Where is Hollywood?: Artificial Intelligence Approach

Shawna Tuli, Lynn Gao, Ethan Song, Jing Huan Ooi
Phanwadee Gift Sinthong, Erik Arriaga, Jucheol Moon

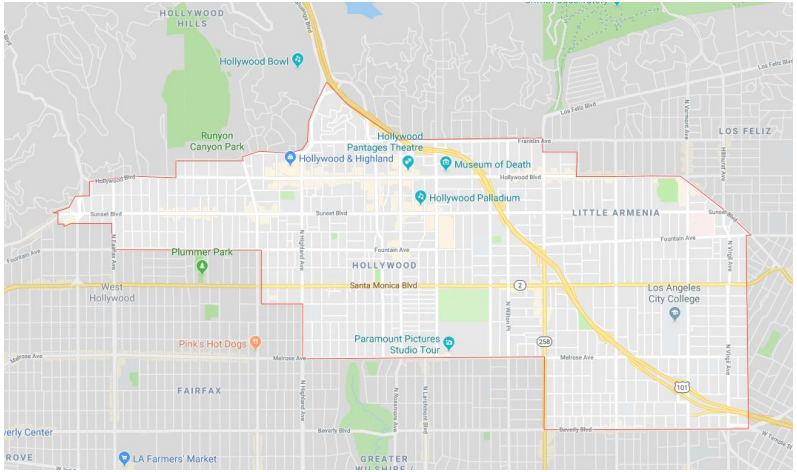
exploreCSR workshop: “Computing to Change the World for the Better”

Context and Motivation

When we want to bring people to visit Hollywood, we are not sure which part of Hollywood is actually considered ‘Hollywood’, so we decided to use tweets and their geolocation to find out where Hollywood truly is.

To achieve this goal, we will use:

- Tweepy
- APIs (twitter, natural language processing, etc.)
- DBSCAN
- google.com/mymaps



Research Design & Methods (JSON)

exampleDictionary =
{"timestamp":"Sat Feb 16 12:46:01 +0000 2019",
{"location":["-68.3191205085327","29.316262917043694"]},
"text":"@TheAn1meMan Can hollywood do hollywood and leave anime alone jeez they cannot get it right. Even James Cameron strug… https://t.co/bbkjLDKVzw"}

This is JSON format.

JSON:
We saved the data into JSON format, allowing us to parse the data more easily and quickly. With JSON, we are able to call what we are looking for using dictionaries, their keys, and their values.

Research Design & Methods (Tweepy)

4 weeks of Tweets scraped using Twitter API

{"timestamp":"Sat Feb 16 12:46:01 +0000 2019",
{"location":["-68.3191205085327","29.316262917043694"]},
"text":"@TheAn1meMan Can hollywood do **hollywood** and leave anime alone jeez they cannot get it right. Even James Cameron strug… https://t.co/bbkjLDKVzw"}

TWEEPY:
In order to access this information, we signed up for a twitter developer account to authenticate ourselves and used Tweepy, a library for Python, to access the tweets and their information. We then collected all the tweets that were in English and wrote them into a file. Through this, we were able to get information about the date (timestamp), geolocation, and twitter text in json format.

- We searched through the data and obtained all the tweets that had ‘ hollywood ’ in the text.
- We found that only 0.2% of all tweets collected mentioned hollywood.

Research Design & Methods (Google NLP)

Sentiment analysis using Google cloud - natural language API.

Positive:
RT @billboard: Congrats @Pink! She received a star on the **Hollywood** Walk Of Fame ⭐
https://t.co/iPDRPqXmgq https://t.co/342tmzc8X2
Score: 0.30000001192092896 Magnitude: 0.6000000238418579

Negative:
@annnay_ so u think u **hollywood** huh
Score: -0.20000000298023224 Magnitude: 0.20000000298023224

Score: how positive or negative the text is.
Magnitude: how much emotion the text has.

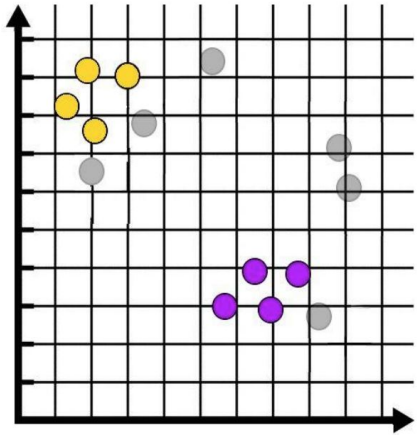
Sentiment	Sample Values
Clearly Positive*	`"score"`: 0.8, `"magnitude"`: 3.0
Clearly Negative*	`"score"`: -0.6, `"magnitude"`: 4.0
Neutral	`"score"`: 0.1, `"magnitude"`: 0.0
Mixed	`"score"`: 0.0, `"magnitude"`: 4.0

Research Design & Methods (DBSCAN)

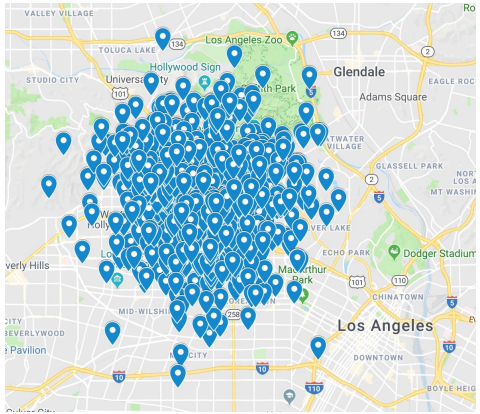
3 weeks of Tweets scraped using Twitter API – Density-based Spatial Clustering of Applications with Noise

