Web Portfolio Github

Master of Science

Portfolio Milestone

Ryan Ondocin
SUID: 83072-6905
April 11, 2020

Overview

The Applied Data Science Program at Syracuse University provides students with the ability to collect, organize, manage, analyze and develop actionable insights from data using a variety of programming languages, statistical tools and techniques. This endows students with the capabilities of becoming successful data science practitioners that can become a valuable asset to any organization they choose.

This presentation will cover the tools and techniques utilized throughout 3 different courses which exemplify successful execution of the critical learning objectives for becoming a data scientist:

We will see examples from:

IST 659: Database Administration

IST 664: Natural Language Processing

IST 718: Big Data Analytics

Learning Objectives

1. Describe a broad overview of the major practice areas in Data Science
2. Collect and Organize Data
3. Identify Patterns in data via visualization, statistical analysis and data mining
4. Develop Alternative strategies based on data
5. Develop a plan of action to implement the business decisions derived from the analyses.
6. Demonstrate Communication skills regarding data and its analysis for relevant professionals in their organization
7. Synthesize the ethical dimensions of data science practice

IST 659: Database Administration

Hospital RDBMS:

Through studying Database Administration under the direction of Professor Hoyos a Hospital database management system was developed to:

- Automate Hospital Operations,
- Improve Planning and data Access regarding clinical data
- Manage/Schedule Appointments
- provide Timely Access to patient/Staff info in a secure manner

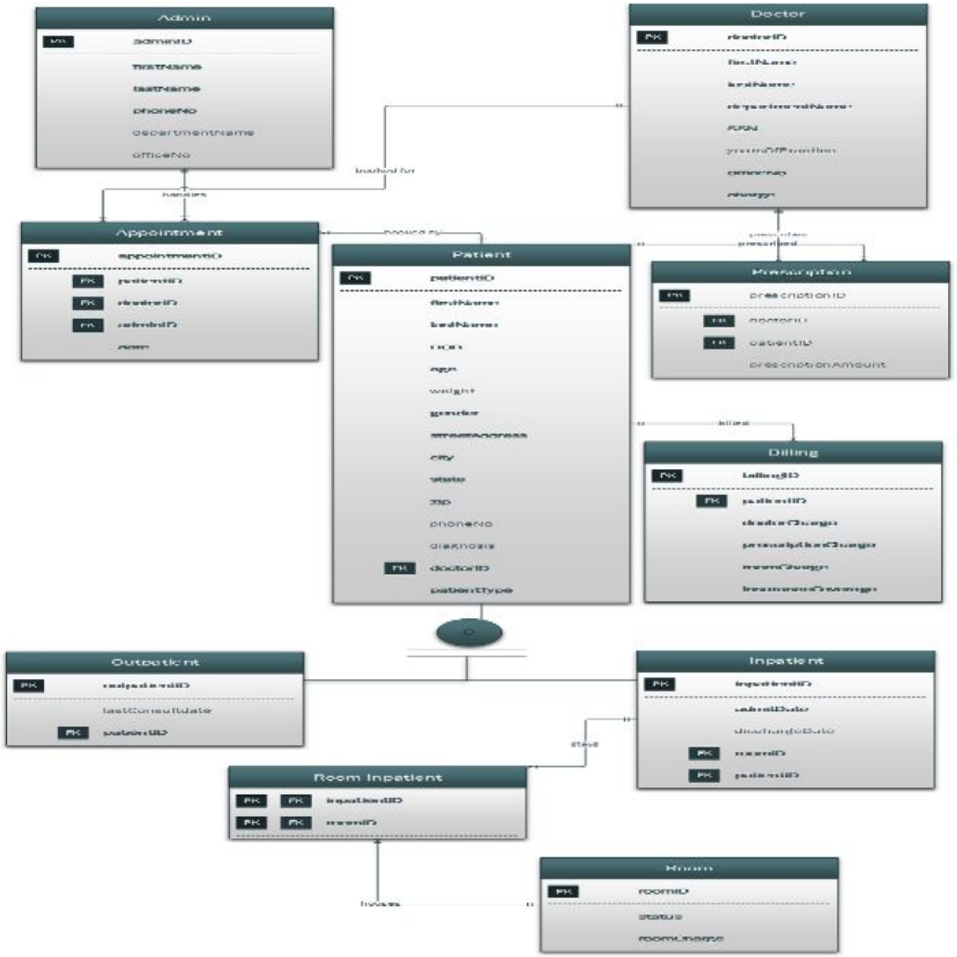
Modeling, Table Creation

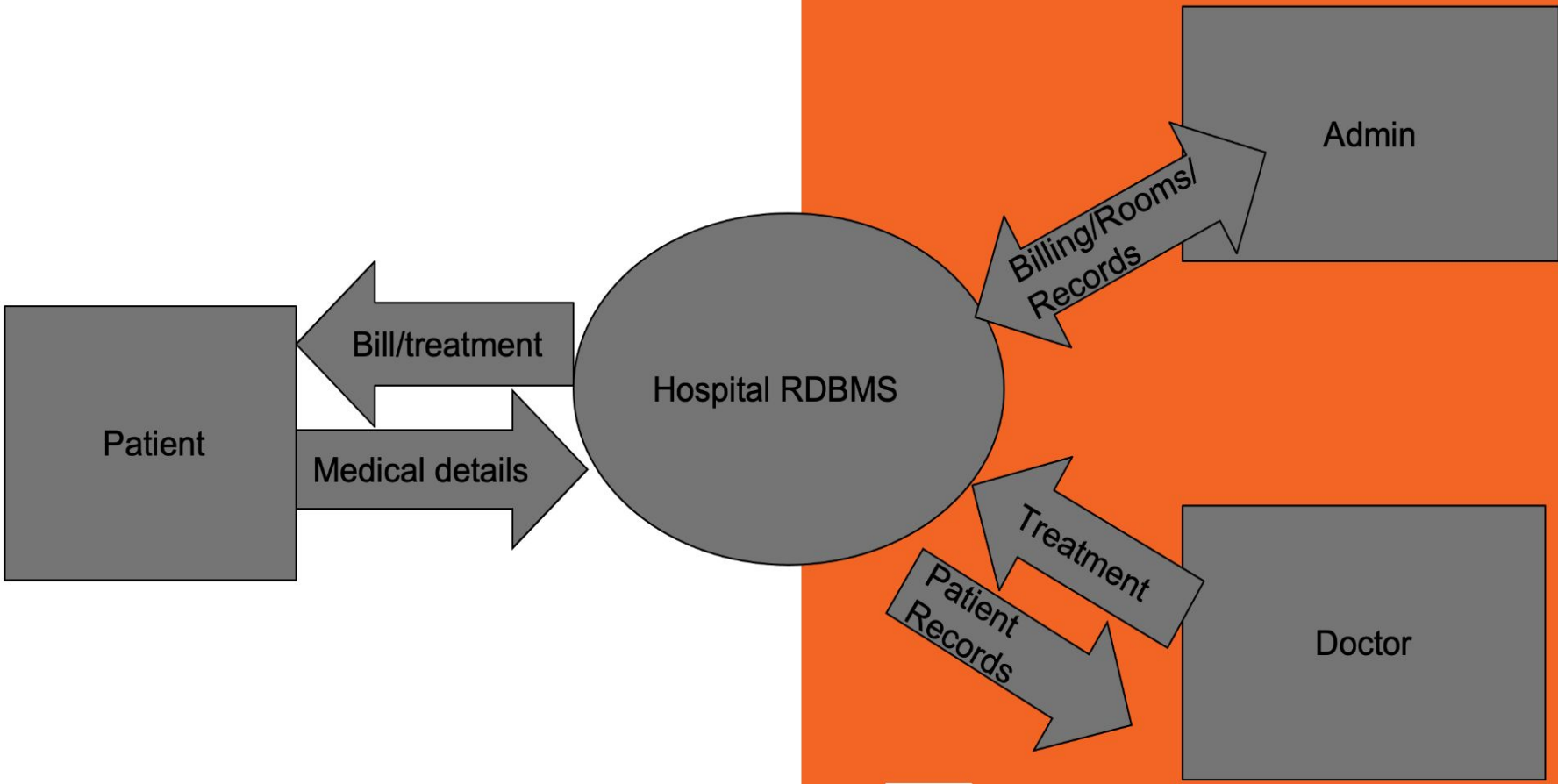
ERDs and Conceptual Models were constructed to organize the relationships between the entities of our database:

Patients, Doctors. Administrators, Rooms , Billing , Diagnoses

SMSS was utilized to populate our tables, While Access was utilized to facilitate reports and user experience

Triggers/Procedures were created to inform staff on various hospital functions




Admin
Billing/Rooms/
Records
Bill/treatment
Hospital RDBMS
Patient
Medical details
Treatment
Patient
Records
Doctor

Major Data Questions:

1. **What is the total cost of a hospital visit for any given patient?**
2. **How many Doctors are in a given department?**
3. **How can administrative assistants ensure that enough rooms are available for all inpatients?**
5. **Which doctor is the most experienced based on years of practice?**
6. **Which diseases are most common?**

What is the total cost of a hospital visit for any given patient?

totalBill_query_report

billingID	patientID	patientType	billingDate	doctorCharge	prescriptionCha	roomCharge	NumberOfDays	talRoomCharge	insuranceCover	TotalBill
100000111	1	I	9/9/2019	250	40	0	1	0	80	58
100000222	2	I	10/6/2019	125	400	100	1	100	60	250
100000333	3	I	10/6/2019	300	80	0	1	0	40	228
100000444	4	I	4/8/2019	530	0	0	1	0	10	477
100000555	5	I	8/18/2019	260	90	100	1	100	0	450
100000666	6	O	5/13/2019	120	0				55	54
100000777	7	O	11/12/2019	1000	15				100	0
100000888	8	O	11/11/2019	400	0				15	340
100000999	9	O	11/11/2019	550	100				45	357.5

Tuesday, December 3, 2019

Page 1 of 1

How many Doctors are in a given department?

NumberOfDoctors_report

departmentName	CountOfdoctor
Cardiology	1
Dermatology	1
Endocrinology	1
Gastroenterology	1
General Internal Medicine	1
Nephrology	1
Oncology	1
Pharmacology	1
Pulmonology	1

How can administrative assistants ensure that enough rooms are available for all inpatients?

dbo_ROOM_report

roomStatus	roomID	roomCharge
Occupied		
	1	100
Vacant		
	2	100
	3	100
	4	100
	5	100
	6	100
	7	100
	8	100
	9	100
	10	100
	11	100
	12	100
	13	100
	14	100
	15	100
	16	100
	17	100
	18	100
	19	100
	20	100

Who is the most experienced Doctor?

