Facility Detection and Popularity Assessment from Large-scale Text Classification of Social Media and Crowdsourced Data

SUMMERY

Goal

ABSTRACT

Advances in technology have continually progressed our understanding of where people are, how they use the environment around them, and why they are at their current location. Having a better knowledge of when various locations become popular through space and time could have large impacts on research fields like urban dynamics and energy consumption. In this paper, we discuss the ability to identify and locate various facility types (e.g. restaurant, airport, stadiums) using social media, and assess methods in determining when these facilities become popular over time. We use natural language processing tools and machine learning classifiers to interpret geotagged Twitter text and determine if a user is seemingly at a location of interest when the tweet was sent. On average our classifiers are approximately 85% accurate varying across multiple facility types, with a peak precision of 98%. By using these methods to classify unstructured text, geotagged social media data can be an extremely useful tool to better understanding the composition of places and how and when people use them.

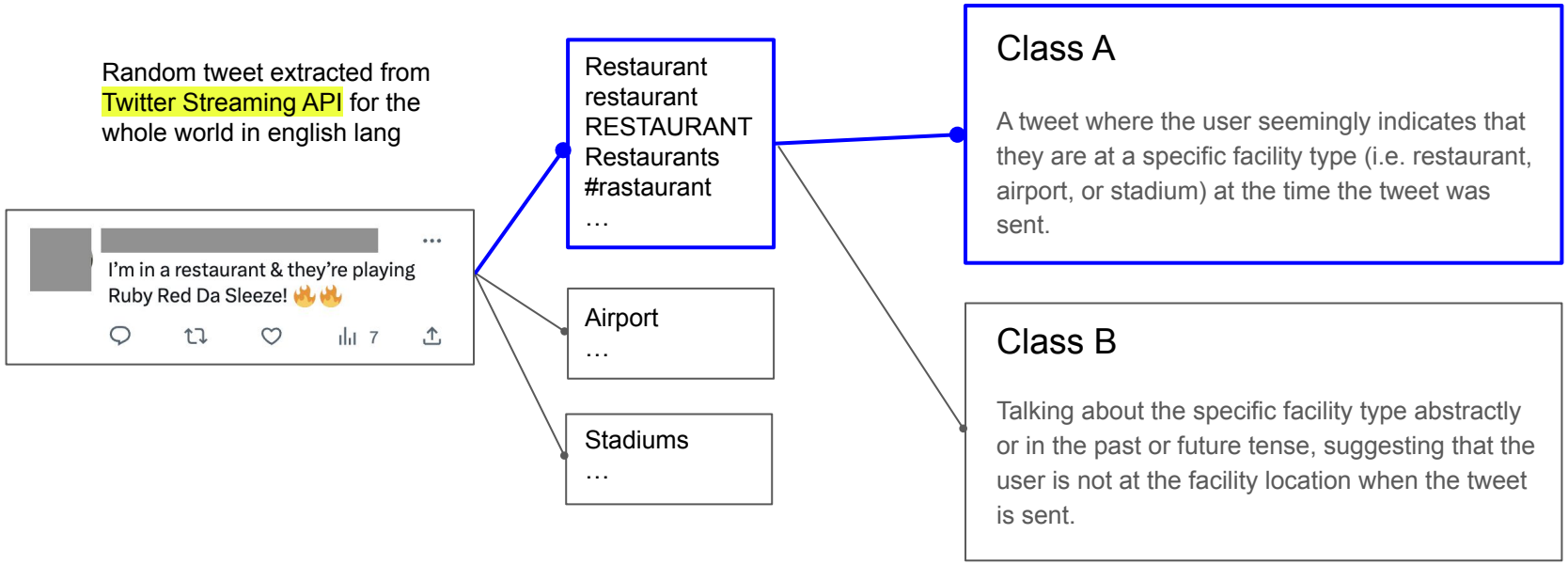
- Identify and locate various facility types (e.g. restaurant, airport, stadiums) using social media, and assess methods in determining when these facilities become popular over time.
- Interpret geotagged Twitter text and determine if a user is seemingly at a location of interest when the tweet was sent.

