SCOPE & APPROACH

DS-670-PHY: Capstone Project in Big Data & Business Analytics

Project plan and management for the selected topic:

Disease State Conversation and Sentiment Analysis: Examination of Digital Community Conversations Within Specific Disease States Via Reddit

~Data Processing & Exploratory Data Analysis~

Week II: Assignment

Due 10/02/2021 – Slide Presentation: 9/29/2021

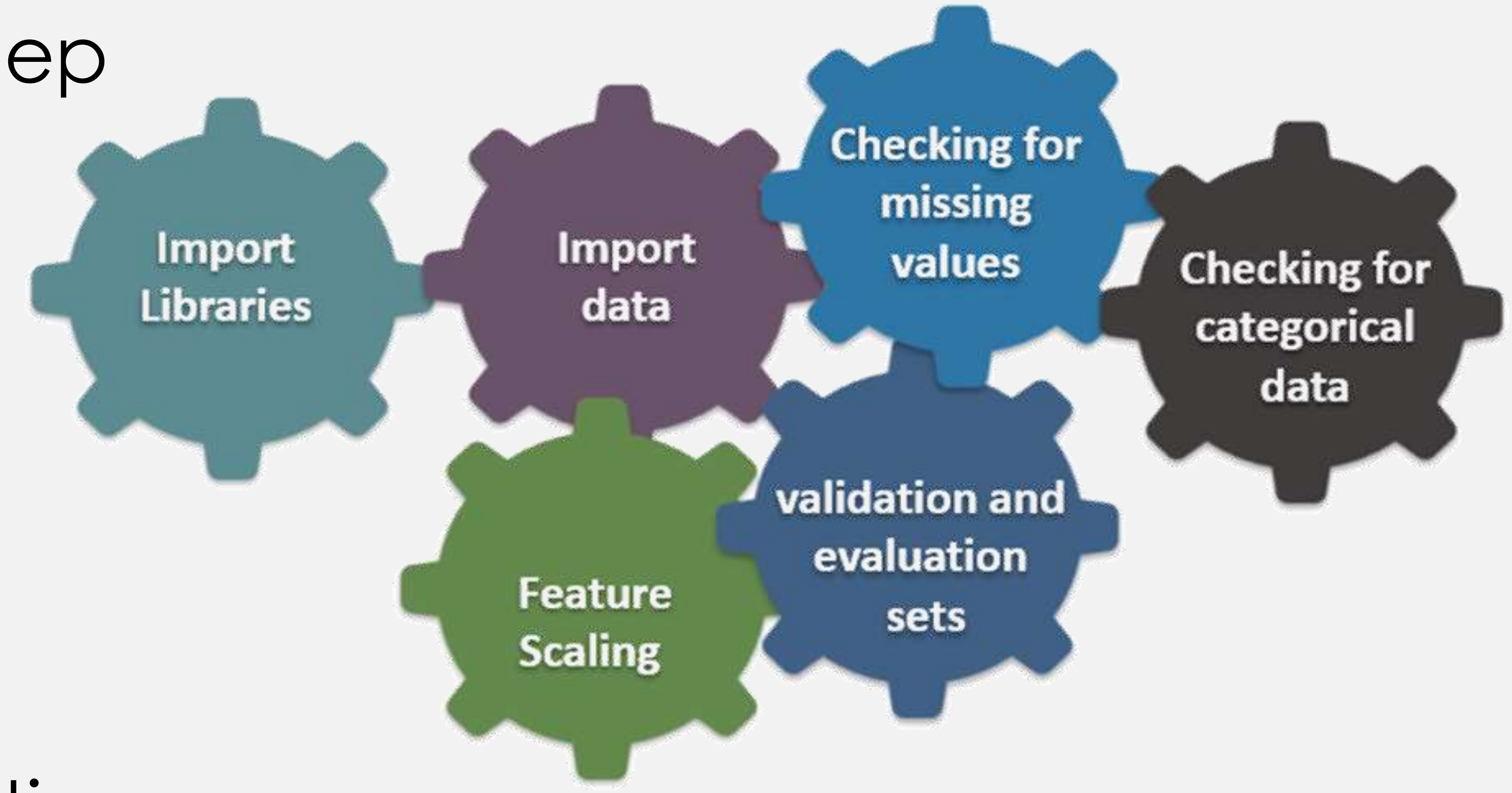
- For the in class presentation 09/29 prepare your progress report, **including tables**, **graphs**, **maps, pictures, code and other material** relevant to the assignment.
- Make sure to add **captions to your graphical output** and to draw **the conclusions / insights** into the next steps of your research.

- Upload your homework assignment in Blackboard no later than Sunday midnight, 10/03.

Introduction: What's Next?

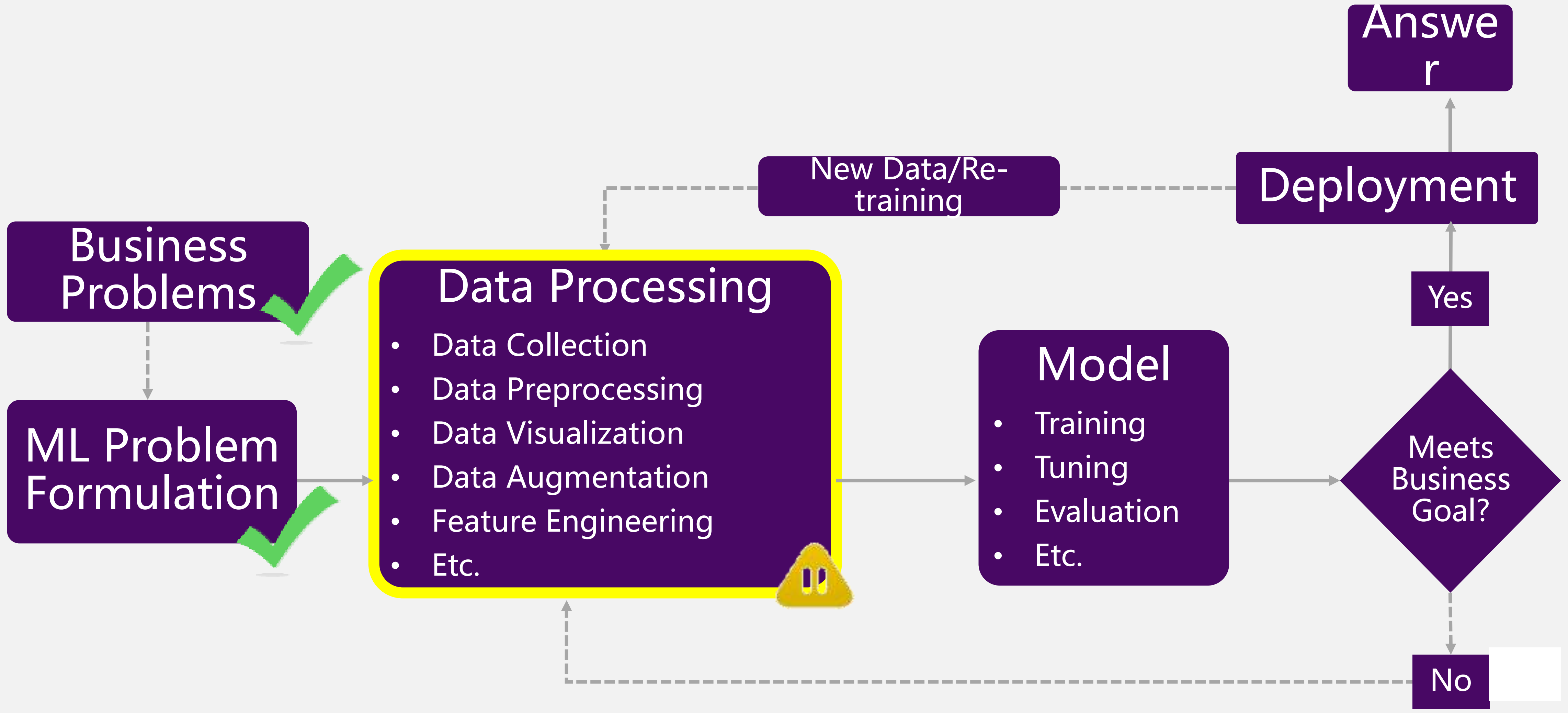
Why data processing for Reddit API?

- Data processing involves converting the data into a more meaningful format which can help the machine learning or deep learning algorithms to perform better.
- The following are the general steps involved:
  - Data Cleaning – Helps in removing inconsistent data and dealing with empty or null values
  - Data Integration – Involves combining data from different sources
  - Data Transformation – Data is transformed using different techniques like normalization, aggregation and generalization


Import Libraries
Import data
Checking for missing values
Checking for categorical data
Feature Scaling
validation and evaluation sets

- The above mentioned steps can be achieved using different techniques like converting into lower case/ upper case, tokenization, removal of punctuations or special characters and stop words, stemming and lemmatization.

Where Are We in the Project Lifecycle?



Business Problems ✓

ML Problem Formulation ✓

Data Processing

- Data Collection
- Data Preprocessing
- Data Visualization
- Data Augmentation
- Feature Engineering
- Etc.

Model

- Training
- Tuning
- Evaluation
- Etc.

Meets Business Goal?

