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

~Literature Review and Analysis~

Reda Mastouri

MSc.in Data Sciences

Week I: Assignment

Due 9/19/2021 – Slide Presentation: 9/15/2021

- Students should review approximately 20-50 literature sources related to their selected research topic, methods/methodologies used by others for this or similar use cases - with their success and failure examples, potential methods that you might use in the Capstone project. This includes peer-reviewed articles, books, dissertations, conference papers etc.
- For the in class presentation 9/15 prepare a structured review and analysis of the literature (not a plane text) - you can make tables, graphs, pictures representing your literature (comparative) analysis and findings in your PowerPoint presentation.
- Make sure to define the project goal/specific tasks, or how you are planning to develop your project (specific steps). Plan to spend 5-10 minutes for your in class presentation, so that each individual in your team has a chance to speak.
- Upload your homework assignment in Blackboard no later than Sunday midnight, 9/19.

Topic n# 4: Disease State Conversation and Sentiment Analysis

Examination of Digital Community Conversations Within Specific Disease States Via Reddit

- **Vision:** 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.

- **Issue:** While Reddit offers robust, open, and community-minded discussions surrounding conditions and disease states, Reddit also provides volumes of unstructured and unclassified data. The development of a repeatable process – that continues to monitor evolving conversations over time – currently requires multiple tools (ex. – tools to scrape threads, tools to analyze keyword content, tools to analyze sentiment, etc.).

- **Method:** After identifying priority conditions and/or disease states with active Reddit communities (ex. – prostate cancer, breast cancer, HIV, etc.), build relational taxonomy (ex. – medicine, treatment, and adherence all have specific topics but have relational discussions) of topical themes addressed within.

- **Potential Output**: Provide use case for healthcare companies on the importance of Reddit as an early source of social indicator of trends and conversational "lexicon" to be used for patient communications and programs.

Team

Who's going to be involved

M.S in Data Sciences:

Reda Mastouri

MS in Data Sciences *from* Saint Peter's University ~ Nov 2021
BA in Applied Mathematics *from* Rutgers University
BS in Computer Sciences *from* NJIT

Solution Architect/ R&D – 2 ½ years
IT Support Specialist – 7 years
STEM Teaching – 4 years

Six Sigma – Yellow belt
CCNA, CompTIA (N+, Security+, Linux+)
AWS Cloud practitioner, CNCF Kubernetes Certified
Azure Solution Architect
IoT Developer

https://www.linkedin.com/in/reda-mastouri/

Introduction: Reddit

- Reddit is a social news website where registered users submit content in the form of links or text posts.
- Users, also known as “redditors”, can then vote each submission “up” or “down” to rank the post and determine its position or prominence on the site’s pages.
- These two attributes associated with a post are referred to as “upvotes” and “downvotes”.
- Redditors can also comment on posts, and respond back in a conversation tree of comments.
- Content entries, that is the posts, are organized by areas of interest or sub-communities called “subreddits”, such as health, politics, programming, or science.

Business Need

Because Reddit is regarded as one of the most effective social network sources for tracking the prevalence of public interests in infectious diseases (e.g., Coronavirus, HIV, and cancer) and controversial health-related issues (e.g., electronic cigarettes and marijuana) over time, reporting on findings derived from social media data nowadays becomes critical for understanding public reactions to infectious diseases.

As a result, we require a faster, more intelligent, and more accurate sentiment analyzer and web scrapper-based engine capable of tracking the latest trends on novel diseases, as well as any conversational "lexicon."
This will serve as a social indicator, providing a collection of use cases for healthcare companies to sensitize consumers through various mediums, communications, and programs to learn about either polemics or significant takeaways from what is happening in social media.

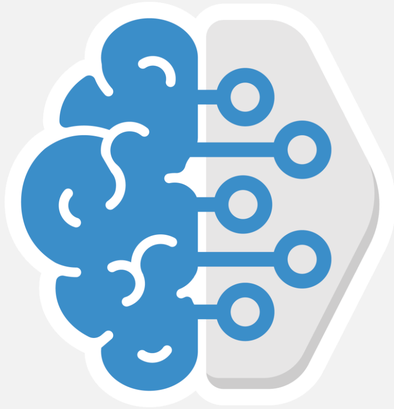
Problem Statement

- We use Reddit to demonstrate social media's potential for public health applications.

1. First, we employ a lexicon-based approach to track the prevalence of keywords indicating public interest in most prevalent diseases such as prostate cancer, breast cancer, HIV, etc.)
2. Second build relational taxonomy (ex. – medicine, treatment, and adherence all have specific topics but have relational discussions) of topical themes addressed within
3. Third, , to better understand the public reactions, we use the Latent Dirichlet Allocation algorithm, to identify either the general themes or motivations for extreme changes in the volume of discussion over time.
4. Lastly, we discuss the implications of our findings for utilizing Reddit data for public health applications.

Summarized Problem Statement

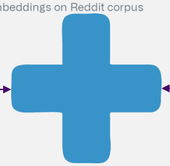
Cognitive Service/ Text Analytics/ Sentiment Analysis/ Key Phrase Extraction



Deep Learning/Machine learning:
- NLP
-

adamyi/
Word2Vec_Reddit

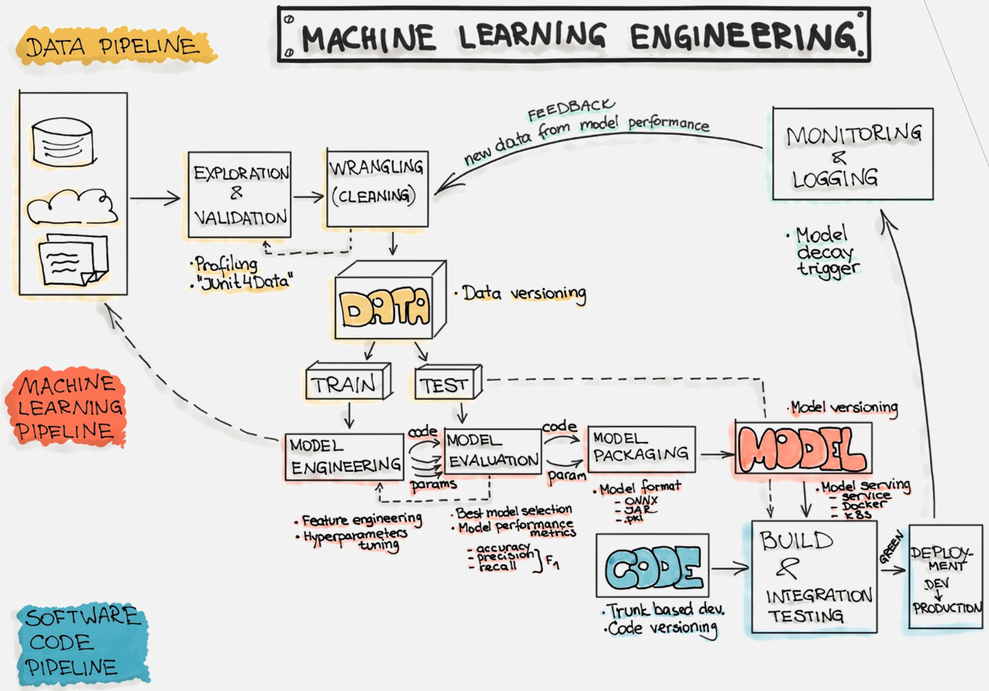
Pretrained word embeddings on Reddit corpus



