space-hospitality

.jpg)

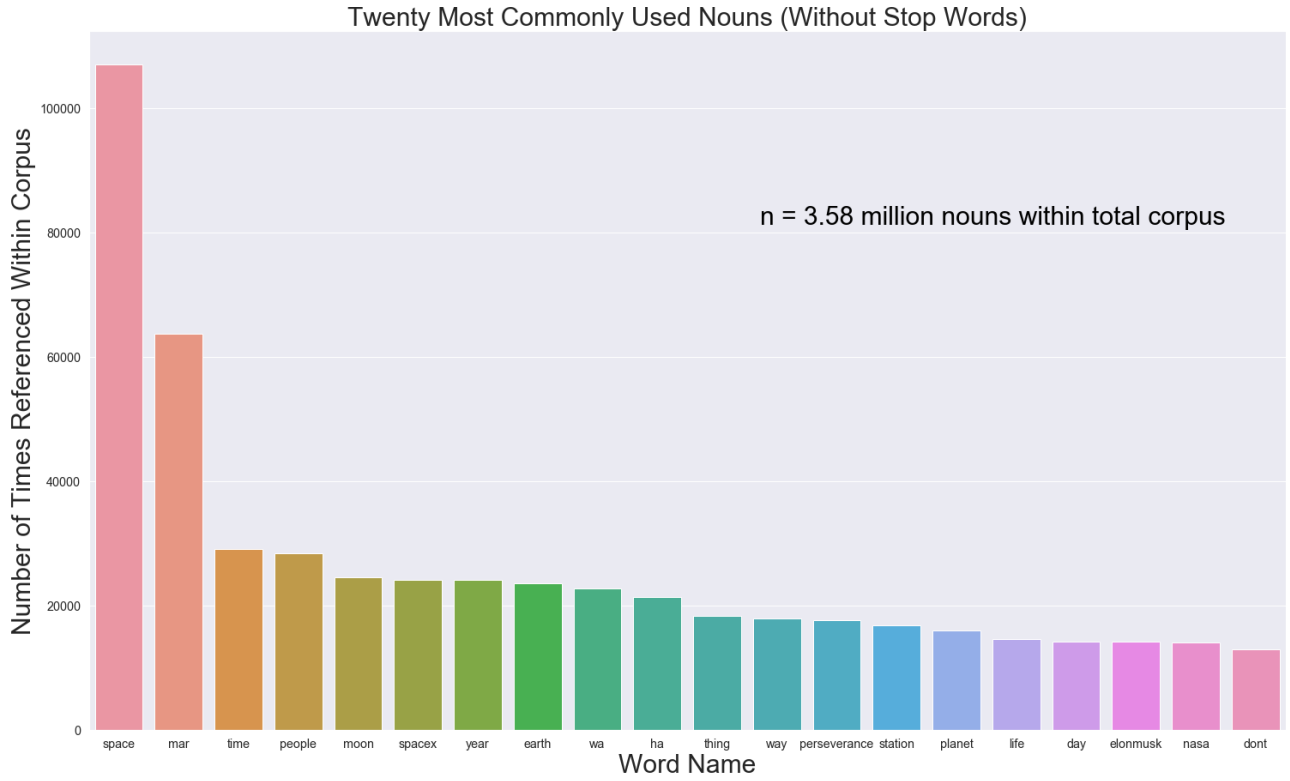
Natural Language Processing and Sentiment Analysis of the Interstellar travel industry