Yes

No

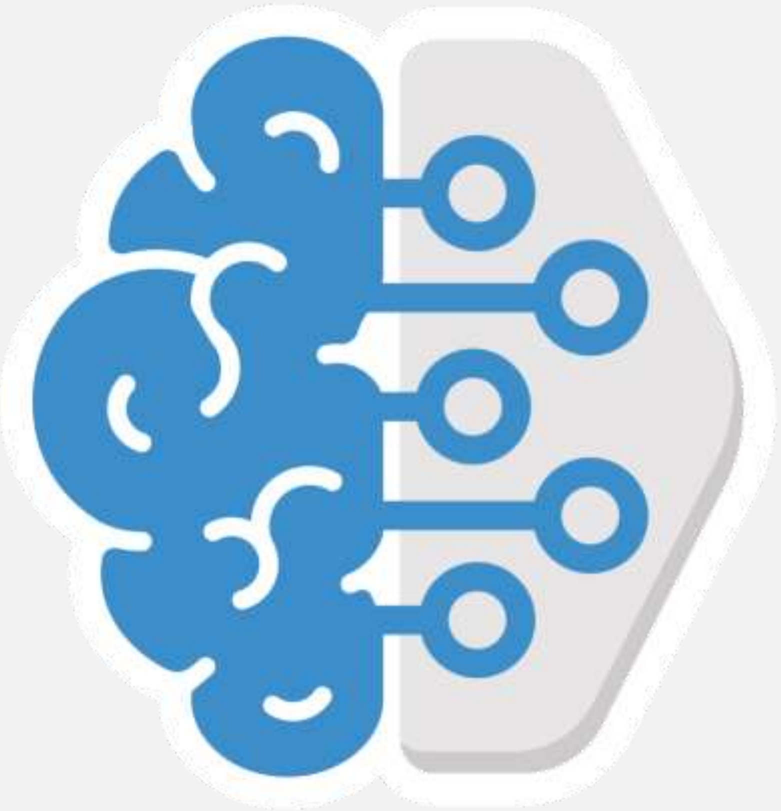
Deployment

New Data/Re-training

Answer

Summarized Problem Statement

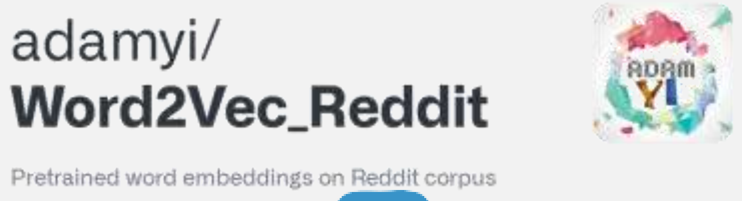
Cognitive Service/ Text Analytics/ Sentiment Analysis/ Key Phrase Extraction



Deep Learning/Machine learning:

- NLP

-



adamyi/
Word2Vec_Reddit

Pretrained word embeddings on Reddit corpus

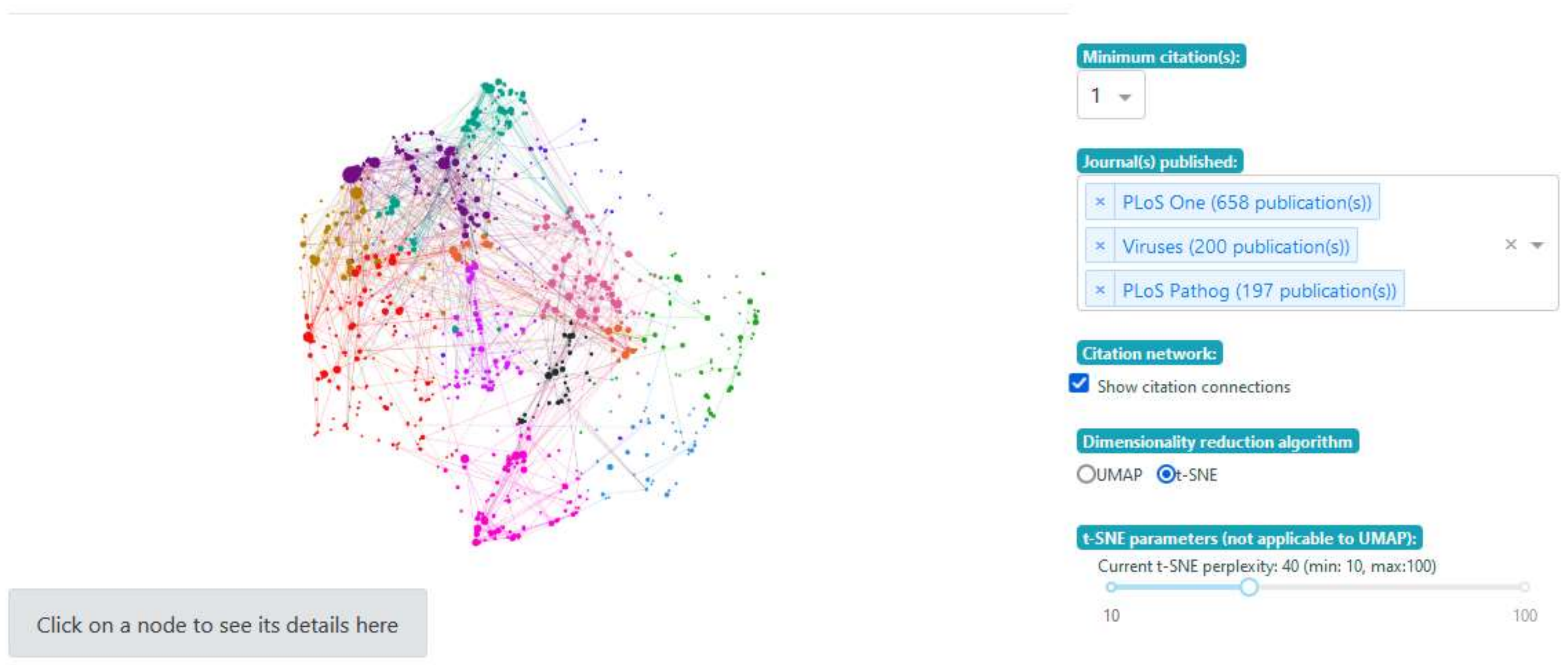
PatientCom (Public Health Trend Collector):
Hybrid Content-Based Collaborative Corpus Filtering Engine with Web Scraping and Sentiment Analysis

Vision

Problem type	Description
Topic Ranking	Helping users find the most trendy subjects on diseases
Recommendation Engine	Giving users the thing they may be most interested in using **NLP** and **LSTM**
Class imbalance (Down Sampling)	Reduce the size of the dominant or frequent class(es).
Reddit Submission Corpus	publicly available Reddit submissions (42 GB)
Disambiguation of words	To understand public reaction we use **LDA**
Sentiment Analysis	Understand the users with Brandwatch's Sentiment Analysis Tool such as **GPT-3**

Final Plotly dash app

PatientCom (Public Health Trend Collector): Hybrid Content-Based Collaborative Corpus Filtering Engine with Web Scraping and Sentiment Analysis



Minimum citation(s): 1

Journal(s) published:
- PLoS One (658 publication(s))
- Viruses (200 publication(s))
- PLoS Pathog (197 publication(s))

Citation network:
☑ Show citation connections

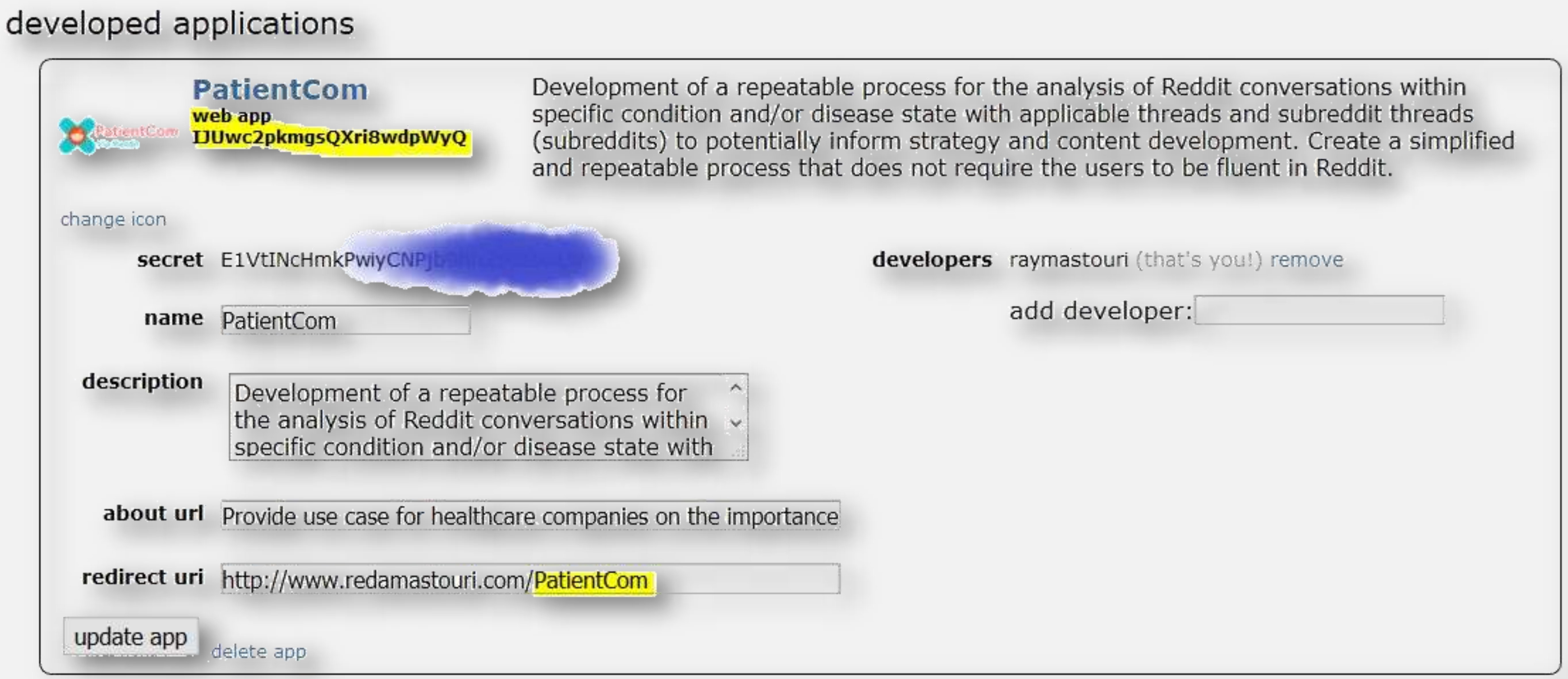
Dimensionality reduction algorithm
○UMAP ●t-SNE

t-SNE parameters (not applicable to UMAP):
Current t-SNE perplexity: 40 (min: 10, max:100)

