Text Mining with Python

Will Stanton: Data Science and Business Analytics Meetup
Feb 26, 2014
Return Path

- Worldwide leader in email intelligence
- Collect and aggregate enormous amounts of email data, including *raw text data*
- Help receivers improve spam filtering with whitelists, blacklist, reputation scoring
- Help senders improve their email sending program
Raw text data

- Unstructured (not in a row-column table form), essentially infinite-dimensional
- Enormous amount of text on the web
  - hundreds of billions of emails sent per day
  - forum posts, articles, even webpage HTML code
What is text mining?

- Uncovering patterns and relationships in text
- Building statistical or machine learning models using text data
  - classification, clustering, predictive models
- Extracting information from text
  - sentiment, subject
Example use-cases

- **Spam detection**
  - Which phrases, subject lines, etc. indicate a spam email?

- **Search**
  - What webpage most closely matches the true meaning of search terms?

- **Literary studies: author identification**
  - Did Shakespeare really write Hamlet?
Example use-cases (cont.)

- **Machine translation**
  - Identifying context: different meanings of same word, “bank on” vs. “bank with” (polysemy)
- **Customer service**
  - Which service request is most urgent?
- **Legal discovery**
  - Which documents are most likely to contain relevant info?
Why Python?

- Python (python.org) is an interpreted, general-purpose programming language
- Readable code
- List comprehensions
- Great data/text mining/presentation libraries (*pandas*, *sci-py*, *sci-kit learn*, *gensim*, *nltk*, *ipython*, *matplotlib*)
Exploratory text mining with *nltk*

See html_explore_presentation_final
LSA with sci-kit learn

see lsa_presentation_final