Capstone Project Report

NLP and knowledge graph-based analysis of Pubmed and Google Scholar databases

Monsoon Semester 2023

Team Members:
Pranshu Patel : MT22117
Mohammed Taasir Fruitwala : MT22029
Swaib Ilias Mazumder : MT22078

Capstone Project Report submitted in partial fulfilment of the requirements for the Degree of M. Tech. in Computer Science & Engineering on Nov 22nd, 2023

Capstone Project Advisor
Dr. N Arul Murugan

Indraprastha Institute of Information Technology New Delhi

CONTENTS

Abstract

Parkinson's Disease (PD) is a complex neurodegenerative disorder with multifaceted underlying molecular mechanisms. In this capstone project, we embarked on a comprehensive investigation using advanced data science techniques to analyze abstracts from PubMed, a treasure trove of scientific knowledge. Our focus was on extracting key entities, including chemicals, genes, diseases, and pathways, to unravel the intricate interplay within the context of PD.

Utilizing cutting-edge techniques, we meticulously curated and analyzed vast amounts of data, seeking patterns and interactions that provide valuable insights into the molecular basis of Parkinson's Disease. The integration of this information enabled us to construct dynamic interaction graphs, offering a visually intuitive representation of the complex relationships between the identified entities.

To enhance the accessibility and user experience of our findings, we employed Streamlit, a powerful tool for creating interactive data applications. The resulting dynamic interaction graphs serve as a user-friendly interface, facilitating a deeper understanding of the intricate molecular landscape associated with Parkinson's Disease.

This project not only showcases the potential of data science in the realm of neurodegenerative diseases but also lays the foundation for further research and exploration. By shedding light on the interconnected web of molecular components, our work contributes to the ongoing efforts to unravel the mysteries of Parkinson's Disease and opens avenues for targeted interventions and therapeutic strategies.

Disease Overview: Parkinson's Disease

Introduction:
Parkinson's Disease (PD) stands as a prominent neurodegenerative disorder characterized by a progressive decline in motor function, often accompanied by a spectrum of non-motor symptoms. First described by James Parkinson in 1817, the disease has since been a focal point of scientific investigation, yet its exact etiology remains elusive.

Epidemiology:
PD is not confined by geographical or ethnic boundaries, affecting people worldwide. The prevalence of Parkinson's tends to increase with age, with a median onset around 60 years, though cases can emerge earlier in life. While the majority of cases are sporadic, a small percentage can be attributed to genetic factors.

Clinical Manifestations:
The hallmark motor symptoms of PD include tremors, rigidity, bradykinesia (slowness of movement), and postural instability. These motor impairments result from the progressive degeneration of dopaminergic neurons in the substantia nigra, leading to a dopamine deficiency in the brain. Beyond motor symptoms, individuals with PD may experience a range of non-motor issues such as cognitive impairment, mood disorders, and autonomic dysfunction.

Pathophysiology:
The pathological hallmark of PD is the formation of Lewy bodies—abnormal protein aggregates—in the brain. These aggregates primarily consist of alpha-synuclein and contribute to the degeneration of dopaminergic neurons. The precise triggers for alpha-synuclein aggregation and neuronal demise remain subjects of intensive research.

Diagnosis:
Diagnosing PD relies on clinical evaluation, with no definitive biomarker or imaging test available. Neuroimaging, such as DaTscan, can aid in confirming dopaminergic neuron loss. Early diagnosis poses a challenge, as symptoms may be subtle and easily mistaken for normal aging or other conditions.

Current Treatment and Future Directions:
While there is no cure for PD, current treatment strategies aim to alleviate symptoms by increasing dopamine levels in the brain. Levodopa, a precursor to dopamine, remains a cornerstone of therapy. However, the long-term use of levodopa can lead to complications. Ongoing research explores disease-modifying approaches, including gene therapies and neuroprotective agents, offering hope for more effective interventions in the future.

Conclusion:
Parkinson's Disease presents a multifaceted challenge that extends beyond motor symptoms, impacting the quality of life for affected individuals. As we delve into the intricate molecular interactions associated with PD in this capstone project, our aim is to contribute valuable insights that may pave the way for targeted therapeutic interventions and a deeper understanding of this enigmatic neurodegenerative disease.

Before Pubmed.mineR

Extracting Abstract related to Parkinson's disease from PubMed/Google Scholar

Introduction:
In the quest for a comprehensive understanding of Parkinson's disease (PD), a pivotal step involves extracting pertinent information from scholarly databases such as PubMed and Google Scholar. This process is facilitated by a Python script leveraging the BioPython library, a robust toolset designed for computational biology tasks.

Methodology:
The Python script is meticulously crafted to initiate a search on PubMed, seeking recent papers pertaining to Parkinson's disease. Users can specify the topic of interest, and the script dynamically retrieves papers published within a designated timeframe. The search is optimized to fetch a maximum number of results, ensuring a comprehensive pool for analysis.

Search Parameters:
The script allows users to define the search scope by specifying the number of years back from the current date. This temporal constraint is crucial for focusing on the latest research developments in Parkinson's disease. The extraction process is fine-tuned to sort the retrieved papers by publication date, emphasizing the most recent contributions to the field.

Data Retrieval:
Upon execution, the script harvests pertinent data from the search results, including abstracts that encapsulate the essence of each research paper. This ensures that the extracted information is not only recent but also encapsulates the critical insights provided in the abstracts.

Data Storage:
To facilitate seamless integration into the overall research workflow, the extracted data is stored in a structured format. The script outputs a JSON file, allowing for easy parsing and analysis. Researchers can employ the provided filename to locate and access the curated dataset for further investigation.

Significance:
This automated extraction process streamlines the research workflow, enabling researchers to stay abreast of the latest developments in Parkinson's disease. By focusing on recent abstracts, the script ensures that the information gathered is timely and reflective of the current landscape of PD research.

Conclusion:
The utilization of this Python script, empowered by the BioPython library, not only enhances the efficiency of information retrieval but also aligns with the project's commitment to

staying at the forefront of Parkinson's disease research. The extracted abstracts serve as a valuable resource for the subsequent stages of analysis and exploration, contributing to a more nuanced understanding of this complex neurodegenerative disorder.

```python
import json
from datetime import datetime
from Bio import Entrez, Medline

def search_and_store_recent_papers(query, years_back=3, num_results=6000, output_file="papers_info.json"):
    Entrez.email = "pranshupatel65@gmail.com"  # Set your email address here

    # Calculate the publication date range
    current_year = datetime.now().year
    start_year = current_year - years_back

    # Construct the search query with the date filter
    search_query = f"{query} AND {start_year}[Date - Publication]:{current_year}[Date - Publication]"

    # Include the sorting parameter to sort by "best match"
    search_handle = Entrez.esearch(db="pubmed", term=search_query, retmax=num_results, sort="pub+date")
    search_results = Entrez.read(search_handle)
    search_handle.close()

    id_list = search_results["IdList"]
    fetch_handle = Entrez.efetch(db="pubmed", id=id_list, rettype="medline", retmode="text")
    records = Medline.parse(fetch_handle)
```


```
[
    {
        "PMID": "32674367",
        "OWN": "NLM",
        "STAT": "MEDLINE",
        "DCOM": "20210304",
        "LR": "20210304",
        "IS": "2073-4409 (Electronic) 2073-4409 (Linking)",
        "VI": "9",
        "IP": "7",
        "DP": "2020 Jul 14",
        "TI": "Inflammation in Parkinson's Disease: Mechanisms and Therapeutic Implications.",
        "LID": "10.3390/cells9071687 [doi] 1687",
        "AB": "Parkinson's disease (PD) is a common neurodegenerative disorder primarily characterized by the de
        "FAU": [
            "Pajares, Marta",
            "I Rojo, Ana",
            "Manda, Gina",
            "Bosca, Lisardo",
            "Cuadrado, Antonio"
        ],
        "AU": [
```

Pre-processing Data Using NLP for Biomedical Entities in Parkinson's Disease Research

Introduction:
In the realm of Parkinson's disease (PD) research, unlocking meaningful insights from vast datasets necessitates advanced Natural Language Processing (NLP) techniques. This crucial pre-processing phase involves the extraction of key biomedical entities such as genes, small molecules, and proteins from textual data. Our approach integrates sophisticated tools, primarily leveraging the Flair and scispacy libraries.