Data Set

Creating an app on Reddit Developer Account

- Go to App Preferences and click create another app… at the bottom.
- Fill out the required details, make sure to select script — and click create app.
- NB: Let's make a note of the personal use script and secret tokens
- Request a temporary OAuth token from Reddit
- Add headers=headers to every request.
- Subscribing in most of the health related subreddits such as disease, breast cancer, prostate cancer, etc.



developed applications

PatientCom
web app
IJUwc2pkmgsQXri8wdpWyQ

Development of a repeatable process for the analysis of Reddit conversations within specific condition and/or disease state with applicable threads and subreddit threads (subreddits) to potentially inform strategy and content development. Create a simplified and repeatable process that does not require the users to be fluent in Reddit.

change icon

secret E1VtINcHmkPwiy[redacted]

developers raymastouri (that's you!) remove

add developer:

name PatientCom

description Development of a repeatable process for the analysis of Reddit conversations within specific condition and/or disease state with

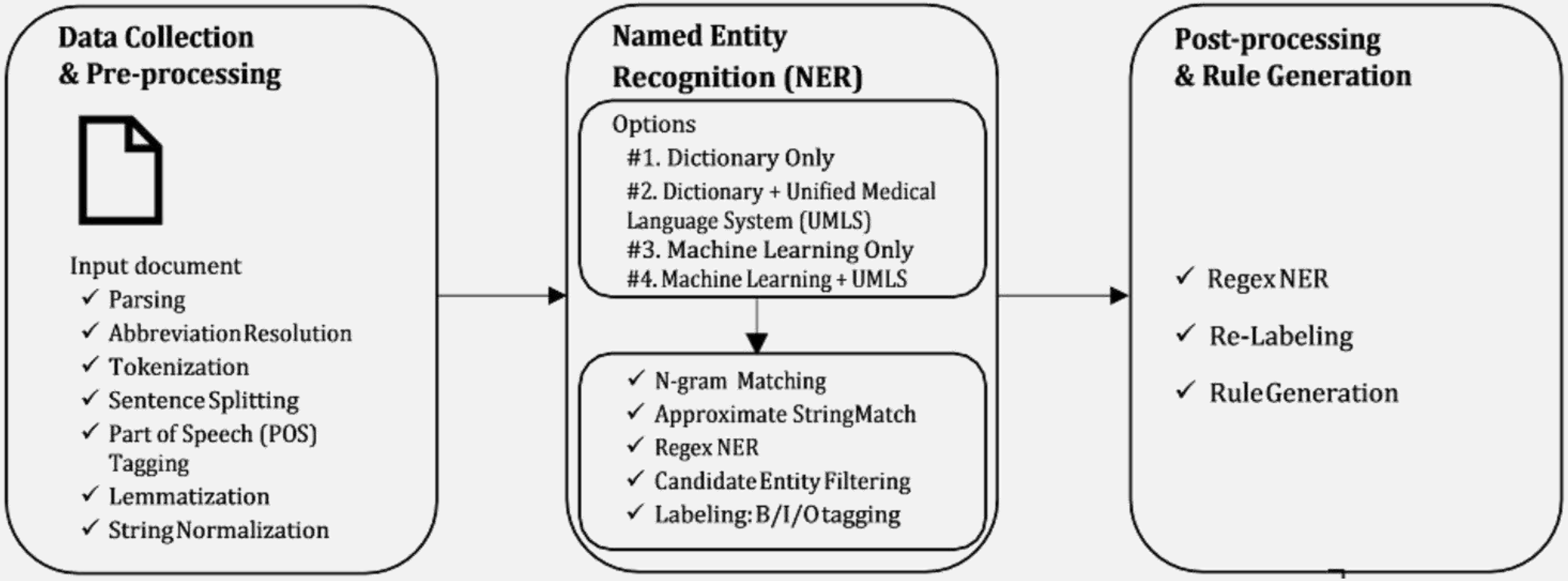
about url Provide use case for healthcare companies on the importance

redirect uri http://www.redamastouri.com/PatientCom

update app delete app

Methodology

What to implement at this stage:



Data Collection & Pre-processing

Input document

- ✓ Parsing
- ✓ Abbreviation Resolution
- ✓ Tokenization
- ✓ Sentence Splitting
- ✓ Part of Speech (POS) Tagging
- ✓ Lemmatization
- ✓ String Normalization

Named Entity Recognition (NER)

Options

#1. Dictionary Only
#2. Dictionary + Unified Medical Language System (UMLS)
#3. Machine Learning Only
#4. Machine Learning + UMLS

- ✓ N-gram Matching
- ✓ Approximate String Match
- ✓ Regex NER
- ✓ Candidate Entity Filtering
- ✓ Labeling: B/I/O tagging

Post-processing & Rule Generation

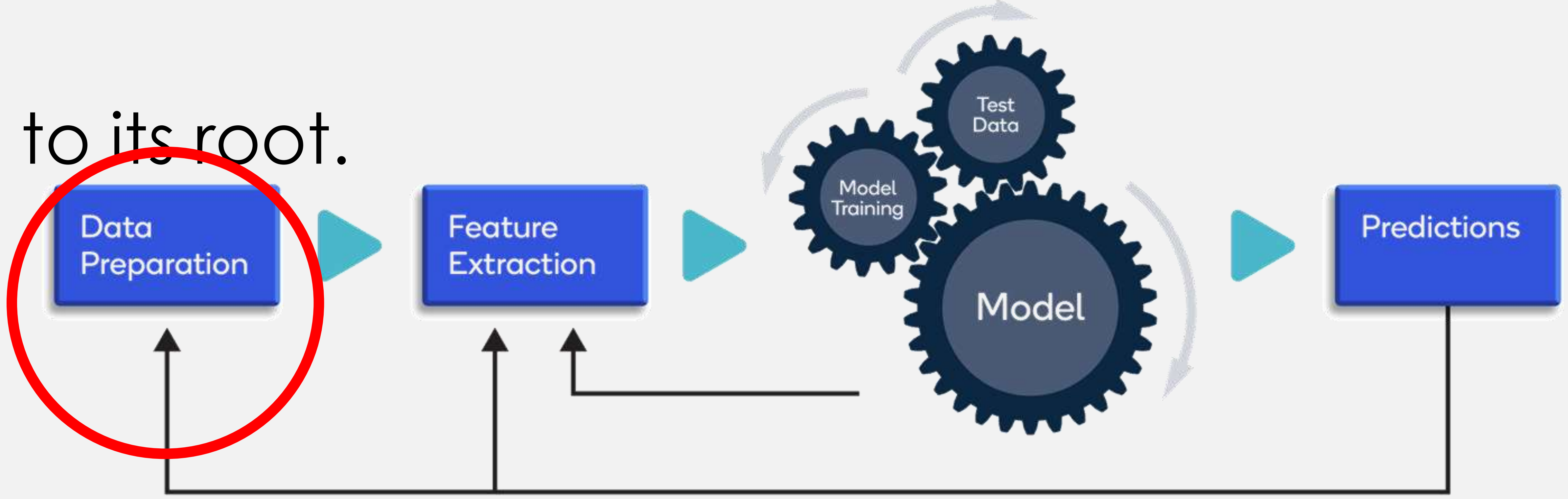
- ✓ Regex NER
- ✓ Re-Labeling
- ✓ Rule Generation

Data Preparation

Steps to data preparation:

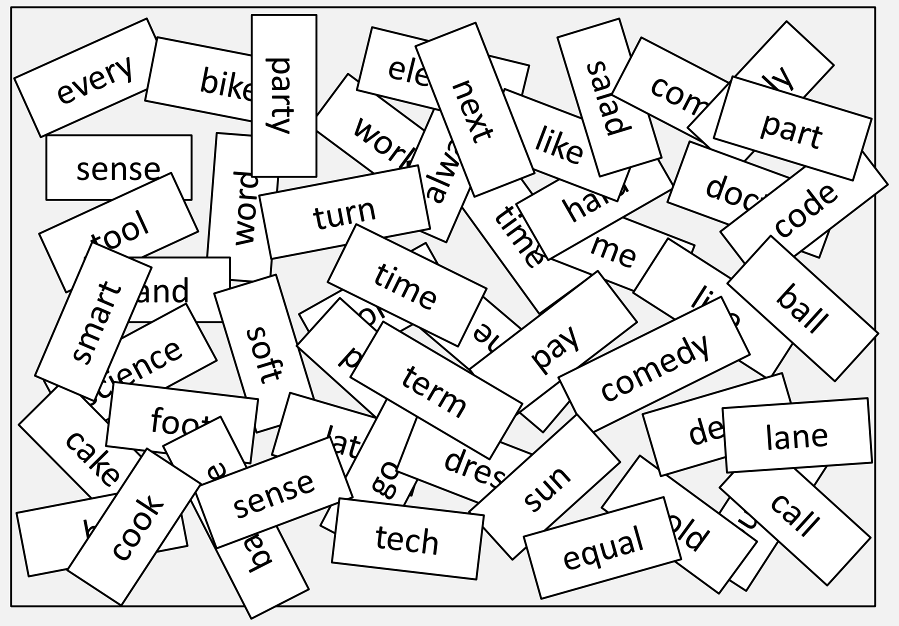
As mentioned earlier, the following are the steps generally involved:

- Loading the text
- Tokenization – Splitting the text into tokens. Depending on the problem statement, token can be a word, sentence or a paragraph
- Converting the word to a single format – lower case or upper case
- Removing punctuations and special characters from the corpus
- Removing stop words
- Stemming and lemmatization – Convert a word to its root.