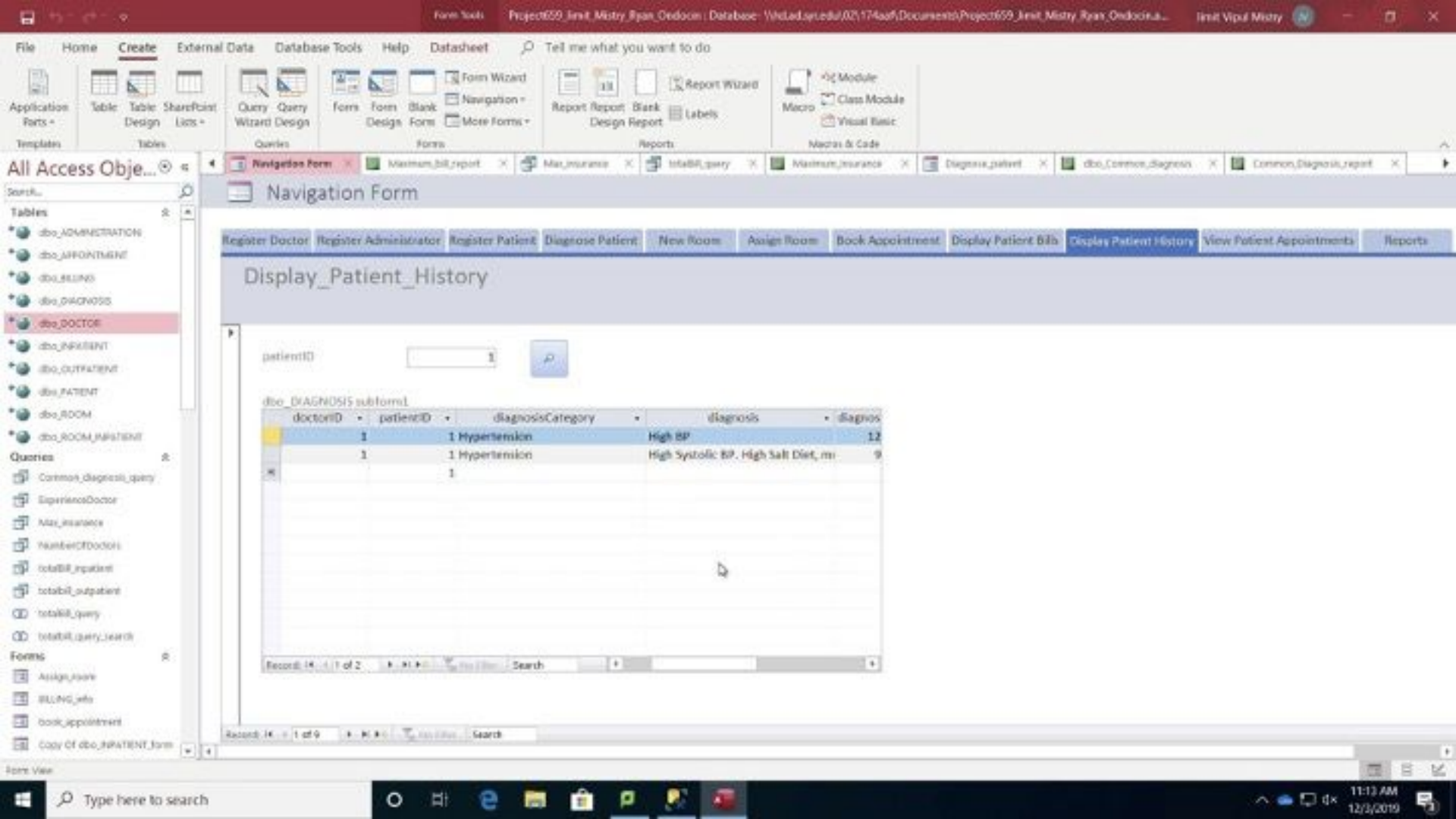
ExperienceDoctor_report1

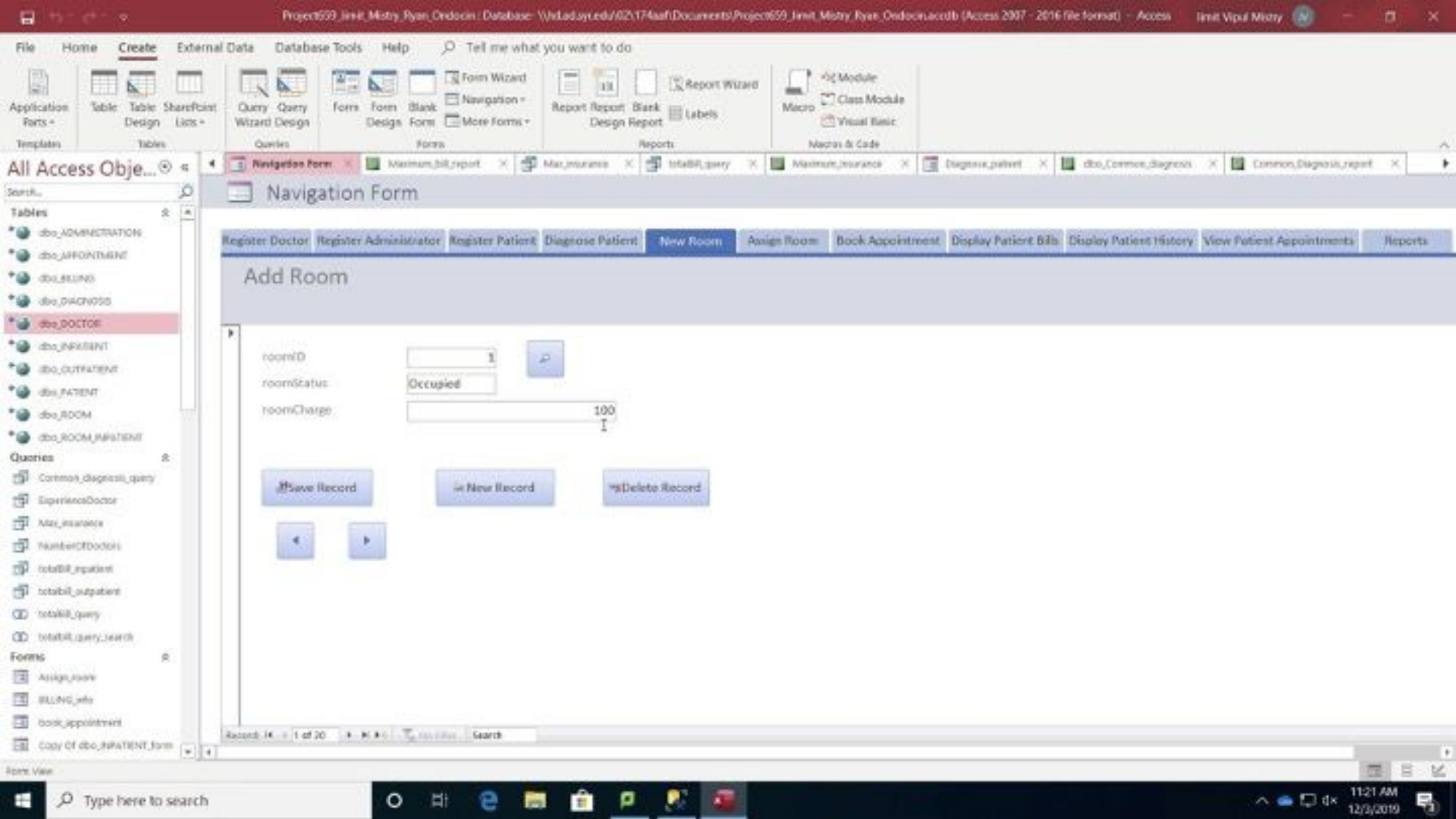
doctorID	firstName	lastName	yearsOfPractice	departmentName
8	Jeffrey	Carpenter	23	Pharmacology
1	Susan	Grey	10	General Internal Medicine
6	Phil	Kinsella	9	Endocrinology
7	Patricia	Smith	8	Pulmonology
4	Beth	Rettinger	7	Oncology
3	John	Noble	4	Dermatology
2	Chris	Billinson	2	Cardiology
5	Amy	Cote	1	Gastroenterology
9	Amanda	Shock	1	Nephrology

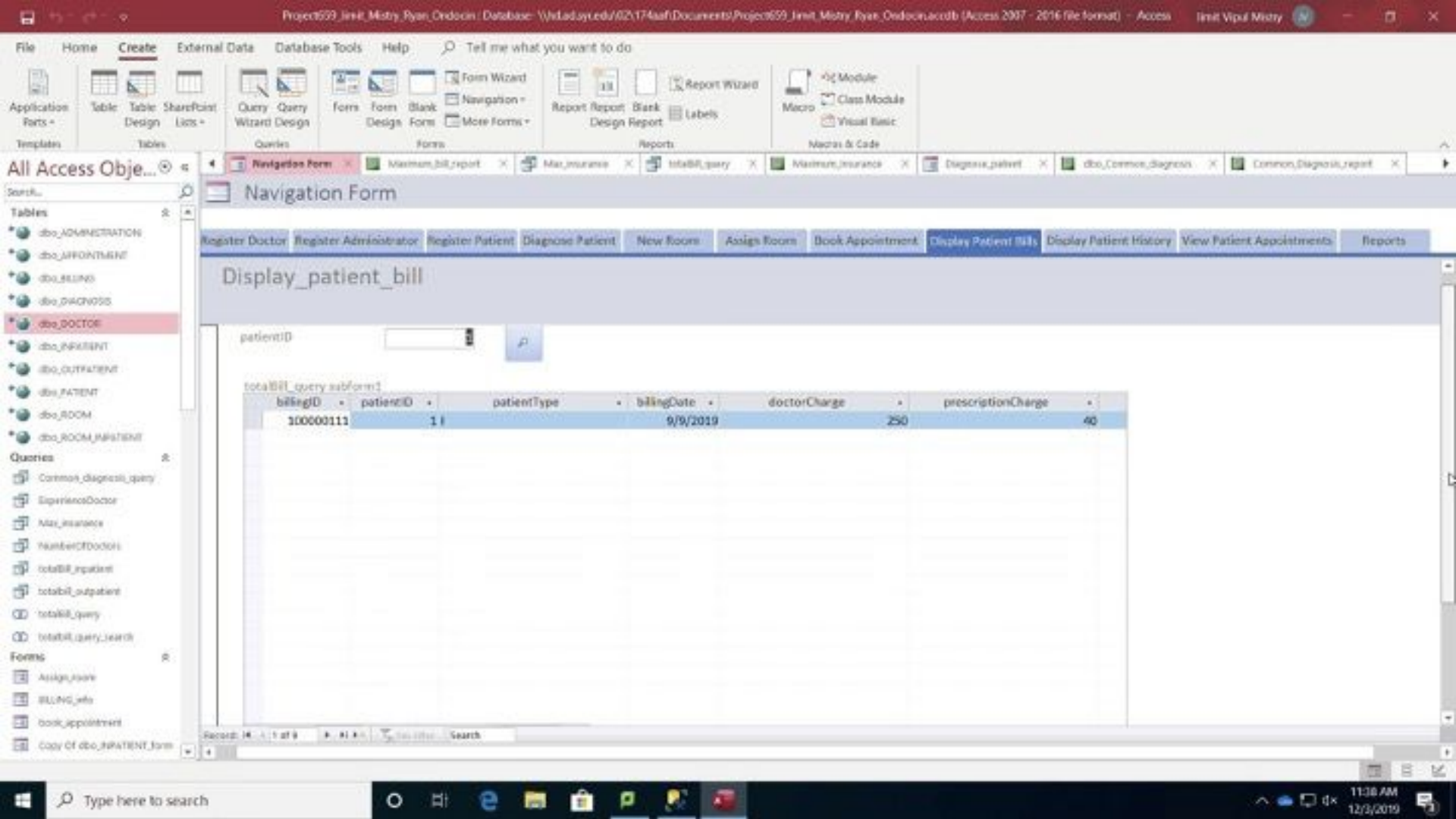
Which diseases are most common?

Common_diagnosis_query_report

CountOfdiagnosisID	diagnosisCategory
2	Hypertension
1	Hypothyroidism
1	Obesity
1	Osteoarthritis
1	Acute bronchitis
1	Allergic rhinitis
1	Anxiety
1	Back Pain
1	Diabetes


File Home Create External Data Database Tools Help Datasheet
Tell me what you want to do
Application Parts
Table Table Design SharePoint Lists
Query Wizard Query Design
Form Form Design Blank Form
Form Wizard Navigation More Forms
Report Report Design Blank Report
Report Wizard Labels
Macro
Module Class Module Visual Basic
Templates Tables Queries Forms Reports Macros & Code
All Access Obje...
Search...
Tables
dbo_ADMINISTRATION
dbo_APPOINTMENT
dbo_BILLING
dbo_DIAGNOSIS
dbo_DOCTOR
dbo_INPATIENT
dbo_OUTPATIENT
dbo_PATIENT
dbo_ROOM
dbo_ROOM_INPATIENT
Queries
Common_diagnosis_query
ExperienceDoctor
Max_insurance
NumberOfDoctors
totalBill_inpatient
totalbill_outpatient
totalbill_query
totalbill_query_search
Forms
Assign_room
BILLING_info
book_appointment
Copy Of dbo_INPATIENT_form
Navigation Form
Maximum_bill_report
Max_insurance
totalBill_query
Maximum_insurance
Diagnosis_patient
dbo_Common_diagnosis
Common_Diagnosis_report
Navigation Form
Register Doctor
Register Administrator
Register Patient
Diagnose Patient
New Room
Assign Room
Book Appointment
Display Patient Bills
Display Patient History
View Patient Appointments
Reports
Display_Patient_History
patientID
1
dbo_DIAGNOSIS subform1
doctorID
patientID
diagnosisCategory
diagnosis
diagnos
1
1 Hypertension
High BP
12
1
1 Hypertension
High Systolic BP, High Salt Diet, m
9
1
Record: 1 of 2
Search
Record: 1 of 9
Search
Form View
Type here to search
11:13 AM
12/3/2019


Project659_Jimit_Mistry_Ryan_Ondocin : Database: \\sdLaJay.edu\02\174aaf\Documents\Project659_Jimit_Mistry_Ryan_Ondocin.accdb (Access 2007 - 2016 file format) - Access
Jimit Vipul Mistry
File
Home
Create
External Data
Database Tools
Help
Tell me what you want to do
Application Parts
Table
Table Design
SharePoint Lists
Query Wizard
Query Design
Form
Form Design
Blank Form
Form Wizard
Navigation
More Forms
Report
Report Design
Blank Report
Report Wizard
Labels
Macro
Module
Class Module
Visual Basic
Templates
Tables
Queries
Forms
Reports
Macros & Code
All Access Obje...
Search...
Tables
dbo_ADMINISTRATION
dbo_APPOINTMENT
dbo_BILLING
dbo_DIAGNOSIS
dbo_DOCTOR
dbo_INPATIENT
dbo_OUTPATIENT
dbo_PATIENT
dbo_ROOM
dbo_ROOM_INPATIENT
Queries
Common_diagnosis_query
ExperiencedDoctor
Max_insurance
NumberOfDoctors
totalBill_inpatient
totalbill_outpatient
totalbill_query
totalbill_query_search
Forms
Assign_room
BILLING_info
book_appointment
Copy Of dbo_INPATIENT_form
Navigation Form
Maximum_bill_report
Max_insurance
totalBill_query
Maximum_insurance
Diagnose_patient
dbo_Common_diagnosis
Common_Diagnosis_report
Navigation Form
Register Doctor
Register Administrator
Register Patient
Diagnose Patient
New Room
Assign Room
Book Appointment
Display Patient Bills
Display Patient History
View Patient Appointments
Reports
Add Room
roomID
1
roomStatus
Occupied
roomCharge
100
Save Record
New Record
Delete Record
Record: 1 of 20
Search
Form View
Type here to search
11:21 AM
12/3/2019


Navigation Form
Register Doctor
Register Administrator
Register Patient
Diagnose Patient
New Room
Assign Room
Book Appointment
Display Patient Bills
Display Patient History
View Patient Appointments
Reports
Display_patient_bill
patientID
totalBill_query subform1
billingID
patientID
patientType
billingDate
doctorCharge
prescriptionCharge
100000111
1
9/9/2019
250
40

IST 659 Reflection:

Developing a Hospital RDBMS taught many valuable lessons in terms of understanding the efficiency of data storage, access, and report generation.

