GeoSkynet

Ken Bier, Anthony Coccia, Aaron Dodson, Matthew Greenfield, & Peter Huang

mobile

modular

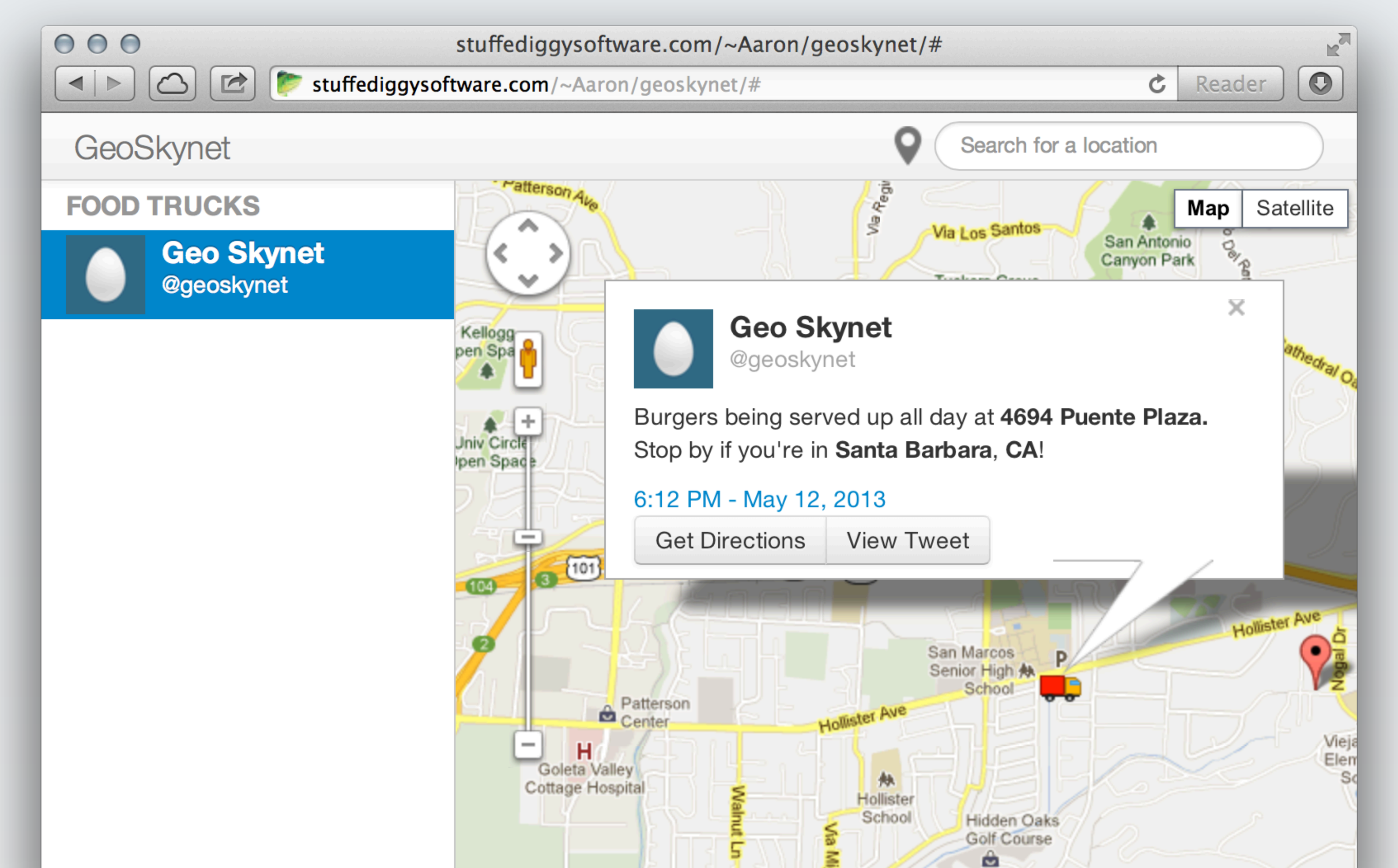
scalable

asynchronous

heuristic-based

natural language

named entity recognition



The Problem

Thousands of food trucks use social media to inform hungry customers of their whereabouts. However, the meaning of these messages can be difficult for computers to extract from “natural language” text.

Although “the corner of State and Haley” uniquely identifies a location to a human, to a computer it’s just another string of characters. Being able to identify these “geostrings” and transform them into a computer-readable representation opens up numerous possibilities for presenting and analyzing location data, and makes possible useful tools for hungry food truck afficianados.

Our Solution

GeoSkynet uses a decoupled, modular architecture to achieve its goal of identifying locations in natural language tweets. A dedicated Twitter sensor module written in Python listens for incoming tweets. As it receives them, it dispatches them to a series of processor modules.

These modules look for addresses, cities, states, zip codes, organizations, and intersections using an extensively tested set of heuristics, as well as Stanford’s Named Entity Recognition engine.