Example tweets

Tweet was made when the user was at an airport

Tweet was made when the user was seemingly not at an airport

Dataset Collection

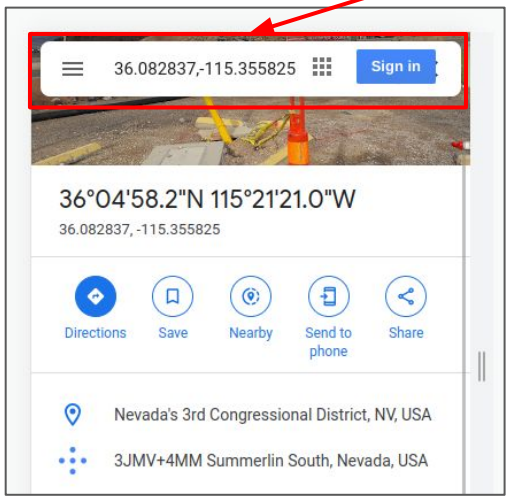


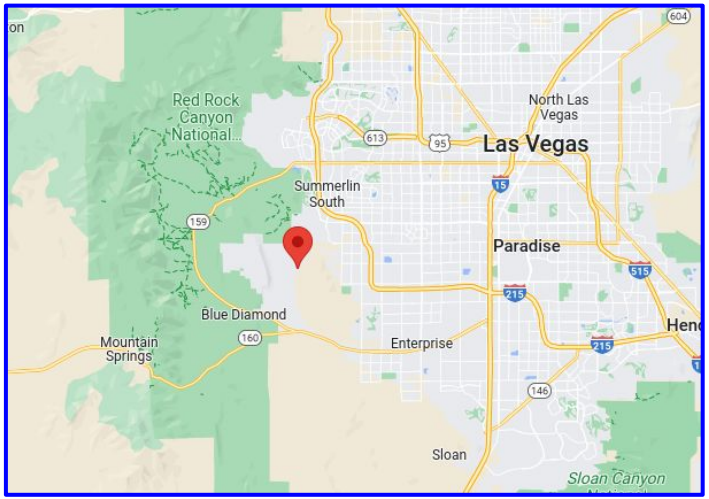
Ground Truth Validation

```
Coordinates: Coordinates(longitude=-115.355825, latitude=36.082837)
Place: Place(id='0134e6167ff7f6ec', fullName='Summerlin South, NV', name
Raw: My gym 💪 and my airport ✈ with the killers and Gwen deep cuts on
Rendered: My gym 💪 and my airport ✈ with the killers and Gwen deep cuts
```

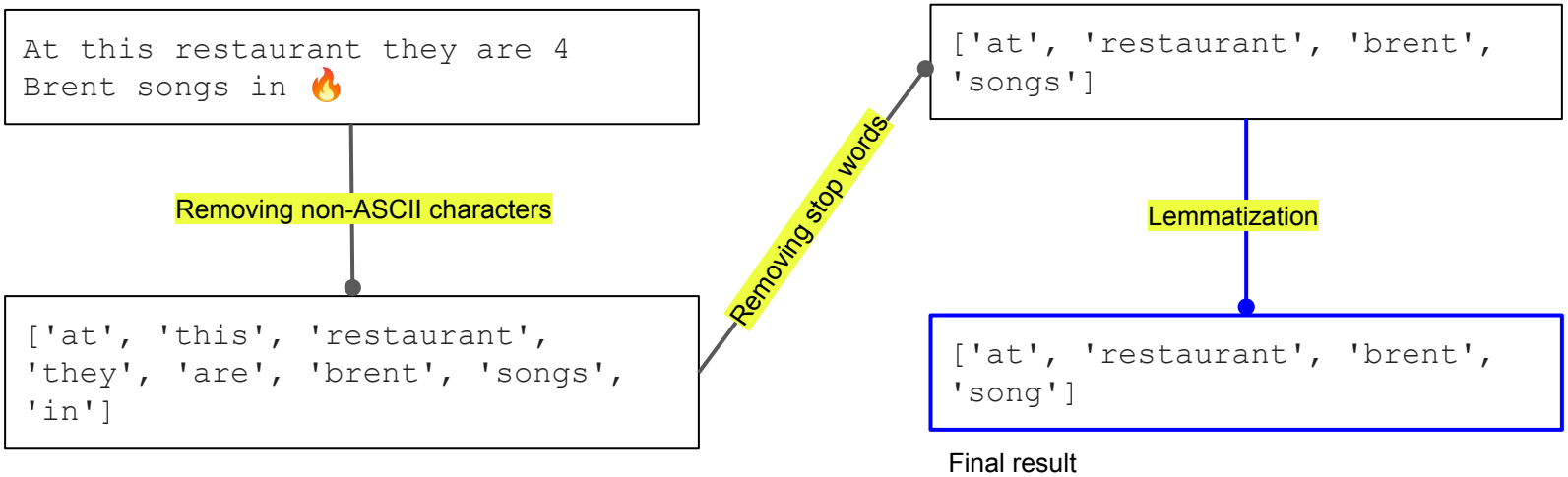
Label: Class A

Label: Class B





Data Sanitation and Cleaning



Vectorization

TF-IDF values

```
(0, 3190)    0.20326458156353006
(0, 6374)    0.259932916090755
(0, 4884)    0.3551754072726634
(0, 4445)    0.27210247658918724
(0, 7544)    0.5097986169293055
(0, 353)     0.04681640826638896
(0, 12030)   0.3651615134188747
(0, 312)     0.2568453514538075
(0, 12560)   0.396181192362104
```

Document Index

Word Index

TF-IDF value

Training Dataset

	content	class
0	update tracked air tag airport man man finally...	0
1	self check at airport passport holder shaawon ...	1
2	stopover at at dubai international airport dubai	0
3	yallll realized forgot wallet got airport tsa ...	0
4	checked at holiday inn express sydney airport ...	1

Classification

Naive Bayes

	precision	recall	f1-score	support
0	0.88	1.00	0.94	448
1	0.88	0.10	0.19	67
accuracy			0.88	515
macro avg	0.88	0.55	0.56	515
weighted avg	0.88	0.88	0.84	515

Support Vector Machine

	precision	recall	f1-score	support
0	0.92	0.99	0.95	448
1	0.87	0.40	0.55	67
accuracy			0.91	515
macro avg	0.89	0.70	0.75	515
weighted avg	0.91	0.91	0.90	515