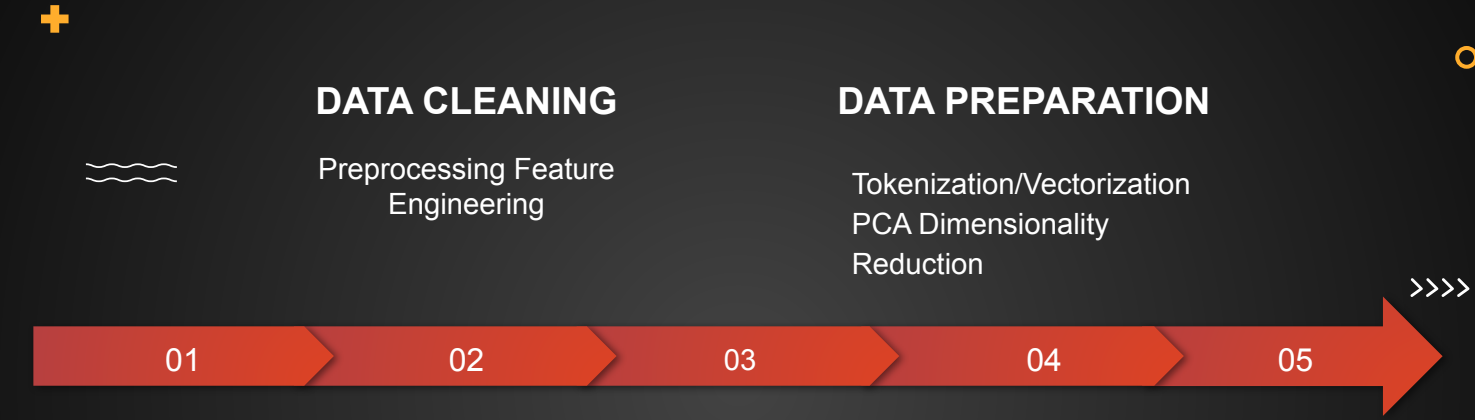
The nature of the project is centralized around dealing with sensitive patient data which means a ton of research was put into HIPAA compliance and User/Admin privileges.
This application is applicable to the field of healthcare analytics often requires pristine information management in order to make critical business decisions

- **Operational Reporting was automated using skills from tools such as Microsoft SQL Server Management Studio and MS Access**

IST 664: Natural Language Processing
Textual Clustering of COVID-19 Related Tweets

- Through studying NLP under the direction of Dr. Lu Xiao textual processing and analysis techniques were covered in order to extract valuable insight from unstructured data.

- In this project, PCA with K-Means and T-SNE to visualize the transmission of information regarding the Coronavirus pandemic on Twitter


DATA CLEANING
Preprocessing Feature Engineering
DATA PREPARATION
Tokenization/Vectorization
PCA Dimensionality Reduction
01
02
03
04
05
Web Scraping
DATA COLLECTION
Popular users /
Most common words
EDA
K-means clustering model
t-SNE for visualization
MODELING

STEP ONE: DATA COLLECTION

---- Web Scraping with Tweepy

[2]
```
accessing credentials file to scrape twitter API
with open(credentials) as cred_data:
        info = json.load(cred_data)
        consumer_key = info['CONSUMER_KEY']
        consumer_secret = info['CONSUMER_SECRET']
        access_key = info['ACCESS_KEY']
        access_secret = info['ACCESS_SECRET']
        #google_api = info['GOOGLE_API']
auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_key, access_secret)
api = tweepy.API(auth, wait_on_rate_limit=True)
```

Adding relvant"coronavirus" search words and setting time frame to tweets since May 1,

[3]
```
search_words = ['coronavirus']
date_since = '2020-05-01'
```

[4]
```
#using tweepy.Cursor to format dataframe
tweets = tweepy.Cursor(
   api.search,
   q = search_words,
   lang = 'en',
   since = date_since,
   tweet_mode='extended' #attempt to retrieve fill_text from truncated tweets
).items(1500)
```

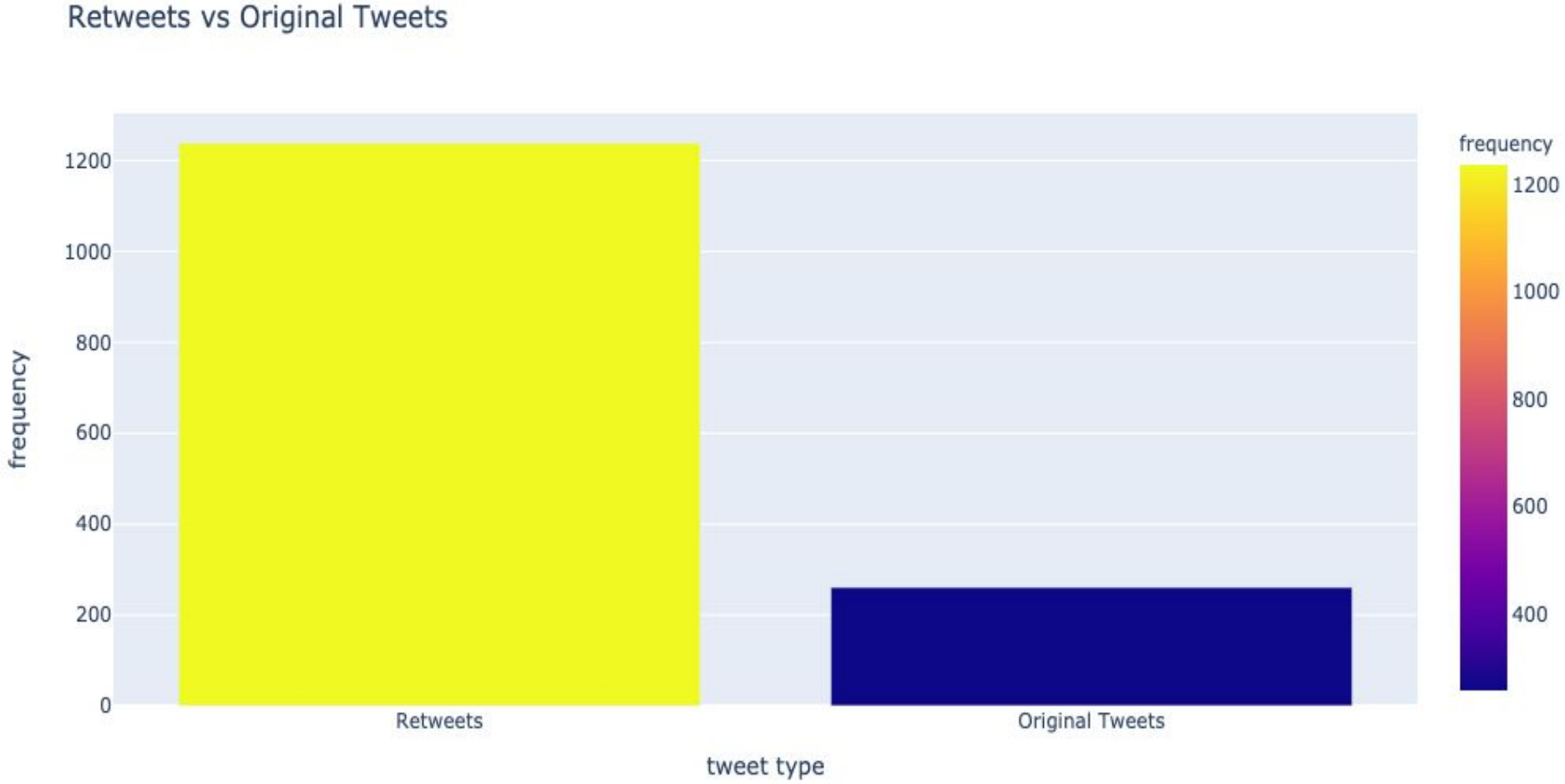
[5]
```
collecting tweet user and location data avaiable (we settled on one dataframe for this project)
tweet_details = [[tweet.full_text, tweet.user.screen_name, tweet.user.location] for tweet in tweets]
tweet_df = pd.DataFrame(data=tweet_details, columns=['text','user','location'])
```

Raw Dataset

	text	user	location
0	RT @mattyglesias: How do you keep people sane?\n\nOutside activity with reasonable precautions looks very safe. Open up the parks and beaches…	CeCe_23Spalding	Brooklyn, NY from Queens
1	RT WAXIEbuzz "RT GreenSeal: Our friends WAXIEbuzz have a #COVID19 prevention guide and info about their disinfectants. Check it out. \n\nhttps://t.co/fri6QXKU18"	HurleySupply	
2	'An Anvil Sitting on My Chest': What It's Like to Have Covid-19 https://t.co/84eslDocVG	Elle2Tha	Los Angeles, CA
3	RT @CNN: An elementary school teacher in Connecticut says she is taking care of a newborn baby boy after getting a phone call from a stude…	Pelleg1Gabriell	Tokyo and Nagoya, Japan
4	The magnitude of human loss from #COVID19 is massively visualized 👇 \nU.S. Mortality: Death Certificates Listing Pneumonia, Influenza, and COVID-19 I CDC\nhttps://t.co/Wd56eeimY4 https://t.co/v0HVP9aSMB	mzkhalil	Cairo - Egypt
...	...	...	...
1495	RT @migov: Stay Home. Stay Safe. Save Lives. The state of Michigan (@migov) & @MichiganHHS report today, May 6, 2020, 657 new COVID-19 case…	MizJeniJonze	
1496	RT @V2019N: A genetic study of samples from more than 7,500 people infected with COVID-19 suggests the new coronavirus spread quickly aroun…	staywoketravel	Los Angeles, CA
1497	RT @Johnrashton47: 🦠 Remember that it's about twice this. If you want the truth check the Financial Times. Follow the money. 🦠 Coronavirus de…	chris_traynor	
1498	Had my Covid-19 antibody test today. Thanks @thesolutioniv for a great easy, quick, and clean experience. Looking forward to getting my results!!! #swissqualitysmile #covid19 #coronavirus #antibodytest… https://t.co/23a4uDHjwL	swissqsmile	Los Angeles, CA
1499	RT @BrianDeLay: In 1847, impoverished Choctaws scraped together sent $170 to send to victims of the Irish potato famine. Some Irish remembe…	ImpeachBDevos	Tell the truth and you shame Trump. - Shakespeare paraphrased

- Locational Inconsistencies
- Twitter data vs. Regular Text

Relying solely on metadata (retweets/mentions/location/time) would deduct from the value of the uniquely textual approach we are attempting.


Retweets vs Original Tweets
frequency
1200
1000
800
600
400
200
0
Retweets
Original Tweets
tweet type
frequency
1200
1000
800
600
400

STEP 2: Data Cleaning via Regular Expressions

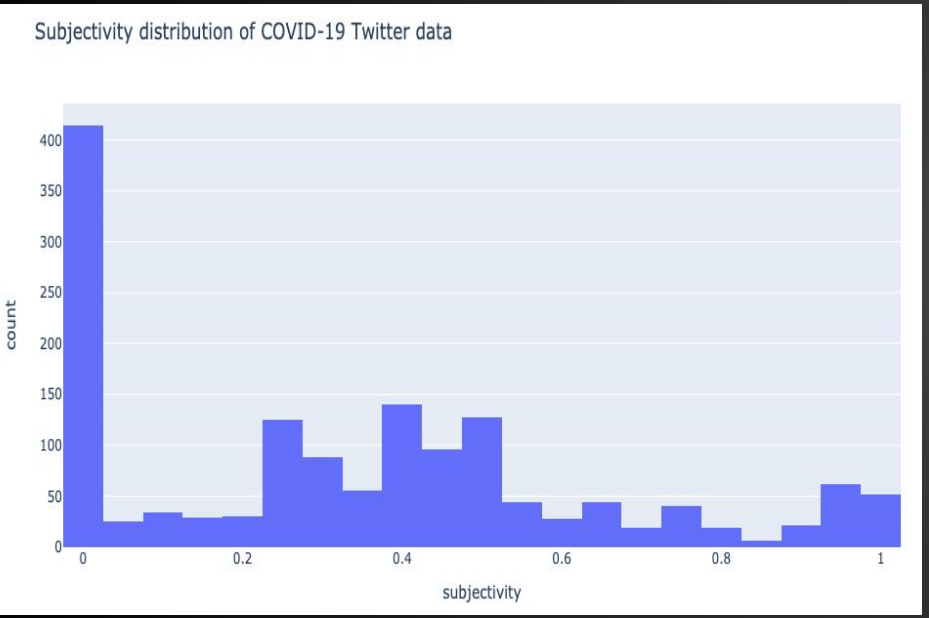
```
# clean tweet
def clean_tweet(tweet):
    # Remove hyperlinks
    tweet = re.sub("https?://[A-Za-z0-9./]*","",tweet)
    # Remove hashtags
    tweet = re.sub(r'#\w*', '', tweet)
    # Remove tickers
    tweet = re.sub(r'\$\w*', '', tweet)
    #@user -> at_user
    tweet = re.sub("@[\w]*","",tweet)
    # To lowercase
    tweet = tweet.lower()
    # Remove Punctuation and split 's, 't, 've with a space for filter
    tweet = re.sub(r'[' + punctuation.replace('@', '') + ']+', ' ', tweet)
    # Remove words: I, a , am, me (2 or less letters)
    tweet = re.sub(r'\b\w{1,2}\b', '', tweet)
    # Remove whitespace (including new line characters)
    tweet = re.sub(r'\s\s+', ' ', tweet)
    # Remove single space remaining at the front of the tweet.
    tweet = tweet.lstrip(' ')
    # Remove emojis or other. special characters
    tweet = ''.join(c for c in tweet if c <= '\uFFFF')
    return tweet
```

