NLP with NUFORC UFO Sightings Data

David Kuralt, Thad Hoskins, Stephanie Buchanan

June 9, 2021

1 Executive Summary

Our goal in this project is to use Web scraping, text cleaning, data visualization and natural language processing (NLP) techniques for gathering and analyzing data. Specifically, we are interested in UFO sighting data from the National UFO Reporting Center (NUFORC). The motivation for this project is the recent media coverage of UFO sightings, and the Congressional initiative to understand and explain these sightings in the interest of national security.

We begin with a detailed description of the Web scraping process, as well as the need to follow links provided in the initial scraping for non-truncated eyewitness accounts. We continue by proposing research questions we wish to answer using the data we collect, such as identifying the most commonly used phrases in eyewitness accounts. We also wish to ascertain how the people providing these accounts feel about what they are describing. Are the accounts objective or subjective? In the Literature Review section of our work, we elaborate on these questions further.

We provide a detailed discussion of the Python packages used in our analysis. Of particular interest to us are those packages that allow us to scrape the NUFORC Web site, to pre-process, or **clean** the data we pull from the Web site, as well as packages that allow us to create visualizations of our data, ascertain the **sentiment** and **objectivity** of the eyewitness accounts, and to determine the most commonly used phrases of the accounts.

We discuss at length the process of cleaning the different attributes of each observation, and our rationale for choosing particular techniques. Then we move on to discuss our data visualizations and how they help us to answer our research questions. Finally, we present our results and conclusions.

2 Introduction

On December 12, 2020, Congress passed a $2.3 trillion Omnibus Appropriations and Coronavirus Relief Package that included the requirement for the Pentagon and other agencies to provide a report of all known information about UAPs (unidentified aerial phenomena) previously known as UFOs (unidentified flying objects). As the 180-day deadline, June 19, 2021, approaches, the amount of sightings and media reports on UFOs have sharply increased. A large volume of these are attributed to pranks and misidentified Starlink satellites or rare aircraft.

This project is aimed at applying Natural Language Processing (NLP) techniques to summary details from UFO sightings reported to the National UFO Reporting Center (NUFORC).

3 Data

The data was scraped from NUFORC's website. The website was setup as a way for people to self-report UFO sightings around the world. There have been a large volume of sightings over the past year. The site attributes a lot of these sightings to the Starlink satellites, and cautions researchers to be discriminating with the information being presented in the reports.

3.1 Scraping the Data

The data is readily available on Kaggle in multiple forms, as well as other sites. Many of these sources have sufficiently cleaned the data for use as Time Series or Location based approaches.

For our analysis, we wanted to use Natural Language Processing, or NLP of the data sets we evaluated. The text of the sighting was truncated, as was the "list view" in the website. To get the full summary of the sighting, one needs to use the detailed summary page. Since this was not available, it was decided to go to the source.

A quick look at the list of sightings by date (NUFORC Index by Month) confirms a large number of records. In this case, they are broken down by month and year. The vast majority of the sightings are from the last 20 years, but that affects the scrape little since we want them all.

The heirarchy of the site is:
month/year $\longrightarrow$ sightings $\longrightarrow$ sighting details
The program would read in the original list of all the months, then navigate to the list of sightings for that month. For each sighting, the program would navigate to the summary page to scrape the full summary.

From the outset, there were some concerns to be addressed and others would emerge.

- Data consistency
- List consistency
- Site limits

3.1.1 Data consistency

Some sites are custom created, which is to say the pages are created manually. In this case, data structures may differ in subtle ways that would mean the program may scrape the wrong data or would error if it cannot find what it is looking for (what is was told to look for).

Fortunately, the data structures and tags were consistent.

3.1.2 List consistency

Similarly, I was concerned about the lists and their links being the same across the entire website. Had they changed 4 years ago (for example) to something different?

Again, the lists and the links were consistent.

3.1.3 Site limits

Knowing there many nearly 100,000 records, this would take time and would be 200,000 (assuming 100,000 records) calls to the website. Would the program be blocked?

There were times in which the connection was refused or timedout. It cannot be said with certainty that the application was blocked but that did throw challenges into the program as it scanned.

3.1.4 Unknown Challenges

- Time
- Errors
- In Process Changes

All three challenges boiled down to one major need for solving the problem of what to do if the program stops? On record #10, this was not a problem. On record 20,000 it is more than a minor inconvenience.

The application would make 200,000 website calls. Time was an issue. A simple version of that was merely keeping the computer going while it scraped. This was a minor issue of a power setting.

Likewise, any errors during the scrape could be catastrophic. One could handle most errors but it is the unforeseen that happen with 99% of the data processed.