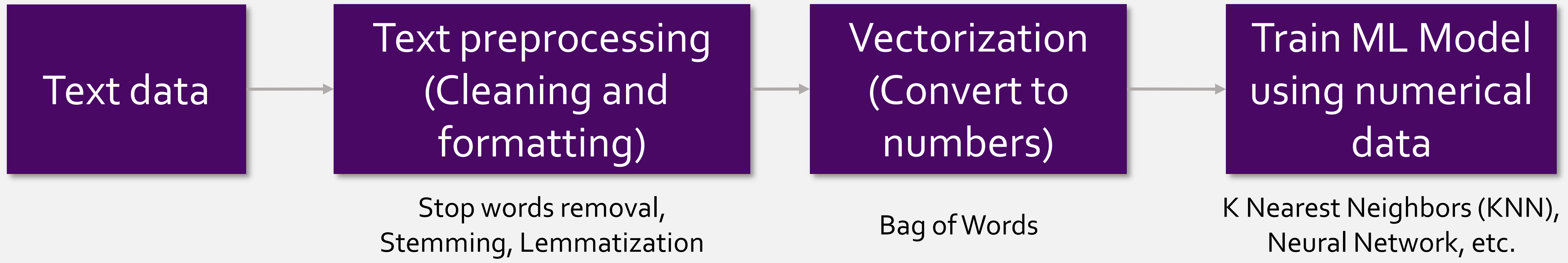
NLP & Text Processing

Token: Words or phrases extracted from documents. Splits text/document into small parts by white space and punctuation.



Feature vector: A numeric array that ML models use for different tasks such as training and prediction

ML models need **well-defined numerical data.**



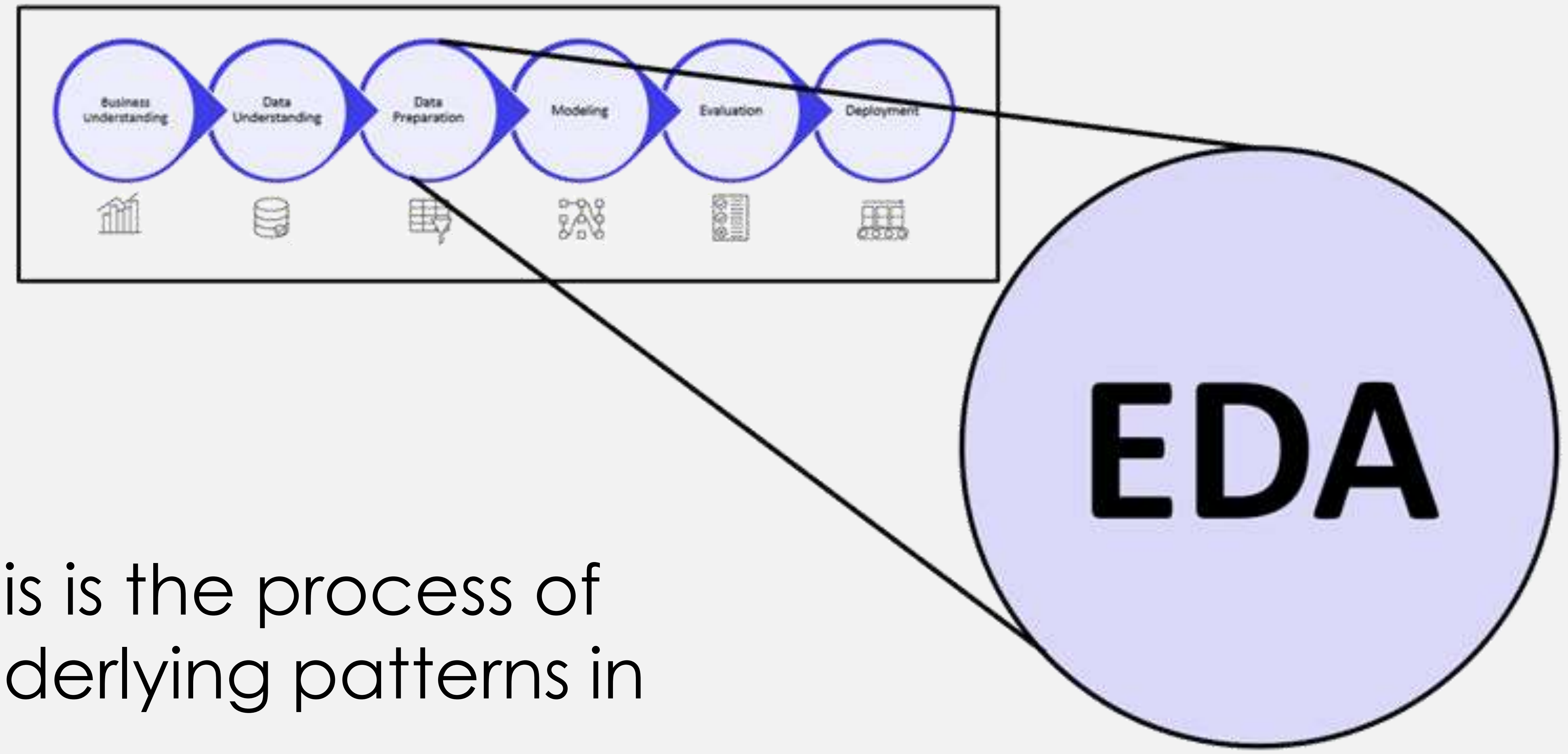
Text data → Text preprocessing (Cleaning and formatting) → Vectorization (Convert to numbers) → Train ML Model using numerical data

Stop words removal, Stemming, Lemmatization

Bag of Words

K Nearest Neighbors (KNN), Neural Network, etc.

EDA Process



What’s EDA in Details

- As the name suggests, Exploratory Data Analysis is the process of exploring data to gain insights and discover underlying patterns in data.
- There is no standard procedure for EDA and it is generally problem-specific.
- EDA typically involves descriptive statistics and visualizations.
- Some common representations in the context of NLP are word clouds (help identify the most frequently occurring words), getting the polarity of the text etc.

Parsing & Spidering

What to opt for?

Parsing:

- It is the process of determining the syntactic structure of a text by analysing its constituent words based on an underlying grammar (of the language).
- The outcome of the parsing process would be a parse tree, where sentence is the root, intermediate nodes such as noun_phrase, verb_phrase etc. have children - hence they are called non-terminals and finally, the leaves of the tree such as the individual words in a sentence.

Spidering:

- It is the process of indexing data on web pages by using a program or automated script. These automated scripts or programs are known by multiple names, including web crawler, spider, spider bot, and often shortened to crawler.
- Web crawlers copy pages for processing by a search engine, which indexes the downloaded pages so that users can search more efficiently. The goal of a crawler is to learn what webpages are about. This enables users to retrieve any information on one or more pages when it’s needed

Tokenization

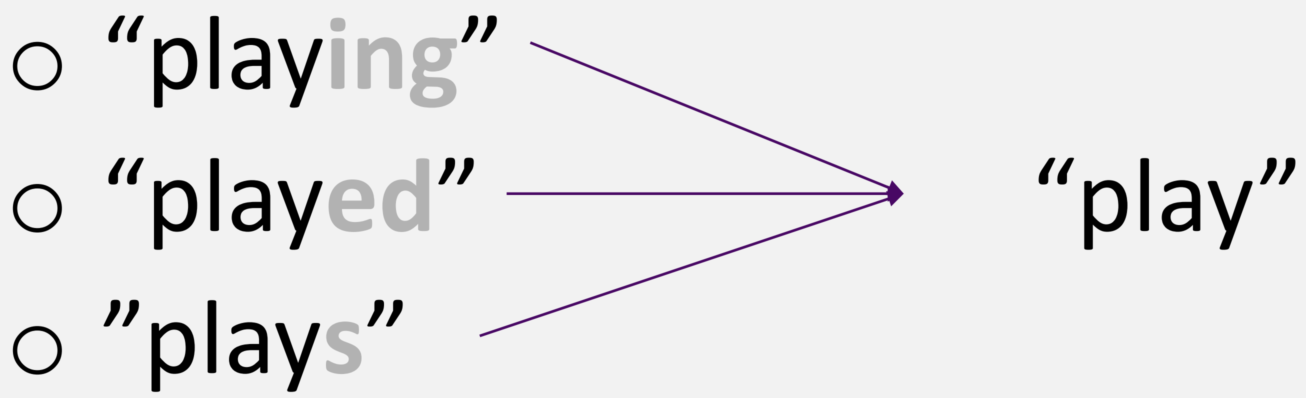
What to opt for?

- Tokenization is essentially splitting a phrase, sentence, paragraph, or an entire text document into smaller units, such as individual words or terms. Each of these smaller units are called tokens
- The tokens could be words, numbers or punctuation marks. In tokenization, smaller units are created by locating word boundaries.
- These are the ending point of a word and the beginning of the next word. These tokens are considered as a first step for stemming and lemmatization

Stemming

Set of rules to slice a string to a substring that usually refers to a more general meaning.

- The goal is to remove word affixes (particularly suffixes) such as “s”, “es”, “ing”, “ed”, etc.



- ○ “playing”
- ○ “played”
- ○ ”plays”

“play”

- The issue: It doesn’t usually work with irregular forms such as irregular verbs: **“taught”**, **“brought”**, etc.

Lemmatization

What to opt for?