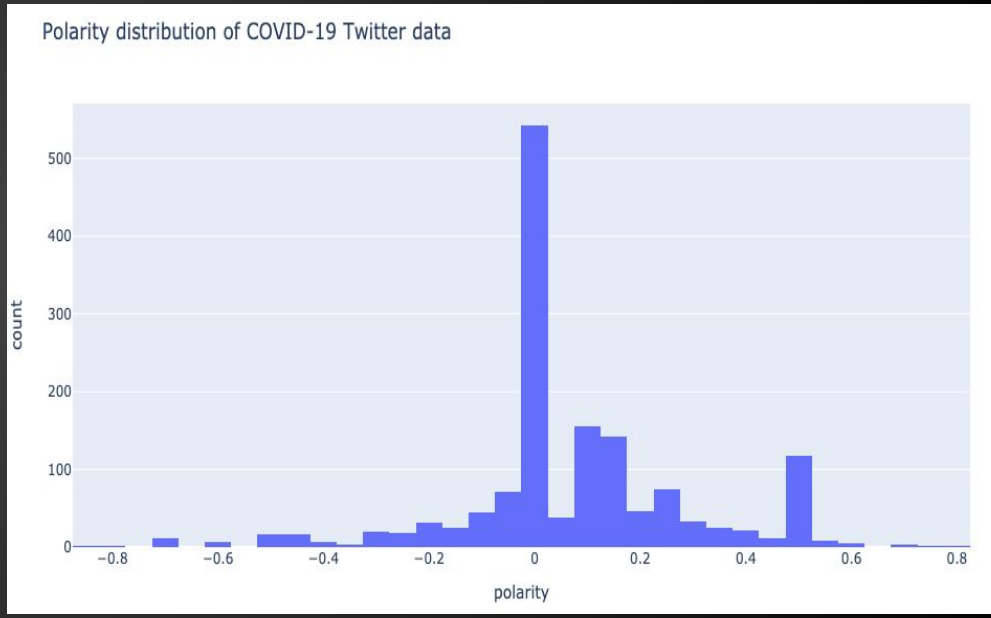
```
[ ] from string import punctuation
    #extracting hashtags as a means of possible cluster validation
    def extraxt_hashtags(text):
      hashtags = re.findall(r'\B#\w*[a-zA-Z]+\w*', text)
      return hashtags
```

```
[ ]  tweet_df['hashtag'] = tweet_df['text'].apply(lambda x: extraxt_hashtags(x))
     tweet_df['clean_text'] = tweet_df['text'].apply(lambda x: clean_tweet(x))
     tweet_df.head()
```

	text	user	location	hashtag	clean_text
0	RT @azcentral: JUST IN: The Arizona Department of Health Services told a team of university experts working on COVID-19 modeling to "pause"…	DesertCarmen	Chandler, AZ	[]	just the arizona department health services told team university experts working covid modeling pause …
1	RT @FOX5Vegas: UPDATE: Drive-thru COVID-19 is underway at the Orleans hotel-casino. @dylankendricktv explains how the appointment-only test…	charlienicole	Nevada, USA	[]	update drive thru covid underway the orleans hotel casino explains how the appointment only test…
2	RT @JenningsK12: @CareSTLHealth has created a new COVID-19 testing site at Fairview Elementary School, 7047 Emma Ave, St. Louis, MO 63136.…	CareSTLHealth	St. Louis	[]	has created new covid testing site fairview elementary school 7047 emma ave louis 63136 …
3	RT @InclusionPhilly: Thanks to @nbcphiladelphia and @mitchreports for a great interview profiling our efforts at @theFPCN to expand testing…	mitchreports	Philadelphia, PA	[]	thanks and for great interview profiling our efforts expand testing…
4	RT @WSJ: Covid-19 is taking workplace surveillance to a higher level, with some employers planning to track movements and gather personal i…	KELLYCLELLAND1	Lothian	[]	covid taking workplace surveillance higher level with some employers planning track movements and gather personal …

Tweet Sentiment via Text Blob





Stopword Removal

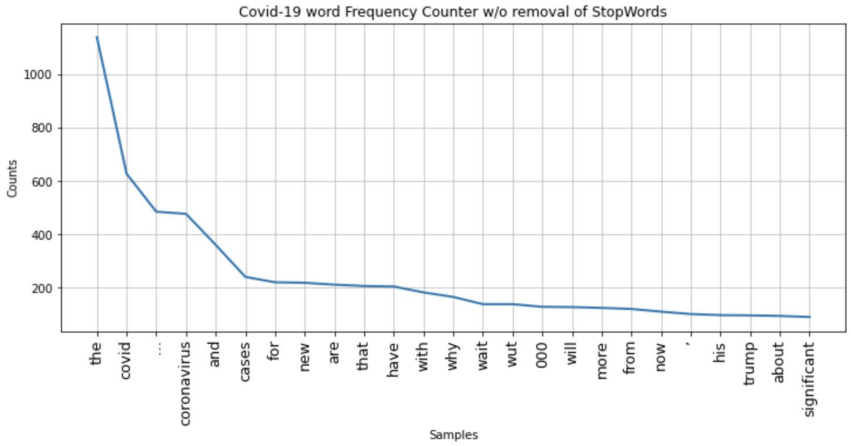
```
##### custom_stop_words removal
stom_stop_words = [ '...',',','000','"', 'the', "-",
  'covid','coronavirus','covid-19','coronavirus','virus','COVID','CoronaVirus','Coronavirus','Covid-19','covid-19']

r w in custom_stop_words:
  if w not in stopwords:
      stopwords.append(w)
```

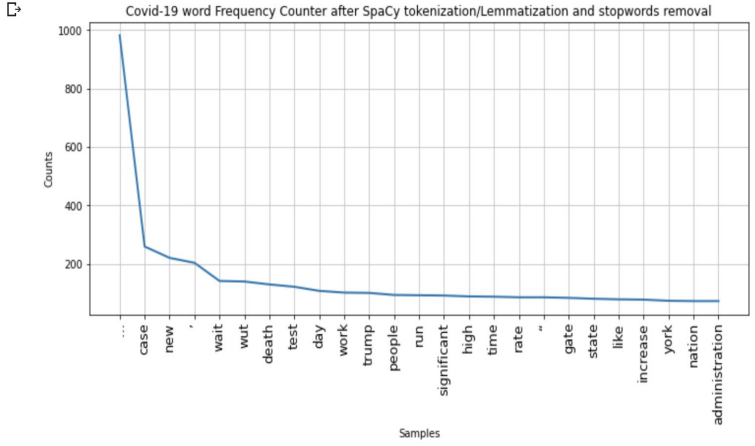
text	user	location	hashtag	cleaned_text	text_word_count	tweet_unique_words	processed_text
Do Hospitals get paid more if patients are listed as #COVID19, on ventilators? Yes, a lot more. Couple this with stats sayin…	ChannelJammer	TOPSIDE OF INDIANA	['#COVID19']	Do Hospitals get paid more if patients are listed as on ventilators Yes a lot more Couple this with stats sayin…	21	20	hospitals pay patient list ventilator yes lot couple stats sayin …
“The biggest conversation for city-builders during the pandemic is the role of streets as a principal public space in ci…	lb_1950	NaN	[]	“The biggest conversation for city builders during the pandemic is the role of streets as a principal public space in ci…	21	20	“ big conversation city builder pandemic role street principal public space ci …
Defense Minister : IIBR revealed a breakthrough #coronavirus vaccine that attacks the virus and neutrali…	maja_la	NaN	['#coronavirus']	Defense Minister IIBR revealed a breakthrough vaccine that attacks the virus and neutrali…	13	13	defense minister iibr reveal breakthrough vaccine attack virus neutrali …
A "much-loved" London-based eye doctor who warned about the dangers of underestimating #coronavirus has died after	rach670	NaN	['#coronavirus']	A much loved London based eye doctor who warned about the dangers of underestimating has died after	18	18	love london base eye doctor warn danger underestimate die contract …

STEP 3: Exploratory Data Analysis (EDA)

```
# most common words in twitter dataset
all_words = []
for line in list(df['clean_text']):
    words = line.split()
    for word in words:
        all_words.append(word.lower())
# plot word frequency distribution of first few words
plt.figure(figsize=(12,5))
plt.xticks(fontsize=13, rotation=90)
plt.title("Covid-19 word Frequency Counter w/o removal of StopWords")
fd = nltk.FreqDist(all_words)
fd.plot(25,cumulative=False)
```

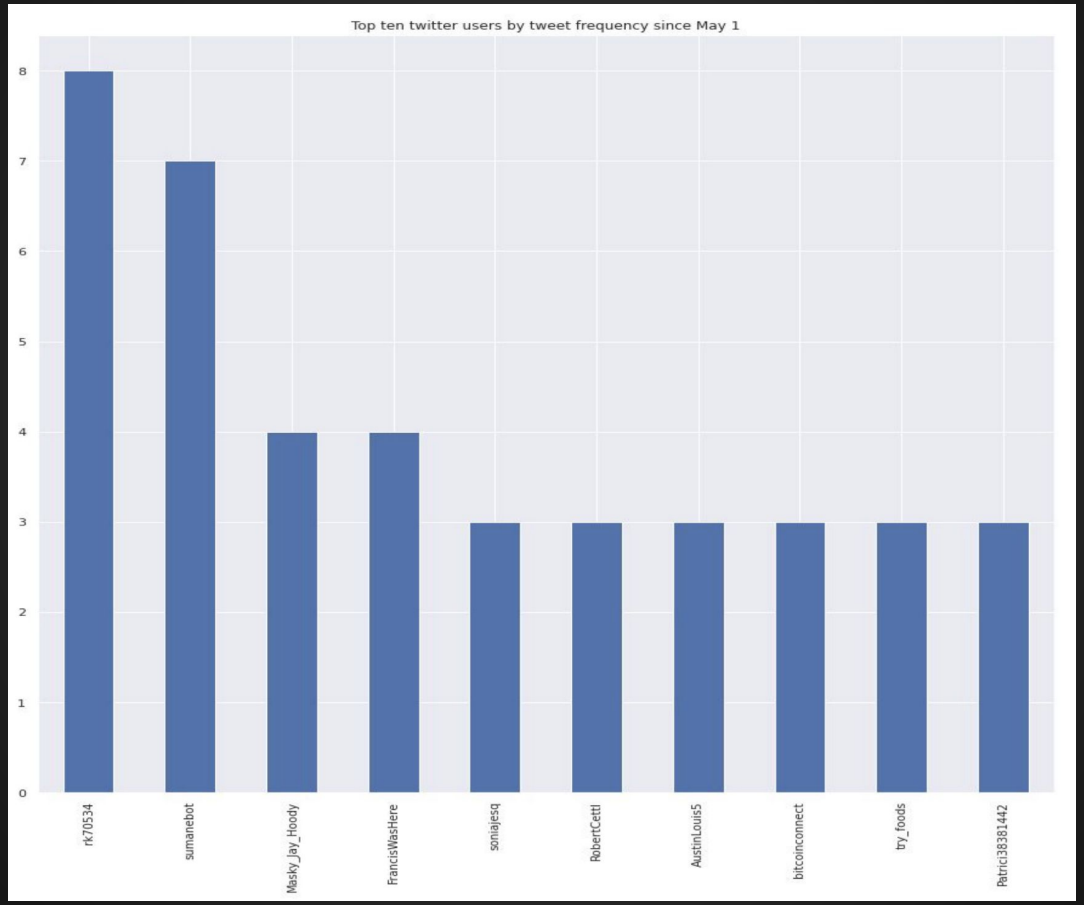


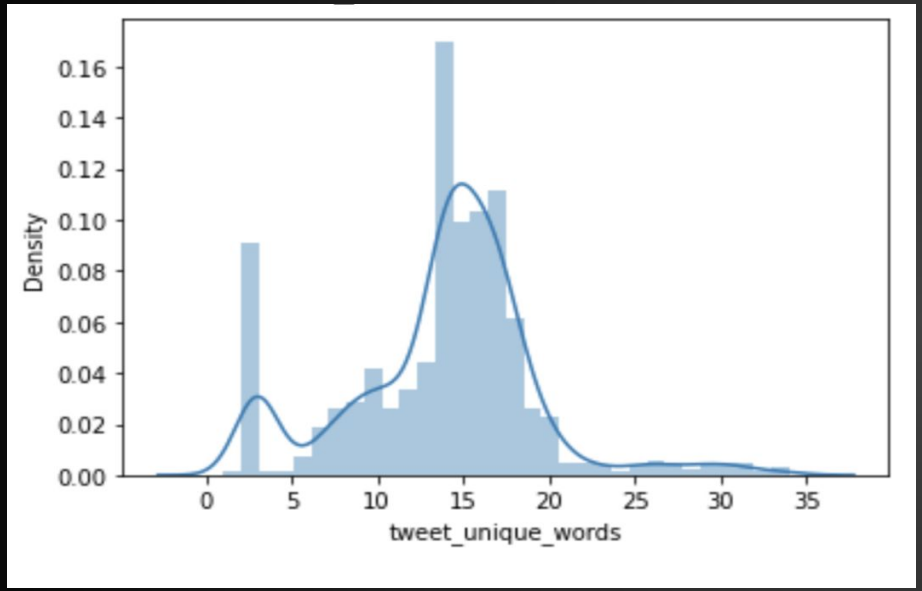

```
# most common words in processed twitter data
all_words = []
for line in list(df['processed_text']):
    words = line.split()
    for word in words:
        all_words.append(word.lower())
# plot word frequency distribution of first few words
plt.figure(figsize=(12,5))
plt.xticks(fontsize=13, rotation=90)
plt.title("Covid-19 word Frequency Counter after SpaCy tokenization/Lemmatization and stopwords removal")
fd = nltk.FreqDist(all_words)
fd.plot(25,cumulative=False)
```



Most Common Words

Top 10 Twitter Users by Tweet Frequency since May 1





- Interesting Bimodal trend observed
- ~3 word we see a peak and also ~ 12 words we see a larger peak. These could possibly be sourced to popular retweeted coronavirus data

