I Want to Believe: Visualization of Linguistic Features in UFO Sighting Reports

Hayley Guillou, guillouh@cs.ubc.ca
Theodore Smith, smithtg@email.arizona.edu

INTRODUCTION

For several decades, the National UFO Reporting Center (NUFORC) has been compiling reports of UFO sightings across North America in a publically-available data set[1]. These reports feature details including the time, location, and duration of each sighting, the categorical shape of the UFO, and a free response description of the encounter. While a small number of parties have produced visualizations from these data, they have predominantly focused on mapping reports based on the fixed response variables such as time and location while seldom providing detailed views of sighting descriptions. This approach presumably stems from the non-uniformity of user-submitted descriptions, which presents a challenge with respect to data aggregation.

The past several years have been marked by increased interest in natural language processing tools capable of extracting relationships between short, irregular segments of text. This trend has largely been driven by access to large data sets derived from social media platforms such as Twitter, featuring concise messages that do not necessarily follow common, language-oriented grammatical structures. We propose that these tools are well suited to the descriptions found in the NUFORC data set, as the descriptions are typically very brief and rarely adhere strictly to the rules of the English language.

The natural language processing tools used in completion of this project will be new to both members of our group. With that in mind, we intend to leverage Hayley Guillou's experience in human-computer interaction and Theodore Smith's experience in geographical data mapping to partition the workload for those tasks, while closely coordinating on the task of visualizing the sighting descriptions.

DATA AND TASK

Data will be derived from a scrubbed version of the NUFORC UFO sighting reports data set found on the machine learning and data science platform Kaggle[2]. This data set represents 80,000 reports from the original data set of roughly 110,000 reports, with cases dropped due to missing location as well as missing or erroneous time. This data set also features a standardized form of the duration of each sighting, with time reported in seconds rather than in the mixed formats found in the original data set. Comments are retained in full-text form without editing. Categorical data including location and shape will be used to guide spatial plotting, while ordinal data including time and duration will be used to temporally filter the map.

This visualization revolves around the separate, but related, tasks of representing the data in a spatiotemporal format as well as representing the linguistic relationships between

isolated report descriptions. The data will be geographically mapped based upon the locations specified in the reports, while the user will have control over the intervals of time and duration they are interested in viewing. Users will also have the option to select a subset of the data based upon geographical location either by a specific area of interest or by a distance threshold from a particular report. A linked-view will present a continually updated summarization of the text content drawn from the report descriptions of the selected data. Ultimately, the primary task will be to allow users to navigate the data from a spatiotemporal perspective in order to observe the impact of selection and filtering on the resultant "sentence tree." The hope is that this will lend insight into the nature of clusters of sightings, based upon the information-rich but poorly standardized comments provided by contributors to the original data set.

PROPOSED INFOVIS SOLUTION

Data will be plotted on a geographic map using shape marks representing the shape specified in each report. The initial view will contain all sightings from the data set (Figure 1, top portion). A slider will be provided to allow users to filter records based upon a specified interval of time. A second slider will be provided in order to allow users to set an interval for sighting durations in order to further filter the data. Users will also have the option to select a sighting from the map view, producing a detailed view of that particular sighting. Upon selecting a report, users will have the option to filter the data based upon a specified distance from the selected report. Finally, users will have the option to select a number of sightings from the map, filtering out all other sightings.

The SentenTree view (Figure 1, bottom portion) will be updated dynamically based upon the data selected in the map view. Our intended scenario of use is that users will explore the data set using the spatiotemporal view in order to observe the effects of their navigation on the output of the NLP view. By hovering over a "sentence path" in the SentenTree, the relevant points in the map view will be highlighted.

As a scenario, imagine you have recently come to the realization that the National UFO Reporting Center is an actual institution and furthermore, you have found out that there have been over 110,000 sighting *reported* to this organization. While this may be shocking, you might be curious where these sightings are happening, what people are reporting to see, and what kind of reports are being filed. It is at this point, you would use our visualization tool to get a aggregated view of where and when these sightings are happening. By using filters and interacting with data points on the map, the NLP-based SentenTree view will update giving a brief aggregation of the textual summary provided with each of the sightings. By highlighting a sentence path, the corresponding data points will be highlighted on the map. By interacting with the map and NLP-based views, you will be able to see highlights of the UFO sighting data with either the intention of learning what is being reported, or for entertainment purposes to playfully mock the frequency of the sightings and the detail provided within each report.

