Module 1 – Natural Language Processing

One-Pager: towards your final assignment

Please read and think about this one-pager before arriving in class for our second session.

We will work on this one-pager during session #2.

1. We will be applying natural language processing methods to user-generated content (UGC)

2. UGC includes online reviews, amazon reviews, blogs, Facebook posts, tweets and others.

    1. In general, UGC includes any kind of content that is spontaneously created by consumers.

    2. UGC reproduces a consumer's voice in its purest form.

amazon

Tripadvisor

twitter

What kind of research questions (RQ) can benefit from UGC?

- UGC can be used to construct independent variables, dependent variables, or both.
  - These variables can then be used in models (e.g., time series, Poisson regression, Anova, etc) to answer interesting questions.

  - I will push you to be ambitious regarding your RQ
    - A few examples of ambitious research questions:
      - What is the effect of tweets on movie box office?
      - Does online chatter affect the stock market ?
      - What number of tweets makes a video (or a product) viral?

Note that there are several restrictions regarding the use of UGC for inference, so extrapolations from the sample data to the population of interest requires careful thinking about selection issues. We will discuss them in class.

Even a 1,000-mile journey starts with a single step…

(Not-Graded) One-pager for the final assignment

Option A: using the movie review dataset from the module 1 Github

- https://github.com/guiliberali/Learning-from-Big-Data-Module-1
  - Reviews_tiny.csv: 1k reviews.
  - Reviews_short.csv: 10k reviews
  - Reviews_medium.csv: 60k reviews (upon request)

Option B: choose your own set of tweets, blogs or reviews

- Choose a domain. Examples: hotels, laptops, massage, rock concerts, songs, chocolate, makeup, your family line of business, your favorite sport, your internship, etc.

Questions

1. Dataset (option A or option B; if option B, please provide the URL)
2. What questions do you want to answer with these data? Why?
3. Time frame of the data you will use and form of aggregation
4. Topics (in option A, 3 topics are given)
5. Training data (in option A, training data provided)