STEP 4&5:

DATA PREPARATION & MODELING

- Tokenize/Vectorize Data
- PCA
- K Means
- Dimensionality Reduction with t-SNE
- Visualize

TF-IDF

Converts string into a measure of how important each word is to the instance out of the tweet as a whole. Word Embedding is a NLP technique for mapping words to vectors of real numbers. It represents words or phrases in vector space with several dimensions.

```
[ ]  from sklearn.feature_extraction.text import TfidfVectorizer #creating a vectorize function for our text data
     def vectorize(text, max_features):

         vectorizer = TfidfVectorizer(max_features=max_features)
         X = vectorizer.fit_transform(text)
         return X
```

$$Term\ Frequency\ (TF) = \frac{Number\ of\ times\ the\ word\ or\ the\ term\ appears\ in\ the\ document}{Total\ number\ of\ words\ or\ terms\ in\ the\ document}$$

$$Inverse\ Document\ Frequency\ (IDF) = \log\left(\frac{Total\ number\ of\ documents}{Number\ of\ documents\ which\ have\ that\ word}\right)$$

Vectorizing Processed Text: Potential overfitting if there are more words than documents (tweets) it's over encompassing so we're not relating back to our original corpus

```
[ ] #Convert a collection of raw documents to a matrix of TF-IDF features.
    #CountVectorizer followed by TfidfTransformer.
    text = df['processed_text'].values
    X = vectorize(text, 1195) #1195 features derived from Interactive plot on Kmeans clustering using Tf-IDF
    X.shape

    (1500, 1195)
```

Following feature_engineering and preprocessing, we need to numerically transform the data into a format that can be handled by our K-means algorithm. These vectors will have to be mapped into a lower dimension via dimensionality reduction(PCA) and t-SNE

PCA - Dimensionality Reduction

```
[ ]  from sklearn.decomposition import PCA

     pca = PCA(n_components=0.90, random_state=42) #90
     X_reduced= pca.fit_transform(X.toarray())
     X_reduced.shape #decreasing n_components or variance in this datframe also led to a decrease in our feature vector.

     (1500, 290)
```

Principal Component Analysis:

Used to reduce dimensions Tf-Idf features while maintaining 90% variance.

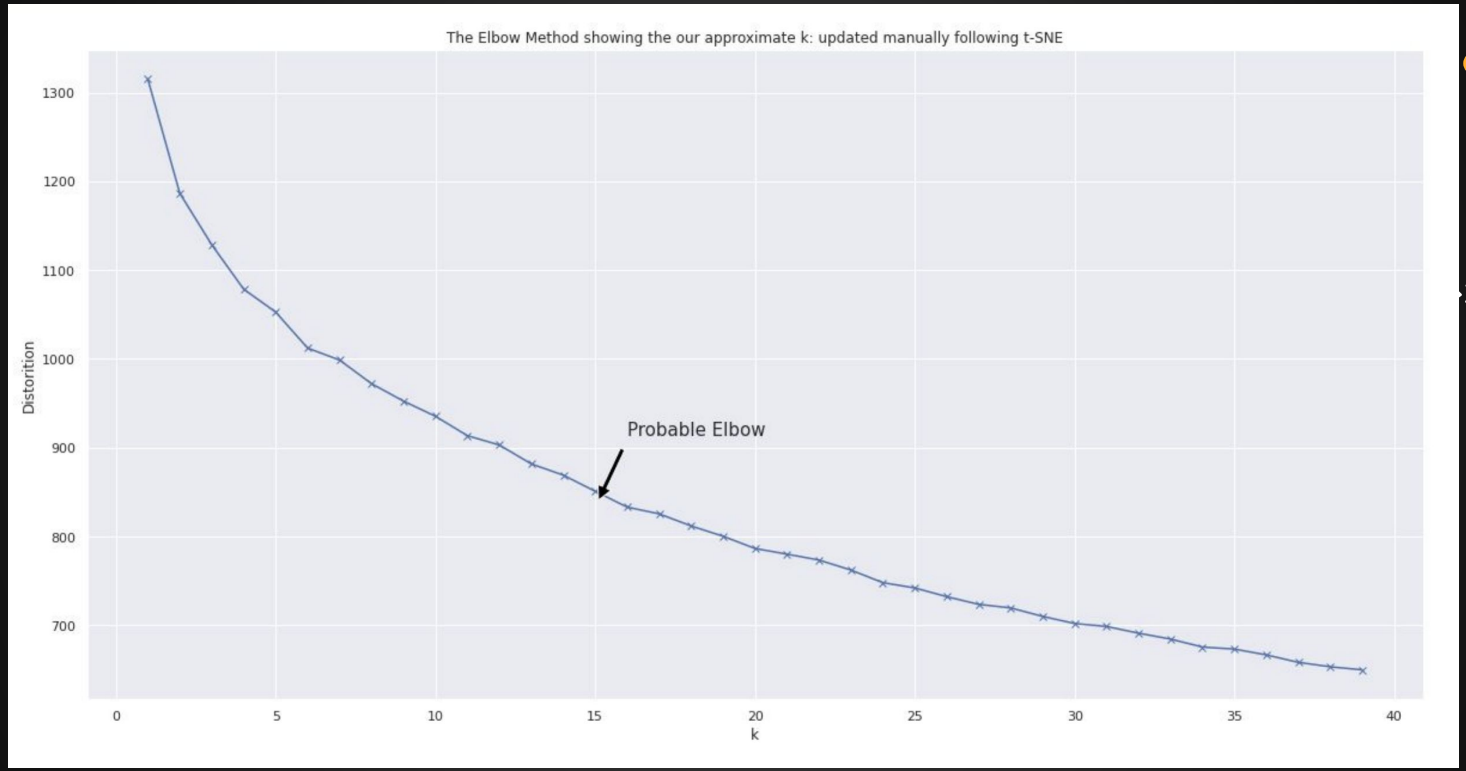
Lowered from 95% in order to obtain a 290 features (opposed to 379).

remove some noisy outliers from our twitter data, to make the clustering process easier for k-means.

K-Means Clustering

This is an unsupervised learning algorithm that will use the elbow curve method to plot distortion vs the number of clusters K. This allows are data to be naturally segregated solely based on our vectorized text. The algorithm is distance-based and tries to simply minimize the SSE (intra or within-cluster variation) to produce coherent clusters.

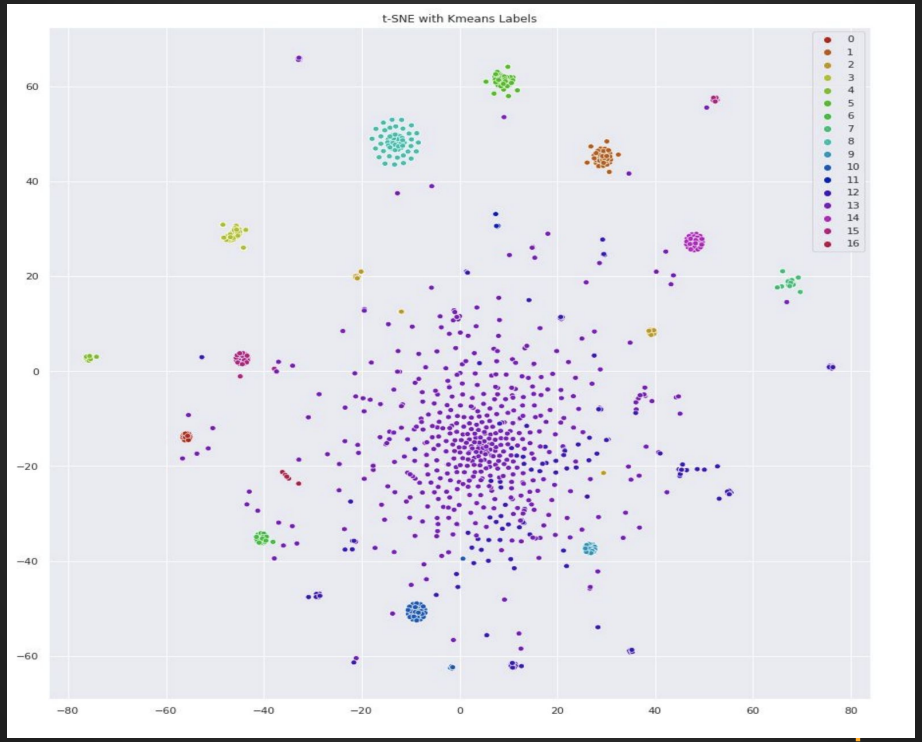
For tweet distinction, k-means will be run on the vectorized text.

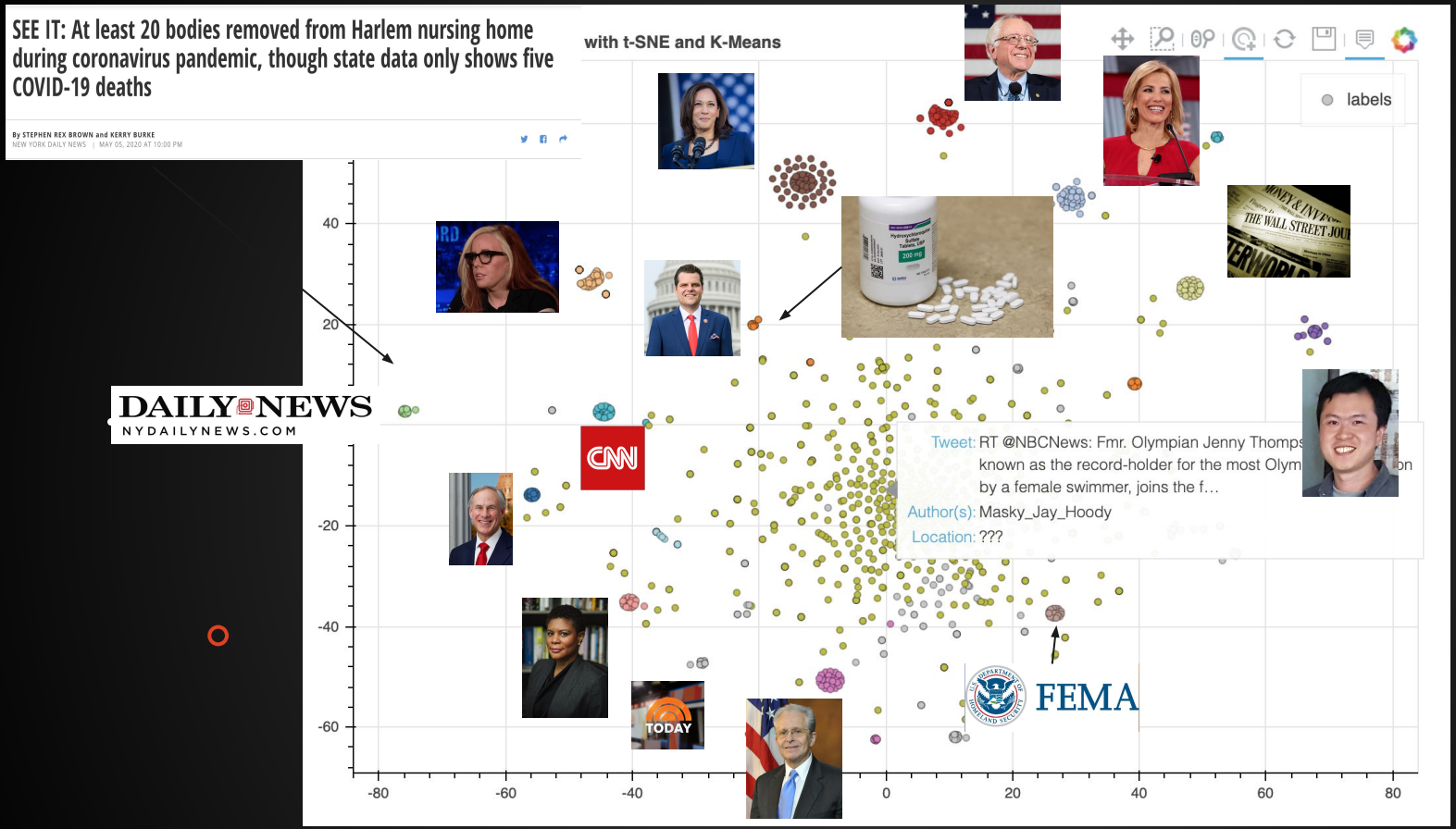

The Elbow Method showing the our approximate k: updated manually following t-SNE
Distortion
k
Probable Elbow
1300
1200
1100
1000
900
800
700
0
5
10
15
20
25
30
35
40

t-Stochastic-Neighbor Embedding

- imported from the sklearn manifold package
- visually map our 290 features onto a 2-dimensional plane. This is done by
- minimizing the KL divergence between joint probabilities of the low-dimensional (1500, 290) and high-dimensional embedding (1500,1195).
- t-SNE's cost function will keep most of the context-based word vectors that we want without being too large to process.
- t-SNE will use the original feature vector X(1500,1195) that was obtained via tf-idf on the preprocessed text. Following this, we have a visual representation of our tweets! Minimization of KL divergence with respect to our observations. It is optimized via gradient descent and the result s a map that shows similarities between our high-dimensional features.

t-SNE with k-Means labels



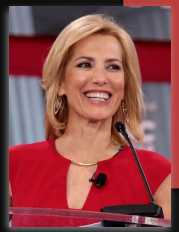


Link to Interactive Bokeh Plot

https://ryanondocin2019.github.io/

@KamalaHarris
Native American tribes like the Navajo Nation are contracting coronavirus at disproportionately high rates. The Treasury must immediately distribute the $8 billion in relief funds for Tribes. They can't wait.