Figure 1: Main aggregate view. The top portion is the map view, which can be filtered according to date, duration, and shape. The bottom portion is the NLP-based SentenTree view, which is responsive based on the filtering of the map.

PROPOSED IMPLEMENTATION APPROACH

We will use pen and paper for our initial, low-fidelity design prototyping. For the final implementation we will produce a browser-based visualization using JavaScript and D3.js. Existing libraries will be used where possible in order to reduce duplication of labor with the intention of allocating a greater proportion of our time to the integration of the NLP-driven view. We primarily plan to utilize SentenTree for visualization of aggregated sighting description content, using visually filtered data from the spatiotemporal view as input to the tool. SentenTree was selected based upon its intended use for social media messages which we perceive to be structurally similar to the comments in the NUFORC data set, as well as on the basis that there is a JavaScript implementation of the tool which simplifies its integration into our own web visualization tool. For geographic mapping, we are likely to use d3.geomap, a library written in JavaScript and built on top of D3.js.

MILESTONES AND SCHEDULE

In Table 1, we have outlined the due dates and time estimations for each milestone of our project. In terms of splitting up the work, Hayley will be working on a large portion of the design process and the visual layout of the tool, and Theodore will be focusing on the back-end

and data binding, along with the geospatial mapping. Up until now, we have been working equally on the writing and research portions but this distribution may change as we go forward.

Task	Deadline	Est. time	Description
Pitches	Oct. 17	2 (x2)	Research topics, create slides, rehearse
Proposal	Nov. 6	14	Meeting with Tamara, researching datasets, reading background materials, creating mockups, writing proposal
Compile data	Nov. 10	2	Determine if scrubbed data on kaggle.com is sufficient, otherwise scrape and clean our own data
Map sightings	Nov. 14	8	Begin preliminary visualization task, mapping the sighting locations geographically;
NLP analysis	Nov. 21	16	Decide on NLP visualization, start preliminary implementation with mock data
Peer Review 1	Nov. 21	4	Prepare slides, prepare preliminary demo
Map View and timeline	Dec. 2	30	Implement map view including fetching data, display/layout, interaction, animation
Peer Review 2	Dec. 6	4	Prepare slides and demo
NLP View	Dec. 10	20	Implement language analysis portion of the visualization including fetching data from map view, display, and animation
Presentation	Dec. 12	10	Create slides, rehearse speaking,
Final Paper	Dec. 15	20	Write paper, create figures and tables

Table 1: Breakdown of milestones

PREVIOUS WORK

The NUFORC dataset that we are using has been used for a number of visualizations, especially on visualization blogs. One blog has an interactive heatmap of the sightings in 2015 using CartoDB[3], another investigates trends in frequency of sightings per capita using histograms[4]. There have also been poster prints showing trends in shape, sighting time, and sightings per capita by John Nelson on IDVSolution's UX Blog[5].

Our inspiration for the NLP visualization is from SentenTree [1], a text visualization technique for summarizing a collection of social media text. The aim of the project was to create a visualization that is cheap to compute and strikes a balance between summarizing the most frequently used words and preserving the sentence structure.

Another work that we seriously considered was TextTile [2], a data visualization tool for structured data and unstructured text. The tool was created to visualize healthcare reviews from Yelp, so the data includes structured ratings and unstructured reviews, quite similar to the UFO sighting reports. However, the textual data we have in the sighting reports matches better with social media text than Yelp reviews and this visualization was found to be overly structured.

REFERENCES

[1] Hu, M., Wongsuphasawat, K., & Stasko, J.: Visualizing social media content with sententree. *IEEE transactions on visualization and computer graphics*, p 621-630, 2017.

[2] Felix, C., Pandey, A. V., & Bertini, E.: TextTile: an interactive visualization tool for seamless exploratory analysis of structured data and unstructured text. *IEEE transactions on visualization and computer graphics*, p 161-170, 2017.

---

[1] http://www.nuforc.org/webreports.html
[2] https://www.kaggle.com/NUFORC/ufo-sightings/data
[3] http://wesmapping.com/blog/mapping-ufo-sightings-in-2015/
[4] http://www.blog.elimak.com/2012/09/hello-ufo-data-visualization-of-the-nuforc-db/
[5] http://uxblog.idvsolutions.com/2015/06/sightings.html