Authors: Fennec C. Nightingale, & Matthew Lipman

Overview

"It's a fixer upper of a planet, but we can make it work." -Elon Musk

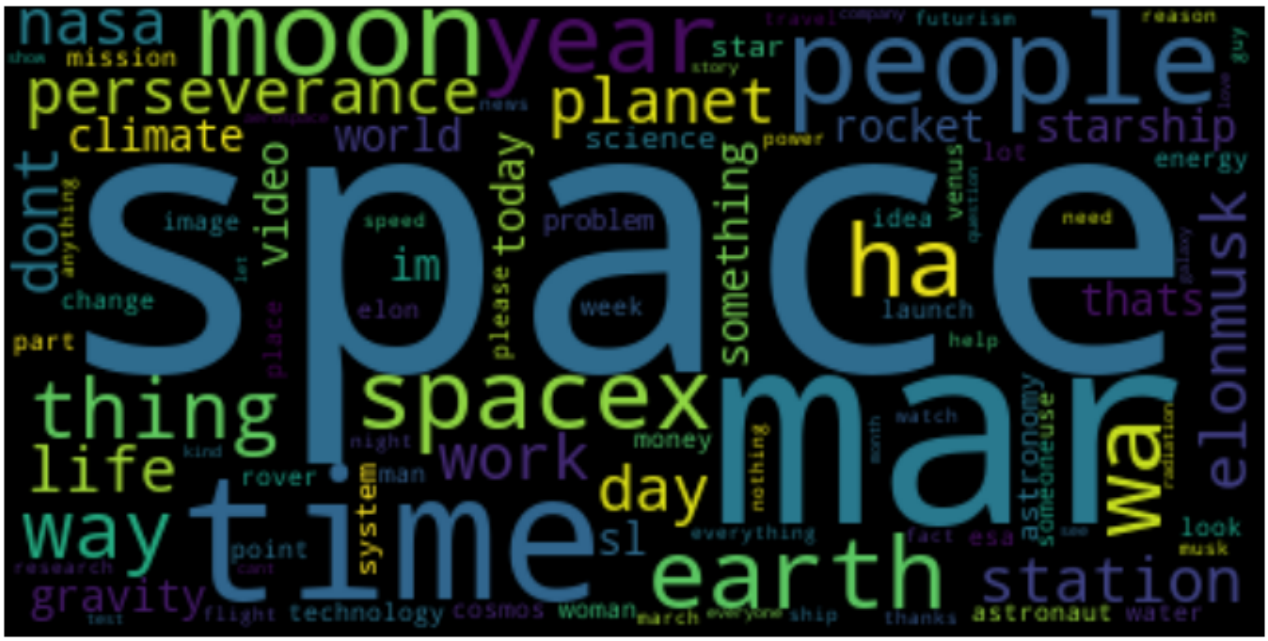
There is a huge arena of both positive and negative discussions about space. To ensure that the leading businesses in the interstellar travel industry drive revenue, they'll need to create a huge demand for and positive discussion around this currently unknown and unexplored industry. This project takes into account the discourse around the space travel industry by analyzing over two million Tweets scraped from Twitter and over two hundred thousand dive YouTube comments from over two hundred different YouTube videos to better understand the sentiment around the most popular words used to discuss the space travel industry. With millions of different statements published by Twitter and YouTube users online and a number of features and attributes about each statement, this dataset provides a strong basis to illustrate what words are used to describe space travel and classify how those words are used in the context of them having an opinion. By understanding, analyzing, and modeling the data, we are able to classify text into three different clusters with with XX.X% accuracy. By accurately classifying this text, we can best inform the industry leaders the way to most appropriately and successfully market space travel to the Twitter and YouTube communities, and beyond.



Business Problem

The interstellar travel industry is a gold mining waiting to have its gold rush. The wealthiest and most successful companies are investing in tremendous resources to create a new way to the people of the world to travel: Interstellar Travel

With this impending bounty waiting to be explored, the industry still has a tremendous gap to cover with many unanswered questions that the public will require in order to be comfortable spending their money on traveling to space and investing in the companies that dedicate their time and resource to providing this soon-to-be-real service.



The word cloud above displays the most commonly used nouns found in the corpus of 3.58 million nouns. As you can see, these words hold substantial relevance and importance. But to dig deeper, our study's guiding question gets at the significance and sentiment of these meaningful words. Through this research and development, I will be answering/addressing the following questions:

1. How do the Twitter and YouTube communities talk about space?
2. Can sentiment around space travel be divided into distinct clusters representing distinct and differing sentiment?
3. How can an understanding of the lexicon of words used to describe space travel help inform future space travel industry leaders how to most effectively market space travel?

