Vision Statement

Problem Statement

With the explosion of user-generated content being produced in recent years as a result of the expanding popularity of social networks and online communication, the amount of attainable knowledge increases every day. At the same time, however, tools to organize and understand this data have not kept pace with its rate of expansion. As a result, the semantic meaning of this data is for the most part accessible only to people.

This situation presents an opportunity for software systems to be developed to make sense of, organize and interpret the data being generated every day by billions of people, and it is this problem that we aim to solve by focusing on the intriguing subproblem of geospatial location data. While some messages, tweets, comments and posts include computer-readable location data, the vast majority do not, even if the textual body contains words or phrases that to a human observer clearly identify a location. By developing a system to extract and analyze location information from plaintext messages, worlds of possible applications are opened – everything from tracking the spread of infectious diseases, identifying trends in real estate and urbanization, putting people in touch with others nearby, and countless more possibilities that depend on a better understanding of location information.

System Goals

We hope to create a system that will be able to take as input unformatted, plaintext, natural language messages from different sources and, in real time, extract meaningful and computer readable geospatial data from the input.

Data Extraction

The first step of the system will be to extract important information out of individual strings of text from social networking sites, blogs, and other websites. Not all text available has useful information, and to store every single character on these massive websites into a single database would require an exorbitant amount of storage space. One of the most difficult tasks of the system will be to parse through text made in ambiguous sentences and various colloquialisms to identify locations, items of interest, dates, and other important information while ignoring unimportant information.

Categorization and Creation of Metadata

After extracting information from the web, the next step will be to organize and store geospatial location, subject, and keyword metadata. Each of this pieces of information will be categorized and tagged based on their source, date, time, topic, and other properties.

Identifying Patterns

By juxtaposing data with similar features, the system will be able to identify and predict the trends.

System Features

Web-Based Interface

Users will be able to view the collected data in raw and organized forms without having to search through the raw databases.

Data Querying and Reporting

Information will be available in multiple formats. The user will be allowed to specify how he or she wants to view the data. A search feature will also allow the ability drill-down into different categories and filter data based on specified tags, dates, and other criteria.

Other Requirements and Constraints

Scalability

As the number and complexity of target data and data sources increase, the efficiency of the extraction and analysis processes should not be affected. Users should be allowed to operate on small datasets in the same manner and time that they would for much larger datasets.

Efficiency

Because the amount of information to be collected and analyzed is virtually endless, the system's algorithms must be designed in an intelligent manner. Simple brute force algorithms would take far too long to execute over such massive amounts of text. In order for the system to be useful, what it offers needs to be accessible quickly. As more and more information is poured into the web, the system should refresh efficiently and in real time.

Technologies

Programming Language: Python
Database: node.JS, MongoDB
Continuous integration: Jenkins
Source Control: Git
Message Broker: RabbitMQ