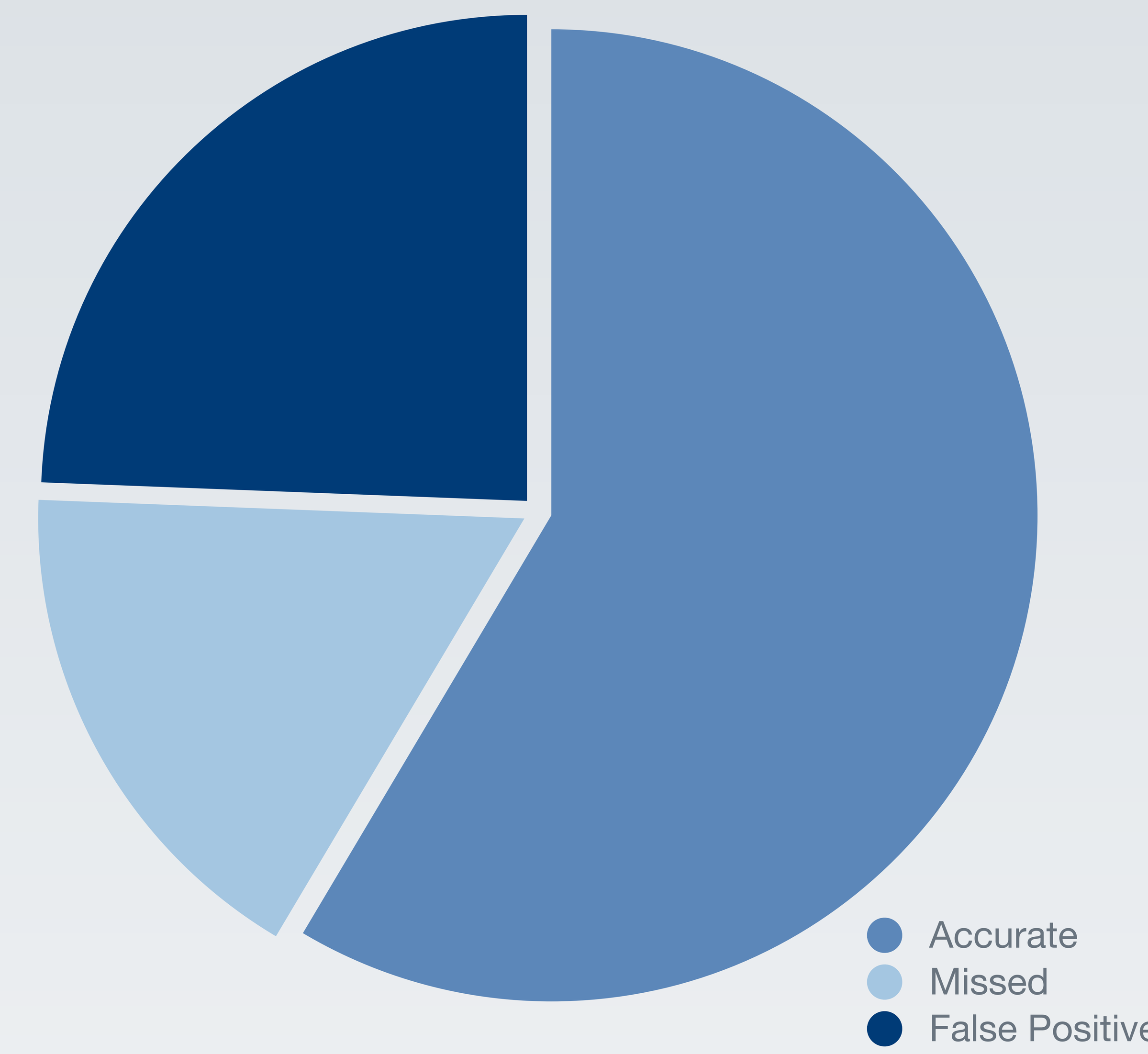
When geostrings are identified, they are sent to the database writer module to be written to MongoDB along with other tweet metadata. From here, they are dispatched to the Geocoder module which uses the Google Maps API to transform them into latitude and longitude coordinates.

The entire process takes place asynchronously on multiple threads with communication facilitated by RabbitMQ, allowing for a scalable, responsive system.

The Results

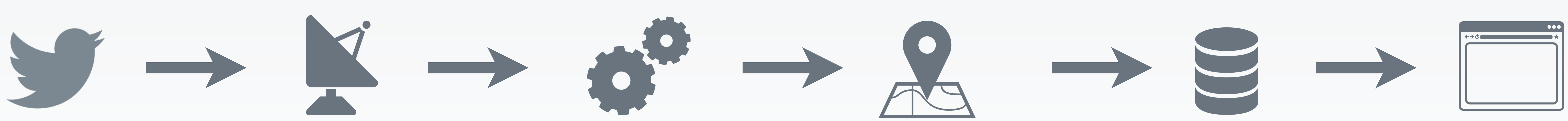
Although the GeoSkynet system does not achieve perfect accuracy, initial results are quite good. Around 60% of geostrings in actual food truck tweets are correctly identified. The overall ratio is heavily dependent on the truck, with some showing much better results.

Both false positives and false negatives account for the remainder. Further refinement of the heuristics or annotating trucks with their general area of operation could yield substantially improved accuracy.



Acknowledgements

This project would not have been possible without the generous support and guidance of our mentors, Setso Metodi and Nehal Desai. We also wish to thank our instructors, Chandra Krintz, Tim Sherwood and Stratos Dimopoulos.



Tweets are posted by food trucks describing their **location**

GeoSkynet **listens** for incoming tweets in **real time**

Tweet text is **analyzed** for **geostrings** by a series of **heuristics**

Locations are **geocoded** to be **transformed** from text into **coordinates**

Coordinates and **tweets** are **written** to the **database**

Nearby food trucks are **queried**, filtered and displayed to the **user**