#Conversations

Industry Sponsor: Twitter

Industry Mentor: Roja Bandari

Introduction

Twitter is regularly used to find and post information on what happens in real-time in the world. When real-time events involve an organization, entity, or vendor, users often interact with the Twitter account of the related organization to ask questions, provide information or feedback, or to get help. For example, you may:

- have a question about a delay in the train schedule or road closure,
- want to give a shout-out to a great food truck,
- want to complain about a cancelled flight,
- want to donate to emergency relief efforts for a recent earthquake,

We can easily see anecdotal examples of this interaction (e.g. here) but we still need accurate measurements of how often this happens as well as deeper analysis. We would specifically like to focus on studying conversations that occur between users and organizations or entities on Twitter. This is important to Twitter because we want people to have a live direct channel to entities and organizations that are relevant to them quickly in the moment that they need it most. What we learn from this project will help Twitter develop features and capabilities most useful to our users.

Below are some of the specific types of analysis we would like to see. Note that answering many of the following questions will not only require technical skills, but also requires thought and intuition about what would be more interesting and useful to learn; based on this intuition you have the freedom to modify what you measure as you make progress.

Characterize conversations with organizations and entities through study of:

a. Distribution of **length** of each conversation (number of Tweets)
b. Distribution of number of conversations initiated/ended by the user vs. the organization
c. **Timeline** of such conversations (Tweets per day) over the duration of the data
d. Distribution of **number of unique users** engaged in each conversation
e. Distribution of **number of followees** in an org's overall conversations (if possible)
f. **Frequency** of these conversations (per organization account and per user account)

g. **Content analysis**
    i. **Initiation and Resolution**: (first and last Tweets in a conversation)
        1. Distributions of the **length** (number of words) for first and last Tweets
        2. **Topic Modeling** and **term frequencies** over aggregate of all initial/final Tweets (aggregated per organization)
        3. Counts of different **hashtags** used in first and last Tweets. What share of the conversation are **questions**, etc.?
        4. More open-ended questions can be explored: Is there **closure**? Does the conversation end with "call us" "dm us" "follow us"?
        5. **Sentiment** of first and last tweets / user and org tweets
h. **Cluster** orgs based on conversations (open-ended and needs brainstorming- time permitting).

Special Requirements

You will need to work with large datasets, including distributed processing tools such as hadoop (or other tools you are familiar with) to process data, extract relevant conversations, group by each organization's account, and compute various desired counts. Some familiarity with computational/statistical tools such as R or Matlab, and ability to process text (using python is easiest) is also required. Finally, you need to browse the conversations of some relevant accounts on Twitter to become familiar with what you are studying.

Data will be provided (most probably) using Gnip. Here is a Gnip data payload example: http://support.gnip.com/sources/twitter/data_format.html#SamplePayloads

Expectations

We would like a full report of results including: Graphs of distributions along with output files with counts of each measured statistic, word frequencies, topic term probability vectors as well as topic probability mixtures per organization account, cluster labels for organization accounts (if clustering is performed).

Recommended Reading

- Review statistics (stdev, confidence intervals, etc), a quick helpful review is videos on khan academy here, especially the first three segments: https://www.khanacademy.org/math/probability/statistics-inferential
- Topic modeling: sections 1 and 2 of https://www.cs.princeton.edu/~blei/papers/BleiLafferty2009.pdf
    - Mallet: http://mallet.cs.umass.edu/ (alternative topic modeling tools are also ok)
- K-means clustering: http://en.wikipedia.org/wiki/K-means_clustering (alternative clustering methods can be used as well)