The clustesr we created aims to classify sentiment into three distinct buckets. The model we created aims to be able to classify whether or not a statement is positive, negative, or neutrals.

Data

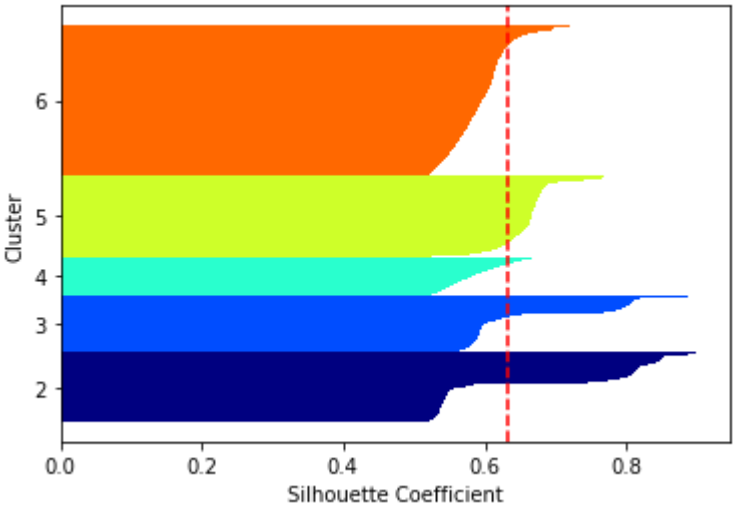
We dug into the online discourse around the subject of space travel. Specifically, our scope of analysis looks at both Twitter and YouTube as platforms by which people converse and pontificate about the industry. We began our work by applying for both Twitter and Google's Application Programming Interface to gain the ability to acquire recent Tweets and YouTube comments along with a few of their important features beyond the text itself such as the favorite count, likes, and repost count. We scraped over two million tweets about space with a focus on travel/tourism and over two hundred thousand comments from over two hundred YouTube videos through the months of March and April 2021 all pertaining to interstellar travel--whether it be NASA's recent rover landing to Mars, SpaceX's launches, or discussion about the future potential businesses and missions yet to occur on our moon, Mars, and beyond. After preprocessing and stripping down our text down to base words, we ended up with about half a million unique text pieces.

Methods

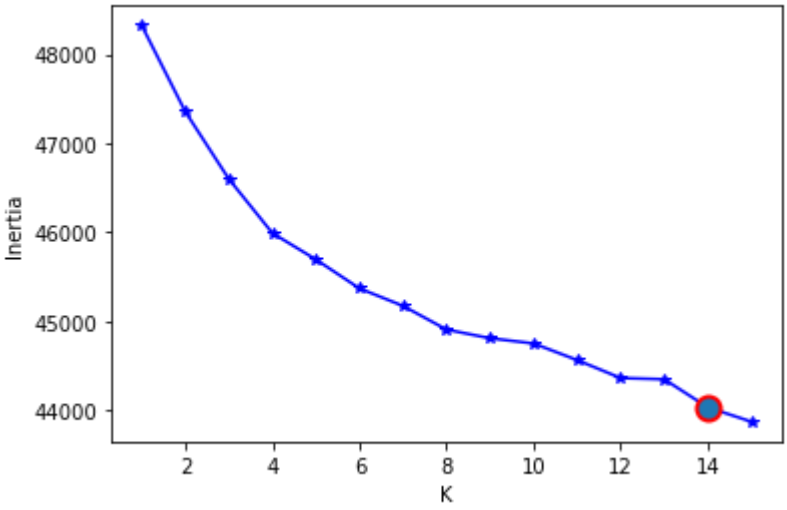
This classification modeling project is in accordance with the CRISP-DM method. After acquiring the data, natural language processing requires certain important preprocessing techniques on the text such as addressing casing, punctuation, stop word removal, lemmatization, tokenization, as well as a number of normalization techniques on the other features. By preprocessing this data, we are able to reduce randomness and dimensionality. Arguably the most significant part of this study was clustering. Since we are working with unsupervised data--meaning data that has not been predetermined as having a specific sentiment--none of our tweets and youtube comments have any predetermined classification. With that said, being able to label and map associations between words is the first major step before identifying and modeling the sentiment of said word. Each subsequent clustering technique had an iterative approach using a number of techniques to address a variety of modeling obstacles such as class imbalances, dimensionality reduction, and inertia. By successfully clustering the text, we were able to streamline the modeling process, and identify a number of evaluation metrics to best analyze our model and selected one metric to strive to minimize classification inaccuracies.

