UFOLOGISTS

BY:

DAVID DECOSTA

KAITLYN NYCE

CAMERON BOWEN

COMP 4450 INTRODUCTION TO
DATA MINING

MAHMOOD HOSSAIN

MARCH 8, 2022

Problem Description

Our group is going to be investigating UFO sightings. There are many sightings of UFO's and people report them. All these reporting's have been gathered into a dataset. We are going to parse through the data to show a clearer picture of what everyone is seeing and where.

Literature Review

[1] This study takes a mass data set of UFO sightings and attempts to legitimize the data. They do this by studying online reports. They look at cluster density and data visualization to search the space of various cluster realizations to divide the best probable clusters that provide them information about the proximity of the UFO sighting. They introduce a forest classifier to identify true events and hoax events, using the best possible features such as region, time, and duration of the sighting.

[2] This study used a text analysis tool to understand data about UFO sightings. Different types of analysis and modeling of textual data was used as well as geographic tools. The combinations of all these tools and methods allowed them to come interesting conclusions about UFOs, such as, UFOs are similar to an aircraft with a peculiar visibility and they disappear within seconds of appearing, among other discoveries.

[3] This study is about processing data collected from alien species. Once the data is harvested it is processed through validation, cleaning, standardization, and geocoding. After the data is processed, it is then moved to a database where they will use the data to map distributions of the alien species.

[4] This study was used on a dataset containing 80,000 records of UFO sightings. Some of the data was missing in the tuples and they had to clean the data to make it easier to analyze. They even chose to leave out one of the attributes. A histogram was constructed to visualize where the most sightings were comparing the United States to the rest of the world. Then did the same with the individual states within the USA. Also, a comparison to see if there was a correlation between the UFO shapes and the time of the sightings.

[5] The study was based off the same CSV file containing 80,0000 rows of data. The initial cleaning of the data stripped weird characters and made the duration and latitude floats. The steps they took were: Importing the data, cleaning up/transforming the data, visualizing the data, splitting the training set and test set, fine tune algorithms, compare accuracy scores, and end up with the best prediction model.

Goal

We are going to perform text parsing and filtering. The most common terms showing up we will classify as its own attribute. These classifiers would be names of the given ships. We then parse through the original data set linking the ship to the sightings appearing most frequently and from that data it will show us where most likely that ship will be seen using predictive analysis. Another feature would be using time-series analysis to predict the next most likely time period for the particular ship to be seen.

Description of the Dataset

The dataset we will be using has 88,876 tuples. The dataset contains date/time, city, state, country, shape, duration, description of the sighting, date posted, latitude and longitude.

References

[1] Harish Krishnamurthy, Anna Lafontant, and Ren Yi. 2019. Localization of geospatial events and hoax prediction in the UFO database. (January 2019). Retrieved March 10, 2022 from https://publications.waset.org/10009995/localization-of-geospatial-events-and-hoax-prediction-in-the-ufo-database

[2] Pradeep Reddy Kalakota, Zabiulla Mohammed , Naresh Abburi, and Dr. Goutam Chakraborty. Alien Nation: Text Analysis of UFO Sightings in the U.S. Using SAS Enterprise Miner 13.1. Retrieved March 10th, 2022 from https://www.researchgate.net/profile/Goutam-Chakraborty-4/publication/279530493_Alien_Nation_Text_Analysis_of_UFO_Sightings_in_the_US_Using_SASR_Enterprise_Miner_131_by_Pradeep_Reddy_Kalakota_Naresh_Abburi_Zabiulla_Mohammed_and_Goutam_Chakraborty/links/55955eb208ae5d8f3930ee84/Alien-Nation-Text-Analysis-of-UFO-Sightings-in-the-US-Using-SASR-Enterprise-Miner-131-by-Pradeep-Reddy-Kalakota-Naresh-Abburi-Zabiulla-Mohammed-and-Goutam-Chakraborty.pdf

[3] Ivan Deriu, Fabio D'Amico, Konstantinos Tsiamis, Eugenio Gervasini, and Ana Cristina Cardoso. 1AD. Handling big data of alien species in Europe: The European Alien Species Information Network Geodatabase. (January 1AD). Retrieved March 10, 2022 from https://www.frontiersin.org/articles/10.3389/fict.2017.00020/full

[4] Saúl Buentello. 2021. Are we alone in the Universe? - data analysis and data visualization of UFO sightings with R. (September 2021). Retrieved March 10, 2022 from https://towardsdatascience.com/are-we-alone-in-the-universe-data-analysis-and-data-visualization-of-ufo-sightings-with-r-42d0798679c3

[5]Markus, Angeliki, Lenka. ML and Big Data 2019. Retrieved March 10, 2022 from https://www.nbi.dk/~petersen/Teaching/AppliedMachineLearning2019.html

[6] (NUFORC), National UFO Reporting Center. "UFO Sightings." Kaggle, 13 Nov. 2019, https://www.kaggle.com/NUFORC/ufo-sightings/version/2.

This is where our dataset is coming from.