OPEN SCIENCE DATA CLOUD

Name: Keval Shah
University: University of Chicago
Adviser: Kyoungsook Kim
University: AIST, Tsukuba, Japan

Visual Data Mining using Social Media data

Abstract

Social media produces large amount of user generated content. Twitter has become a rich resource for analysis, producing a average of 58 million tweets per day. This unstructured data contains a lot of information about the social aspects of it's users. Researchers have become interested in using this data to structure, cluster and find any pattern in the data.

Problem Statement

Social media data contains heterogeneous data types like relationships, time, text and spatial-temporal information, which poses interesting and challenging tasks for visual data mining. The heterogeneous nature of the data opens up numerous possibilities of understanding real world phenomena. In order to harness this opportunity, we need to develop a data visualization tool that encompasses geo-spatial, time and topic data and enables finding relationships and patterns among the data.

Topic modeling

Text mining using R

Text Mining Steps:

Create a corpus.

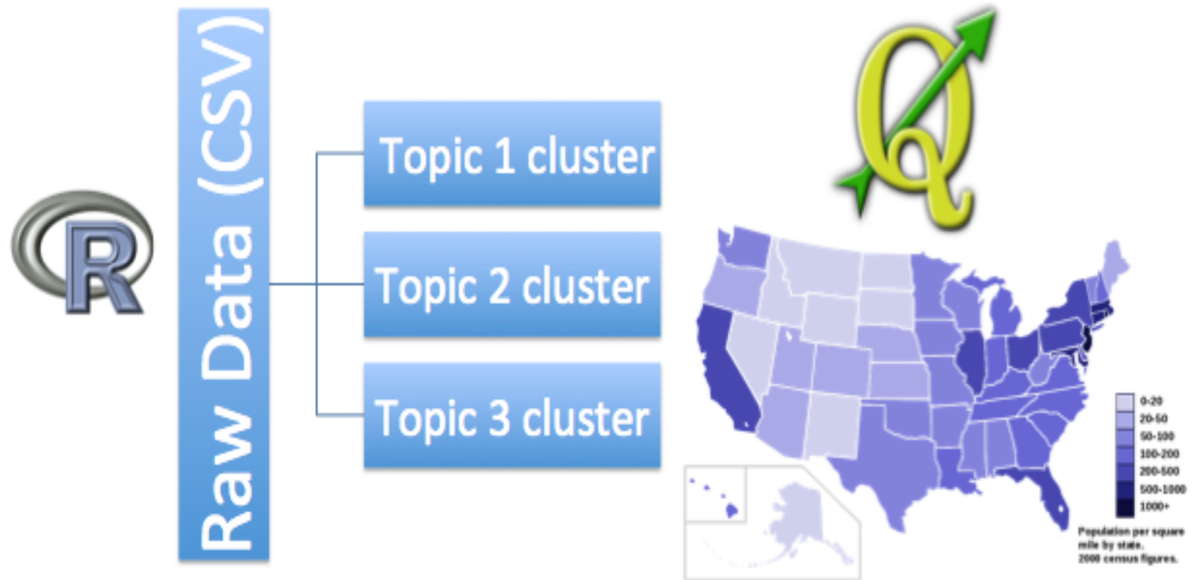
Create a term document matrix.

Extract high frequency keywords.