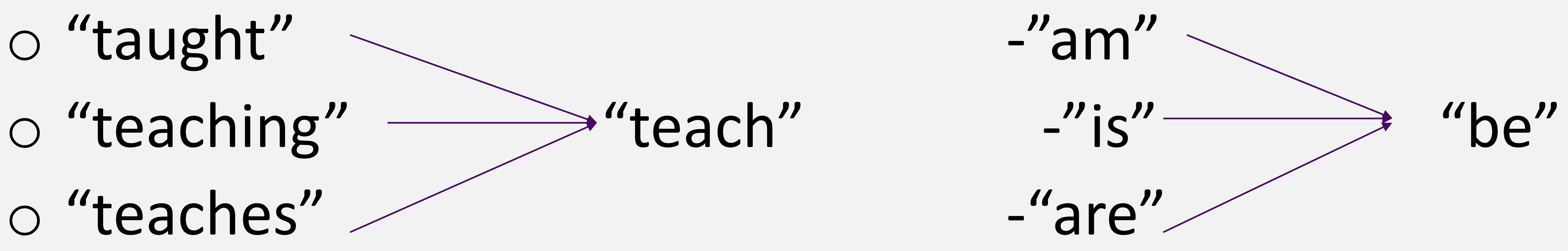
- For grammatical reasons, documents are going to use different forms of a word, such as organize, organizes, and organizing. Additionally, there are families of derivationally related words with similar meanings, such as democracy, democratic, and democratization. In many situations, it seems as if it would be useful for a search for one of these words to return documents that contain another word in the set.
- The goal of both stemming and lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form. For instance: am, is, are => be
- Lemmatization usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma . If confronted with the token saw, stemming might return just s, whereas lemmatization would attempt to return either see or saw depending on whether the use of the token was as a verb or a noun.

Lemmatization – part 2

What to opt for?

Similar to stemming, but more advanced. It uses a look-up dictionary.

- Handles more situations and usually works better than stemming.



○ “taught”

○ “teaching”

○ “teaches”

“teach”

-”am”

-”is”

-“are”

“be”

- For the best results, correct word position tags should be provided: “adjective”, “noun”, “verb” etc.

Sentence Splitting

What to opt for?

- Sentence splitting is the process of dividing text into sentences.
- A sentence splitter is also known as a sentence tokenizer, a sentence boundary detector, or a sentence boundary disambiguator.
- It involves splitting documents into sentences via a set of rules.
- Some example include using full stop(.), newline (/n) characters, considering the whole document as a sentence

String Normalization

What to opt for?

- Text normalization is the process of transforming a text into a canonical (standard) form. For example, the word “gooood” and “gud” can be transformed to “good”, its canonical form. Another example is mapping of near identical words such as “stopwords”, “stop-words” and “stop words” to just “stopwords”.
- Text normalization is important for noisy texts such as social media comments, text messages and comments to blog posts where abbreviations, misspellings and use of out-of-vocabulary words (oov) are prevalent.
- Unfortunately, unlike stemming and lemmatization, there isn’t a standard way to normalize texts. It typically depends on the task. For example, the way you would normalize clinical texts would arguably be different from how you normalize sms text messages.

Text Vectorization

Bag of Words (BoW) & Term Frequency (TF)

- **Bag of Words method converts text data into numbers.**
- **It does this by**
  - Creating a **vocabulary** from the words in all documents
  - Calculating the **occurrences** of words:
    - **binary** (present or not)
    - **word counts**
    - **frequencies**

	a	cat	dog	is	it	my	not	old	wolf
“It is a dog.”	**1**	0	**1**	**1**	**1**	0	0	0	0
“my cat is old”	0	**1**	0	**1**	0	**1**	0	**1**	0
“It is not a dog, it a is wolf.”	**2**	0	**1**	**2**	**2**	0	**1**	0	**1**

Term frequency (TF): Increases the weight for **common** words in a document.

$$tf(term, doc)= \frac{number\ of\ times\ the\ term\ occurs\ in\ the\ doc}{total\ number\ of\ terms\ in\ the\ doc}$$

	a	cat	dog	is	it	my	not	old	wolf
“It is a dog.”	**0.25**	0	**0.25**	**0.25**	**0.25**	0	0	0	0
“my cat is old”	0	**0.25**	0	**0.25**	0	**0.25**	0	**0.25**	0
“It is not a dog, it a is wolf.”	**0.22**	0	**0.11**	**0.22**	**0.22**	0	**0.11**	0	**0.11**

Text Vectorization

Inverse Document Frequency (IDF) & TF-IDF

term	idf
a	log(3/3)+1=**1**
cat	log(3/2)+1=**1.18**
dog	log(3/3)+1=**1**
is	log(3/4)+1=**0.87**
it	log(3/3)+1=**1**
my	log(3/2)+1=**1.18**
not	log(3/2)+1=**1.18**
old	log(3/2)+1=**1.18**
wolf	log(3/2)+1=**1.18**

Inverse document frequency (IDF): Decreases the weights for **commonly** used words and **increases** weights for **rare** words in the vocabulary.

$$idf(term) = log\left(\frac{n_{documents}}{n_{documents\ containing\ the\ term} + 1}\right) + 1$$

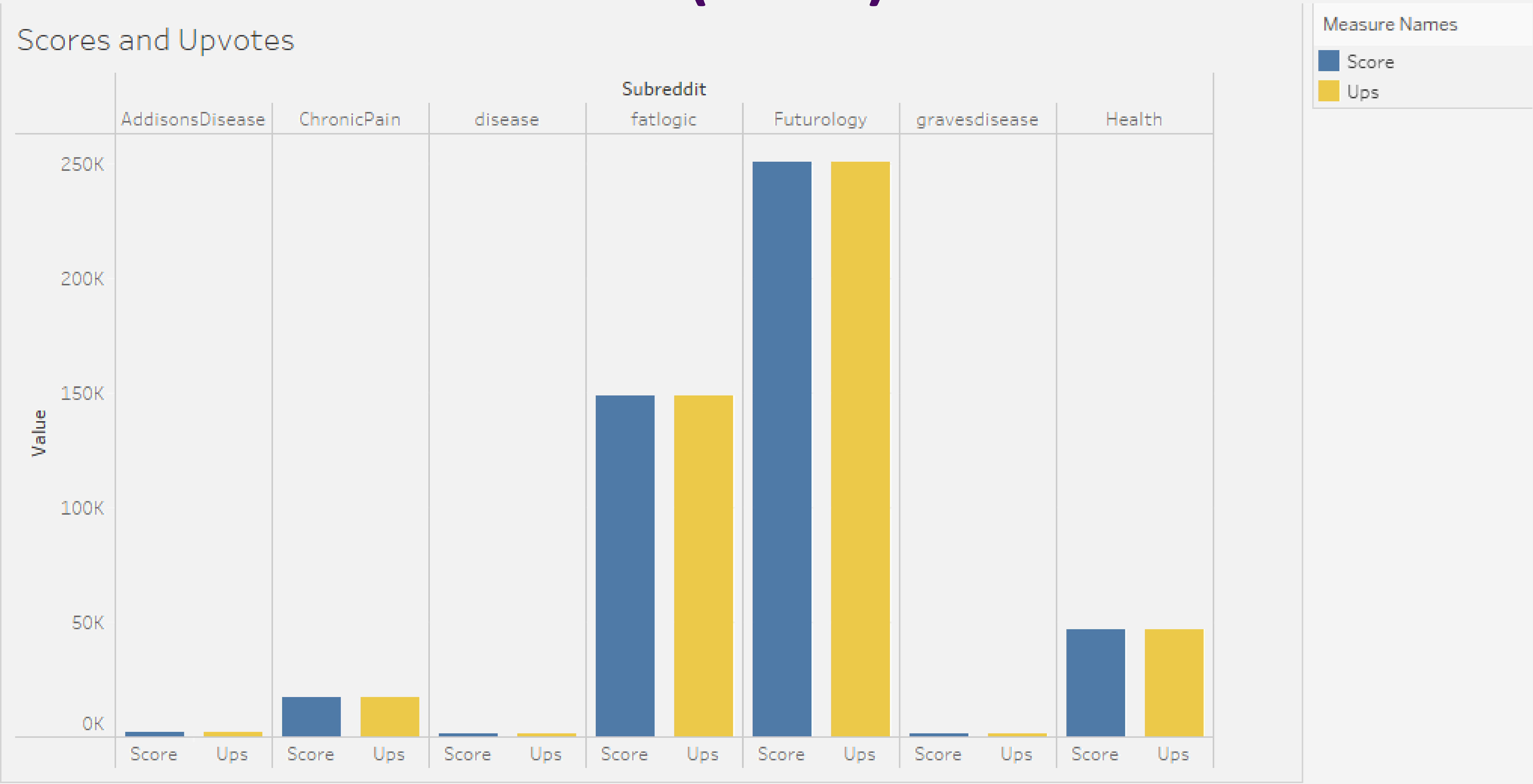
$$e.g.\ idf(\text{"}cat\text{"}) = 1.18$$

Term Freq. Inverse Doc. Freq (TF-IDF): Combines term frequency and inverse document frequency.

$$tf_{idf}(term, doc) = tf(term, doc) * idf(term)$$

	a	cat	dog	is	it	my	not	old	wolf
“It is a dog.”	**0.25**	0	**0.25**	**0.22**	**0.25**	0	0	0	0
“my cat is old”	0	**0.3**	0	**0.22**	0	**0.3**	0	**0.3**	0
“It is not a dog, it a is wolf.”	**0.22**	0	**0.11**	**0.19**	**0.22**	0	**0.13**	0	**0.13**

Tableau: Data Visualization (Part I)



©fulcrumdigital

Tableau: Data Visualization (Part II)

