Text Mining / NLP on Twitter

...

2017-02-11 at McCoy College of Business Administration

Mario Ezekiel Hernandez, mario@m-ezekiel.com

Information retrieval & domain ontologies

- Dewey Decimal System
  - Used in 200,000 libraries across 135 countries
  - Criticism: it is a static ontology based on Eurocentric academic disciplines and whose content is limited by the storage, publication and access resources.
- Digital content has far exceeded static classification systems in size and diversity.
  - Blogs, tweets, product reviews, images, emoticons, etc.
- Fundamental 21st century computational problem:
  - How do we classify and organize digital content for human retrieval and consumption?

Good news: Humans are experts at understanding language.

Bad news: Computers are bad at it.

Linguistic alchemy: turning words into numbers

Basic text transformations for document classification:

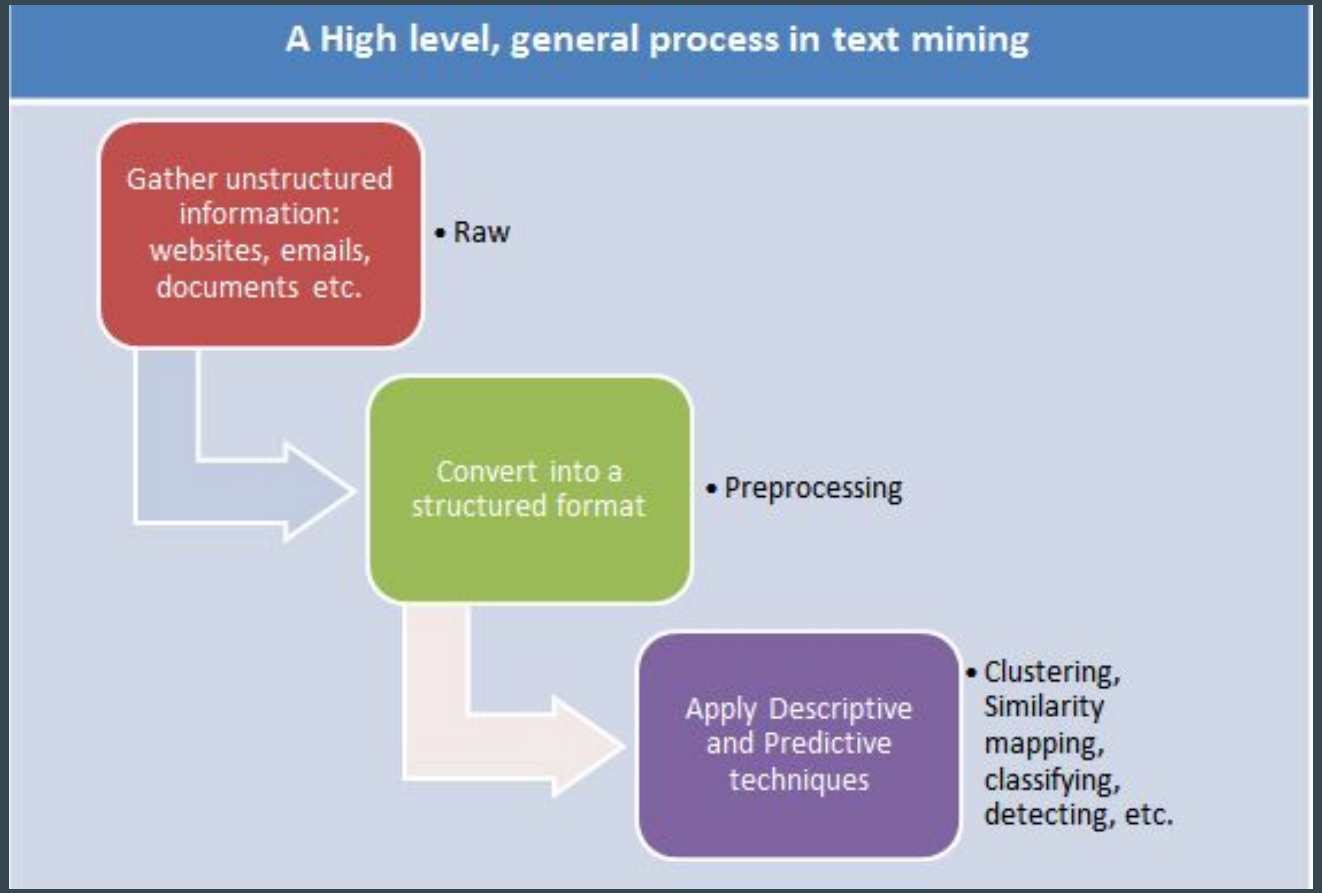
- Tokenization
  - Splitting text corpora into word elements for lexical analysis.
- Word form normalization
  - Stemming, lemmatization, unified case, alternate spellings, punctuation removal.
- Stopword removal
  - Removal of words which lack direct information about the content of a work.
  - Zipf's law: the frequency of any word is inversely proportional to its rank in the frequency table.
  - Can be problematic when trying to extract emotional tone (see Pennebaker)