Lastly, there were in Process Changes that needed to happen, mainly in the first half of processing.

All of those issues required solutions to the "restart" issue. Stopping it easy. Restarting is not.

The solution iteratively changed and finally settled.

- Save the data to a file
- Pickup where it left off

Save the data to a file Saving the data to a file was simple enough, but there had to be a balance. Many file I/O calls would slow the program. Too few saves would mean unacceptable amounts of data would be lost and need to be reprocessed.

In the end, saving the file after each month seemed to balance that. At most, the program would need to reprocess a month of data. As was the nature of the data, this problem grew smaller as more months were processed since the sightings are heavily weighted to the present.

Pickup where it left off Because the program would save the data to a file, it was then necessary to figure out where the program stopped.

The solution was to load the data from the file into a Dataframe, then count the rows. The application would then begin processing each month. The number of records for that month was known for the search. That number was substracted from the total. If the remaining number was greater than zero, the program went to the next month. Continue that until the number was less than zero.

At that point, the program would begin processing the month again. Because the program only saved to the file once the month was complete, any records processed would not be saved if the program was stopped manually or on error.

3.1.5 Missing Detailed Summary

During web-scraping, some pages failed to get detailed summaries (133 in total). The final step of the web-scraping program is a check for those pages to see why they failed to get a detailed summary.

That check resulted in all 133 returning a 404 page not found error. A random sampling (manually) was used to manual check the pages, confirming the detail page did not exist.

In the data cleanup, the truncated sighting summary was used to fill in the missing data.

4 Research Question

The goal of this study is to answer the following questions using NLP:

- What are the most common n-grams in the summaries for UFO sightings?
- Does the sentiment of the detailed summary change with respect to location in the United States?
- Does the objectivity of the summary change with respect to location in the United States?

5 Methods

NLP is described as a branch of artificial intelligence with roots in computational linguistics[1]. These techniques help computers understand, analyze and manipulate human languages. For example, NLP makes it possible for a computer to read text, interpret it, determine the most common phrases, measure sentiment, and measure objectivity of the text.

The main platform used in the analysis of the UFO sightings dataset was the Natural Language

[1] source: https://www.sas.com/en_us/insights/analytics/what-is-natural-language-processing-nlp.html

Toolkit[2] This toolkit provides a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning.

An 'n-gram' is defined as a sequence of n words in some text. The n-grams that interest us are common word groupings in the UFO sighting summaries.

Sentiment analysis can help determine the ratio of positive to negative engagements in a group of words, in this case, detailed summaries of UFO sightings. Algorithms are used to classify text into positive and negative categories. For this project, NLTK's pre-trained sentiment analyzer called VADER (Valence Aware Dictionary and sEntiment Reasoner) was used.

VADER is best suited for short sentences with some slang and abbreviations like the text commonly found in social media. For this reason, the summaries were split into sentences and analyzed separately. The output from VADER is a dictionary of different scores: negative, neutral, positive, and compound scores. The negative, positive and neutral scores all add up to 1 and can't be negative. The compound score is similar to an average, but calculated differently and can range between -1 and 1.

The TextBlob library also provides a function to determine the sentiment of a sentence along with the objectivity of a sentence. The output of the TexBlob sentiment function is a tuple of two values: polarity and subjectivity. Like the compound score from NLTK's sentiment analyzer, polarity is a float value within the range [-1.0 to 1.0] where 0 indicates neutral, +1 indicates a very positive sentiment and -1 represents a very negative sentiment.

Subjectivity is a float value within the range [0.0 to 1.0] where 0.0 is very objective and 1.0 is very subjective. A subjective sentence will express personal feelings, views, beliefs, opinions, allegations, desires, beliefs, suspicions, and speculations. In contrast, an objective sentence is factual.[3]

6 Literature Review

There are several groups that have historically analyzed the NUFORC data. One of them is the Ruza, Poopalasingam, Weber group that asked the questions: Are UFO sightings explainable? and Is there any relation between shape and emotion?. [4]

To answer the first question, the group plotted the number sightings per day from January 2004 to December 2007. They then correlated the spikes to various events such as the missile launches, meteor events, ISS transits and rocket fuel dumps. The conclusion from this analysis was while there were some peaks that did not correlate to any known space or weather event, most of the sightings did correlate to a known event.

The second question was answered by looking at the percentages of the detailed summaries that contained words correlating to known emotions such as risk, anger, sadness, and reward. They then plotted the nets of shape with the percentages of all the emotions used for the shape

[2]Bird, Steven, Edward Loper and Ewan Klein (2009), Natural Language Processing with Python. O'Reilly Media Inc.
[3]https://medium.com/@rahulvaish/textblob-and-sentiment-analysis-python-a687e9fabe96
[4]https://ada-nuforc-analysis.github.io/

groups. The conclusion was they could not infer any meaningful differences between the popular shapes.