Results

Clustering



We used a series of different clustering techniques with a number of evaluation metrics to best identify the right clusters. The goal here is to find the ideal number of classifications to best describe the sentiment. The idea with finding the best clusters is a matter of striking a balance between finding distinct classification while having a similarity and association of words inside each cluster. Instead of forcing the sentiment into a binary classification of positive versus negative, we looked into leveraging a multi classification model to establish more of a gradient of sentiment. Using a number of techniques, we were able to pin down five clusters to best classify the text.



Modeling

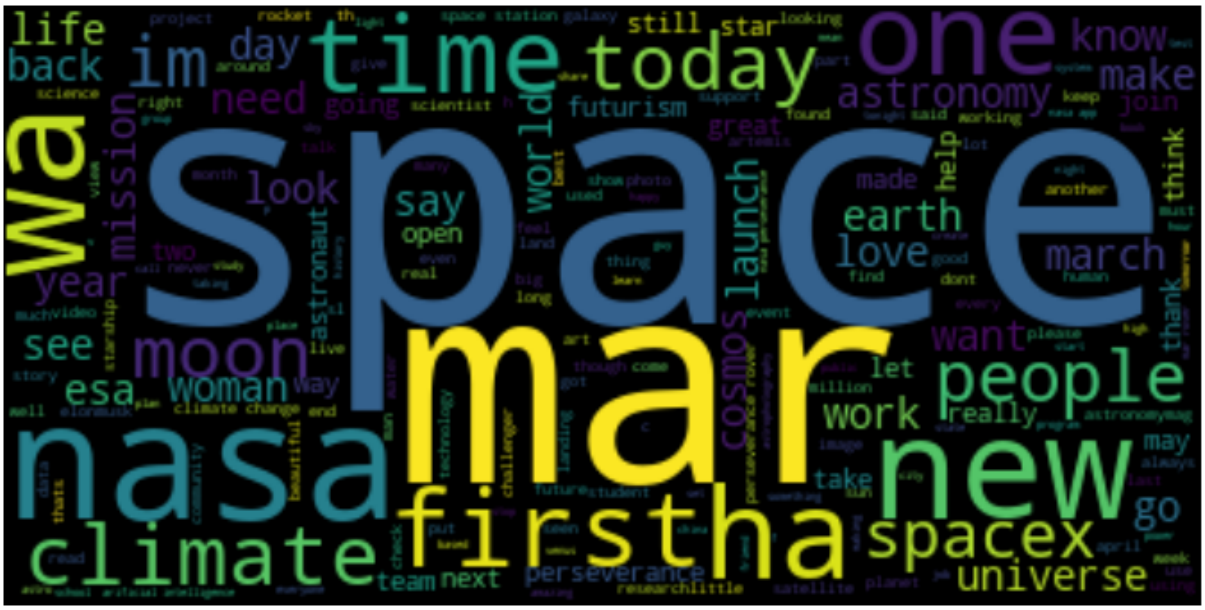
After clustering, we used the Naive Bayes classifier to be able to forecast future data. The algorithm we developed was able to accurately classify 92% of the data. While two clusters demonstrated some degree of false positives and false negatives, this model did an exceptional job at classifying the text with three of the five clusters. In those three--labeled clusters 0, 1, and 2, virtually no text was misclassified! By honing in on accuracy, the model ensures that we capture a holistic approach where a both false positive and false negative misclassifications are taken into account. By focusing on accuracy, this model avoids both types of errors.


