ML1010 - APPLIED MACHINE LEARNING

Project Proposal - UFO Sightings

Prepared for: Annie Lee, PHD, MMATH
Prepared by: Jairo Melo, Ing. MSC
March 27, 2019
Proposal number: 001

Motivation

“The universe is a pretty big place. If it's just us, seems like an awful waste of space.”

— Carl Sagan, Contact

Since my childhood, I have been fascinated with idea we’ve might been visited from other planets. NUFORC or, National UFO Reporting Center, has collected ufo reports for close to a century, 80,000 plus observations. This dataset contains US/Canada geolocated and time standardized, shape of the UFO, duration in text, and a short description of the sighting.
Full acknowledgement to the NUFORC organization as they present this data to the world is acknowledged in my project.

A	B	C	D	E	F	G	H	I	J	K
10/10/1949	san marcos	tx	us	cylinder	2700	45 minutes	This event took place in early fall around 1949-50. It occurred after a Boy Scout meeting in the B	4/27/2004	29.8830556	-97.941111
10/10/1949	lackland afb	tx		light	7200	1-2 hrs	1949 Lackland AFB&#44 TX. Lights racing across the sky & making 90 degree turns on a dir	12/16/2005	29.38421	-98.581082
10/10/1955	chester (uk/england)		gb	circle	20	20 seconds	Green/Orange circular disc over Chester&#44 England	1/21/2008	53.2	-2.916667
10/10/1956	edna	tx	us	circle	20	1/2 hour	My older brother and twin sister were leaving the only Edna theater at about 9 PM&#44...we ha	1/17/2004	28.9783333	-96.645833
10/10/1960	kaneohe	hi	us	light	900	15 minutes	AS a Marine 1st Lt. flying an FJ4B fighter/attack aircraft on a solo night exercise&#44 I was at 5	1/22/2004	21.4180556	-157.80361
10/10/1961	bristol	tn	us	sphere	300	5 minutes	My father is now 89 my brother 52 the girl with us now 51 myself 49 and the other fellow which	4/27/2007	36.595	-82.188889
10/10/1965	penarth (uk/wales)		gb	circle	180	about 3 mins	penarth uk circle 3mins stayed 30ft above me for 3 mins slowly moved of and then with the bl	2/14/2006	51.434722	-3.18
10/10/1965	norwalk	ct	us	disk	1200	20 minutes	A bright orange color changing to reddish color disk/saucer was observed hovering above power	10/2/1999	41.1175	-73.408333
10/10/1966	pell city	al	us	disk	180	3 minutes	Strobe Lighted disk shape object observed close&#44 at low speeds&#44 and low altitude in Oct	3/19/2009	33.5861111	-86.286111
10/10/1966	live oak	fl	us	disk	120	several minutes	Saucer zaps energy from powerline as my pregnant mother receives mental signals not to pass	5/11/2005	30.2947222	-82.984167
10/10/1968	hawthorne	ca	us	circle	300	5 min.	ROUND &#44 ORANGE &#44 WITH WHAT I WOULD SAY WAS POLISHED METAL OF SOME KIND	10/31/2003	33.9163889	-118.35167
10/10/1968	brevard	nc	us	fireball	180	3 minutes	silent red /orange mass of energy floated by three of us in western North Carolina in the 60s	6/12/2008	35.2333333	-82.734444
10/10/1970	bellmore	ny	us	disk	1800	30 min.	silver disc seen by family and neighbors	5/11/2000	40.6686111	-73.5275
10/10/1970	manchester	ky	us	unknown	180	3 minutes	Slow moving&#44 silent craft accelerated at an unbelievable angle and speed.	2/14/2008	37.1536111	-83.761944

https://gengo.ai/datasets/16-strange-datasets-for-machine-learning/
https://github.com/planetsig/ufo-reports

Project Plan

Through CRISP-DM, I will provide the steps to complete the Sentiment analysis for the UFO sightings dataset.
Data understanding
Data preparation
Modeling
Evaluation

Deployment

Machine Learning Approach

We will approach this analysis using unsupervised learning algorithms. First, we will identify the TF-IDF, term frequency inverse document frequency, is a method of quantifying article word counts.

"TFIDF, short for term frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.[1] It is often used as a weighting factor in searches of information retrieval, text mining, and user modeling. The tf–idf value increases proportionally to the number of times a word appears in the document and is offset by the number of documents in the corpus that contain the word, which helps to adjust for the fact that some words appear more frequently in general. Tf–idf is one of the most popular term-weighting schemes today; 83% of text-based recommender systems in digital libraries use tf–idf.[2]"
Taken from: https://en.wikipedia.org/wiki/Tf%E2%80%93idf

When TF-IDF matrix of vectors is obtained, we can compare these similarities to each other by using clustering methods; for example, k-means or Hierarchical clustering.

Finally, we will implement a sentiment analysis on the description of the event, to identify the emotion and subjective in the language used in the reports. We could used a NLP library for example TextBlob to identify the type of sentiment of the reports; which could be associated to the sightings.