7 Data Cleaning

The columns in the original dataset before cleaning were "Date_Time", "City", "State", "Shape", "Duration", "Summary", "Posted", "Detail_Link", and "Detail_Summary". The "Detail_Summary" column was populated from the "Detail_Link" if available, and if not, was populated from the "Summary" column.

7.1 Date_Time

We encountered an issue when converting the "YY" format of the year to a "YYYY" format. All years in the 1900's up to 1967 were correctly formatted with "19" as the first two digits. But from 1968 on, the years were formatted with "20" as the first two digits. We applied a function to this column to subtract 100 from any year greater than 2021. Finally, all times were normalized to UTC.

7.2 Location

The original dataset contained City and State fields. The data was "mess" to say the least. The end goal was to have Latitude and Longitude, not just a City, State, Country combination to narrow to the United States. This would be useful for maps later.

7.2.1 City, State, Country

First, after exploring the data, many city fields contain information was additional to the city name. Some contained parenthetical descriptions that included Country, comments, hotel names, etc. Likewise for the state which contain state, regional, country, oceans, etc.

Next, an empty Country field was added.

Then, the parenthetical were segmented. Using the parenthetical, in many cases, state and country information could be added. Many of these were manual, while some were categorical and could update many records at once.

That data was merged back into the main dataset. That data was then corrected where city and state combinations, matched against known databases, could be assigned to the United States. Other knowns, such as Canada were also assigned.

Knowing we would limit our scope to the United States, anything that remained would be discarded. It is likely there are US sightings in that data, but with 65,000 remaining records the cost to benefit ratio pointed to letting them go.

Finally, using ARCGIS, the city, state, country combinations returned Latitude and Longitude.

Initially, this routine ran for 14 hours since it was getting data from an API. Having to run that more than once, the routine needed to be optimized.

The solution which yielded a significant performance increase was to group by city, state, country combinations. Search the database yielding 18,000 unique combinations. A single call then updated multiple rows. This is a classic example of "brute force" to start, then changing when time needs to be cut. This change brought the time below 4 hours.

Initially, this was in an apply() statement. However, the need to stop and check progress made a for loop advantageous. This came up in the spell check routine, as well. Mild performance loss saved time when the routine stopped, wither from error or manually. It could then pick up where it left off. Some time lost to potential inefficiency was gained in not losing hours of work before it was stopped.

7.2.2 Summary Text

Because our primary application is NLP, cleaning the text is important. We followed several guides for this, which included removing stop words, characters, etc.

We relied heavily on the NLTK library for the cleanup.

The order of the operations matters for which to put in and when. The order could be optimized further, primarily to optimize the Spell Check, i.e., put as few words through the check as possible.

The data was first made lower case to standardize on lower. Next, the text was tokenized to separate individual words for processing.

Directly after tokenizing, the string was crammed back together but with Stop words removed. At this point, a WordCloud was a created. Comparing the WordCloud before and after removing stop words gives instant justification for their removal. Important words in phrasing rise to the top.

Late in the project, a spell check was added. The vast majority of the data remains unaffected, but there are misspellings. Unfortunately, the spell check routine ran for 16 hours before stopping, failing to yield benefit. The routine was optimized slightly, making sure Stop Words were removed to eliminate some words. As was stated in the Location cleanup, a loop was utilized allowing for stopping and starting to track progress.

To gain an understanding of the scope of the Spell Check, allowing for US only, there are more than 67,000 records. Some with as few as 5 words but some as many about 30 or more. Each word is checked for spelling.

Then we used regular expressions to newline characters, punctuation and special symbols. We replaced hyphens with white spaces, then stripped any white spaces from the beginnings and ends of each string. Then we looked for all instances of multiple consecutive white spaces in the summaries and replaced these with a single white space.

Next was to expand contractions. Again, an NLTK library was used.

Again, stop words were removed. With the spell check and contraction removal, we may have more unnecessary words.

Lastly, we Lemmatized the words, bringing the differences of words down as much as possible, e.g., changing run, ran, running to run. This would help in our later WordCloud and further NLP analysis.

The new Summary "phrases" were saved to an added column in order to preserve the original for other uses, if necessary.

7.3 Shape

Shape is an attribute we read directly from the NUFORC data. We standardized similar names, such as "triangular" to "triangle," "rectangular" to "rectangle," etc.