Matrix

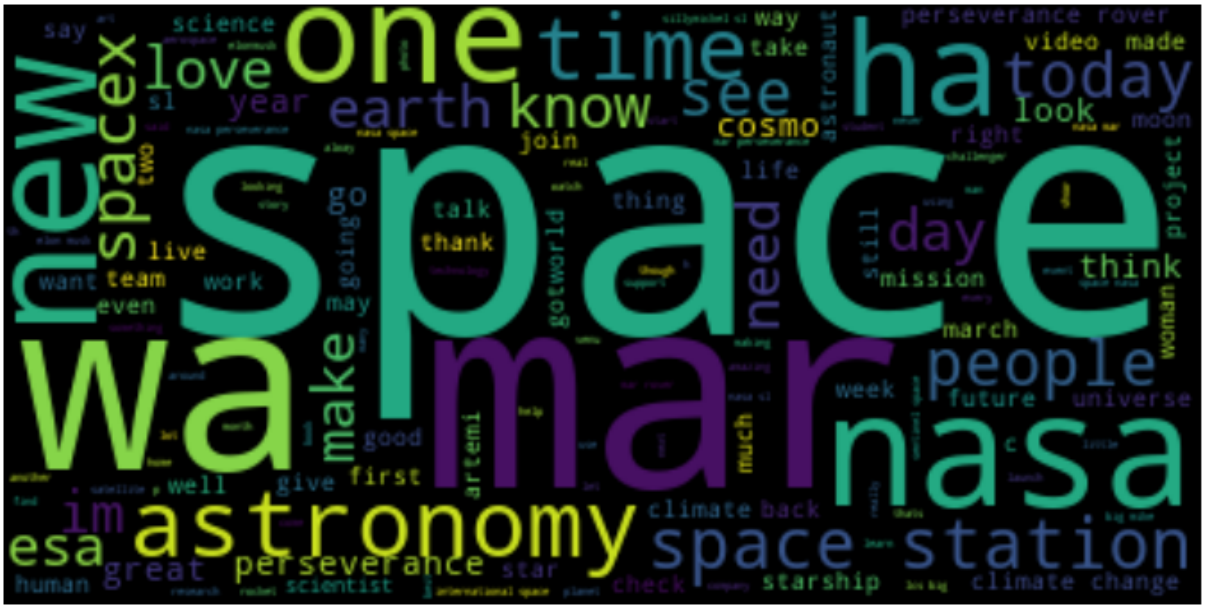
In addition, this Naive Bayes Classifier illustrates:

1. 92% Accuracy
2. Clusters 0, 1, 2 -> virtually 100% accuracy
3. Cluster 3 -> false positives/negatives present, a cluster accuracy of 76%
4. Cluster 4 -> some false positives/negatives, a cluster accuracy of 87%

Below are the word clouds of the most commonly used words within the five clusters