**Regular Expressions are extremely helpful when developing corpus-specific heuristics.
http://regexr.com

Linguistic alchemy: turning numbers into insights

- Lexical analysis / word frequencies
  - High level overview of document content and keyword extraction
  - Algorithms: Word2Vec, Latent Dirichlet allocation, TF-IDF
- Syntactic parsing
  - Part of speech tagging and syntax trees for sentiment and named-entity extraction
  - http://cogcomp.cs.illinois.edu/page/demo_view/pos
- Word-sense disambiguation
  - Based on probabilistic model of word collocation and topic clusters
  - Open problem in computational linguistics- computers still suck at this.
- Typing dynamics & keystroke biometrics
  - Keystroke dynamics and syntactic parsing (academic paper Oct. 2016)
  - Journal of Writing Research

A brief overview of TM workflow

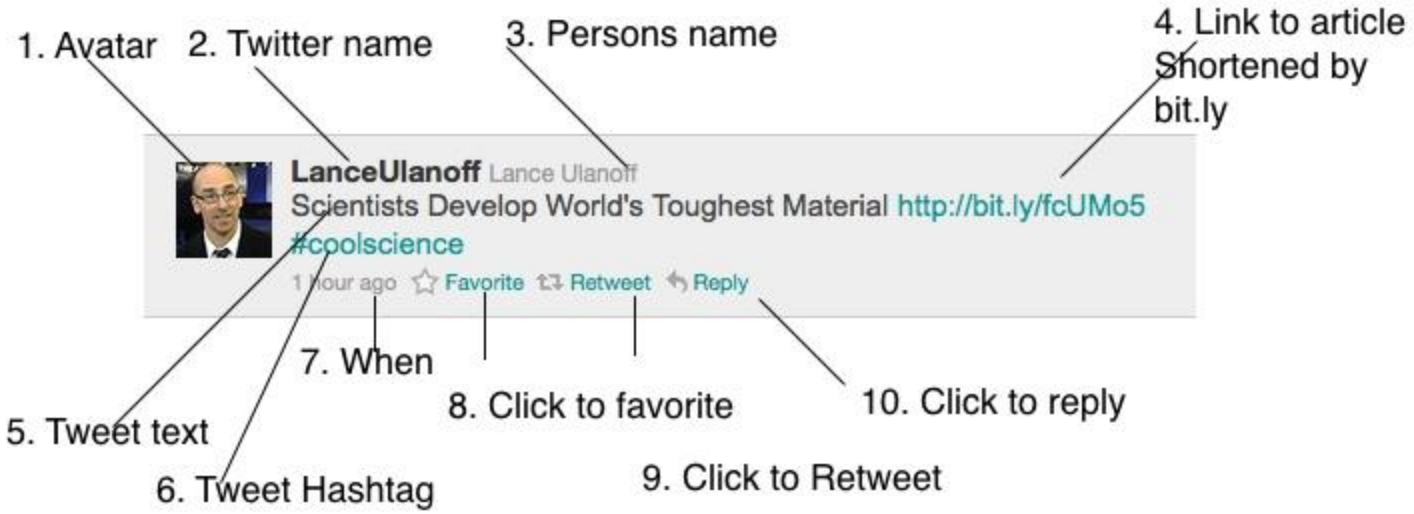


Twitter: a near-perfect text mining playground

- A cheap, abundant, and highly structured real-time data source
  - Large # of observations is like the Rolls Royce of experimental design*
  - Most tweets are pre-tagged with topic hashtags and interaction symbols (@...)
- R-packages: twitteR, tm, wordcloud, stringr, dplyr
  - twitteR requires the whole OAUTH dance and developer account privileges.
  - tm uses the TF-IDF algorithm
  - wordcloud makes visualizations based on token frequencies
  - Stringr provides functions for string operations
  - dplyr for dataframe manipulation (from developer/celebrity Hadley Wickham)
  - tidytext for sentiment analysis

**A note on experimental design: ALWAYS be aware of biases in your sampling method.

Tweet Anatomy 101



What are we looking for in text data?

