Look! UFO

Jin Yan (jyan16), Jing Qian (jqian7), Yao Yao (yyao17), Lei Tian (ltian1)
{jin_yan, jing_qian, yao_yao, lei_tian}@brown.edu

February 13, 2017

Abstract

This is the pre-proposal of our final project for CS1951A.

1. VISION

1.1. Big Idea

- Predicting when and where UFO will appear, and the trend of its occurrence.
- Figuring out which of these UFO reports are fake by Natural Language Processing and statistical analyzing.
- Drawing UFO charts based on data of their reports with corresponding weather and GIS information.

1.2. What will we find

- Predicting the probability of the appearance of UFO at certain location and time.
- Finding the relationship between UFOs features and characteristics with the local weather and GIS where it appears.
- Figuring out which UFO reports may not be true.
- Visualizing UFO reports data, weather and GIS data, and drawing charts.

2. DATA

2.1. Dataset to use

We plan to use dataset from GIS, weather Data, UFO-reports.

2.2. How to collect

We can download/scrap all data from online sources about UFO, GIS and weather.

2.3. How big is it

UFO data: ($\approx 100,000$)

data/time, city, state, shape, duration, summary, posted date

Weather data: ($\approx 693,000$):

Date, Time, Sky Condition, Visibility, Rel Humid, Wind Speed, Wind Dir, Pressure

GIS data: Google Map API

2.4. How to clean

- For UFO-report data, we plan to drop the witness data too early, because of the lack of weather data and the change of geography circumstance.
- Quantization of different fields from different data sources, like time for UFO-report and weather dataset.
- Using PCA to decrease high dimension dataset, in order to facility model

3. METHODOLOGY

3.1. what we plan to do

Predicting about UFO needs correlation of many sources of data, including the UFO-report data, geography information, and the weather. UFO Reports is the basic data of our system. There are 5 steps regarding our data:

- Creating database of GIS, UFO-Reports and weather.
- Using NLP technology to analyze summary data in the UFO Reports data. To extract features about UFO. Creating chart (style, color, characteristics) of UFO.

- Indexed by date and location data of UFO Reports, scraping the corresponding GIS and weather data, and integrating them together.
- Using machine-learning or neural network to train our data, figure out patterns and correlation between UFO witness with GIS and weather at that time.
- Combining the information entered by user and weather forecasting data to do prediction In order to predict the possibility to see an UFO in the future.

3.2. Technologies to analyze

MongoDB for storing our data. Machine Learning, Neural Network, NLP to build data analysis models. MapReduce as computing method.

3.3. Technologies to visualize

D3, CG to create charts that show the features we extract from the witness summary data. D3 creates visualization of Big Data analysis results. Map to show the distribution of UFO witnesses, as well as visualizes their relation with weather and GIS. All of the above can be shown on a website.

4. CAPSTONE

None of our group members need CAPSTONE to fulfill our requirement. However, we will consider the following features since they are integral part of our project.

4.1. Database

We will be using either MongoDB or mySQL depends on our finalized data scheme.

4.2. Data pipeline

We will create a system based in python that does Natural Language Processing, Classification and generating statistical results. With UFO-witness summaries, we are able to extract high probability features of UFO occurrence, such as shape, color, movement, brightness. We would then use these data to generate charts of all documented UFO characteristics. With time, location and weather data we could used to predict future events of UFO. By user inputting desired destination, time and weather condition, our system will generate a probability one would encounter UFO appearance. We will also build an interactive front-end for users to change parameters and see results.

4.3. Front-end

We will incorporate HTML5 + CSS3 with Python Server to create a modern interactive web application. Users will be able to access using a computer or smartphone. By selecting parameters such as weather, location, and time, users will see the likelihood of an UFO occurrence. Users could also see visualizations of statistical data by selecting criteria such as time of the day, day of the month, month of the year, places, shapes, weather conditions etc. Finally we will provide a text-area where users could enter UFO reports, and our application will give probability of whether such report is a hoax or not based on analyzing the existing language contents of UFO events.

5. GET STARTED

5.1. Befor TA check-in

By the first TA check-in, we will first accomplish data collection and cleaning. Then, for the part of data integration, we plan to combine the UFO data with information of GIS and weather to for further prediction. Additionally, a database management system will be built for supporting before first check-in.

5.2. Midterm report

By the midterm report, we will provide a summary of NLP as well as corresponding Charts of UFO and visualization. Also, we plan to partially complete the prediction and HCI part of our project.