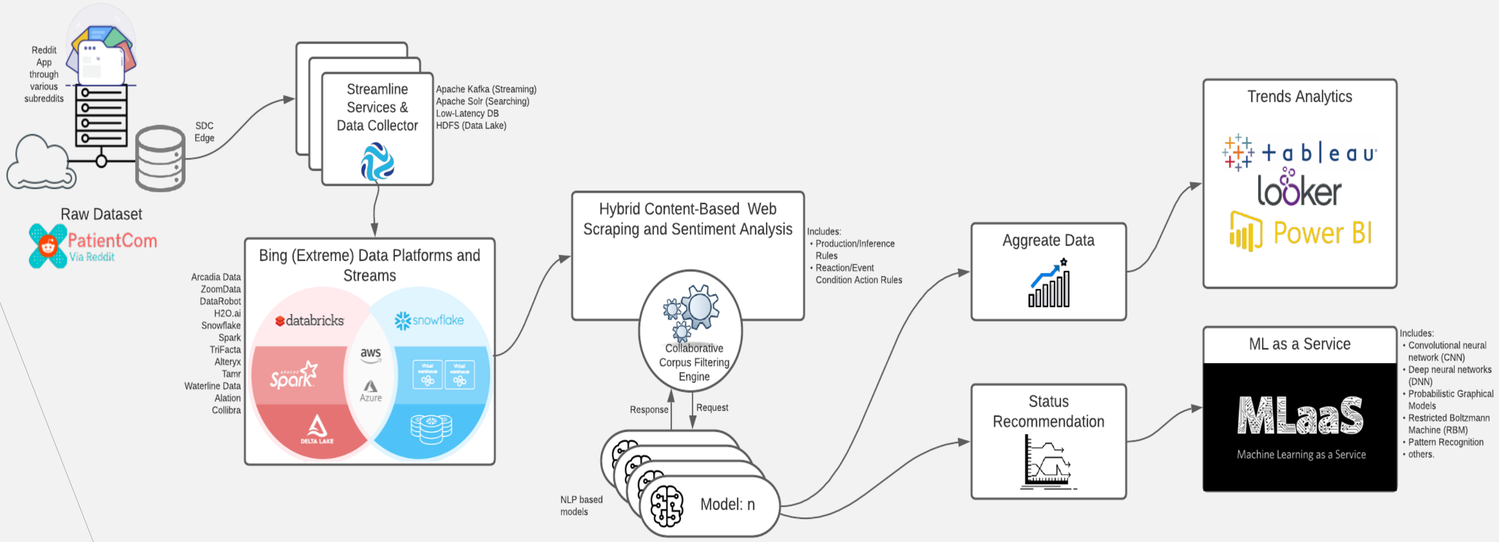
***PatientCom* (Public Health Trend Collector):**
Hybrid Content-Based Collaborative Corpus Filtering Engine with Web Scraping and Sentiment Analysis

ARCHITECTURE

HIGH LEVEL SYSTEM ARCHITECTURE



Theoretically



Practically

Technology Stack

WHAT IS TO BE IMPLEMENTED

Content Engine

We need to develop a Natural Language Processing (NLP) algorithm to extract mentions of medical conditions from text, trained it on a reddit app dataset that we labeled through crowdsourcing, and apply it on geo-referenced Reddit at scale

Data Processing & Reuse Strategy

1 –NLP medical entity extractor

2 – Geo-located Reddit posts

3- Medical conditions in social media

4 – Clustering the networks of medical conditions

5 – Set Composition (Rule grouping and precedence)

6- Taxonomy of medical conditions from Reddit

This is arranged in two levels, with some examples of individual conditions. The names of the level-1 and level-2 categories were assigned by the authors after manual inspection. We can manually arrange the top-level categories into six coherent themes. (B) A selection of the most frequent conditions that belong to multiple categories. (C) The list of top-level categories from International Classification of Diseases (ICD-11) by the World Health Organisation (WHO).

.

Machine Learning Models

Extract nouns using Python Natural Language Toolkit (NLTK) package.

Use Latent Dirichlet Allocation (LDA) for the time period of our interests

LSTM: **Long Short Term Memory** networks – usually just called “LSTMs” – are a special kind of RNN, capable of learning long-term dependencies.