8 Visualizations

The data was explored before analyzing by NLP for any trends seen in the time of day, seasonality or year. As seen in Figure 1, the number of reported UFO sightings in NUFORC's databases drastically increase in 2006. NUFORC was started in September 1974 with a 24-hour hotline for people to self-report sightings. In 2005, NUFORC expanded its operation to include a fax machine and the online site. The addition of the option for online reporting most likely contributed to the rise in reported sightings in 2006.

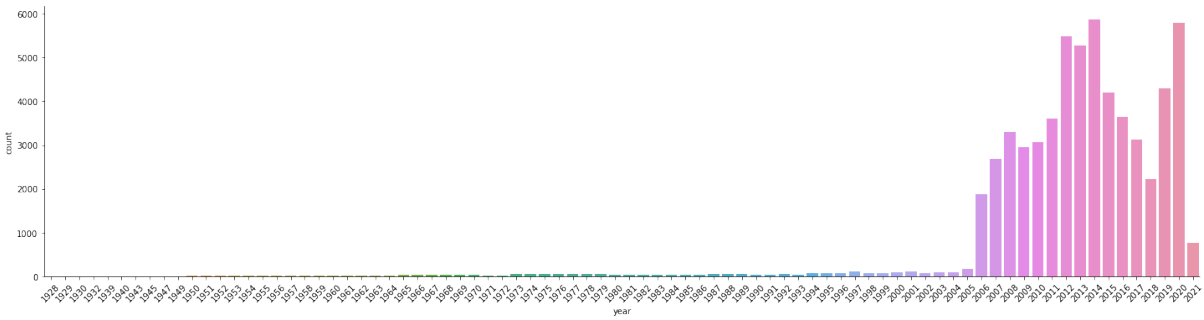

Figure 1: Count of UFO sightings per year from January 1928 up to March 2021

As seen in Figure 2, there are days of the year that are more likely for UFO sighting reports: day 1, New Year's Day, and day 185: the 4th of July.

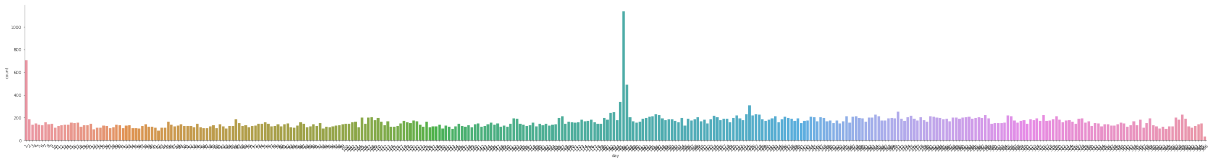

Figure 2: Count of UFO sightings per day of the year.

Figure 3 confirms the most likely time to witness a UFO and then report is from 11p to around 6a. This is most likely due to the darkness enabling observations.

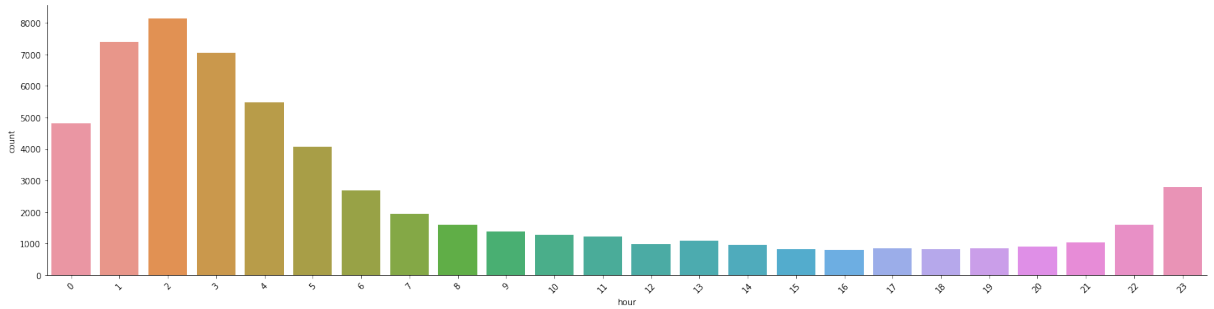

Figure 3: Count of UFO sightings per hour of the day.

Figures 4 and 5 show the median values for sentiment and subjectivity, respectively, per year. As seen in the figures, the majority of the years have a median value of 0.1 for sentiment, and 0.3 for subjectivity. This indicates that the majority of the reports have a slightly positive tone, and tend to be more objective. More noise was observed in the early years, and this increase in noise is attributed to the lower number of reports in the years prior to 2006.

The sentiment values tended to be more positive in the years prior to 2006, most likely due to the fact the report was made some time after the sighting. The subjectivity values were higher in the years 1928-1929. This is probably due to the length of time between the actual sighting and making the report.