@BernieSanders
At a time when his own administration is predicting a significant increase in COVID-19 cases, our "stable genius" president is winding down his coronavirus task force. This is a true American tragedy. Thousands will unnecessarily die because of Trump's contempt for science.



@IngrahamAngle
Gates should run for president. He wants to run your lives...New York will work with the Gates Foundation to develop a blueprint to "reimagine education" in the COVID-19 era.

@WSJ
"Brazil is already the global epicenter of the coronavirus." One study concludes the country might have more Covid-19 cases than the U.S.

@triblelaw
FEMA now projects 200,000 new coronavirus cases a day by the end of the month (it's now about 25,000 new cases a day), and by June 1, about 3,000 deaths daily from Covid-19. Trump is making BS excuses for suppressing those awful numbers while he urges people to go back to work.

@cnni
The virus that causes Covid-19 has been circulating in people since late last year, and must have spread extremely quickly after the first infection, a new genetic analysis shows

@alondra
'We Are Not Essential. We Are Sacrificial.'
I'm a New York City subway conductor. At least 98 of my co-workers have died of Covid-19.

IST 664 Reflection:

This project provided the opportunity for the collection and structuring of externally-sourced data via Twitter's API.

identification of patterns between clusters of tweets helped us develop insights into the transmission of information on social media platforms regarding COVID-19. These insights provide a birds-eye view of what's happening on your social media feed to get a sense of how misinformation could spread and who are the key players in the network

- The privacy of twitter Users was considered as we only scraped relevant textual data (opposed to geolocation, or demographic information about users).
- With such a large volume of data being generated each day, it's imperative for companies to effectively analyze and understand the ecosystem of social networking platforms.

2020 Election Sentiment

Group 9

Jenny Cao

Prathamesh Datar

Ryan Ondocin

Shripad Laddha

Project Overview

- Role of social media in elections
- Perform sentiment analysis for 2020 US election
- Supervised and Unsupervised machine learning algorithms
- NLP techniques on twitter data to deduce inferences

Inference, Prediction and other Goals

- Which words were the most important in predicting sentiment label?
- How do the frequency of hashtags provide a sentimental snapshot of the platform or the election?
- Is there any clear polarity in tweets to conclude presence of communities?
- Can we find users that differ significantly from other users?

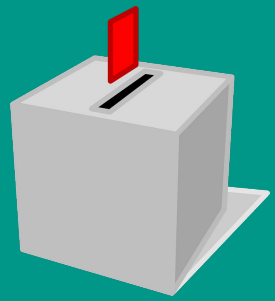
Dataset

- This project contains two datasets: Data before election result (97,200 rows and 7 columns) and data after election result (45370 rows and 7 columns)
- Used Tweepy for extracting election data

Tweepy Limitation

- Twitter's standard API allows users to retrieve tweets within the last Seven days and is limited to scraping 18,000 tweets within every 15 minutes.

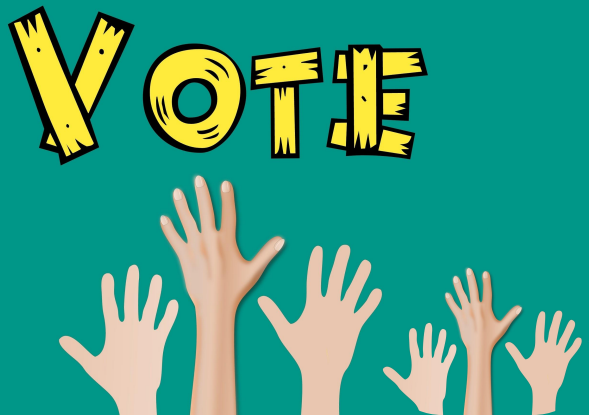
Data Cleaning and Feature Engineering



- Removing retweeted data
- Generating hashtags column
- Removing special signs from tweets
- Sentiment analysis TextBlob
- Preprocessing and feature engineering
- Tokenization
- Vectorization(TF-IDF)

word	idf
t	0.521332621019889
https	0.5758754084260893
biden	0.7784479456907105
election	0.7955633165449632
trump	0.810699050180526
s	0.9523054931144004
vote	1.0246000062483462
joe	1.1326859154296876
donaldtrump	1.2586016929042994
joebiden	1.3032014500718612
harris	1.3286843932760455
realdonaldtrump	1.3551450579531341
debate	1.3658638331909358
kamalaharris	1.39276706295098
amp	1.4627216755602244
covid	1.5109496449902888
just	1.5156384298795302
president	1.5248695823090688
bidenharris	1.5358418159678173
people	1.5593057639887036

