INFO 290T - Data Mining Final Project Proposal

Team members

- Irina Lozhkina, Janine Heiser, Xavier Malina

Project Goals

In the past five years, information available online has bombarded users in greater and greater quantities. As more users adopt social media, the rate at which information is shared has increased. This increased velocity in information sharing is changing social norms; when an specific event happens social network users can discuss the event with one another in real time. The goal of this project is to understand how policy and cultural effects percolate through social networks, as well as understand how information diffuses across geography. Another goal of this project is to understand rate and subjects that go viral across social networks.

Project Data:

We plan to use the trending data generated by scraping web for trending items on sites like Twitter, Google, Instagram, Facebook, Yahoo and Youtube. Using an ETL application that runs on an amazon ec2 server, Janine was able to collect and organize this trend data in real time. The ETL consists of a several of automated crontab jobs that set off a a series of automated OO-ruby scripts that make API calls or scrapes trend data from different social networking websites. The automation allows internet trend data to be collected 24/7. Each data item collected has attributes such as the subject of the trend, the number of views/likes, searches, latitude/longitude coordinates, and the time when the trend trend appeared. The chart below details the specific data sources used and the data that is collected.

Social Network	Data Location	Data Collected
Twitter	Twitter API resource, GET trends/place	• trend location, trend title, time of trend, href link to trend on twitter • data available for 412 different geographic locations
Google	http://www.google.com/trends/hottrends	• trend location,trend title, trend rank, search count, link to trend image, href link to website about trend, time of trend • data available for 40+ countries
Youtube	http://www.youtube.com/trendsdashboard	• trend location, trend title, number of views, link to trending video, time of trend • data available for 40+ countries
Instagram	Instagram API resource, GET Most Popular	• trend location, trend/photo caption, trend/photo tags, href link to trending photo, likes count, time of trend • data available for a variety of lat/long coordinates
Facebook	www.facebook.com facebook homepage (located in the upper right corner of the user's homepage	• trend rank, trend title, description of trend, href link to photo of the trend, href link to trend news story, time of trend • specific location data not available
Yahoo	www.yahoo.com (trends located on right side of page)	• trend rank, trend title, time of trend • specific location data not available

Project questions, assumptions

We will try to answer the following questions:

- How similar or dissimilar are trending terms and topics on each of the platforms (Google, Facebook, YouTube, Instagram, Twitter)? Are there matching trends?
- Depending on the geographical location, are users searching for similar terms and topics? Is there a time lag between locations? What locations have similar preferences?
- Do topics/terms remain in the top 10 across platforms, for similar time frames? Do some platforms trend together?
- Based on the geographical locations, we would like to check if people from the same continent have similar tastes.

We are planning to use explore the following datamining techniques:

- cosine similarity between trend pattern ‘fingerprints’ → we will write a mr. job job to do this.
- -fuzzy clustering of trend types