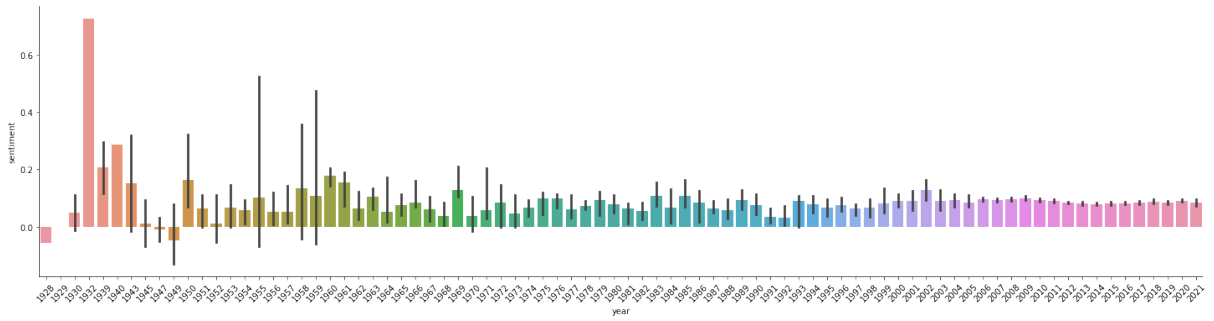

Figure 4: Median sentiment value per year

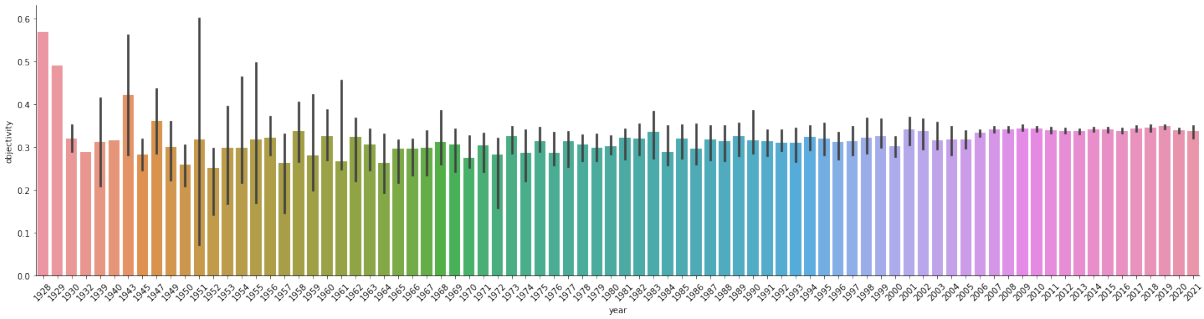

Figure 5: Median objectivity value per year

Figure 6 shows a WordCloud image from the UFO shapes attribute in the dataset. The words used most frequently to describe the shape are 'Light', 'Circle', 'Triangle' 'Sphere', and 'Fireball'. 'Light' was the most frequently used word, and could indicate that for a large number of sightings in this dataset, no craft was observed, only an unidentified light.

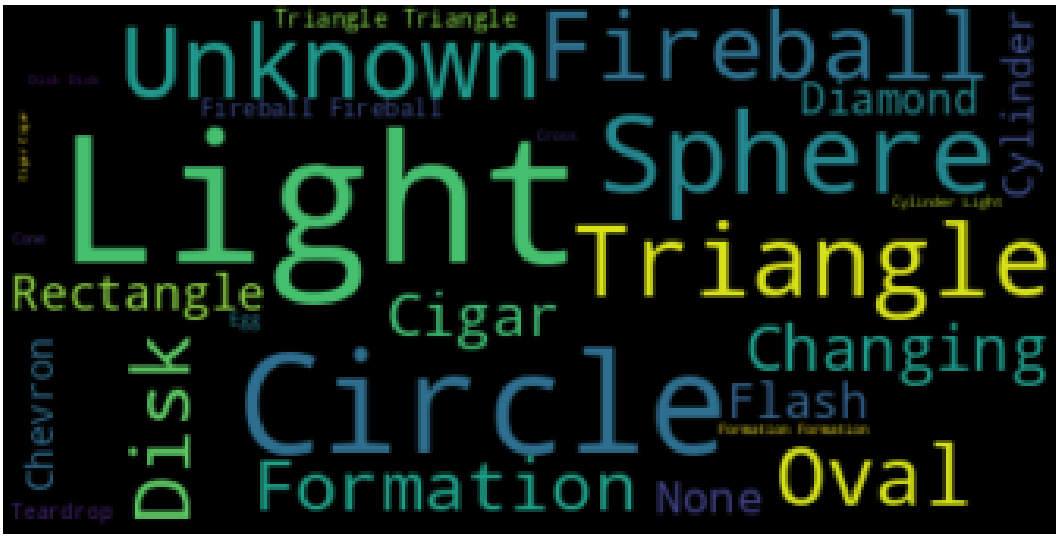

Figure 6: Word cloud focusing on shapes

Figure 7 shows a WordCloud image from the cleaned text, Detail_Summary_nltk, attribute in the dataset. The words most frequently used in the summaries are 'white', 'light', 'bright', 'looked', and 'sky'. 'Red' and 'orange also stood out as common words in the summaries which indicate white, red and orange are common colored lights that are unidentifiable in the night sky.