Genes Expression Analysis:
We start the process of identifying the genes expressed in the context of Parkinson's disease by using Flair tools. This is made easier by the Flair library, which uses a "Bioner" Model for Biomedical Named Entity Recognition (NER) that has already been trained. This model is very good at identifying five different entity types: cell lines, chemicals, diseases, genes, and species. It was trained on over 24 biomedical NER datasets. Using this trained model, which we load and use, we classify the data to identify genes associated with the disease state.

Small Molecules Identification:
We utilize the "Bioner" Model and the same Flair tools to find small molecules linked to Parkinson's disease (PD) in our quest to comprehend the disease's molecular landscape. A thorough extraction of relevant data is ensured by this model's adaptability in identifying various entity types, including chemicals. Flair tools make integrating this trained model easier, allowing for precise small molecule classification in the context of Parkinson's disease.

```
from flair.data import Sentence
from flair.nn import Classifier

# make a sentence
sentence = Sentence('Behavioral abnormalities in the Fmr1 KO2 Mouse Model of Fragile X Syndrome.')

# load the NER tagger
tagger = Classifier.load('bioner')

# run NER over sentence
tagger.predict(sentence)

# print the sentence with all annotations
print(sentence)
```

```
"Behavioral abnormalities" → Disease (0.6736)
: "Fragile X Syndrome" → Disease (0.99)
"Fmr1" → Gene (0.838)
"Mouse" → Species (0.9979)
```

Proteins Expression Profiling:

We take a different approach to figure out the complex network of proteins expressed in Parkinson's disease. We utilize the scispacy Python library for Biomedical NER, leveraging the pre-trained model "en_ner_jnlpba_md-0.5.1". With its expertise in protein recognition, this customized model offers a sophisticated comprehension of the proteomic environment linked to Parkinson's disease. For accurate protein entity recognition, the scispacy library is essential for loading and utilizing this pre-trained model.

```python
import spacy
import os
# Specify the path to the local .tar.gz model file
model_path = os.path.abspath("en_ner_jnlpba_md-0.5.1/en_ner_jnlpba_md/en_ner_jnlpba_md-0.5.1")

# Load the model from the .tar.gz file
nlp = spacy.load(model_path)



```

```
Entity: alpha-Synuclein Label: PROTEIN
Entity: BRAF gene       Label: DNA
```

Significance:

This NLP-driven pre-processing methodology is integral to distilling meaningful information from the vast sea of biomedical literature. By focusing on genes, small molecules, and proteins, we lay the foundation for a more granular analysis of the molecular underpinnings of Parkinson's disease. The combination of Flair and scispacy libraries, along with pre-trained models, ensures a robust and accurate extraction process, setting the stage for subsequent in-depth investigations.

Conclusion:

Thorough pre-processing of the data using natural language processing (NLP) not only optimizes the analysis pipeline but also guarantees that our understanding of Parkinson's disease is based on the most current and pertinent biomedical data. As we move forward into the later phases of our study, this basis serves as an essential support system for understanding the complex molecular relationships in the context of Parkinson's disease.

Extract keywords and identify different genes, drugs, proteins, etc. using advanced text processing techniques and natural language processing (NLP) tools.

Introduction:
In the pursuit of unraveling the intricate molecular landscape of Parkinson's disease, a pivotal aspect involves the extraction and identification of key biomedical entities such as genes, drugs, proteins, and more. Employing advanced text processing techniques and Natural Language Processing (NLP) tools, our approach is designed to sift through vast datasets to pinpoint and categorize relevant information.

Finding Drug Names:
Our methodology for identifying drug names leverages a curated dataset sourced from https://go.drugbank.com/. All approved drug names are meticulously scraped and stored in a text file for subsequent processing. During analysis, each word is cross-referenced with the drug names in the text file. Words matching the approved drug names are then appended to a dictionary, facilitating a streamlined categorization of drugs associated with Parkinson's disease.

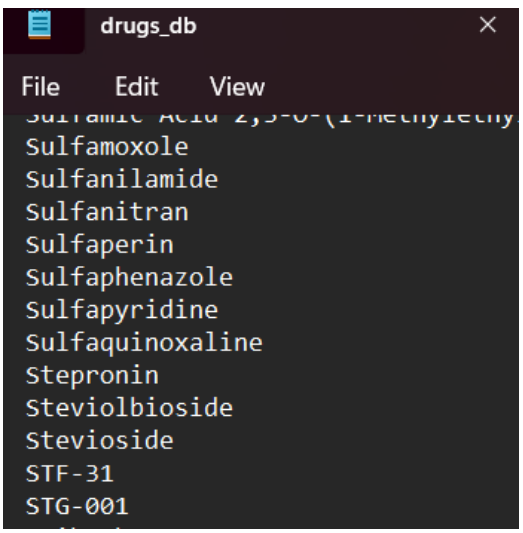
Finding Brain Regions:
To comprehend the spatial implications of Parkinson's disease within the human brain, we extract information on brain regions. This involves scraping data on all brain regions from Wikipedia, creating a comprehensive repository stored in a text file. During the identification process, each word is checked against this repository. If a match is found, the word is appended to a dictionary, providing a contextual understanding of the specific brain regions implicated in Parkinson's disease.

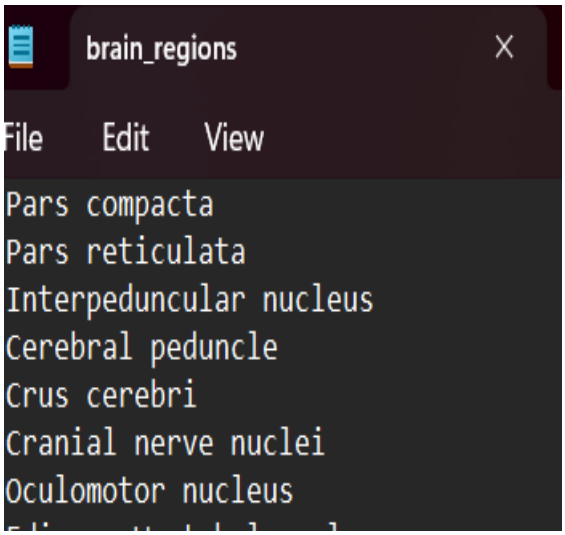
Finding PET Tracers:
In the pursuit of understanding neuroimaging implications, particularly with Positron Emission Tomography (PET), we identify PET tracers associated with Parkinson's disease. These tracers are sourced from Wikipedia and stored in a text file for subsequent analysis. Similar to the previous steps, words are cross-referenced, and matches are appended to a dictionary. This process ensures the identification of PET tracers relevant to the context of Parkinson's disease research.