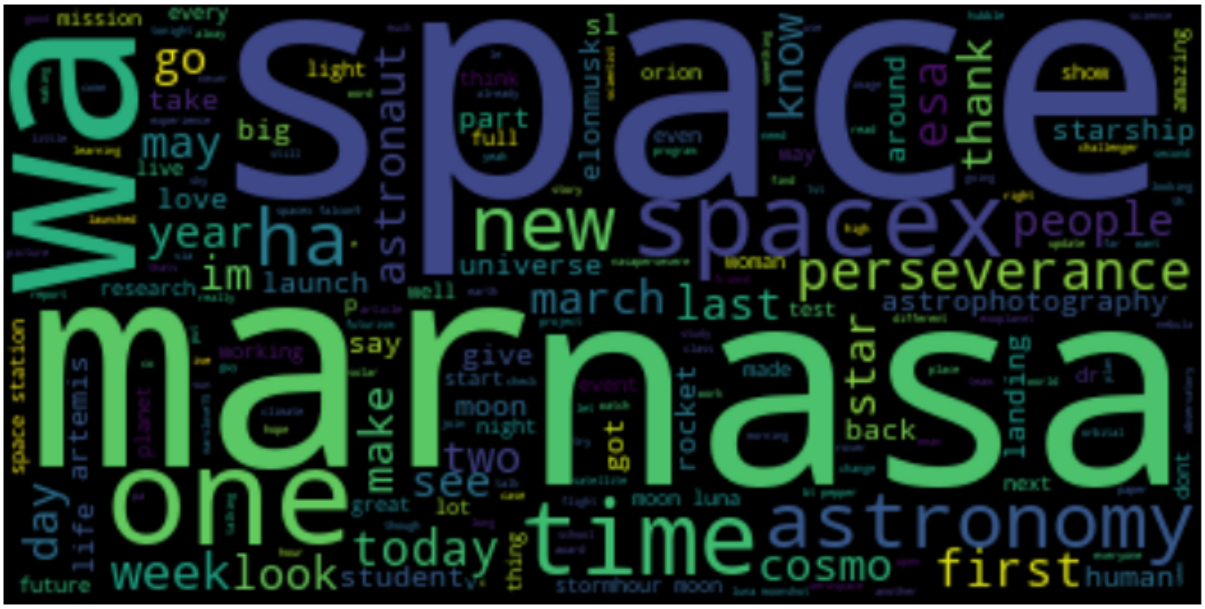


- Cluster 0

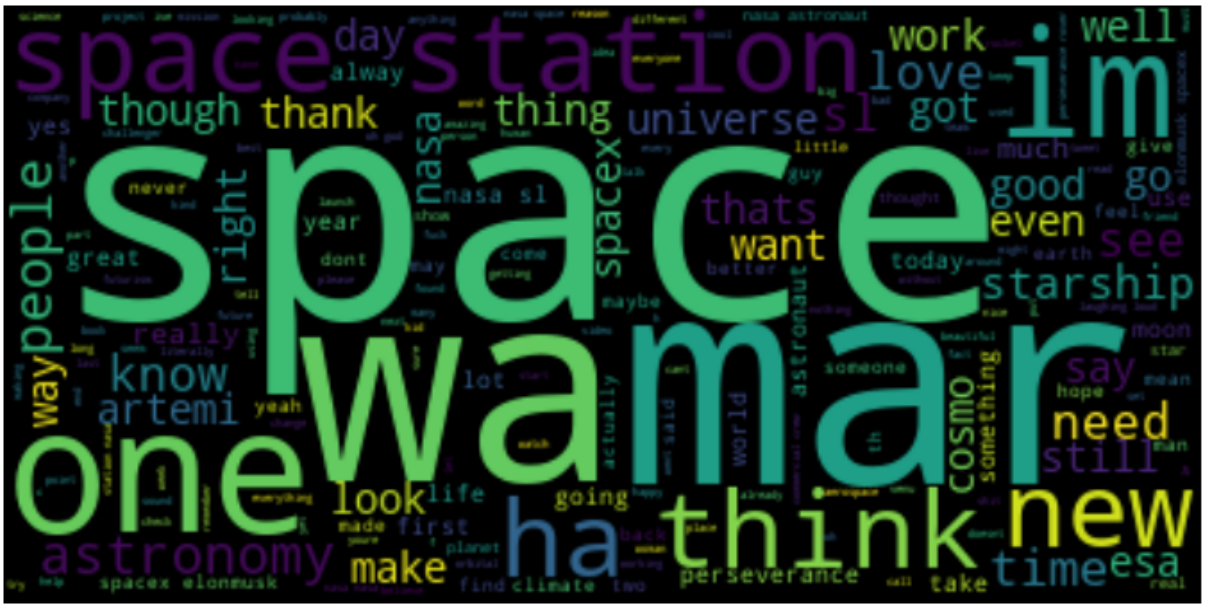


- Cluster 1

- Cluster 2



- Cluster 3



- Cluster 4

Conclusions

This analysis leads to three recommendations:

- **ONE.** Leverage the use of the word "Space." Space is commonly found within the discourse, but in congruence with other words will add significance.
- **TWO.** Partner and collaborate with current industry leaders. Institutions like NASA and SpaceX not only have scientific headway, but they have built reputations,

high esteem, and clout. Partnering with organizations of this scale will grow positive marketing schemes.