Figure 7: Word cloud from the summaries

Choropleth maps were used to help visualize trends by state seen in the number of sightings, median sentiment and median objectivity values by year. At first, the un-normalized values were plotted, but produced maps of a single color. Normalizing the values by 2020 state population helped identify real trends without the effect of population density since states with high populations are expected to have higher numbers of reported sightings. The assumption was made that state populations across the US have not significantly changed over the years analyzed in proportion to each other, and is a good surrogate to use. The researchers acknowledge that using the state population according to the year of the sighting would be a better approach in future analyses.

Figure 8 shows the number of UFO sightings per capita by state. Vermont stands out as having the highest number of sightings normalized to state population. Other states with high number of sightings normalized to state population include New Hampshire, Maine, Washington, Oregon, Idaho, Montana, and Alaska. The Pacific Northwest and the Northeast stand out as regions with high number of sightings. This could be due to lower levels of light pollution in these areas.

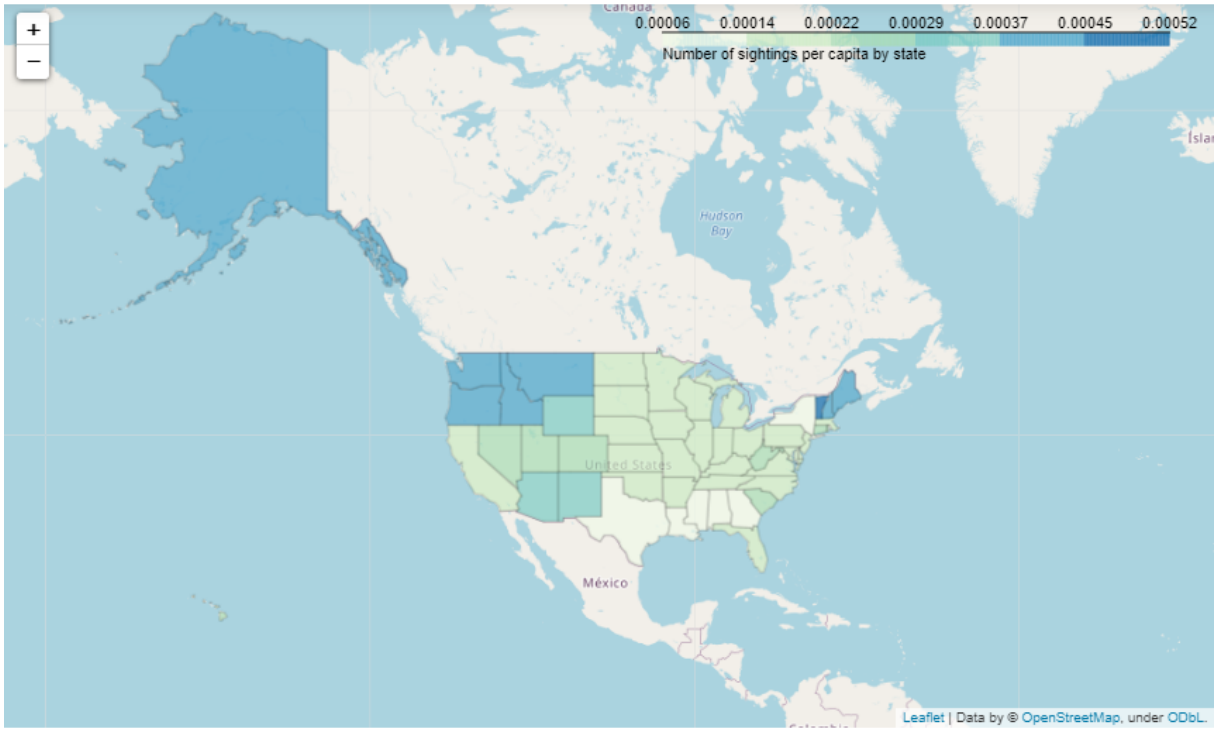

Figure 8: UFO sightings per capita by state

Figure 9 shows a choropleth map of the median sentiment value across the states normalized to the state population. The median was selected as the measure of central tendency because of the lower sensitivity to outliers. When the mean was used, the sentiment values were +/- 0.01 across all the states and did not produce informative maps.