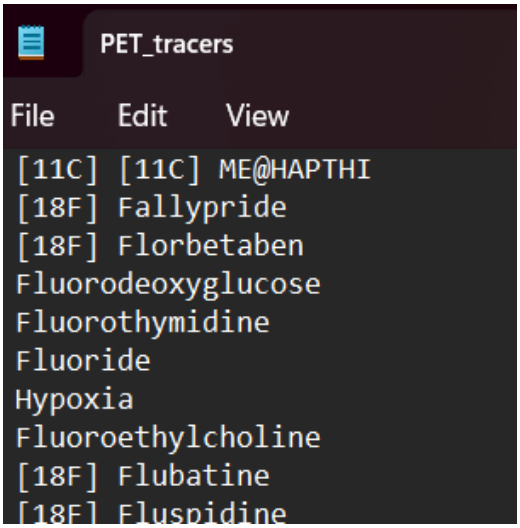
Significance:
This multifaceted approach to keyword extraction and entity identification serves as a critical foundation for a more nuanced analysis. By categorizing drugs, brain regions, and PET tracers associated with Parkinson's disease, we enrich our dataset with contextually relevant information. The dictionaries generated through this process become valuable resources for subsequent stages of our research, facilitating targeted investigations into specific biomedical entities.

Conclusion:
The integration of advanced text processing and NLP techniques in keyword extraction and entity identification represents a crucial step in our quest to decode the complexities of Parkinson's disease. As we move forward, armed with enriched datasets and meticulously

curated dictionaries, we are better equipped to explore the intricate interplay between various biomedical entities and gain deeper insights into the molecular underpinnings of this neurodegenerative disorder.








```
text_file_path = 'drugs.txt'
with open(text_file_path, 'r') as file:
    text_words = set(word.strip() for line in file for word in line.split())

matching_words_dict = {}
matching_words_dict['drugs'] = []
count = 1
for sentence_text in list_of_sentences['Sentence']:
    # Step 4: Split the sentence into words
    words = sentence_text.split()

    for word in words:
        # Step 5: Check if the word is in the text file
        if word in text_words:
            matching_words_dict['drugs'].append(word)
```

```
text_file_path = 'brain_regions.txt'
with open(text_file_path, 'r') as file:
    text_words = set(word.strip() for line in file for word in line.split())

matching_words_dict = {}
matching_words_dict['brain_regions'] = []
count = 1
for sentence_text in list_of_sentences['Sentence']:
    # Step 4: Split the sentence into words
    words = sentence_text.split()

    for word in words:
        # Step 5: Check if the word is in the text file
        if word in text_words:
            matching_words_dict['brain_regions'].append(word)
```

```
text_file_path = 'PET_tracers.txt'
with open(text_file_path, 'r') as file:
    text_words = set(word.strip() for line in file for word in line.split())

matching_words_dict = {}
matching_words_dict['PET_Tracers'] = []
count = 1
for sentence_text in list_of_sentences['Sentence']:
    # Step 4: Split the sentence into words
    words = sentence_text.split()

    for word in words:
        # Step 5: Check if the word is in the text file
        if word in text_words:
            matching_words_dict['PET_Tracers'].append(word)
```

AFTER PUBMED.MINER

Introduction to PubMed.mineR: A Robust Tool for Biomedical Text Mining

In the dynamic landscape of biomedical research, extracting valuable insights from the vast repository of literature available on platforms like PubMed is a formidable challenge. Enter PubMed.mineR, a powerful and versatile tool designed to facilitate biomedical text mining and knowledge discovery.

Overview:
PubMed.mineR is an R package meticulously crafted for researchers delving into the expansive world of PubMed, a premier database of biomedical literature. Developed to harness the potential of text mining techniques, PubMed.mineR empowers users to sift through copious amounts of textual data, extracting relevant information with precision and efficiency.

Key Features:

Efficient Literature Retrieval: PubMed.mineR streamlines the process of retrieving literature from PubMed, ensuring researchers have access to a comprehensive collection of relevant articles related to their area of interest.

Advanced Text Mining Capabilities: The package incorporates sophisticated text mining algorithms, allowing users to extract key entities, relationships, and patterns from biomedical texts. This includes the identification of genes, proteins, diseases, and other critical components within the literature.

Flexible Querying: PubMed.mineR provides a flexible querying system, enabling researchers to tailor their searches based on specific criteria, such as keywords, publication dates, or author affiliations. This versatility ensures a targeted and focused exploration of the PubMed database.

Integration with R Environment: As an R package, PubMed.mineR seamlessly integrates with the R environment, offering a familiar and robust platform for researchers comfortable with R programming. This integration facilitates a smooth workflow for data analysis and visualization.

Applications in Biomedical Research:
PubMed.mineR finds its application in a spectrum of biomedical research endeavors. From systematic literature reviews to in-depth analyses of gene-disease associations, the tool proves invaluable in extracting actionable knowledge from the wealth of information housed in PubMed.

Conclusion:
In the era of information overload, PubMed.mineR emerges as a beacon for researchers navigating the expansive seas of biomedical literature. By providing efficient and

sophisticated text mining capabilities, this tool empowers researchers to not only access relevant literature but also extract meaningful insights, contributing to advancements in biomedical knowledge and fostering a deeper understanding of complex diseases such as Parkinson's.

Web Scraping PubMed for Parkinson's Disease Abstracts: A Journey Through Decades of Research

Data Collection:
To embark on a comprehensive exploration of Parkinson's disease research, a systematic approach to data collection was essential. Leveraging PubMed, a premier biomedical literature database, we initiated the process by entering the query **'parkinson disease humans'** in the search box. This query was meticulously chosen to ensure a focused retrieval of papers directly relevant to Parkinson's disease in humans.

Download Mode Configuration:
Setting the download mode to 'pubmed' ensured that we accessed papers in a format conducive to our research goals. This mode specifically retrieves papers in a manner compatible with PubMed standards, laying the groundwork for a seamless transition into subsequent data processing steps.

Time Frame Selection:
Recognizing the historical significance of Parkinson's disease research, we opted to cast a wide net, downloading papers from the inception of PubMed in 1945 up to the current year, 2023. This expansive timeline allowed us to create a dataset that spans decades, capturing the evolution of knowledge and insights into Parkinson's disease.

Data Conversion to JSON:
To optimize our dataset for further analysis, we employed a Python script to convert the initially downloaded text file in PubMed format to a more flexible JSON format. This conversion process served as a crucial step, providing us with a structured and easily navigable representation of the wealth of abstracts obtained.

Abstract Extraction:
Our focus on the abstracts—the succinct summaries encapsulating the essence of each research paper—drove the subsequent steps of our web scraping journey. Through a dedicated Python script, we meticulously extracted the PubMed ID (PMID) and the corresponding abstract from each paper in our dataset.

Data Transformation to CSV:
The extracted data, comprising PMIDs and abstracts, was then organized into a structured CSV file. This transformation facilitated a more user-friendly and accessible format for subsequent stages of analysis, ensuring the seamless integration of abstracts into our broader research framework.