- **THREE.** Evoke emotion. The power building an emotional response adds tremendous value to your marketing capabilities. Customers think with their hearts, and by utilizing words like “love,” you are likely to create a positive sentiment around your product.

Future Work

Further analyses could yield additional insights to substantiate behavior leading to more effective marketing.

- **Incorporate more social media platforms:** We recognize that our current study involved the Twitter and YouTube communities, which may have their own biases and perspective. By bringing in more platforms such as Reddit, Facebook, and Instagram, this study will improve in comprehensiveness tremendously.
- **Time series analysis:** Looking at change of space sentiment after major events could help companies best adapt to substantial changes in requirements to the safety and understanding of how customers will travel. In addition, it will be effective to study how specific words have changed in sentiment over time to forecast how certain features may become more important in the future as well as products to invest more/less in for future development.

For More Information

See the full analysis in our Jupyter Notebook or review our Presentation.

For additional info, contact us here: Fennec C. Nightingale, & Matthew Lipman,

Repository Structure

```
data too large to push to github, files note how they're stored locally
├──.ipynb_checkpoints
├──.gitignore
├──data
    ├──3.20.youtube
    ├──3.21.youtube
    ├──3.26.youtube
    ├──3.28.youtube
    ├──3.29.youtube
    ├──4.6.youtube
    ├──3.26.twitter
```

```
    ├──matts_twts
├──Images
    ├── Appendix.png
    ├── fig_ confusionmatrix.png
    ├── fig_cloud_cluster0.png
    ├── fig_cloud_cluster1.png
    ├── fig_cloud_cluster2.png
    ├── fig_cloud_cluster3.png
    ├── fig_cloud_cluster4.png
    ├── fig_elbowmethod.png
    ├── fig_histogram.png
    ├── fig_silhouette_best.png
    ├── fig_silhouette_better.png
    ├── fig_silhouette05.png
    ├── fig_silhouette07.png
    ├── fig_silhouette09.png
    ├── fig_silhouette11.png
    ├── fig_silhouette_better.png
    ├── fig_wordplot_corpus.png
    ├── image_blastoff.png
    ├── k4.png
    ├── k5.png
    ├── k6.png
    ├── tfidf.png
    ├── wordcloud_cluster0_old.png
    ├── wordcloud_cluster1_old.png
    ├── wordcloud_cluster2_old.png
├── pdfs
    ├──Clustering Sklearn.pdf
    ├──Cleaning.pdf
    ├──Evaluating Clustering Sklearn.pdf
    ├──FAISS and Bayes Final Modeling.pdf
    ├──Github Repo.pdf
    ├──Presentation.pdf
    ├──Sentiment Analysis.pdf
    ├──Twitter Scraping.pdf
    ├──Visualizing Corpus
    ├──Word Clouds.pdf
    ├──YouTube Scraping.pdf
├──Cleaning.pdf
├── Cluster_Evaluation_Sklearn.ipynb
├── Cluster_NLP_Sentiment_Analysis.ipynb
├── Clustering_Sklearn.ipynb
├── Data_Cleaning.ipynb
├── Data_Collection-Twitter_Scraping.ipynb
├── Data_Collection-YouTube_Scraping.ipynb
├── Final_Modeling.ipynb
├── README.ipynb
├── Clustering_Sklearn.ipynb
├── Data_Cleaning.ipynb
```

```
├── Visualizing-Corpus.ipynb
└── Visualizing-Wordclouds.ipynb
```