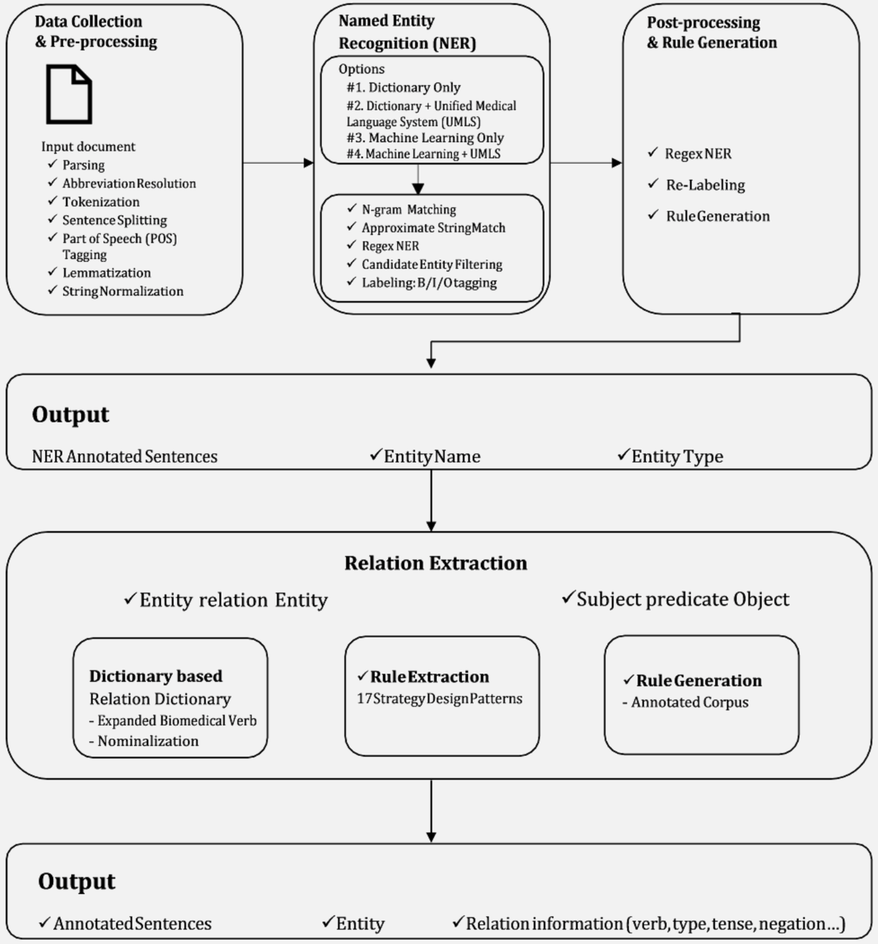
Methodologies

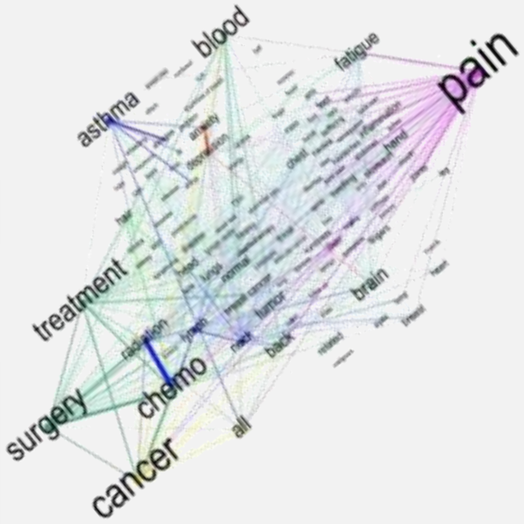
What is to be implemented (Part I)

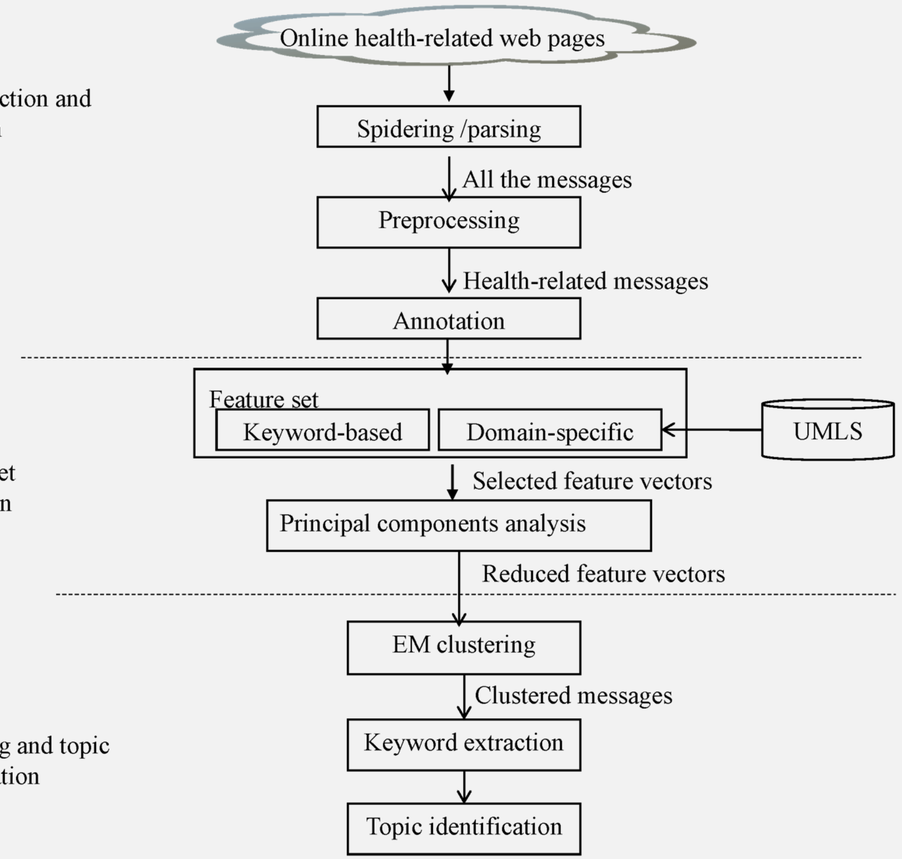
Methods	Description	Pros and Cons
Statistical technique	Using negative binomial regression as our prediction method because both our dependent variables, karma and comments are counts, and negative binomial regression is typically well-suited to handle over-dispersed count outcome variables.	Using a measure called deviance to evaluate goodness of fit, since this model has no direct analog of the proportion of variance explained by the predictors (R2) in linear regression. Deviance is a measure of the lack of fit to the data in a negative binomial regression model—lower numbers are better. It is calculated by comparing a model with the saturated model—a model with a theoretically perfect fit
natural language processing (NLP) and language modeling LDA	The extracted nouns were used to create language models—a set of topics generated from document-level word co-occurrences for a given set of documents—using Latent Dirichlet Allocation47 (LDA) for the time period of our interests. We elected to use LDA, an unsupervised algorithm, due to the lack of a ground truth dataset. We considered each post and its associated comments as a single document	One advantage of using LDA as opposed to other unsupervised clustering techniques is that the algorithm considers each document with multiple topics.
Investigating mediated public engagement with science on the “science” subreddit: From the participants’ perspective	ordinal logistic regression of users’ demographics on the perceived effect of r/science exciting the public about science.	No effects were found regarding whether or not the participants believed in r/science AMAs’ positive influence on communicating science for focused attention
Deep Learning tool for Natural Language Processing that extracts mentions of virtually any medical condition or disease from unstructured social media text.	With that tool at hand, we can process Reddit and Twitter posts, analyzed the clusters of the two resulting co-occurrence networks of conditions, and discovered that they correspond to well-defined categories of medical conditions.	This resulted in the creation of the first comprehensive taxonomy of medical conditions automatically derived from online discussions. We validated the structure of our taxonomy against the official International Statistical Classification of Diseases and Related Health Problems (ICD-11), finding matches of our clusters with 20 official categories, out of 22. Based on the mentions of our taxonomy’s sub-categories on Reddit posts geo-referenced in the U.S., we were then able to compute disease-specific health scores. As opposed to counts of disease mentions or counts with no knowledge of our taxonomy’s structure, we found that our disease-specific health scores are causally linked with the officially reported prevalence of 18 conditions.