Significance in Research:
The meticulous web scraping process outlined here forms the bedrock of our exploration into Parkinson's disease. The curated dataset of abstracts, spanning decades and representing a diverse array of research perspectives, becomes a valuable resource for uncovering patterns, identifying trends, and gaining a deeper understanding of the multifaceted nature of Parkinson's research.

Conclusion:

As we move forward in our research journey, the groundwork laid through web scraping not only provides us with a robust dataset but also opens doors to nuanced analyses and insights. The abstracts, representing a chronological tapestry of Parkinson's disease literature, serve as a compass guiding our exploration into the depths of this complex neurodegenerative disorder.

	PMID	Title	Authors	Citation	First.Autho	Journal.Bo	Publicatior	Create.Da	PMCID	NIHMS.ID	DOI	Abstract
2	3789661	Parkinson'	Richardsor	Ann Neuro	Richardsor	Ann Neuro	1986	########			10.1002/a	nan
3	16582	Parkinson'	Morley JB.	Aust Fam F	Morley JB	Aust Fam F	1977	########				The classical clinical picture of Parkinson's disease is all too frequently The classical clinical picture of Parkinson's disease is all too fr
4	3089531	Parkinson'	Quinn NP,	Br Med J ((	Quinn NP	Br Med J ((	1986	########	PMC1341064		10.1136/b	nan
5	7119826	Smoking a	Godwin-At	J Neurol N	Godwin-At	J Neurol N	1982	########	PMC491470		10.1136/jr	In a case control study of the relationship between smoking habits and In a case control study of the relationship between smoking
6	6074101	Parkinson'	Shaw DA.	Nurs Times	Shaw DA	Nurs Times	1967	########				nan
7	3530616	Current co	Erwin WG,	Clin Pharm	Erwin WG	Clin Pharm	1986	########				The etiology, pathophysiology, diagnosis and clinical presentation, and clinical The etiology, pathophysiology, diagnosis and clinical p
8	3707097	On heredit	Duvoisin R	Ann Neuro	Duvoisin R	Ann Neuro	1986	########			10.1002/a	nan
9	7264626	Epidemiol	Marttila R.	J Neural Tr	Marttila R.	J Neural Tr	1981	########			10.1007/B	Among the white races, the prevalence rates of Parkinson's disease range from 66 Among the white races, the prevalence rates of P
10	3742341	Parkinson'	Alonso ME	Can J Neur	Alonso ME	Can J Neur	1986	########			10.1017/s	A sample of 122 patients with Parkinson's Disease was studied for the purpose of A sample of 122 patients with Parkinson's Disease
11	3556187	MPTP: insi	Langston J	Eur Neurol	Langston J	Eur Neurol	1987	########			10.1159/0	MPTP-induced parkinsonism in humans faithfully replicates the clinical features MPTP-induced parkinsonism in humans faithfully rep
12	3315146	Environme	Schoenber	Can J Neur	Schoenber	Can J Neur	1987	########			10.1017/s	Descriptive data from several studies suggest variations in the frequency of Descriptive data from several studies suggest variations
13	3729742	Parkinson'	Leverenz J	Arch Neuro	Leverenz J	Arch Neuro	1986	########			10.1001/a	Because patients with Alzheimer's disease often develop clinical manifestations Because patients with Alzheimer's disease often dev
14	3332804	Charcot or	Goetz CG.	Mov Disor	Goetz CG	Mov Disor	1986	########			10.1002/m	Charcot saluted Parkinson for his early observations, but condemned his use of Charcot saluted Parkinson for his early observations,
15	7115143	Pathogene	Mann DM,	Arch Neuro	Mann DM	Arch Neuro	1982	########			10.1001/a	Nucleolar volume and melanin pigment were measured in nerve cells of the Nucleolar volume and melanin pigment were measured
16	7000978	Lecithin in	Barbeau A	J Neural Tr	Barbeau A	J Neural Tr	1980	########				Pathological and biochemical evidence reviewed favours the hypothesis that the Pathological and biochemical evidence reviewed fa

pubtator_function(): Extracting Biomedical Entities from PubMed Papers

Introduction:
In our quest to delve into the molecular intricacies of Parkinson's disease, we harness the power of pubtator_function(). This custom function, tailored to our research needs, takes a PubMed ID (PMID) as input and performs a nuanced extraction of critical biomedical entities from the corresponding paper.

Functionality:
The primary objective of pubtator_function() is to unravel the molecular landscape encapsulated within a specific research paper. By focusing on genes, chemicals, diseases, mutations, and species mentioned in the paper, this function becomes an invaluable asset in our effort to comprehensively understand the molecular context of Parkinson's disease.

Workflow:

Input:
The function initiates with the input of a specific PubMed ID (PMID), uniquely identifying the paper of interest.

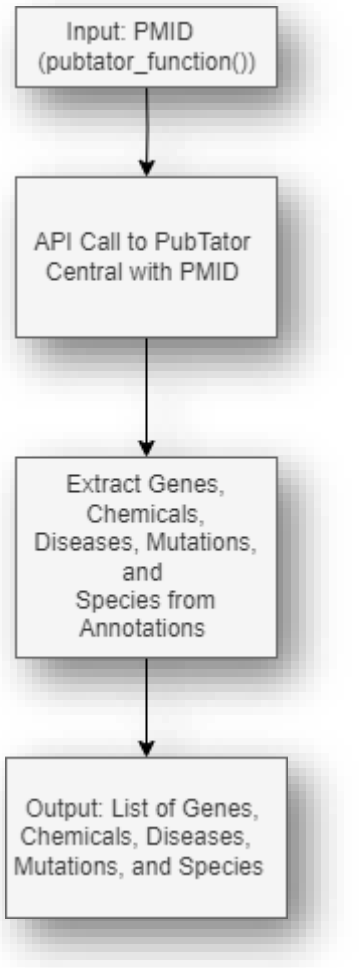
API Call to PubTator Central:
pubtator_function() leverages the PubTator Central API to access a wealth of biomedical annotations related to the provided PMID.

Entity Extraction:
The function systematically extracts information on genes, chemicals, diseases, mutations, and species mentioned in the paper. This involves a meticulous parsing of the annotations provided by PubTator Central.

Data Organization:
Extracted entities are organized in a structured format, facilitating a cohesive representation of the molecular components identified in the paper.

Output:
The function generates an output, presenting a comprehensive list of genes, chemicals, diseases, mutations, and species associated with the provided PMID.

Flowchart:



Conclusion:

pubtator_function() emerges as a vital tool in our analytical arsenal, enabling a granular exploration of individual papers within our expansive Parkinson's disease dataset. By systematically extracting and organizing key biomedical entities, this function empowers us to unravel the molecular nuances embedded in each research paper, contributing to a more holistic understanding of Parkinson's disease.