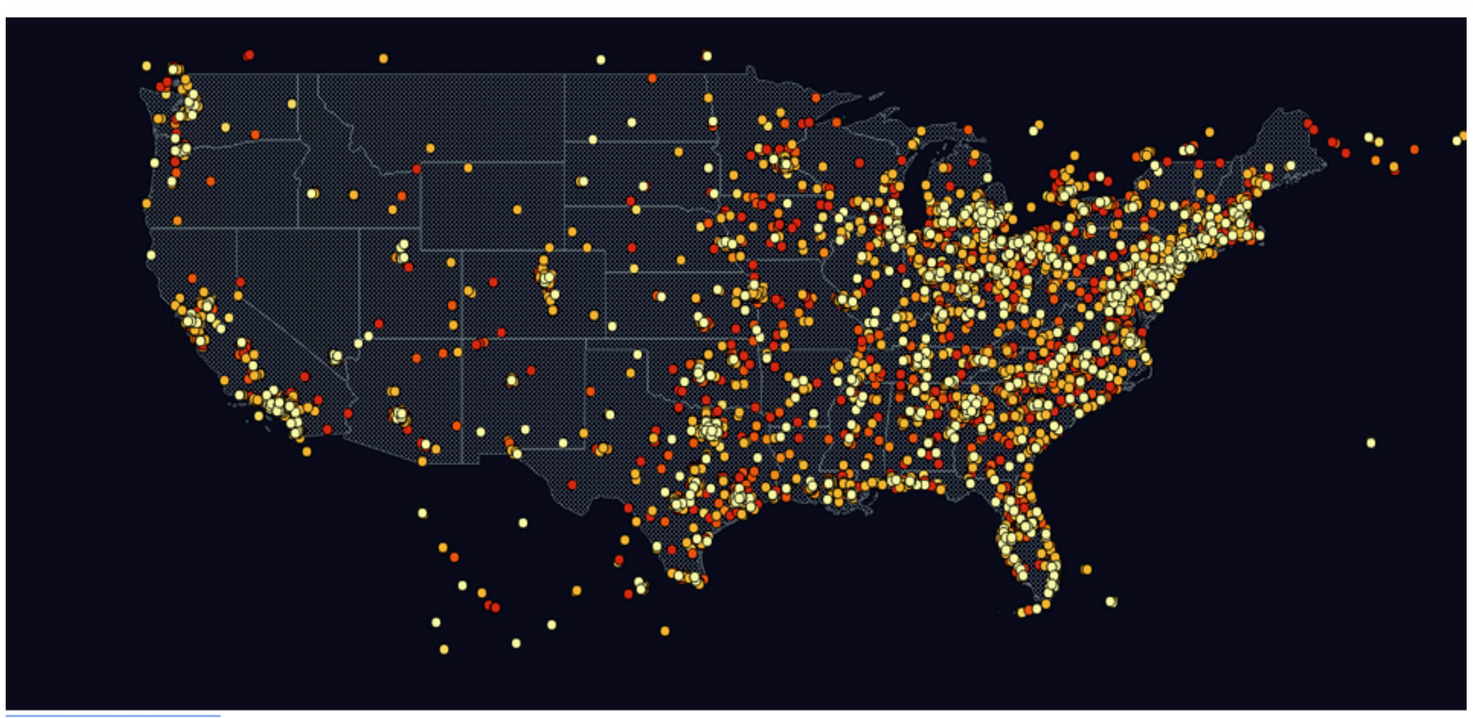
Cluster data based on keywords. Find tweets text that match the keyword.

Assign topics to each row in the cluster. (Data preparation)

Pipeline



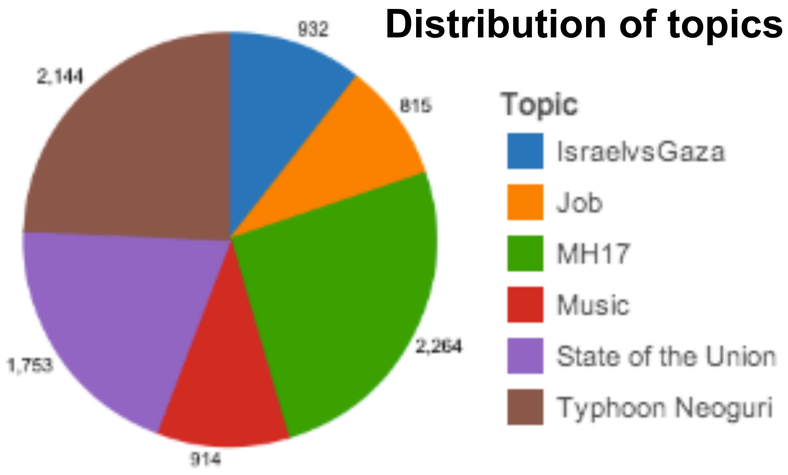
Visualization



The version I of the tool plots geo-coordinates from the dataset on to the U.S States shape file. *Qgis* merges the geo-encoded shape file and imports the geo-coordinates from the dataset. The tool enables visualization of heterogeneous data and finding patterns based on location and topic clusters.

The proposed visualization enables data discovery using visual data mining techniques, not only answers questions or verify hypothesis but also raises intriguing questions. Also, if there are any relationships between users in different geographies or any patterns seen over a time interval. The tool creates an interesting picture for understanding real world phenomena and taking the discoveries further into deeper analysis.

Distribution of topics



References

- https://www.census.gov/geo/maps-data/data/tiger.html
- Kumar, Shamanth, Morstatter, Fred, and Huan Liu. *Twitter Data Analytics*. Springer 2013.

Dataset

Our raw dataset contains numerous attributes regarding the tweet. It includes both user-generated content such as tweet text, hastags etc and metadata about the tweet such as language, geo-coordinates, time. For the purpose of our analysis, we're interested in text, geo-coordinates, time information from the raw twitter data.

Future Work

Add user interaction –
1. Filter by location
2. Filter by time
3. Filter by topic

- Fetch twitter data using Streaming API
- Compute measures such as degree centrality, Eigen vector centrality, network analysis.
- Sentiment analysis
- Visualizing in 3 Dimensional Space.