Methodologies

What is to be implemented (Part II)







Potential Challenges

Regarding our approach: **Serial Pipeline Deployment**

- In this project, I aspire to partly tackle the potential issues by:
    1. automatically discovering the medical taxonomy present in online discussions
    2. based on the discovered taxonomy, proposing social media health metrics for a variety of conditions that can be blended with official data
    3. computing each condition's metric based on the limited set of symptoms related to that condition without over-fitting
    4. proposing broader health metrics, making it possible to examine multiple conditions simultaneously

Disadvantage

- The complexity grows as the number of subreddits grows.
- Understanding the extent to which the greater reddit population engaged in discussing emerging diseases is valuable from a gener[illegible]n perspective.

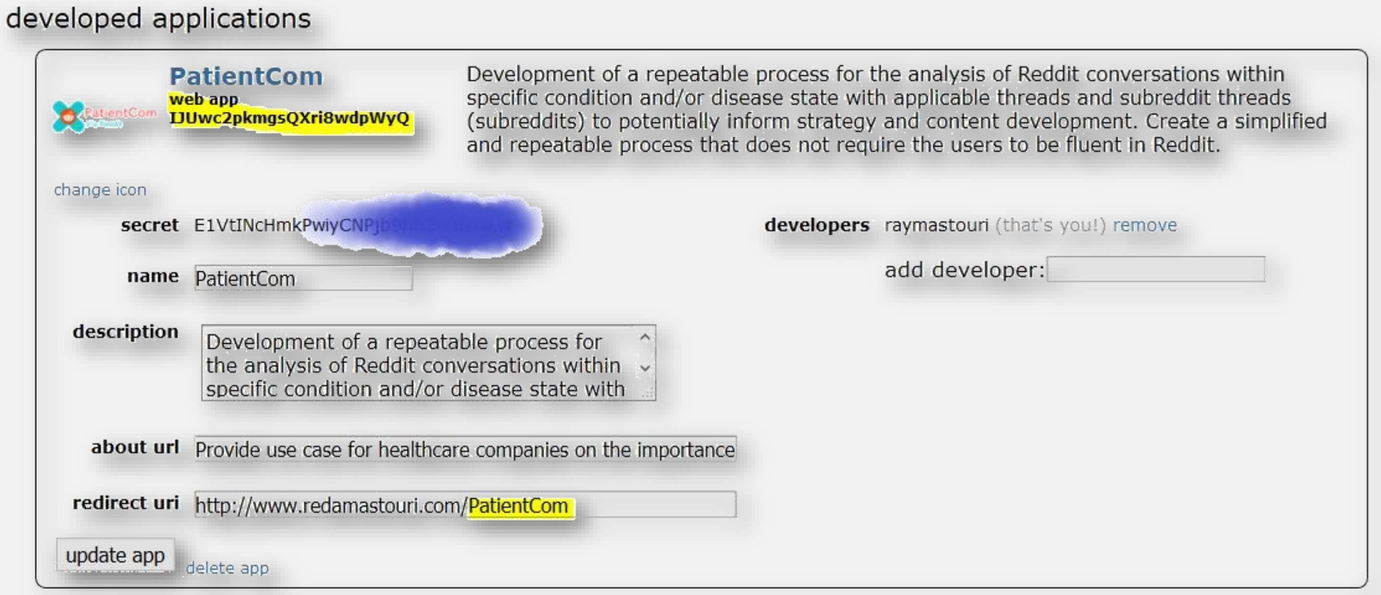
Advantage

- Tackling all the potential limitations lies on the following assumptions:
    - Periodicity of the published content on subreddits
    - Obsolescence of the content within a defined frequency
    - Data dimensionality
    - The need for broad health measures rather than measures tailored to specific diseases.