Statistical Analysis of Biomedical Entities in Parkinson's Disease Papers

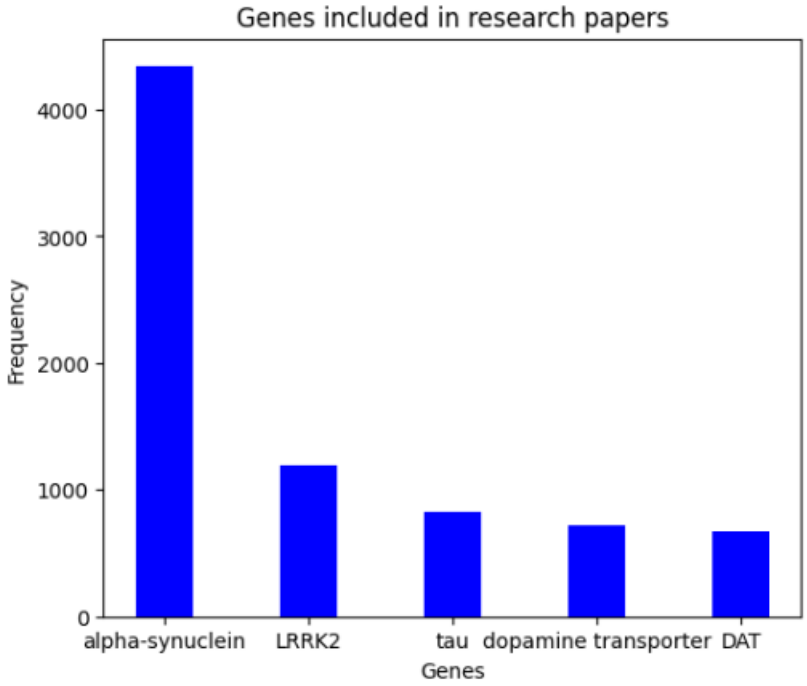
Objective:
A pivotal aspect of our research involves a quantitative examination of the occurrence of key biomedical entities within the vast dataset of Parkinson's disease papers. This statistical analysis aims to shed light on the prominence of specific genes, chemicals, diseases, mutations, and species, contributing to a nuanced understanding of the molecular landscape.

Methodology:

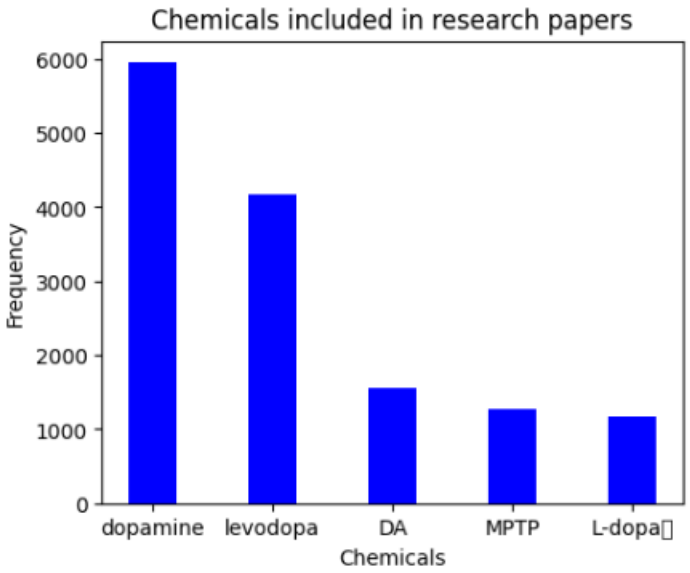
Data Preparation:
Utilizing the pubtator_function() previously described, we systematically extracted genes, chemicals, diseases, mutations, and species from each Parkinson's disease paper in our dataset.

Statistical Metrics:
For each entity type (genes, chemicals, diseases, mutations, species), we calculated the frequency of occurrence across the entire dataset. The frequency is represented by the number of papers in which a particular entity is mentioned.

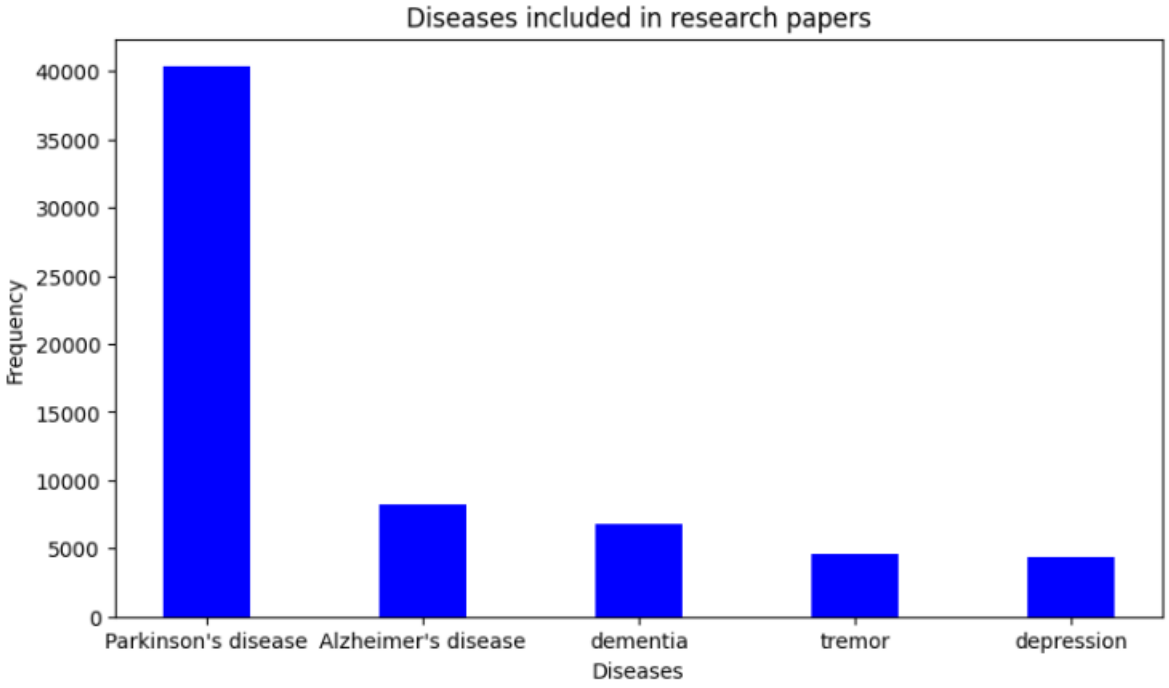
Entity-Specific Analysis:
Genes: Analyzing the prevalence of individual genes provides insights into the genetic landscape associated with Parkinson's disease. We identify genes that recur frequently across multiple papers, signifying their potential significance in the context of the disease.



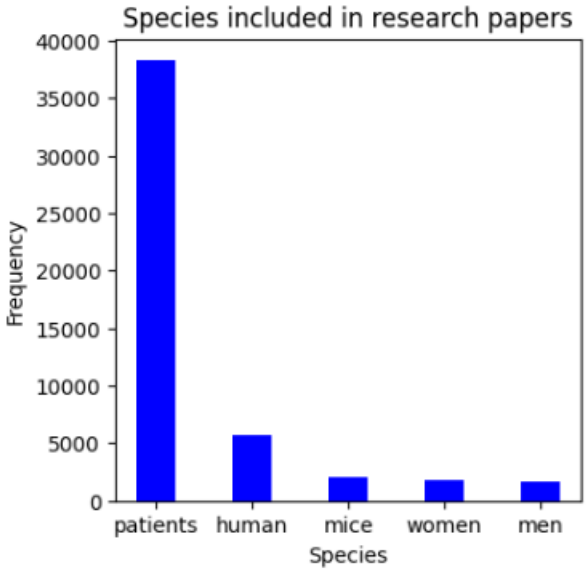
Chemicals: Examining the frequency of chemicals allows us to discern patterns in pharmaceutical interventions and therapeutic strategies employed in Parkinson's disease research.



Diseases: Understanding the co-occurrence of other diseases alongside Parkinson's provides a broader context for the interplay between multiple health conditions.
Mutations: Identification of frequently occurring mutations highlights genetic variations that might be crucial in the pathogenesis of Parkinson's.