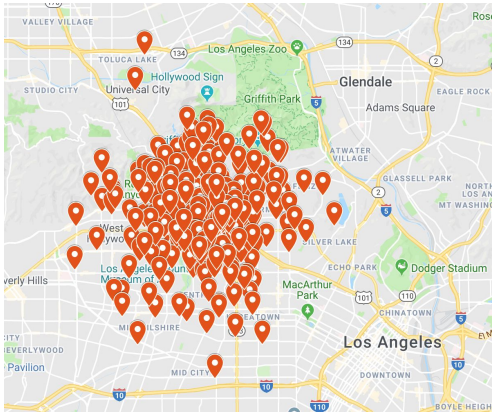
{"timestamp":"Sat Feb 16 12:46:01 +0000 2019",
{"location":["-68.3191205085327","29.316262917043694"]},
"text":"@TheAn1meMan Can hollywood do **hollywood** and leave anime alone jeez they cannot get it right. Even James Cameron strug… https://t.co/bbkjLDKVzw"}

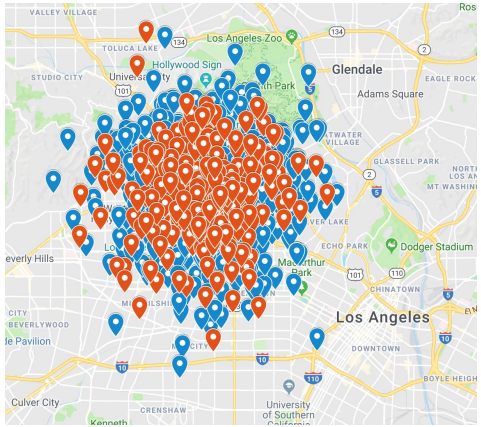
Noise is outside of the core and edge points in the clusters, which have been identified using a hyperparameter (radius).



Results, Observations & Insights: Sentiment Analysis

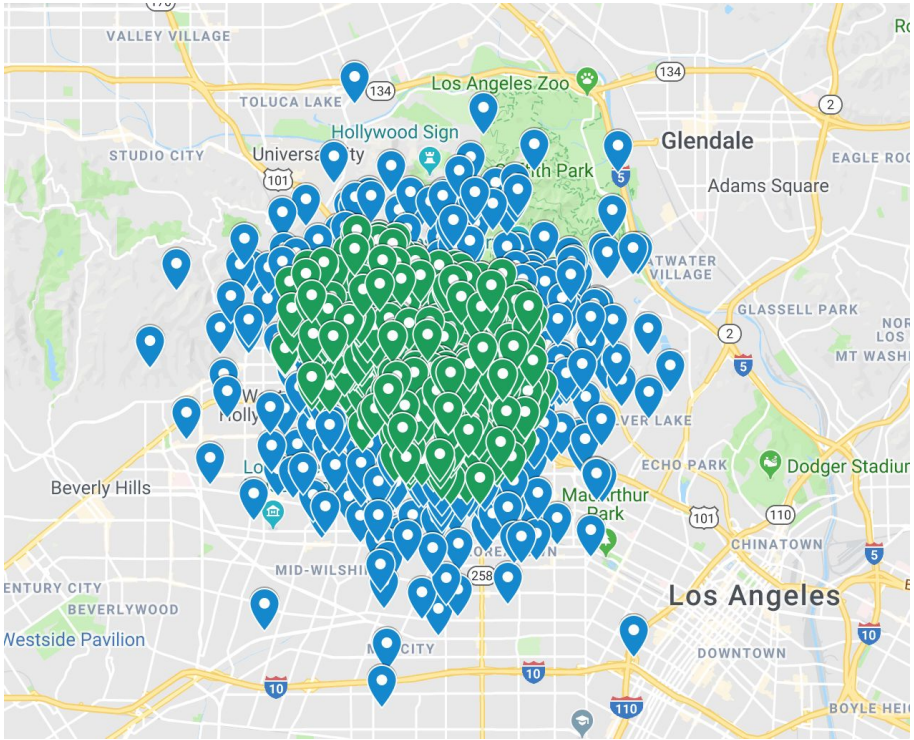






- We were able to scan a large database of tweets.
- We then filtered tweets that pertained to Hollywood, and plotted where these tweets were written with geolocation.
- Sentiment analysis was used to differentiate between tweets painting Hollywood positively (pins in blue) and negatively (pins in red).

Results, Observations & Insights: DBSCAN



- Before Google's DBSCAN, the plot of Hollywood tweet locations contained significant noise (blue).
- After using DBSCAN clustering algorithm, we were able to eliminate outliers (green).
- This allows us to determine which tweets originated from within Hollywood with greater accuracy

Conclusions & Future Plans

We found Hollywood!

Lynn: I learned a lot more about APIs, DBSCAN algorithm, and Natural Language Processing. The whole project was very fun and allowed me to broaden my horizons.

Jing Huan As a EE and BME student,I have gained a lots of computing and data collection knowledge. The project open my eyes to the methods available out there to collect meaningful information using AI.

We will change the world for the better!

Ethan: Working with Google's Sentiment Analysis and DBSCAN algorithms were able to analyze given tweet data fast and efficiently. It greatly helped me learn more about data analytics and machine learning

Shawna: The DB clustering algorithm is necessary in distinguishing the significant data points and was super cool to learn about :)