Practically: Data Set

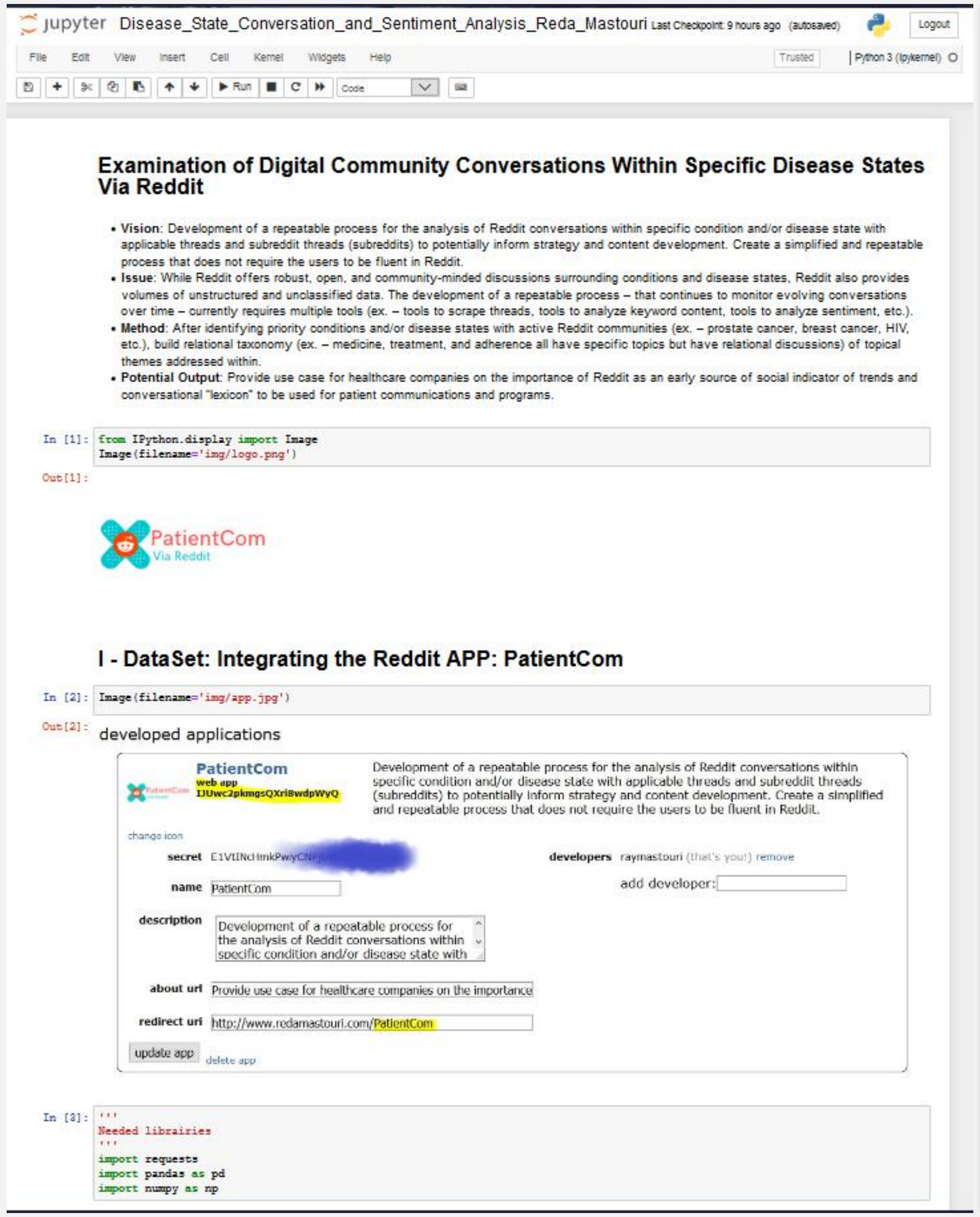
Creating an app on Reddit Developer Account

- 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.



Data Set on Python Jupyter Notebook

Integrating the Reddit App



jupyter Disease_State_Conversation_and_Sentiment_Analysis_Reda_Mastouri Last Checkpoint: 9 hours ago (autosaved)

Examination of Digital Community Conversations Within Specific Disease States Via Reddit

- **Vision**: 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.
- **Issue**: While Reddit offers robust, open, and community-minded discussions surrounding conditions and disease states, Reddit also provides volumes of unstructured and unclassified data. The development of a repeatable process – that continues to monitor evolving conversations over time – currently requires multiple tools (ex. – tools to scrape threads, tools to analyze keyword content, tools to analyze sentiment, etc.).
- **Method**: After identifying priority conditions and/or disease states with active Reddit communities (ex. – prostate cancer, breast cancer, HIV, etc.), build relational taxonomy (ex. – medicine, treatment, and adherence all have specific topics but have relational discussions) of topical themes addressed within.
- **Potential Output**: Provide use case for healthcare companies on the importance of Reddit as an early source of social indicator of trends and conversational "lexicon" to be used for patient communications and programs.

In [1]:
```
from IPython.display import Image
Image(filename='img/logo.png')
```
Out[1]:

PatientCom
Via Reddit

I - DataSet: Integrating the Reddit APP: PatientCom

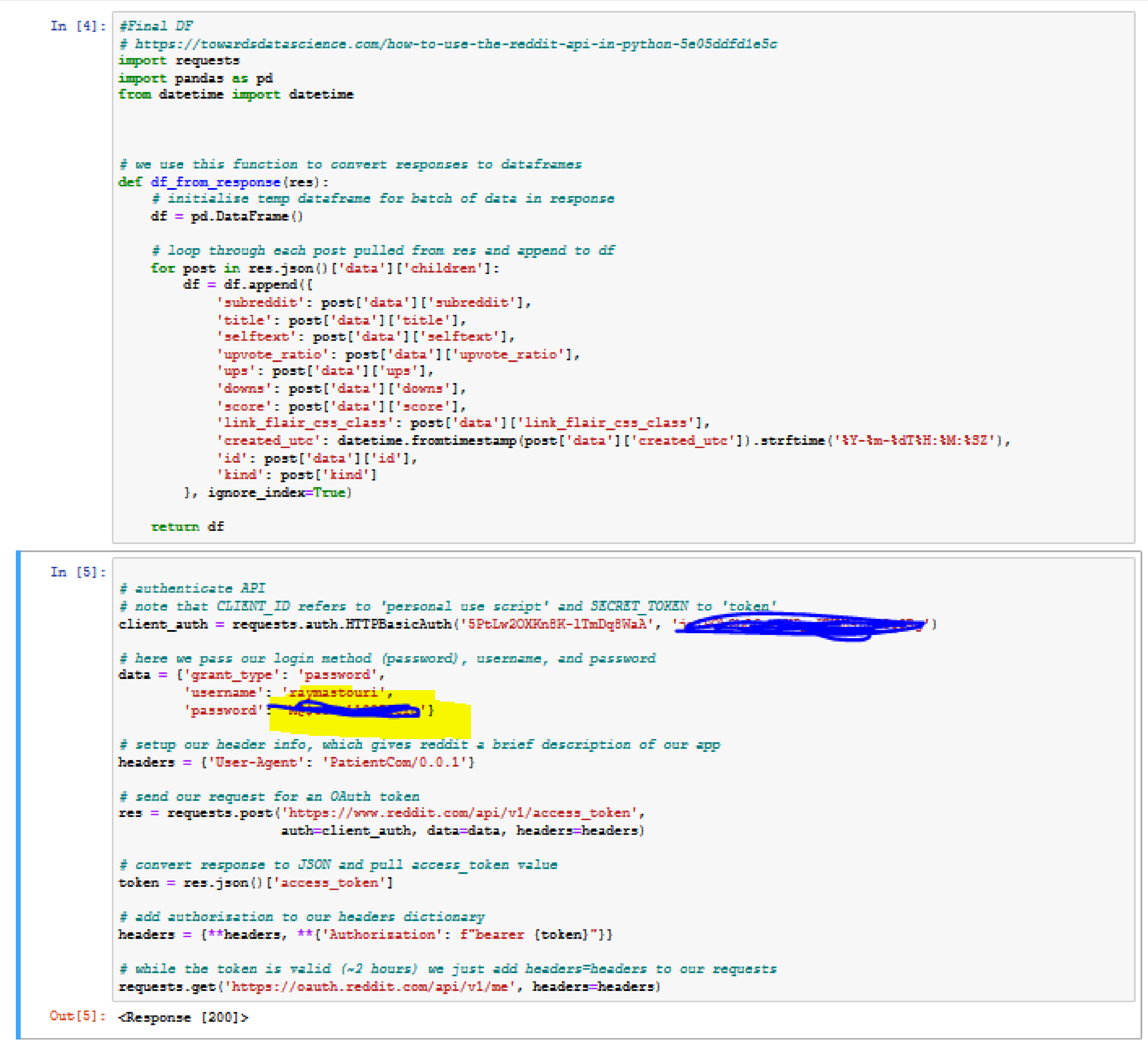
In [2]:
```
Image(filename='img/app.jpg')
```
Out[2]: developed applications