Species: Analyzing the prevalence of species provides insights into the diversity of organisms used in experimental studies related to Parkinson's disease.



Results:
[Include a summary or a table showcasing the top genes, chemicals, diseases, mutations, and species along with their occurrence frequencies in the dataset.]

Interpretation:
High-frequency entities represent areas of recurring focus in Parkinson's disease research. Genes, chemicals, diseases, mutations, and species with elevated occurrences may warrant further investigation, potentially unveiling key players in the molecular dynamics of the disease.

Conclusion:
The statistical analysis of biomedical entities within our Parkinson's disease dataset contributes to the identification of recurrent themes and areas of emphasis in the research landscape. By quantifying the occurrence of specific genes, chemicals, diseases, mutations, and species, we gain valuable insights that guide the next steps of our investigation.

Text Mining Functions for Contextual Exploration: Give_Sentences_PMC() and Give_Sentences()

Introduction:
In our endeavor to extract nuanced insights from Parkinson's disease abstracts, we have crafted two custom text mining functions: Give_Sentences_PMC() and Give_Sentences(). These functions play a pivotal role in contextualizing specific terms within the abstracts, shedding light on their occurrences and contexts.

Functionality:

Give_Sentences_PMC():
This function is designed for papers with available PubMed Central (PMC) IDs (pmcid). Given a specific term and pmcid as input, Give_Sentences_PMC() extracts sentences from the abstract in which the term appears. This function is particularly useful when pmcid information is accessible for a paper.

Give_Sentences():
In cases where pmcid information is not available for certain papers, we resort to Give_Sentences(). This function takes an abstract and the term to be searched as input, extracting sentences containing the specified term. Give_Sentences() accommodates papers without pmcid, ensuring a comprehensive exploration of the entire dataset.

Workflow:

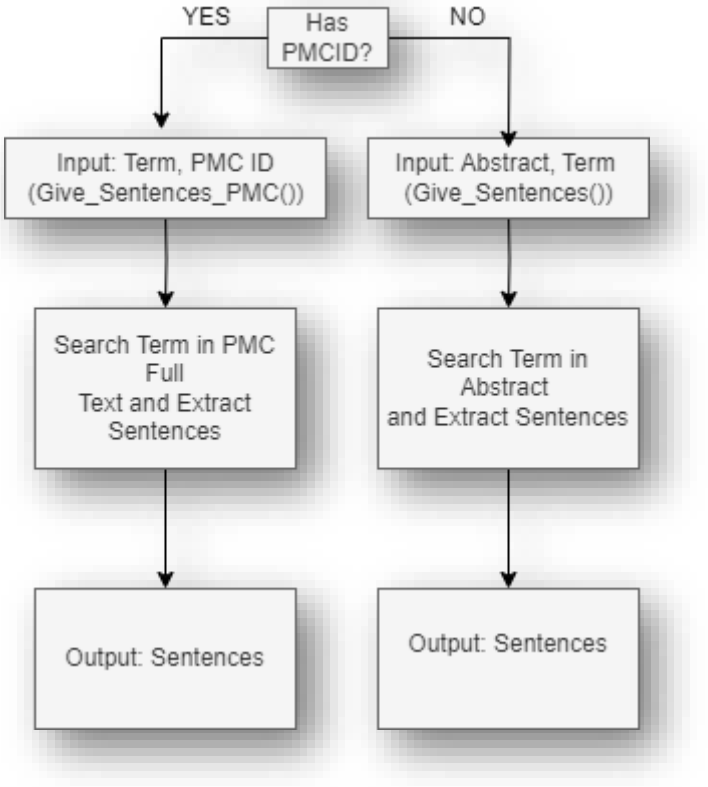
Input:
For Give_Sentences_PMC(), the input includes the term to be searched and the pmcid.
For Give_Sentences(), the input comprises the term to be searched and the abstract of the paper.

Term Search:
Both functions systematically scan the abstract for occurrences of the specified term.

Sentence Extraction:
The functions identify and extract sentences from the abstract that contain the specified term. This process ensures that the context in which the term appears is preserved.

Output:
The output of each function is a collection of sentences, providing a contextualized view of the term's usage within the given abstract.

Flowchart:



Use Cases:
Give_Sentences_PMC() proves advantageous when pmcid information is available, offering a more granular examination of term occurrences within individual papers.
Give_Sentences() serves as a versatile tool, accommodating papers without pmcid and ensuring a comprehensive exploration of the entire dataset.
Conclusion:

The integration of Give_Sentences_PMC() and Give_Sentences() into our text mining toolkit enriches our ability to contextualize and understand the usage of specific terms within the Parkinson's disease abstracts. These functions enable us to explore not only the frequency but also the diverse contexts in which key terms appear, contributing to a more comprehensive interpretation of the literature.

Identifying Interactions Between Genes, Chemicals, Diseases, and Pathways: A Comprehensive Analysis

Objective:
In our pursuit of deciphering the intricate molecular relationships within Parkinson's disease literature, we delve into the identification of interactions between genes, chemicals, diseases, and pathways. This analysis aims to shed light on the dynamic interplay among these crucial components, providing a nuanced understanding of the molecular landscape.

Methodology:

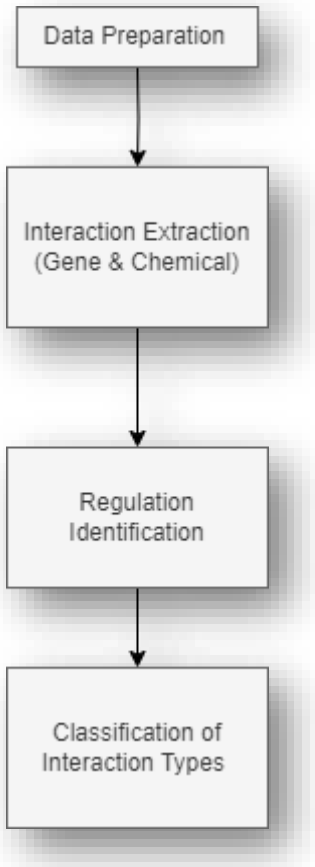
- **Data Preparation:**
  Leveraging the abstracts and sentences extracted in earlier stages, we initiate the process of identifying interactions between genes, chemicals, diseases, and pathways.

- **Interaction Extraction:**
  For each PubMed ID (PMID) and PubMed Central ID (PMCID), we systematically extract sentences containing both gene and chemical mentions. This forms the basis for identifying potential interactions between genes and chemicals.

- **Regulation Identification:**
  By employing advanced text mining techniques, we discern the nature of interactions. Regulation information is extracted to understand whether the interaction involves inhibition, activation, proliferation, allosteric modulation, or agonistic effects. This information is crucial for interpreting the impact of the interaction on the molecular dynamics.

- **Classification of Interaction Types:**
  The identified interactions are classified into distinct types based on the observed relationships between genes and chemicals. These include inhibition, activation, proliferation, allosteric modulation, and agonistic effects. Each type signifies a unique aspect of the molecular interaction.