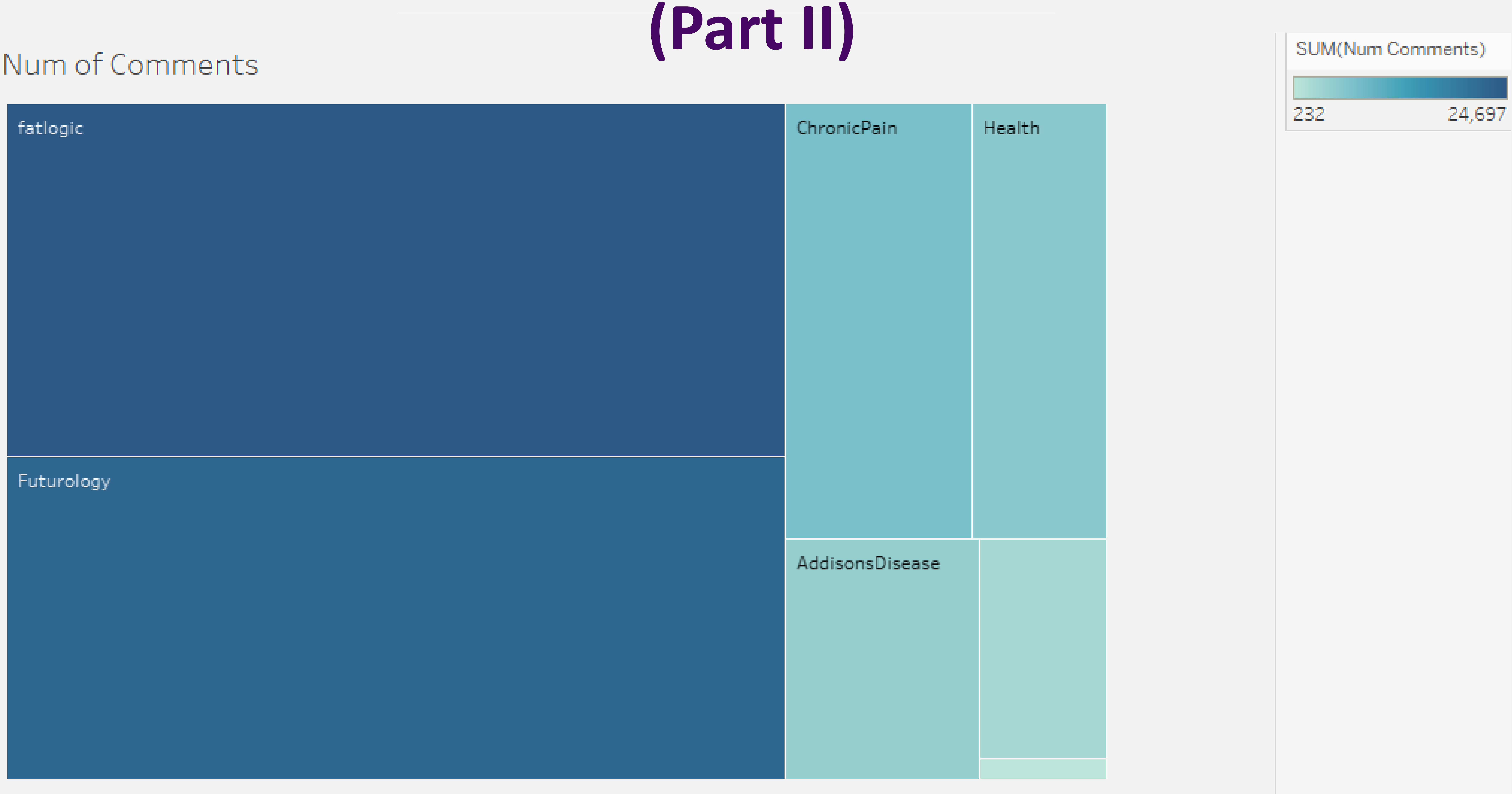
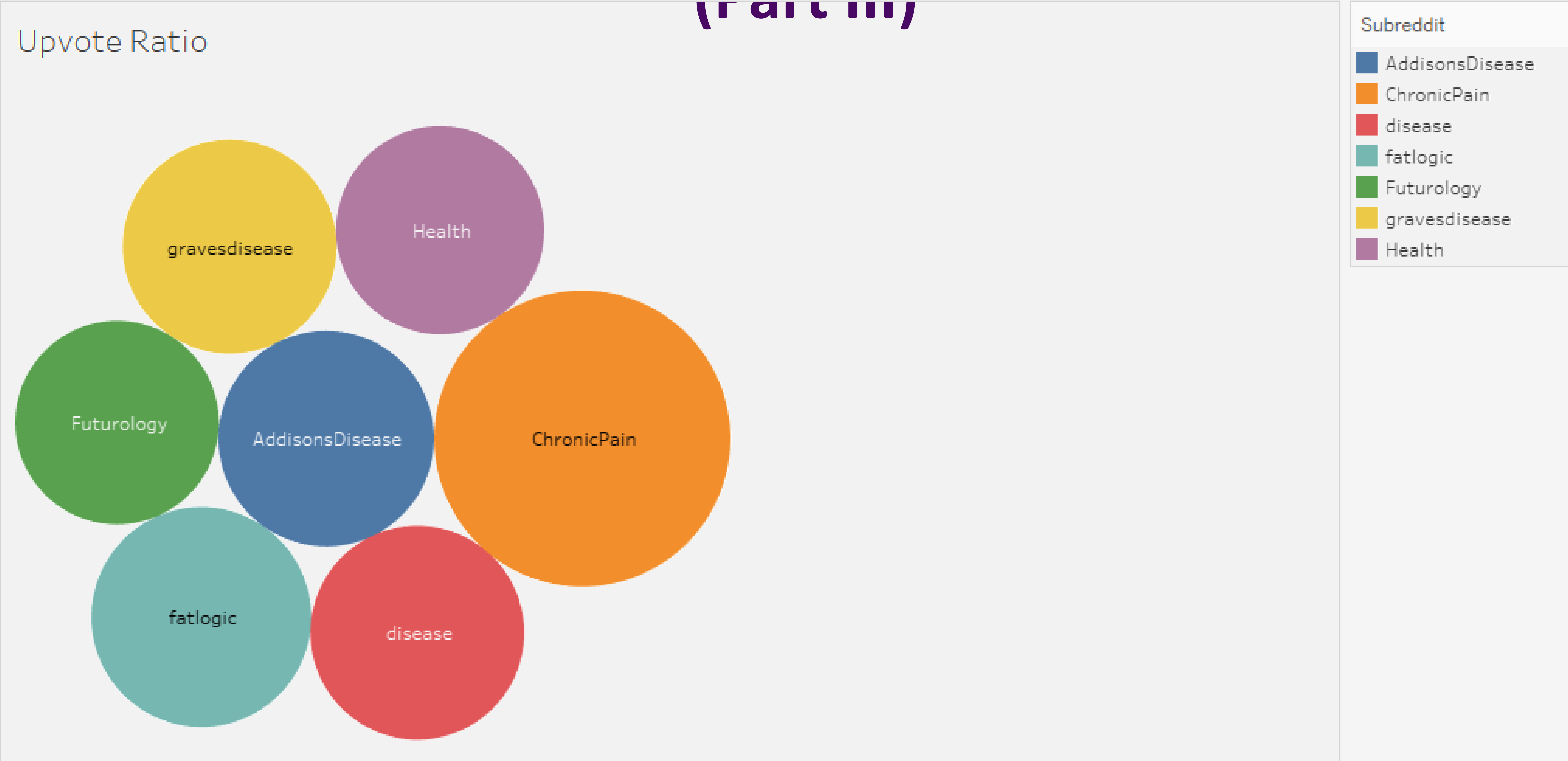
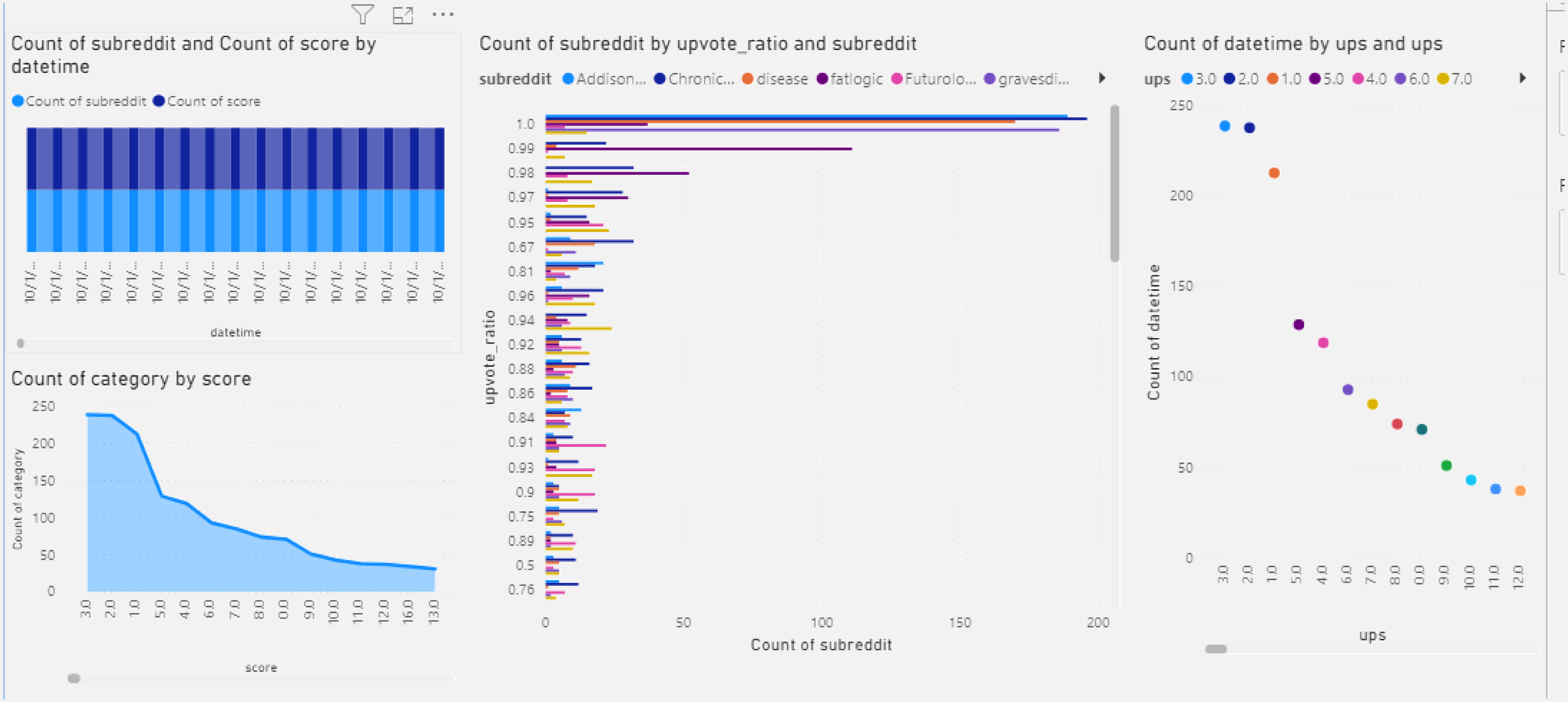