The normalized sentiment values across all states ranged from 0.00 to 0.11, and most states showed very neutral summaries. The states that had a positive sentiment tone to the summary were California and Texas and Florida.

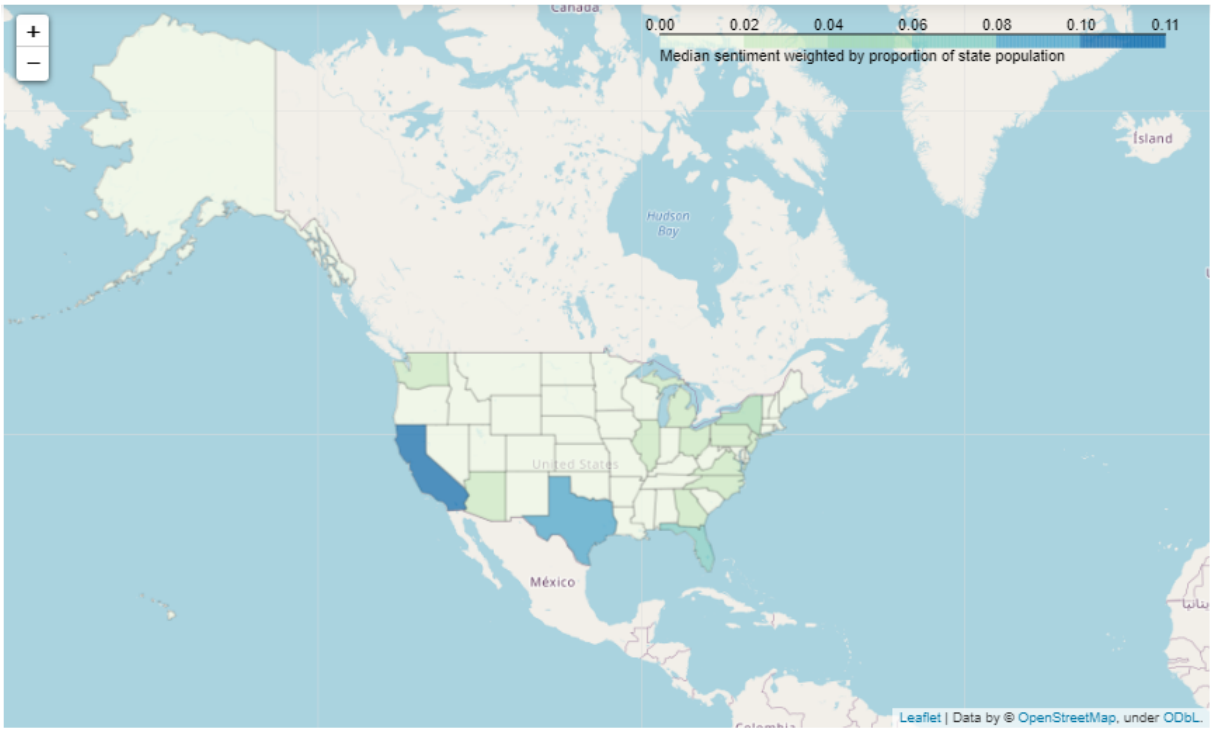

Figure 9: Median sentiment weighted by proportion of state population

Figure 10 shows a choropleth map of the median objectivity value across the states normalized to the state population. The median was again selected as the measure of central tendency for the same reason noted above.

The normalized objectivity values across all states ranged from 0.001 to 0.040, and all states showed very objective summaries. Some states stand out as having slightly more subjective tone to the summary, and these states include California, Texas, and Florida. These were the same states that stood out as having a more positive tone to the summary, and could indicate a positive correlation between sentiment and subjectivity.