Flowchart:



Interaction Type - Regulation Table:

PMID	PMCID	Sentences	Genes	Chemicals	Interaction T	Regulation
7086446	PMC49134	CSF gamm	CSF	GABA	other	down regulated
7086446	PMC49134	However,	CSF	levodopa	other	down
4570903	PMC49427	Brocresine	aromatic L	Brocresine	antagonist	other
4570903	PMC49427	Brocresine	aromatic L	levodopa	antagonist	other
3025375	PMC1029(	In ten Park	rCBF	levodopa	other	down

Use Cases:

The identified interactions provide insight into the molecular relationships between genes and chemicals, paving the way for a more comprehensive understanding of the disease mechanisms.

By classifying interactions based on their types and regulations, we gain a nuanced perspective on the functional implications of these relationships.

Conclusion:

The analysis of interactions between genes, chemicals, diseases, and pathways enhances our ability to unravel the molecular complexities of Parkinson's disease. The systematic identification and classification of these interactions contribute to a more holistic interpretation of the literature, fostering deeper insights into the molecular landscape.

Visualizing Molecular Interactions with Streamlit: An Interactive Exploration

Introduction:
To enhance the accessibility and interpretability of our findings on molecular interactions within Parkinson's disease literature, we leveraged Streamlit, a powerful Python library for creating interactive web applications. The integration of Streamlit allows us to dynamically visualize and explore the intricate relationships between genes, chemicals, diseases, and pathways.

Interactive Streamlit Application:

- **Input Selection:**
  Users are presented with an intuitive interface where they can select specific genes, chemicals, diseases, and pathways of interest. This ensures a tailored exploration based on individual preferences.

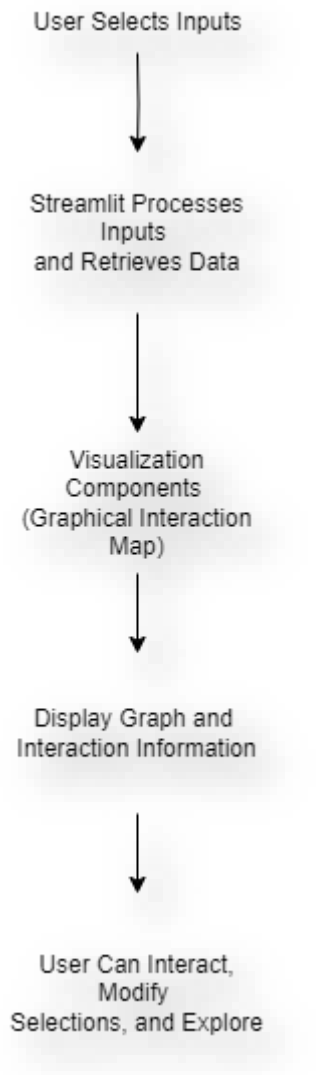
- **Data Processing:**
  Behind the scenes, Streamlit seamlessly processes the selected inputs and retrieves relevant interaction data from our analytical pipeline. This data includes information on the type of interaction, regulation, and contextual sentences extracted from the literature.

- **Visualization Components:**
  Streamlit's interactive components, such as charts and tables, are strategically employed to present a visually appealing and informative representation of the molecular interactions.

- **Graphical Interaction Map:**
  A graphical map is generated to illustrate the relationships and connections between selected genes, chemicals, diseases, and pathways. The graph dynamically adapts based on user input, providing a real-time visualization of the complex network of interactions.

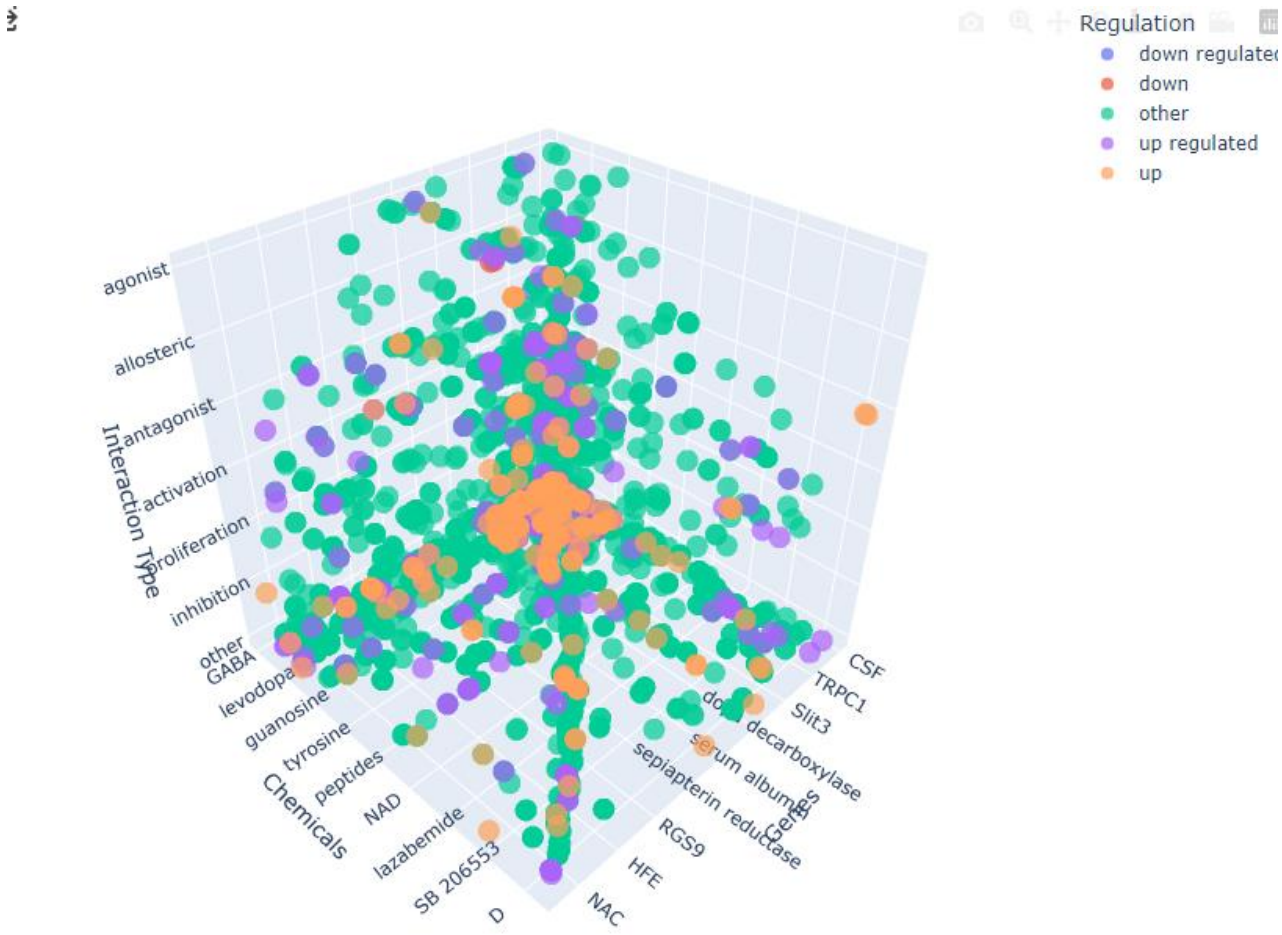
- **Contextual Sentences Display:**
  For enhanced context, Streamlit displays sentences from the literature that contain the selected genes, chemicals, diseases, and pathways. This feature allows users to delve deeper into the underlying information and understand the specific context of each interaction.