Tableau: Data Visualization (Part III)

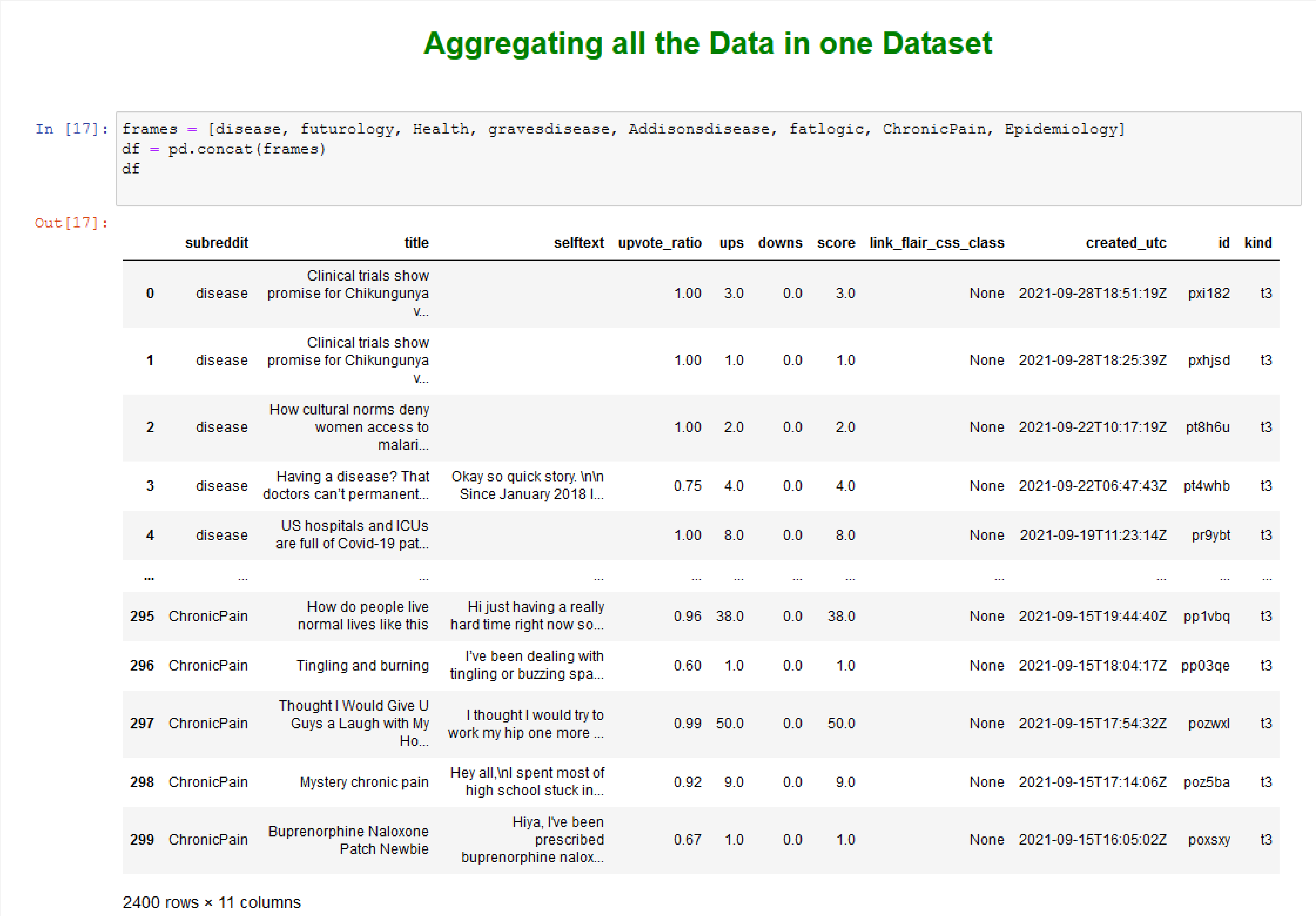


PowerBI: Data Visualization



Data Set on Python Jupyter Notebook

EDA of the Reddit App



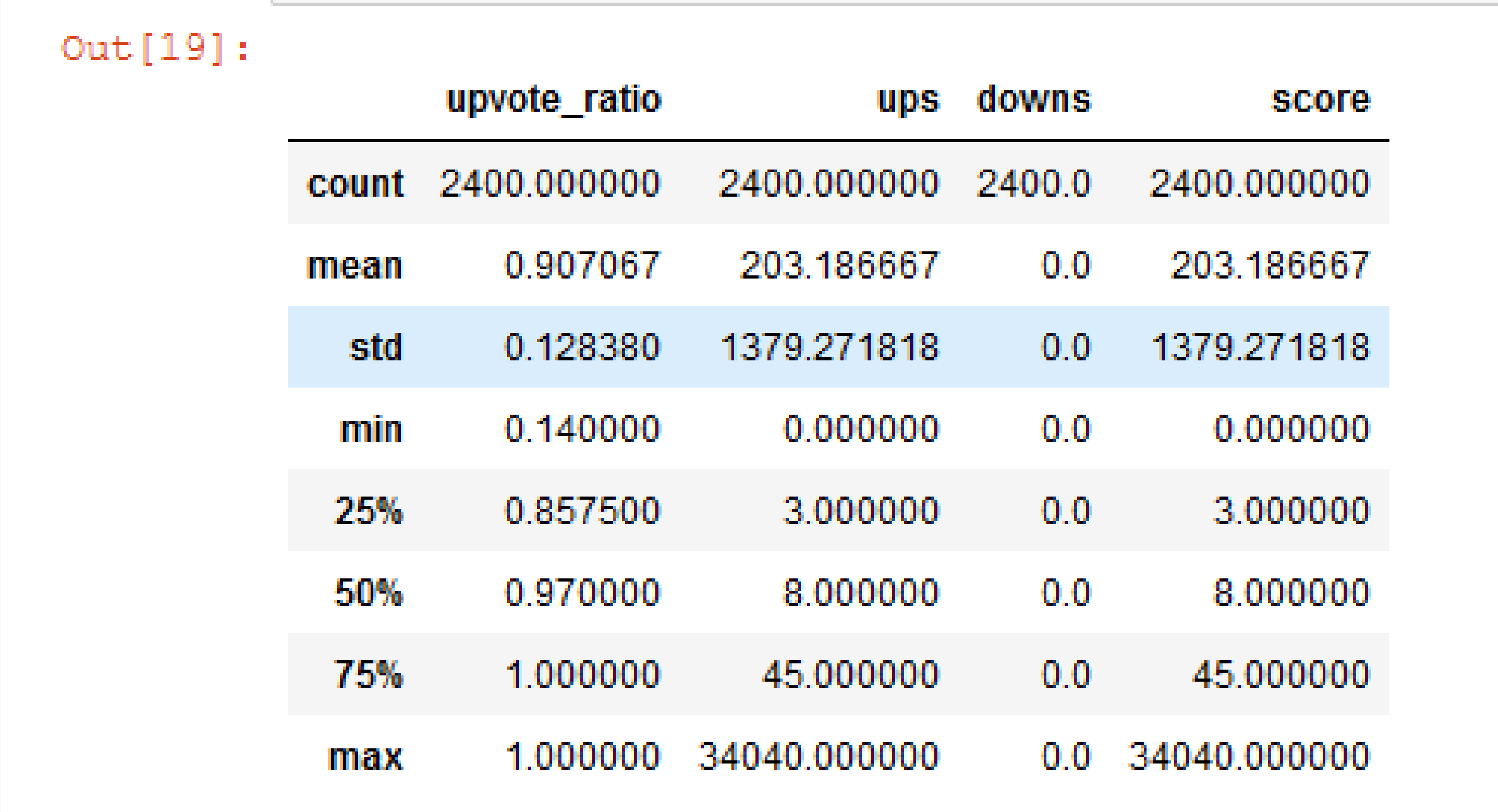
Aggregating all the Data in one Dataset

```
In [17]: frames = [disease, futurology, Health, gravesdisease, Addisonsdisease, fatlogic, ChronicPain, Epidemiology]
         df = pd.concat(frames)
         df
```

Out[17]:

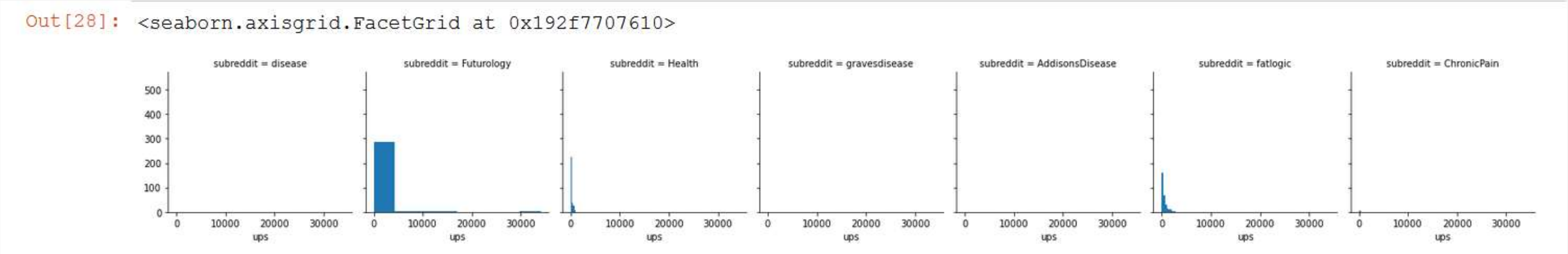
	subreddit	title	selftext	upvote_ratio	ups	downs	score	link_flair_css_class	created_utc	id	kind
0	disease	Clinical trials show promise for Chikungunya v...		1.00	3.0	0.0	3.0	None	2021-09-28T18:51:19Z	pxi182	t3
1	disease	Clinical trials show promise for Chikungunya v...		1.00	1.0	0.0	1.0	None	2021-09-28T18:25:39Z	pxhjsd	t3
2	disease	How cultural norms deny women access to malari...		1.00	2.0	0.0	2.0	None	2021-09-22T10:17:19Z	pt8h6u	t3
3	disease	Having a disease? That doctors can't permanent...	Okay so quick story. \n\n Since January 2018 I...	0.75	4.0	0.0	4.0	None	2021-09-22T06:47:43Z	pt4whb	t3
4	disease	US hospitals and ICUs are full of Covid-19 pat...		1.00	8.0	0.0	8.0	None	2021-09-19T11:23:14Z	pr9ybt	t3
...	...	...	...	...	...	...	...	...	...	...	...
295	ChronicPain	How do people live normal lives like this	Hi just having a really hard time right now so...	0.96	38.0	0.0	38.0	None	2021-09-15T19:44:40Z	pp1vbq	t3
296	ChronicPain	Tingling and burning	I've been dealing with tingling or buzzing spa...	0.60	1.0	0.0	1.0	None	2021-09-15T18:04:17Z	pp03qe	t3
297	ChronicPain	Thought I Would Give U Guys a Laugh with My Ho...	I thought I would try to work my hip one more ...	0.99	50.0	0.0	50.0	None	2021-09-15T17:54:32Z	pozwxl	t3
298	ChronicPain	Mystery chronic pain	Hey all,\nI spent most of high school stuck in...	0.92	9.0	0.0	9.0	None	2021-09-15T17:14:06Z	poz5ba	t3
299	ChronicPain	Buprenorphine Naloxone Patch Newbie	Hiya, I've been prescribed buprenorphine nalox...	0.67	1.0	0.0	1.0	None	2021-09-15T16:05:02Z	poxsxy	t3

2400 rows × 11 columns



Out[19]:

	upvote_ratio	ups	downs	score
count	2400.000000	2400.000000	2400.0	2400.000000
mean	0.907067	203.186667	0.0	203.186667
std	0.128380	1379.271818	0.0	1379.271818
min	0.140000	0.000000	0.0	0.000000
25%	0.857500	3.000000	0.0	3.000000
50%	0.970000	8.000000	0.0	8.000000
75%	1.000000	45.000000	0.0	45.000000
max	1.000000	34040.000000	0.0	34040.000000



```
Out[28]: <seaborn.axisgrid.FacetGrid at 0x192f7707610>
```

Conclusion and Insights

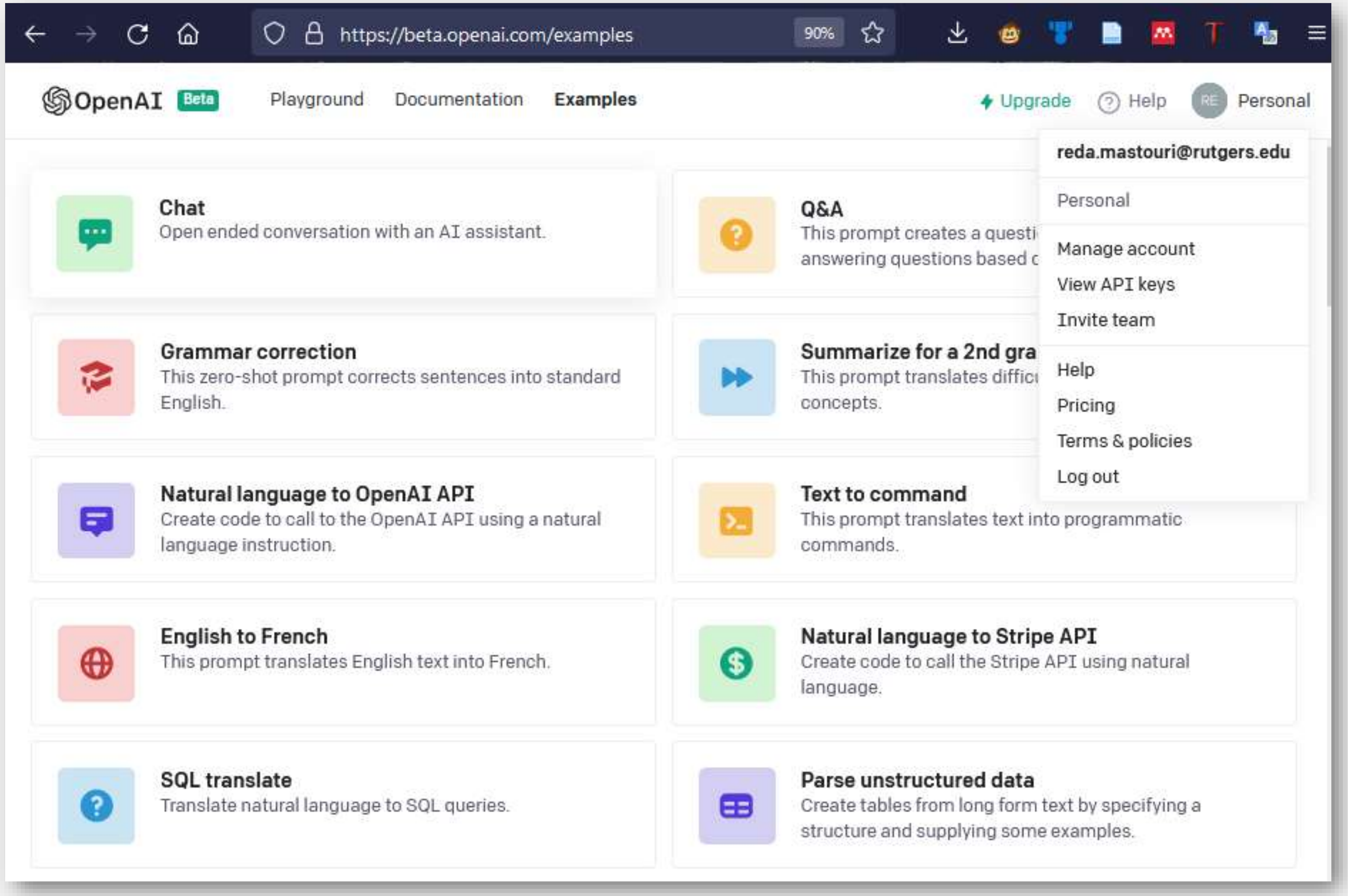
What we have learnt

- The Reddit API makes it extremely easy to compile a lot of news data fairly quickly. It's definitely worth the time and effort to enhance the data collection steps since it's so simple to get thousands of rows of disease opinions to use for further analysis and prediction.

- There's still a lot that could be engineered in regards to data mining, and there's still a lot to do with the data retrieved. The next step we will continue our analysis by merging to live-feed dataset to construct and train a recommendation engine.

Next Step

- Modeling: LSTM, LDA
- Implementing the openAI GPT-3 API for advanced sentiment analysis



https://beta.openai.com/examples 90%

OpenAI Beta Playground Documentation Examples Upgrade Help RE Personal

reda.mastouri@rutgers.edu

Personal

Manage account

View API keys

Invite team

Help

Pricing

Terms & policies

Log out

Chat
Open ended conversation with an AI assistant.

Q&A
This prompt creates a questi answering questions based c

Grammar correction
This zero-shot prompt corrects sentences into standard English.

Summarize for a 2nd gra
This prompt translates diffic concepts.

Natural language to OpenAI API
Create code to call to the OpenAI API using a natural language instruction.

Text to command
This prompt translates text into programmatic commands.

English to French
This prompt translates English text into French.

Natural language to Stripe API
Create code to call the Stripe API using natural language.

SQL translate
Translate natural language to SQL queries.

Parse unstructured data
Create tables from long form text by specifying a structure and supplying some examples.

☎

+ 1 (856) 882-9056
+ 1 (201) 985-4032

✉

rmastouri@saintpeters.edu
kpavuluri@saintpeters.edu

🌐

https://www.linkedin.com/in/reda-mastouri/
https://www.linkedin.com/in/kalyani-pavuluri-30416519

THANK YOU!

Saint Peter's
UNIVERSITY
The Jesuit University of New Jersey