The 20 least important IDF scores



Methodology

Exploratory Analysis

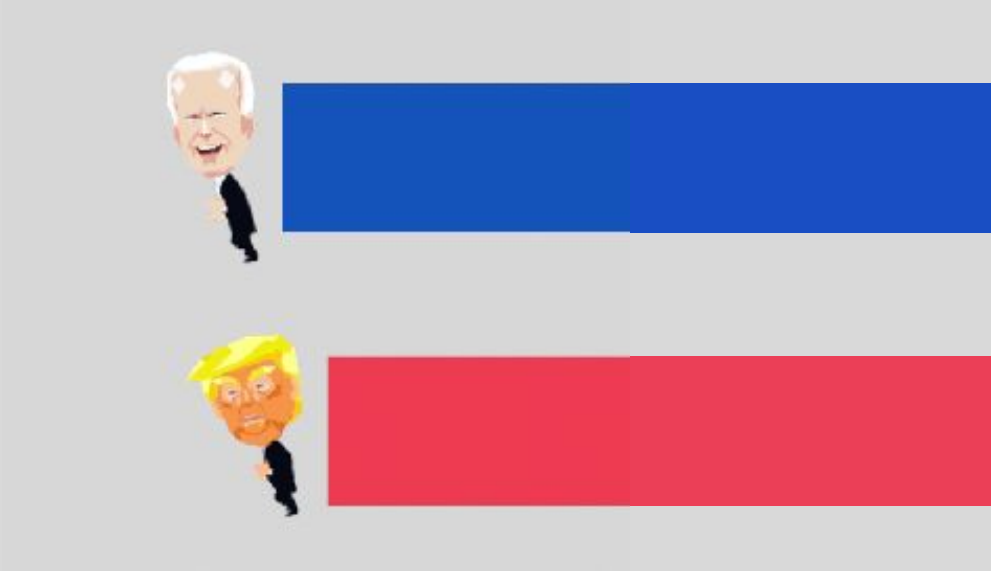
- Hashtag Analysis

Supervised learning

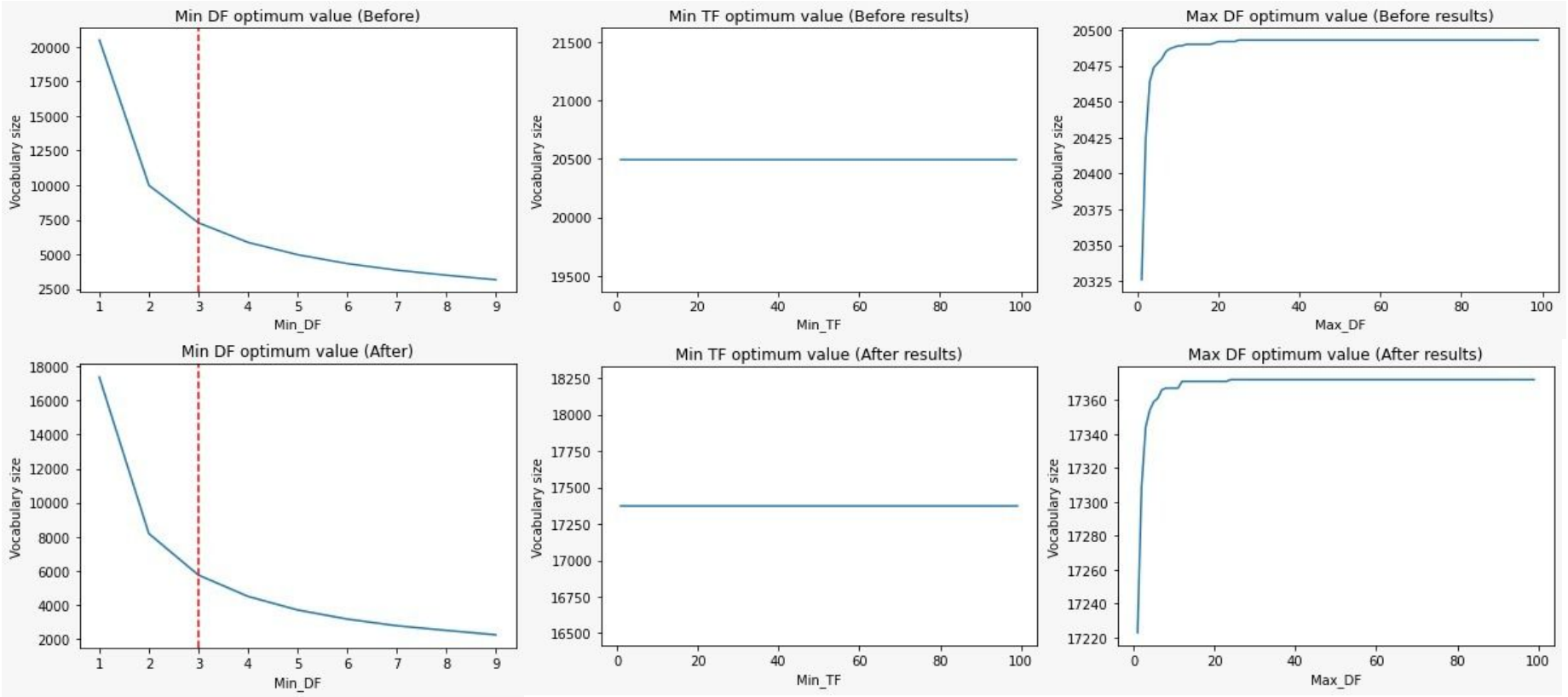
- Logistic Regression
- Random Forest

Unsupervised learning

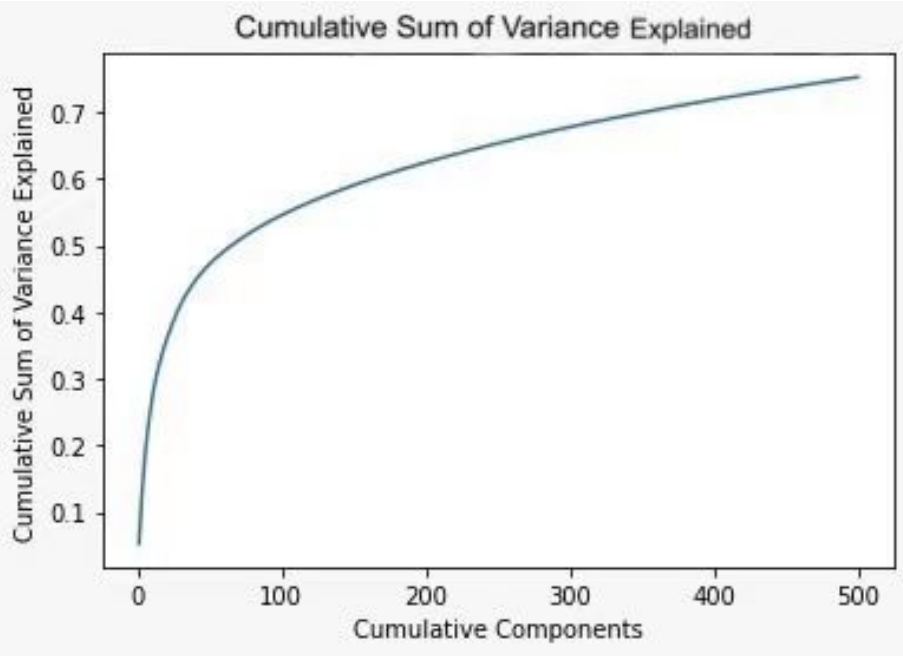
- K Means
- PCA

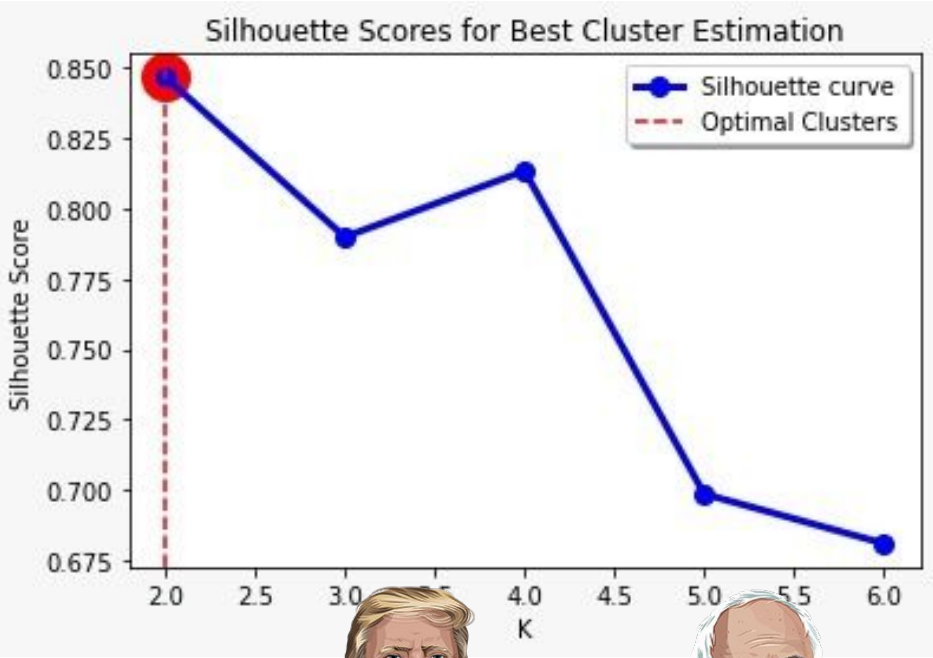


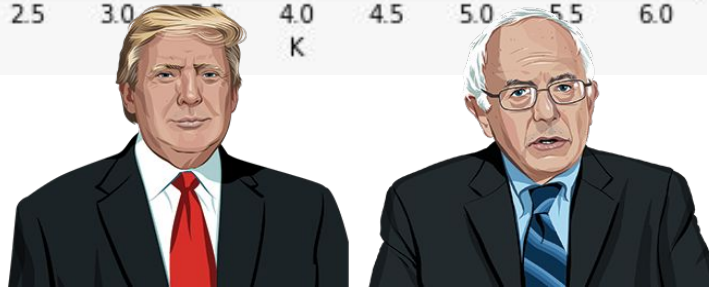
Countvectorizer Hyperparameter Tuning

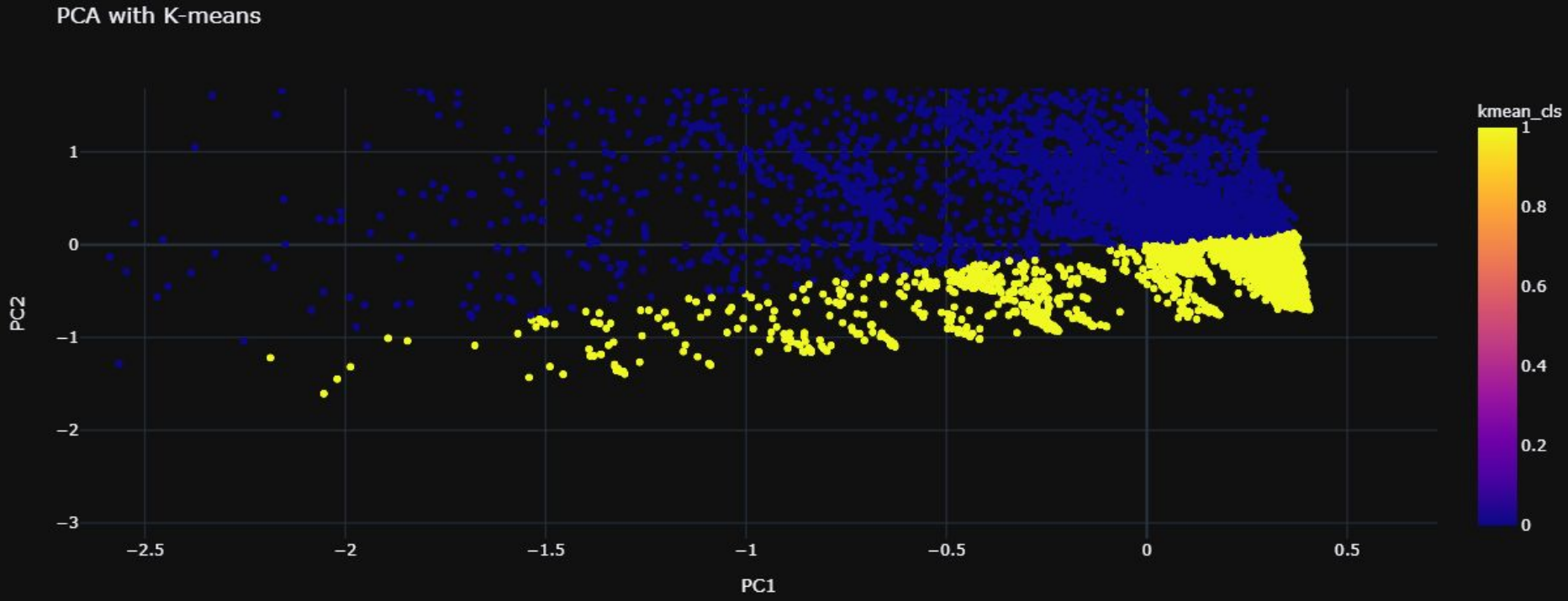


PCA & KMeans (Before)

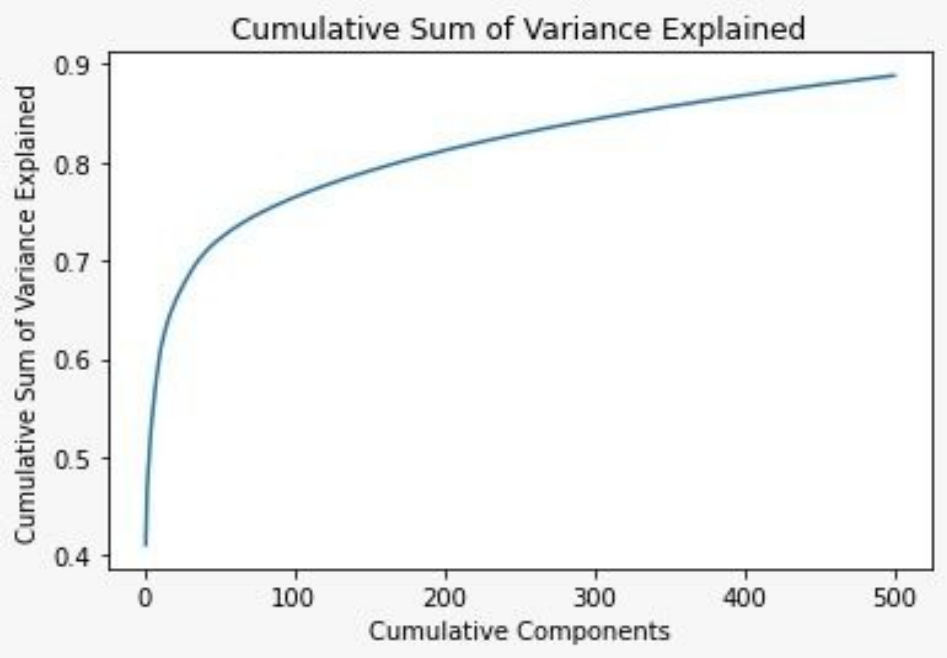


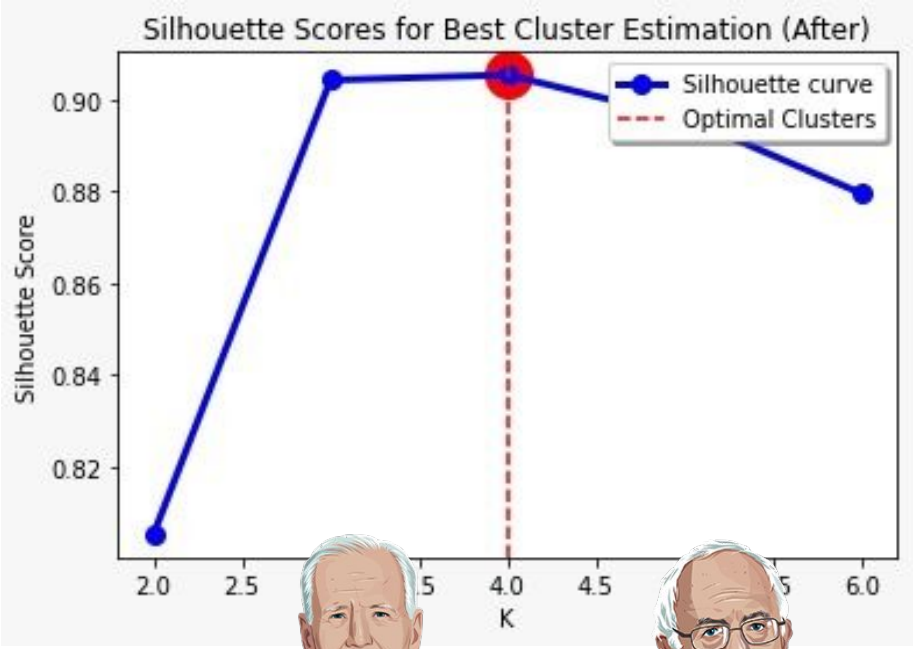


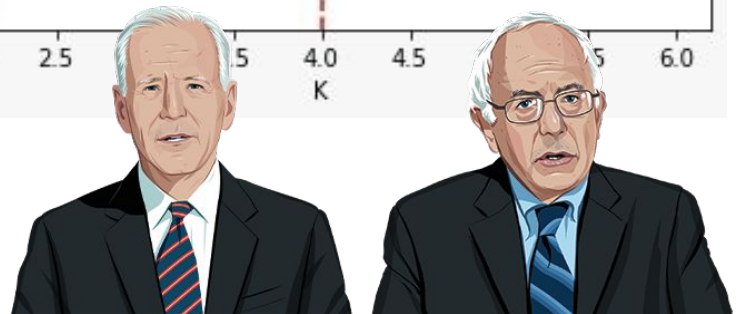


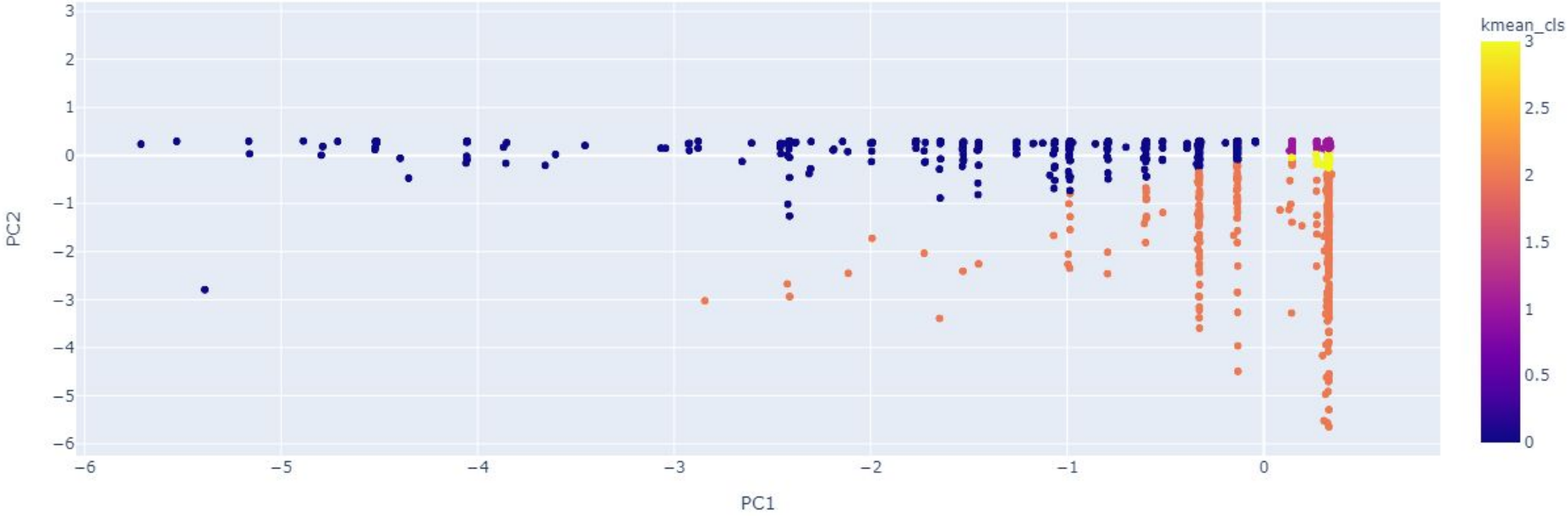

PCA with K-means
kmean_cls
1
0.8
0.6
0.4
0.2
0
PC2
1
0
−1
−2
−3
−2.5
−2
−1.5
−1
−0.5
0
0.5
PC1

PCA & KMeans (After)



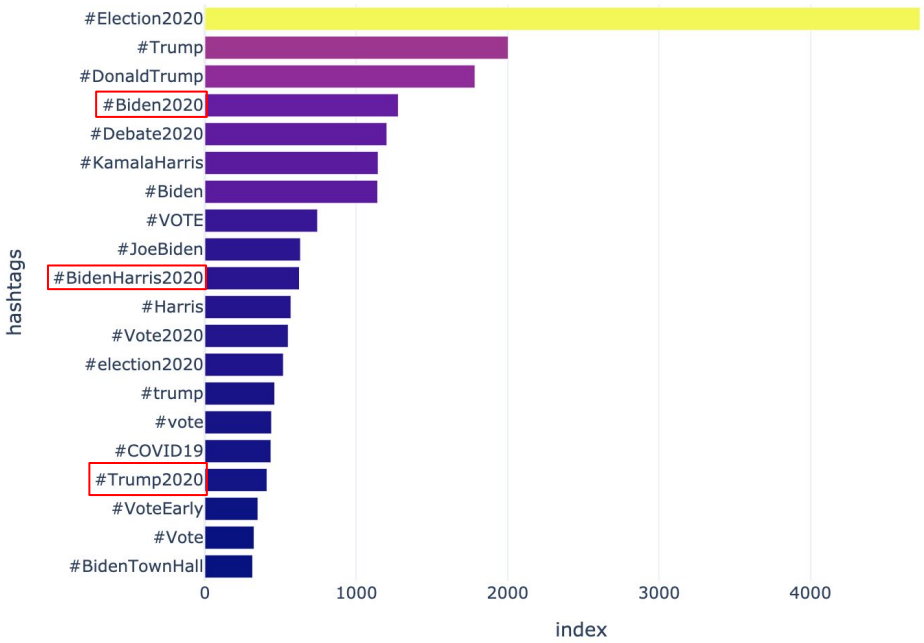




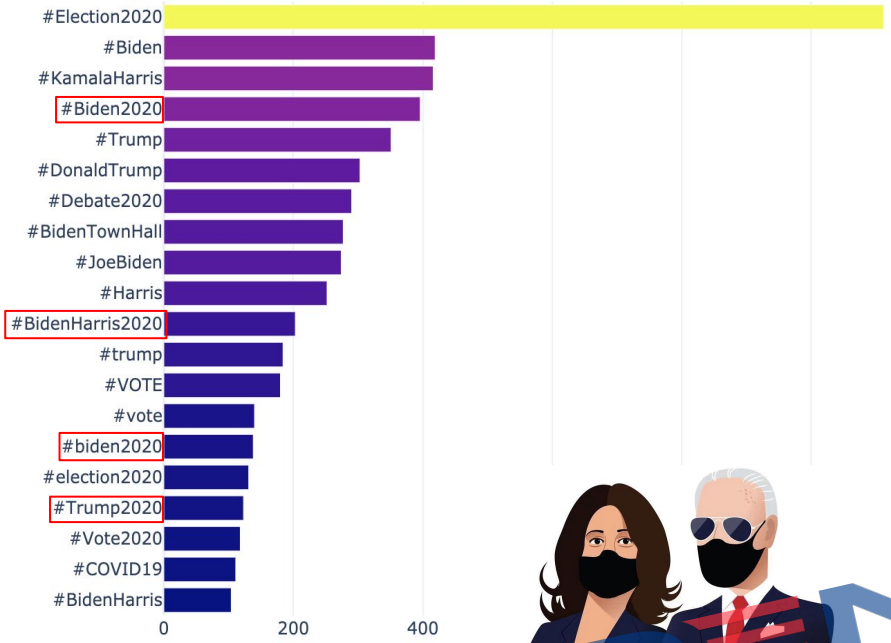

PCA with K-means (After)
PC2
PC1
kmean_cls
3
2
1
0
−1
−2
−3
−4
−5
−6
−6
−5
−4
−3
−2
−1
0
3
2.5
2
1.5
1
0.5
0