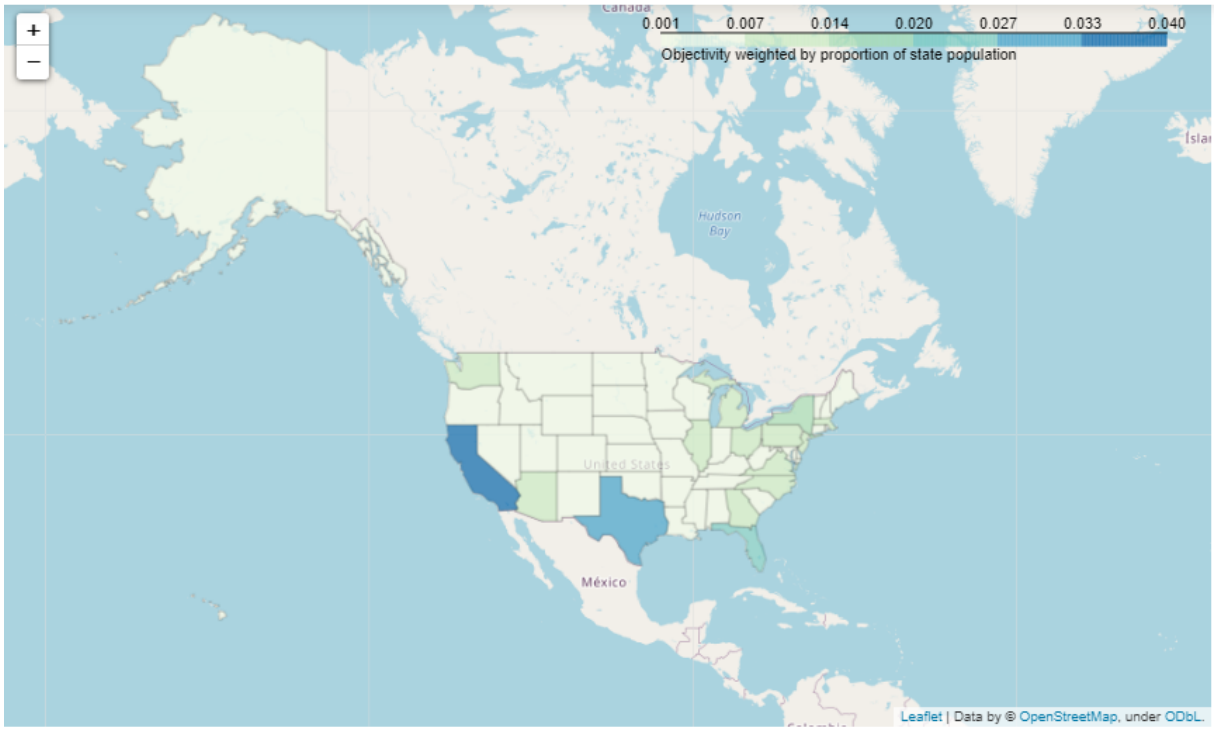

Figure 10: Objectivity weighted by proportion of state population

9 Results

The results of this analysis showed the most common n-grams are as follows: The five most common 2-grams are **looked like**, **white light**, **could see**, **red light**, and **bright light**. The five most common 3-grams are **bright white light**, **never seen anything**, **seen anything like**, **high rate speed**, and **bright orange light**. The five most common 4-grams are **never seen anything like**, **i never seen anything**, **light moving across sky**, **moving high rate speed**, and **saw bright white light**. The most common element of all of these reported sightings is that of light moving through the sky, suggesting that most sightings are at night.

As expected, the summaries tended to be fairly neutral and objective. This is expected as a report of an event is, by definition, neutral and objective. The states with slightly more positive and subjective tone to their summaries were California, Texas, and Florida. One thing these states have in common are NASA and SpaceX launch locations. The more positive and subjective tone may be due to the awareness and excitement of launches in those areas. This would make sense as excitement could be viewed as both a positive and subjective emotion. This observation also indicates a possible correlation between sentiment and subjectivity.

Some other key observations unrelated to the research questions were

- The number of sightings reported to NUFORC drastically increased in 2006.
- Popular days to view and then report UFO sightings are New Year's Day and the 4th of July.

- The most likely time to view a UFO is between 11p and 6a.
- Lights are the most common UFO seen in the skies, and not physical crafts.
- White, red, and orange are common colored lights that cannot be identified in the night sky.
- The Pacific Northwest and the Northeast stand out as regions with high number of sightings per capita.

10 Conclusions

NLP was used in this study to identify trends observed in the language used to report UFO sightings across the United States. Significant time was spent scraping the data from the NUFORC website and cleaning the summaries. The spelling was corrected, contractions expanded and words lemmatized for all the summaries. Other text cleaning included removing punctuation, and handling repeated spacing.

A bag of words model enabled determination of the most common n-grams in the summaries. WordCloud images were also used to visualize the most frequent words in the summaries.

Sentiment and subjectivity analysis confirmed the UFO sightings reports across the United States were very neutral and objective. California, Texas, and Florida have slightly more positive and subjective tones to their summaries indicating some extra awareness and excitement about UFO sightings in those areas.

In conclusion, NLP enabled interesting insights to be extracted from the detailed UFO sighting summaries. The observations made could be due to a variety of factors that are not known to the researchers, but do provide a starting place for deeper analysis in the future.