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.

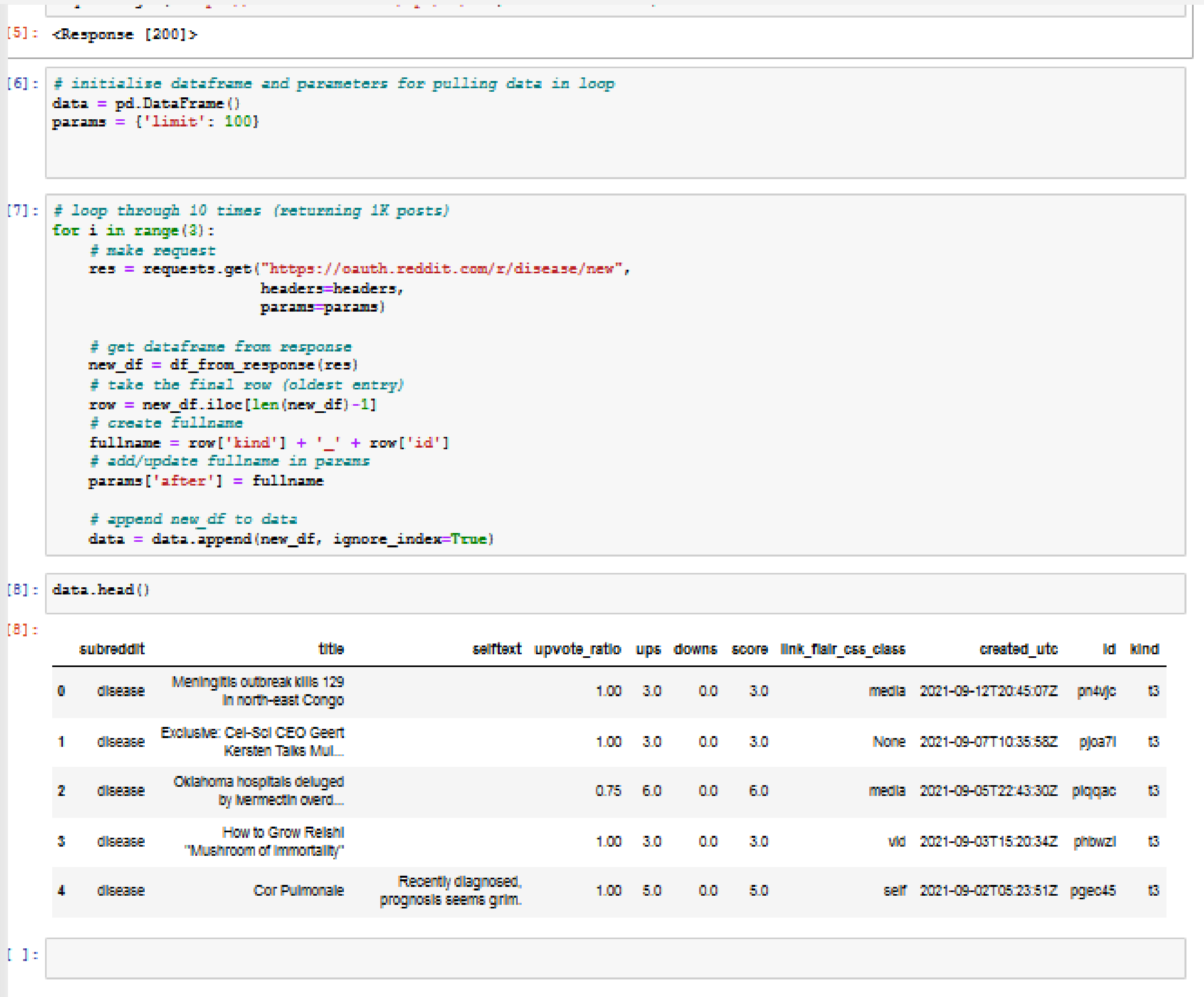
change icon
secret [redacted]
developers raymastouri (that's you!) remove
name PatientCom
add developer:
description Development of a repeatable process for the analysis of Reddit conversations within specific condition and/or disease state with
about url Provide use case for healthcare companies on the importance
redirect uri http://www.redamastouri.com/PatientCom
update app delete app

In [3]:
```
'''
Needed librairies
'''
import requests
import pandas as pd
import numpy as np
```



In [4]:
```
#Final DF
# https://towardsdatascience.com/how-to-use-the-reddit-api-in-python-5e05ddfd1e5c
import requests
import pandas as pd
from datetime import datetime


# we use this function to convert responses to dataframes
def df_from_response(res):
    # initialize temp dataframe for batch of data in response
    df = pd.DataFrame()

    # loop through each post pulled from res and append to df
    for post in res.json()['data']['children']:
        df = df.append({
            'subreddit': post['data']['subreddit'],
            'title': post['data']['title'],
            'selftext': post['data']['selftext'],
            'upvote_ratio': post['data']['upvote_ratio'],
            'ups': post['data']['ups'],
            'downs': post['data']['downs'],
            'score': post['data']['score'],
            'link_flair_css_class': post['data']['link_flair_css_class'],
            'created_utc': datetime.fromtimestamp(post['data']['created_utc']).strftime('%Y-%m-%dT%H:%M:%SZ'),
            'id': post['data']['id'],
            'kind': post['kind']
        }, ignore_index=True)

    return df
```

