Content Analytics

Dr. Sumeet Gupta

Outline

- Using Text as Data
- Conducting Various Analysis

Using Text as Data

- Until now, our data has typically been
  - Structured
  - Numerical
  - Categorical
- Tweets are
  - Loosely Structured
  - Textual
  - Poor spelling, non-traditional grammar
  - Multilingual

Text Analytics

- People care about textual data, but how do we handle it?
- Humans can't keep up with Internet-scales volumes of data
  - ~ 500 millions tweets per day
- Even at a small scale, the cost and time required may be prohibitive

How can Computers Help?

- Computers need to understand text
  - This field is called Natural Language Processing
  - The goal is to understand and derive meaning from human language
- In 1950, Alan Turing proposes a test of machine intelligence: passes if it can take part in a real-time conversation and cannot be distinguished from a human

History of Natural Language Processing

- Some progress: "chatterbots" like ELIZA
- Initial focus on understanding grammar
- Focus shifting now towards statistical, machine learning techniques that learn from large bodies of text
- Modern "artificial intelligences": Apple's Siri and Google Now

Why is it Hard?

- Computers need to understand text
- Ambiguity
  - "I put my bag in the car. It is large and blue"
  - "It" = bag? "It" = car?
- Context:
  - Homonyms, metaphors
  - Sarcasm

Sentiment Analysis - Example

- Apple is a computer company known for its laptops, phones, tablets, and personal media players
- Large numbers of fans, large number of "haters"
- Apple wants to monitor how people feel about them over time, and how people receive new announcements.
- Challenge: Can we correctly classify tweets as being negative, positive, or neither about Apple?

Sentiment Analysis - Example

- Emails contain large amount of unstructured data
- Can we analyze emails for identifying the possibility of fraud?

Sentiment Analysis - Procedure

- Creating the Dataset
  - E.g., Twitter Data
- Need to construct the outcome variable for Tweets
  - Can be done automatically using dictionaries
  - Or need to manually code the tweets (e.g., Amazon Mechanical Turk)
- A Bag of Words

Sentiment Analysis - Procedure

- Cleaning up Irregularities
  - Text data often has many inconsistencies that will cause algorithms trouble
  - Apple, APPLE and ApPLE will be counted differently
  - Convert all these words to lower/upper case
  - Remove Punctuation
  - Removing unhelpful terms (such as the, is at, which etc.)
  - Stemming (removing the endings of most words)

Sentiment Analysis - Procedure

- A Bag of Words approach
  - Fully understanding text is difficult
  - Similar Approach
    - Count the number of times each word appears in a line
    - "This course is great. I would recommend this course to my friends."
  - Remove sparse terms
  - Build a data frame
- Perform Data Mining Tasks
  - Association Mining
  - Classification Modeling
  - Clustering
  - Prediction etc.

Downloading Twitter Data

- Login into your Twitter Account
- Go to http://apps.twitter.com
- Make a new app
- Give Name, Description and Placeholder URL
- Leave the callback URL blank
- Use the 'R' Code to handshake with Twitter and Download Tweets