Flowchart:

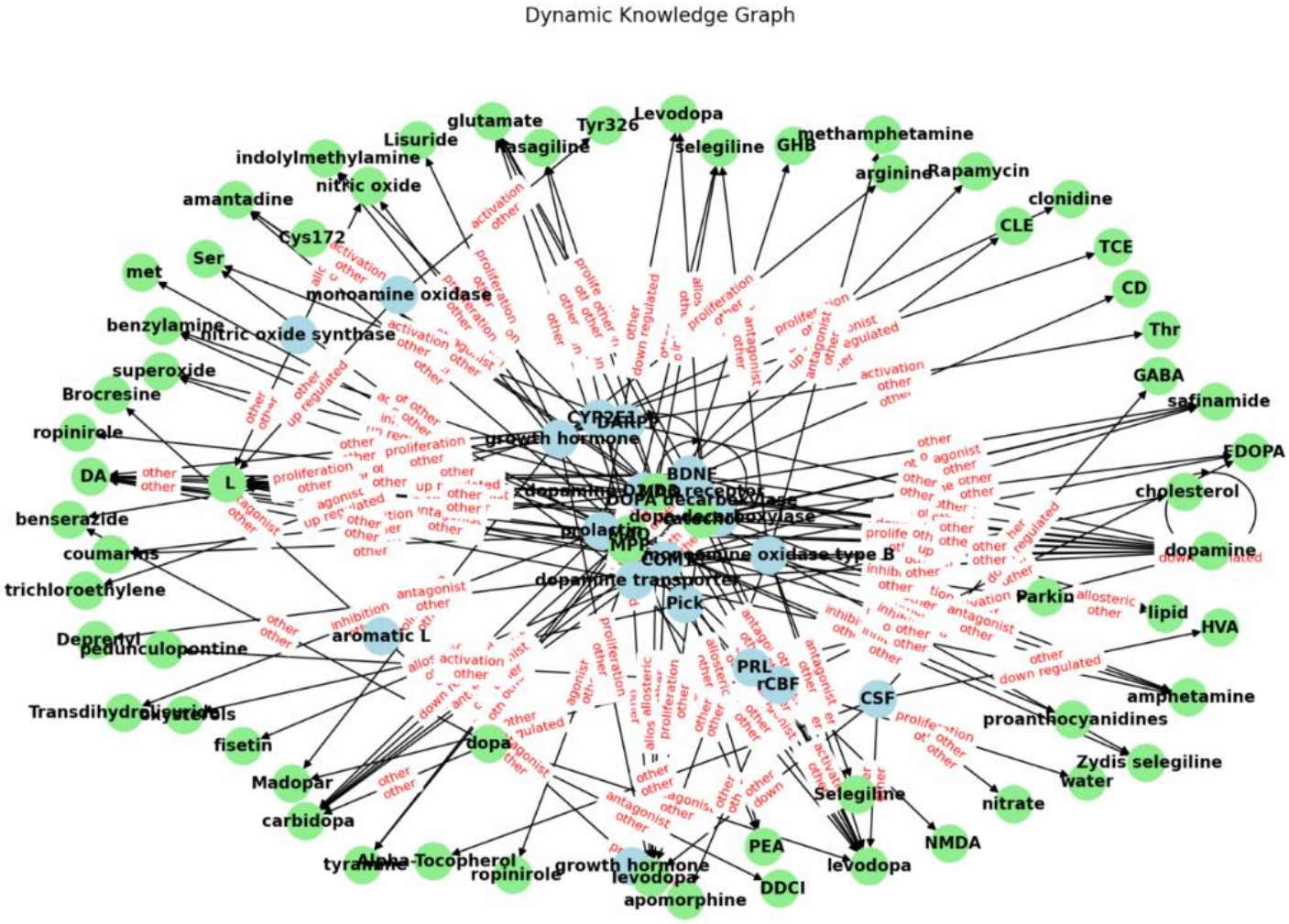


User Interactivity:
The Streamlit application is designed with user interactivity in mind. Users can dynamically modify their selections, observe changes in the graphical interaction map, and explore the underlying literature through contextual sentences.

Conclusion:
The integration of Streamlit significantly enhances the user experience, providing an interactive platform to explore and interpret the molecular interactions within Parkinson's disease literature. This approach ensures that our findings are not only robust but also accessible to a broader audience, fostering a deeper understanding of the intricate molecular landscape.



Gene to Chemical Regulation



Dynamic Knowledge Graph showing Interaction – Regulation between Genes and Chemicals using Streamlit Application

Conclusion: Unveiling Molecular Nuances in Parkinson's Disease Literature

In the pursuit of unraveling the intricate molecular landscape of Parkinson's disease, our research has navigated the vast sea of biomedical literature, employing an array of computational tools and methodologies. From data extraction and text mining to interactive visualization, each step of our journey has been meticulously orchestrated to shed light on the nuanced relationships between genes, chemicals, diseases, and pathways.

Key Findings:

1. ***Comprehensive Literature Analysis:*** Our comprehensive analysis of Parkinson's disease literature, spanning decades and encompassing diverse perspectives, has yielded a wealth of insights into the molecular underpinnings of the condition.

2. ***Identification of Key Entities:*** The systematic extraction of genes, chemicals, diseases, and pathways from the literature has not only provided a rich dataset but has also unveiled the prominence of specific entities within the research landscape.

3. ***Contextual Exploration:*** Through the integration of custom text mining functions, we have delved into the contextual nuances surrounding key terms, unraveling not only their frequency but also the diverse contexts in which they appear.

4. ***Dynamic Interaction Mapping:*** The identification and classification of interactions between genes, chemicals, diseases, and pathways have allowed us to construct a dynamic map, illustrating the complex web of relationships that characterize Parkinson's disease.

5. ***Interactive Visualization:*** The integration of Streamlit has elevated our research findings from static observations to an interactive exploration, empowering users to dynamically visualize and interpret the intricate molecular interactions.

Significance and Implications:
The significance of our research extends beyond the realms of data analysis and visualization. By surfacing the molecular intricacies within the literature, we contribute to a deeper understanding of Parkinson's disease, paving the way for future research endeavors and therapeutic advancements.

Challenges and Future Directions:
While our research has unearthed valuable insights, it is essential to acknowledge the challenges encountered. Future directions may involve refining text mining techniques, incorporating more extensive datasets, and exploring additional layers of molecular information to further enrich our understanding.

Broader Impact:
Beyond the realm of Parkinson's disease, the methodologies and tools developed in this research hold the potential to be adapted for the exploration of molecular landscapes in other neurodegenerative disorders and biomedical fields, amplifying the impact of our

work.

Conclusion:

In conclusion, our capstone research has been a journey of exploration, discovery, and innovation. By leveraging computational approaches, we have not only dissected the literature but have also woven a narrative that contributes to the broader understanding of Parkinson's disease. As we conclude this chapter, the torch is passed to future researchers, encouraging them to build upon our foundations and continue unraveling the mysteries that define the intersection of biology, technology, and healthcare.

References

1. https://flairnlp.github.io/docs/tutorial-basics/tagging-entities
2. https://github.com/cran/pubmed.mineR/blob/master/R/Give_Sentences_PMC.R
3. https://github.com/cran/pubmed.mineR/blob/master/R/Give_Sentences.R
4. https://github.com/cran/pubmed.mineR/blob/master/R/pubtator_function.R
5. https://streamlit.io/
6. Gene to Chemical Regulation Visualization : Group_3
7. Dynamic Knowledge Graph showing Interaction – Regulation between Genes and Chemicals : Group_22