In [5]:
```
# authenticate API
# note that CLIENT_ID refers to 'personal use script' and SECRET_TOKEN to 'token'
client_auth = requests.auth.HTTPBasicAuth('5PtLw2OXKn8K-1TmDq8WaA', [redacted])

# here we pass our login method (password), username, and password
data = {'grant_type': 'password',
        'username': 'raymastouri',
        'password': [redacted]}

# setup our header info, which gives reddit a brief description of our app
headers = {'User-Agent': 'PatientCom/0.0.1'}

# send our request for an OAuth token
res = requests.post('https://www.reddit.com/api/v1/access_token',
                    auth=client_auth, data=data, headers=headers)

# convert response to JSON and pull access_token value
token = res.json()['access_token']

# add authorization to our headers dictionary
headers = {**headers, **{'Authorization': f"bearer {token}"}}

# while the token is valid (~2 hours) we just add headers=headers to our requests
requests.get('https://oauth.reddit.com/api/v1/me', headers=headers)
```
Out[5]: <Response [200]>



[5]: <Response [200]>

[6]:
```
# initialize dataframe and parameters for pulling data in loop
data = pd.DataFrame()
params = {'limit': 100}
```

[7]:
```
# loop through 10 times (returning 1K posts)
for i in range(3):
    # make request
    res = requests.get("https://oauth.reddit.com/r/disease/new",
                       headers=headers,
                       params=params)

    # get dataframe from response
    new_df = df_from_response(res)
    # take the final row (oldest entry)
    row = new_df.iloc[len(new_df)-1]
    # create fullname
    fullname = row['kind'] + '_' + row['id']
    # add/update fullname in params
    params['after'] = fullname

    # append new_df to data
    data = data.append(new_df, ignore_index=True)
```

[8]:
```
data.head()
```
[8]:

	subreddit	title	selftext	upvote_ratio	ups	downs	score	link_flair_css_class	created_utc	id	kind
0	disease	Meningitis outbreak kills 129 in north-east Congo		1.00	3.0	0.0	3.0	media	2021-09-12T20:45:07Z	pn4vjc	t3
1	disease	Exclusive: Cel-Sci CEO Geert Kersten Talks Mul...		1.00	3.0	0.0	3.0	None	2021-09-07T10:35:58Z	pjoa7l	t3
2	disease	Oklahoma hospitals deluged by ivermectin overd...		0.75	6.0	0.0	6.0	media	2021-09-05T22:43:30Z	piqqac	t3
3	disease	How to Grow Reishi "Mushroom of Immortality"		1.00	3.0	0.0	3.0	vid	2021-09-03T15:20:34Z	phbwzl	t3
4	disease	Cor Pulmonale	Recently diagnosed, prognosis seems grim.	1.00	5.0	0.0	5.0	self	2021-09-02T05:23:51Z	pgec45	t3

[ ]:

Literature Review: References (Part I)

1. 30 Facts & Statistics On Social Media And Healthcare [WWW Document], 2017. . ReferralMD. URL https://getreferralmd.com/2017/01/30-facts-statistics-on-social-media-and-healthcare/ (accessed 9.15.21).
2. Bazzaz Abkenar, S., Haghi Kashani, M., Mahdipour, E., Jameii, S.M., 2021. Big data analytics meets social media: A systematic review of techniques, open issues, and future directions. Telematics and Informatics 57, 101517. https://doi.org/10.1016/j.tele.2020.101517
3. Carelle, N., Piotto, E., Bellanger, A., Germanaud, J., Thuillier, A., Khayat, D., 2002. Changing patient perceptions of the side effects of cancer chemotherapy. Cancer 95, 155–163. https://doi.org/10.1002/cncr.10630
4. Chen, H., Hara, N., McKay, C., 2021. Investigating mediated public engagement with science on the "science" subreddit: From the participants' perspective. PLoS One 16, e0249181. https://doi.org/10.1371/journal.pone.0249181
5. Choudhury, M.D., De, S., n.d. Mental Health Discourse on reddit: Self-Disclosure, Social Support, and Anonymity 10.
6. Foufi, V., Timakum, T., Gaudet-Blavignac, C., Lovis, C., Song, M., 2019. Mining of Textual Health Information from Reddit: Analysis of Chronic Diseases With Extracted Entities and Their Relations. J Med Internet Res 21, e12876. https://doi.org/10.2196/12876
7. Fox, S., Purcell, K., 2010. Chronic Disease and the Internet. Pew Research Center: Internet, Science & Tech. URL https://www.pewresearch.org/internet/2010/03/24/chronic-disease-and-the-internet/ (accessed 9.15.21).
8. Gonzalez, G., Sarker, A., Oconnor, K., Savova, G., 2017. Capturing the Patient's Perspective: a Review of Advances in Natural Language Processing of Health-Related Text. Yearbook of Medical Informatics 26, 214–227. https://doi.org/10.15265/IY-2017-029
9. Hara, N., Abbazio, J., Perkins, K., 2019. An emerging form of public engagement with science: Ask Me Anything (AMA) sessions on Reddit r/science. PLoS One 14, e0216789. https://doi.org/10.1371/journal.pone.0216789
10. Kim, Y.H., Beak, S.H., Song, M., n.d. Constructing Linguistic Verb Source for Relation Extraction 4.

Literature Review: References (Part II)

