraging cloud computing and big data technologies, understanding domain-specific requirements, and careful planning,

testing, and iteration. By considering these factors, the pipeline can be optimized to deliver valuable insights and drive data-driven decision making.

Conclusion

In conclusion, using the Airflow and Spark automation pipeline provides an end-to-end solution in the medical NLP domain. The implementation of NER in the medical domain can provide valuable insights that can be used in various ways, such as carrying out medical research. In this project, we worked on abstracts and can further extend it to the whole PDF files. We can also begin the process of relation extraction and finally build knowledge graphs. These steps can help healthcare professionals and researchers gain a better understanding of medical data, leading to improved patient care and better treatment outcomes.

References:

1. Spacy documentation, https://spacy.io/usage/embeddings-transformers
2. https://pubmed.ncbi.nlm.nih.gov/?term=non+small+cell+lung+cancer%5BJournal%5D
3. {S}cispa{C}y: {F}ast and {R}obust {M}odels for {B}iomedical {N}atural {L}anguage {P}rocessing: "Neumann, Mark and King, Daniel and Beltagy, Iz and Ammar, Waleed",