Hashtags Analysis

Hashtags and number of times it has been used (Cluster 1)

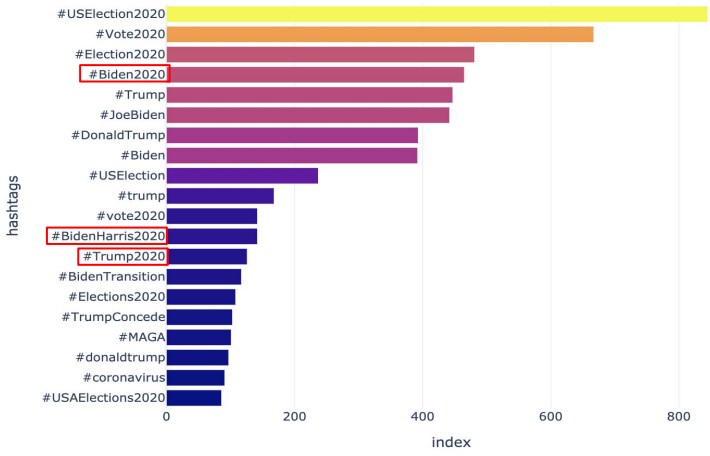


Hashtags and number of times it has been used (Cluster 2)

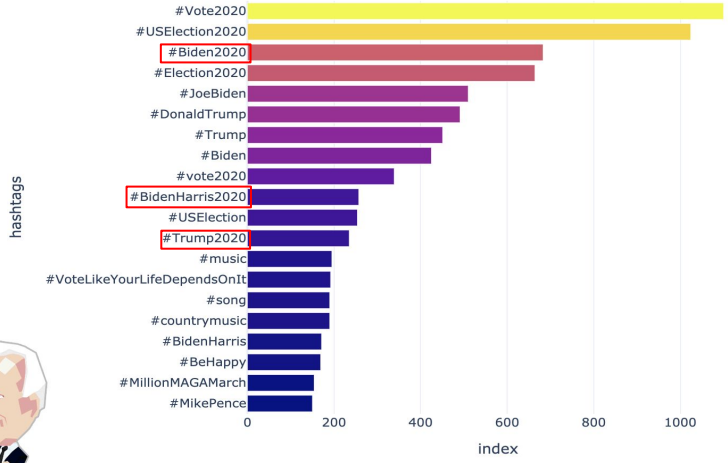


BEFORE

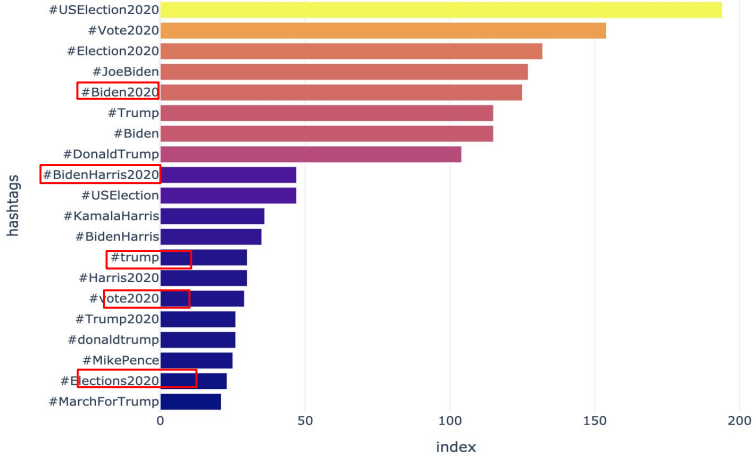
Hashtags and number of times it has been used (Cluster 1)



Hashtags and number of times it has been used (Cluster 2)

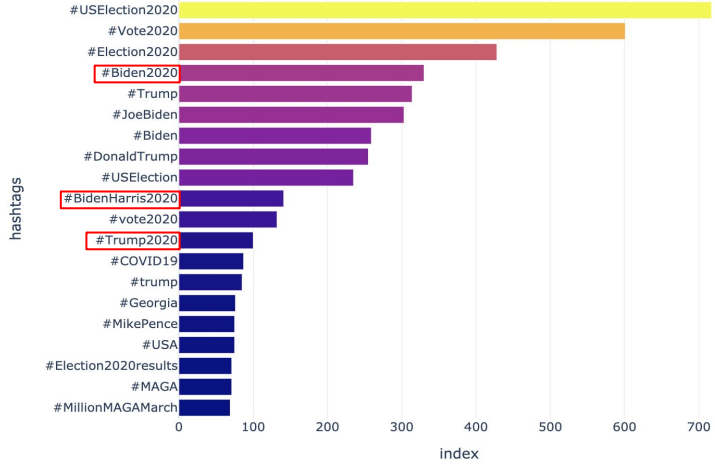


Hashtags and number of times it has been used (Cluster 3)



AFTER

Hashtags and number of times it has been used (Cluster 4)



Positive/Negative words for Logistic Regression

BEFORE Election

POSITIVE

	word	score
0	proud	0.072794
1	right	0.072888
2	nice	0.074059
3	lol	0.076070
4	good	0.078195
5	better	0.081684
6	win	0.087409
7	love	0.089324
8	best	0.096129
9	great	0.101759

NEGATIVE

	word	score
0	worst	-0.144012
1	corrupt	-0.132141
2	outrageous	-0.128522
3	boring	-0.127508
4	insane	-0.126952
5	pathetic	-0.124976
6	terrible	-0.122350
7	disgusting	-0.122205
8	disappointed	-0.121327
9	crap	-0.121073

AFTER Election

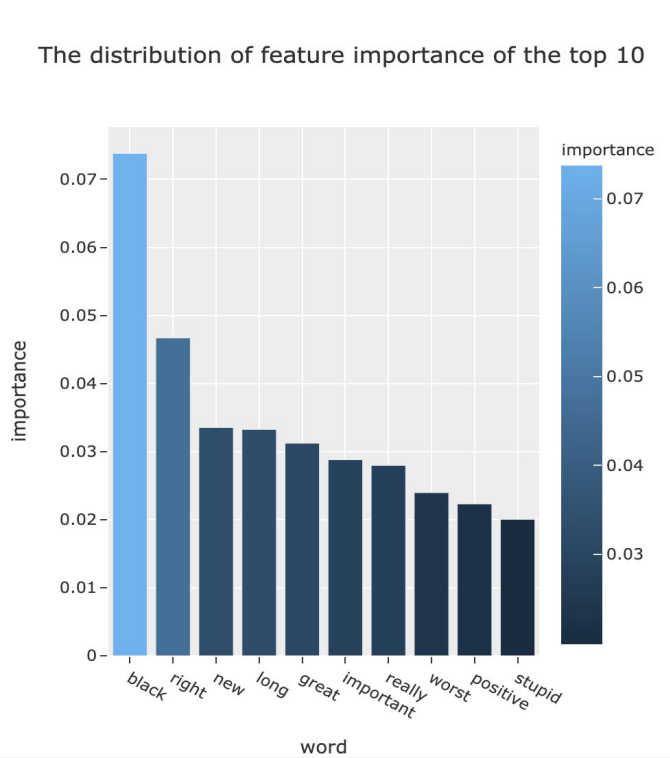
POSITIVE

	word	score
0	funding	0.062866
1	nice	0.063788
2	surely	0.063897
3	important	0.064522
4	banned	0.064879
5	happy	0.067130
6	excellent	0.067677
7	interesting	0.067723
8	elect	0.069356
9	good	0.069526

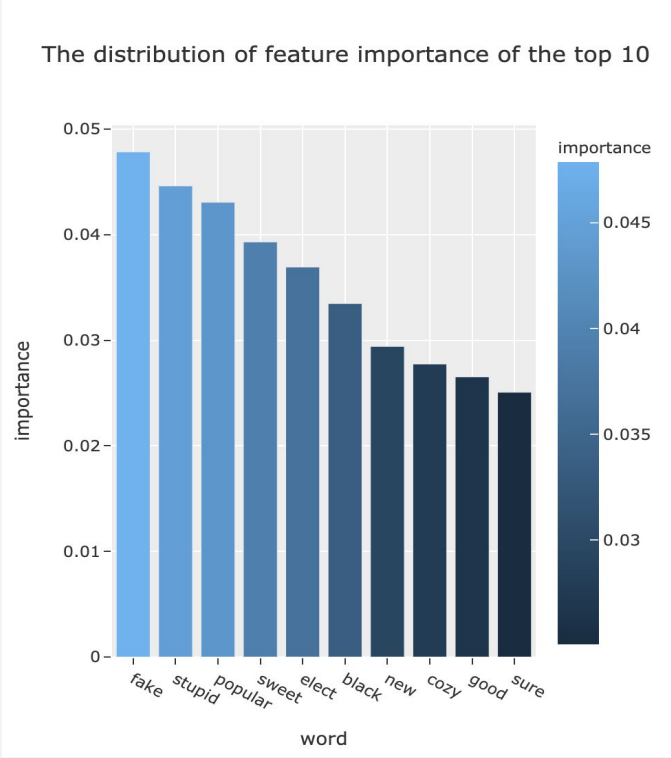
NEGATIVE

	word	score
0	moron	-0.137150
1	unlikely	-0.129946
2	disgusting	-0.127548
3	cold	-0.125458
4	plain	-0.123157
5	terrifying	-0.116187
6	crap	-0.115854
7	tragic	-0.115005
8	pathetic	-0.114434
9	bad	-0.112800

Feature Importance for Random Forest



BEFORE

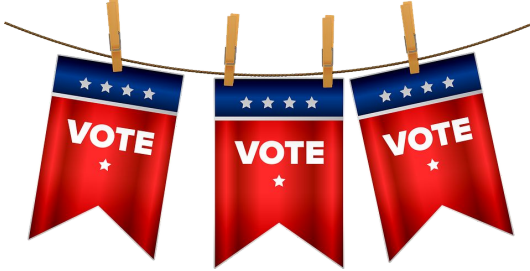


AFTER

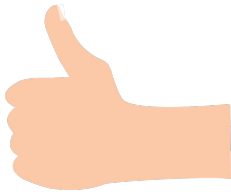
Problems Encountered

- Discerning Sarcasm from the hashtag analysis
- Label Generation using TextBlob
- Use of Emojis for sentiment analysis
- Data Collection

Summary of prediction & Inference goals



- Hashtag frequency and support vs sarcasm
- Important predictors: black, right → different contexts
- No clear polarity between the two datasets but clear transitions between optimal cluster
- PCA outliers



IST 718: Reflection

Learning Objectives: Reflection

- This presentation demonstrated the successful execution and implementation of each of the seven learning objectives required from this program

  Data of various types were collected and organized using modern day programming tools and database solutions. Statistical methods and ML/Data mining tools were utilized to the highest degree in each of the following projects:

  IST 659 Created, collected and managed data for a Hospital RDBMS

  IST 664 Clustered Covid-19 related tweets using PCA, K Means and T-SNE

  IST 718: 2020 Sentiment Analysis of Election related Twitter Data

Thank you.

References:

Eren, E. Maksim. Solovyev, Nick. Nicholas, Charles. Raff, Edward, COVID-19 Literature Clustering, 2020,April, location = University of Maryland Baltimore County (UMBC), Baltimore, MD, USA, https://github.com/MaksimEkin/COVID19-Literature-Clustering

Ganesan, K. (2019, April). All you need to know about text preprocessing for NLP and Machine Learning. Retrieved November 30, 2020, from: https://www.kdnuggets.com/2019/04/text-preprocessing-nlp-machine-learning.html

Ondocin, R. J. (n.d.). (2021) IST 659: Database Administration. Retrieved from https://github.com/Ryanondocin2019/MSADS_PortfolioMilestone/tree/main/IST659/IST659

Ondocin, R. J. (n.d.). (2021) IST 664: Natural Language Processing. Retrieved from https://github.com/Ryanondocin2019/MSADS_PortfolioMilestone/tree/main/IST664/IST664

"Sklearn.decomposition.PCA¶." Scikit, scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html.