1. Low, D.M., Rumker, L., Talkar, T., Torous, J., Cecchi, G., Ghosh, S.S., 2020. Natural Language Processing Reveals Vulnerable Mental Health Support Groups and Heightened Health Anxiety on Reddit During COVID-19: Observational Study. Journal of Medical Internet Research 22, e22635. https://doi.org/10.2196/22635
2. Lu, Y., Zhang, P., Liu, J., Li, J., Deng, S., 2013. Health-Related Hot Topic Detection in Online Communities Using Text Clustering. PLOS ONE 8, e56221. https://doi.org/10.1371/journal.pone.0056221
3. Nzali, M.D.T., Bringay, S., Lavergne, C., Mollevi, C., Opitz, T., 2017. What Patients Can Tell Us: Topic Analysis for Social Media on Breast Cancer. JMIR Medical Informatics 5, e7779. https://doi.org/10.2196/medinform.7779
4. Pandrekar, S., Chen, X., Gopalkrishna, G., Srivastava, A., Saltz, M., Saltz, J., Wang, F., 2018. Social Media Based Analysis of Opioid Epidemic Using Reddit. AMIA Annu Symp Proc 2018, 867–876.
5. Park, A., Conway, M., 2018. Tracking Health Related Discussions on Reddit for Public Health Applications. AMIA ... Annual Symposium proceedings. AMIA Symposium 2017, 1362–1371.
6. Park, A., Conway, M., Chen, A.T., 2018. Examining Thematic Similarity, Difference, and Membership in Three Online Mental Health Communities from Reddit: A Text Mining and Visualization Approach. Comput Human Behav 78, 98–112. https://doi.org/10.1016/j.chb.2017.09.001
7. Scepanovic, S., Aiello, L.M., Zhou, K., Joglekar, S., Quercia, D., 2021. The Healthy States of America: Creating a Health Taxonomy with Social Media. arXiv:2103.01169 [cs].
8. Sharma, R., Wigginton, B., Meurk, C., Ford, P., Gartner, C.E., 2017. Motivations and Limitations Associated with Vaping among People with Mental Illness: A Qualitative Analysis of Reddit Discussions. International Journal of Environmental Research and Public Health 14, 7. https://doi.org/10.3390/ijerph14010007
9. Silveira Fraga, B., Couto da Silva, A.P., Murai, F., 2018. Online Social Networks in Health Care: A Study of Mental Disorders on Reddit, in: 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI). Presented at the 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI), pp. 568–573. https://doi.org/10.1109/WI.2018.00-36
10. Sobkowicz, P., Sobkowicz, A., 2021. Agent Based Model of Anti-Vaccination Movements: Simulations and Comparison with Empirical Data. Vaccines (Basel) 9, 809. https://doi.org/10.3390/vaccines9080809
11. Song, M., Kang, K., An, J.Y., 2018. Investigating drug–disease interactions in drug–symptom–disease triples via citation relations. Journal of the Association for Information Science and Technology 69, 1355–1368. https://doi.org/10.1002/asi.24060
12. Turcan, E., McKeown, K., 2019. Dreaddit: A Reddit Dataset for Stress Analysis in Social Media, in: Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019). Presented at the Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019), Association for Computational Linguistics, Hong Kong, pp. 97–107. https://doi.org/10.18653/v1/D19-6213

Project Plan

What is to be delivered?

Project Estimated Completion Date: 10/27th/2021

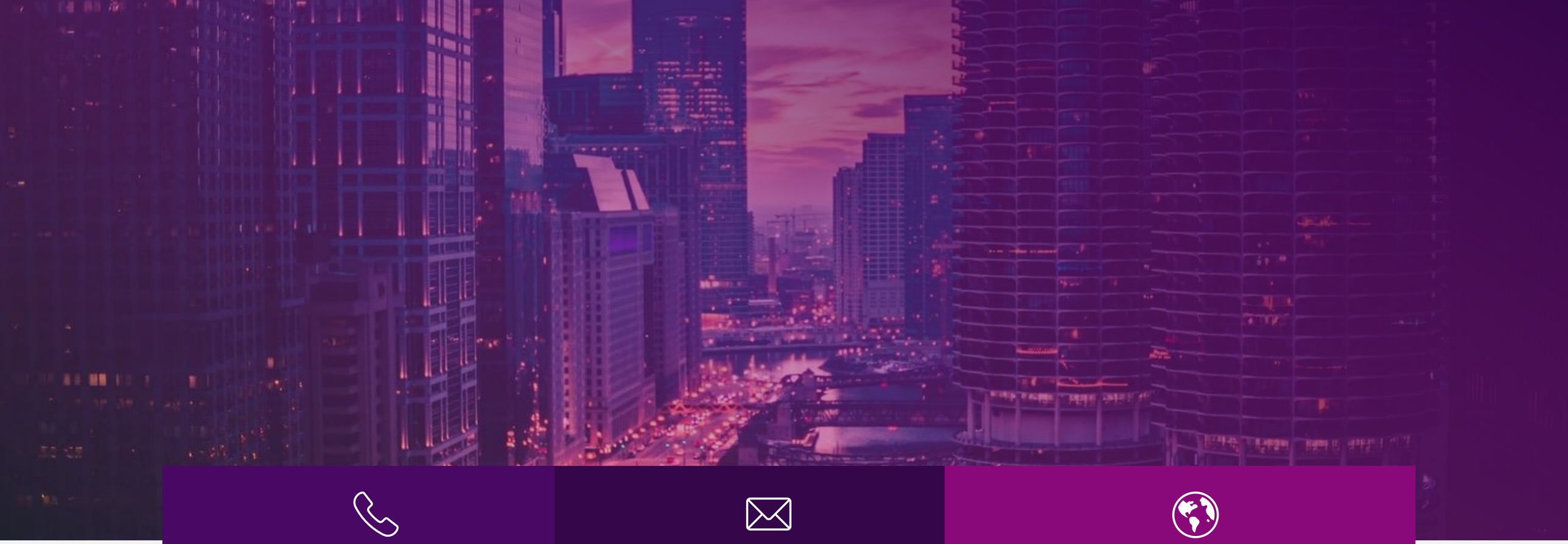
1 App Engine:
- Create an Account and commit all the code contribution to gitlab on a daily basis.
- Create app on Reddit with key credential to automate the web scrapping
- Build dataframes per subreddits and merge them into final dataframe
- Use LSTM and NLP algorithm along with LDA to build the sentiment analysis core
- Operationalize the app as service and deploy it to the cloud
- UAT and model performace testing for the end-to-end pipeline

2 Best practices:
- While using Visual studio as IDE and Connect to the project in SonarQube or SonarCloud, let's make sure we're applying the same rules that will be used during analysis

3 Coding artifacts:
Making sure that gitlab repository has all the databricks notebooks or AzureML notebooks being pushed to the gitlab repository:
e.g. https://community.cloud.databricks.com/

4 Documentation:
- Use Zotero or Mendeley tools to index all the peer-reviewed articles, books, dissertations, conference papers related to this topic
- Prepare 10 minutes of summary notes (or pptx/ Demo) for the completion of the allocated working epic.

Nb: Use these notes for the final article



+ 1 (856) 882-9056

rmastouri@saintpeters.edu

https://www.linkedin.com/in/reda-mastouri/

THANK YOU!

Saint Peter's UNIVERSITY

The Jesuit University of New Jersey