- Keywords / topics
  - Used in search optimization, recommendation engines
- Sentiment / emotional tone
  - Used in assessing product performance, or informal opinion polling in politics.
  - Stanford NLP Sentiment Analyzer
- Psycholinguistic indicators of personality or cultural identity
  - Used in psychological profiling for marketing or open-source intelligence.
  - Ex. MTBI personality types (controversial)
- Language identification
  - Including regional dialects and other idiomatic subtypes.

twitteR package demo

- Basic profile analysis

  https://github.com/m-ezekiel/twitterProfileAnalysis
  - Use # and @ as symbolic markers for text processing.
- Hashtag collocation

  https://github.com/m-ezekiel/twitterProfileAnalysis/blob/master/hashtags_fXn.R
  - Parse hashtags and reduce dimensionality by displaying only unique entries.
- Mining follower data for potential clients

  https://github.com/m-ezekiel/textMining_sketches/blob/master/mining_followerData_sketch.R
  - Personal pronouns in acct. description may be an indicator of non-commercial human operator.
- Assessing language usage in organizational accounts

  https://github.com/m-ezekiel/textMining_sketches/blob/master/ivyTweets_linguistic_analysis_sketch.R
  - Tokenizing content and running queries across aggregated databases.
- Markov-based language generator

  https://github.com/m-ezekiel/textMining_sketches/blob/master/markov_Fxn_and_corpusToVector.R
  - Demonstration of syntax-naive probabilistic computer-generated “speech”

Text Mining and Account Security

- Account access requires at least two things: account name and password.
- Account names
  - Social footprints are all over the web and people tend to reuse account names.
  - People are particularly in love with the firstname.lastname@mail.com pattern.
- Passwords
  - Humans are incredibly lazy and will do everything possible to avoid truly random and long passwords/passphrases. Even “made-up language” human-created passwords often obey phonological patterns such as CVC.
  - People commonly use num3r1c or ch@r@cter substitutions, or simply append numbers to names or dictionary words. These are all terrible and easily exploitable practices.
- Email is the gold standard for account escalation. Most of us try to keep it private.
  - Account escalation is gaining access to a low security account in order use those credentials to gain further access through account resets (which often only require email access).

[image set removed]

...but it was about reverse engineering private account information based on publicly available data.

Text Mining and Account Security

- Once again, email is the gold standard of account(s) access.
- Solutions
  - Two-factor authentication
  - Add linguistic noise: first.last+something@email.com or f.i.r.s.t.last@email.com
  - Don't reuse passwords or account names
  - Avoid sending sensitive information through insecure channels
  - Use end-to-end encrypted messaging services such as Signal, Wire, or Protonmail.
  - Learn about security

Additional Resources and Case Studies

- SPAM detection and evasion
  - Data poisoning and trend manipulation
- Polygraph: Rappers sorted by vocabulary
- Trump tweets vs. staff tweets
- Detecting alzheimer's / cognitive decline with lexical analysis
- Health and emergency alerts, epidemiology
  - Fundamental problem of technological access which biases ALL twitter data
- Hypothesis induction in scientific literature
- Typing dynamics and self-censorship at Facebook
- Finding the most depressing Radiohead song w/sentiment analysis

Questions

Markov Language Generator

- Probabilistic language model which is devoid of syntactic or semantic context.
- The degree and order of the model dictate how much it may deviate from the training corpus.
  - Overfitting yields verbatim training text
  - Simple models yield high novelty but are the semantic equivalent of selecting nearly random word combinations.
- Advanced models use conditional probabilities to generate more natural-sounding text (higher degree).
- Used extensively in chatbots.