Making Sense of Web Data with Natural Language Processing

Cork Big Data & Analytics, 2017-11-13

Image: https://markovikj.com/assets/img/wclouds/research.png

About Me

- Johannes Ahlmann
- fluquid.com
  - Sales & Client Intelligence
  - Intelligent Lead Generation
  - Large-scale web crawls
  - Gathering and Enriching Web Data
- webdata.org
  - Share Libraries and Best Practices
  - Bring Data Scientists and SME Companies together
  - ForDevelopers
  - AwesomeAvailableDatasets
- Contact:
  johannes@fluquid.com

Data is Noisy

- Data is noisy (typos, free text, etc.) ("Mnuich", " Munich", "munich")
- Data can vary syntactically ("12.00", 12.00, 12)
- Many ways to represent the same entity ("Munich", "München", "Muenchen", "Munique", "48.1351° N, 11.5820° E", "zip 80331–81929", "[ˈmʏnçn̩]", "Minga", "慕尼黑")
- Entity representations are ambiguous

  <Munich City, Germany>

  <Munich County, Germany>

  <Munich, North Dakota>
- Wikipedia disambiguation

Natural Language Processing

1. Content Extraction
2. Parsing
3. Named Entity Extraction,
4. Topic Modelling
5. Sentiment Analysis

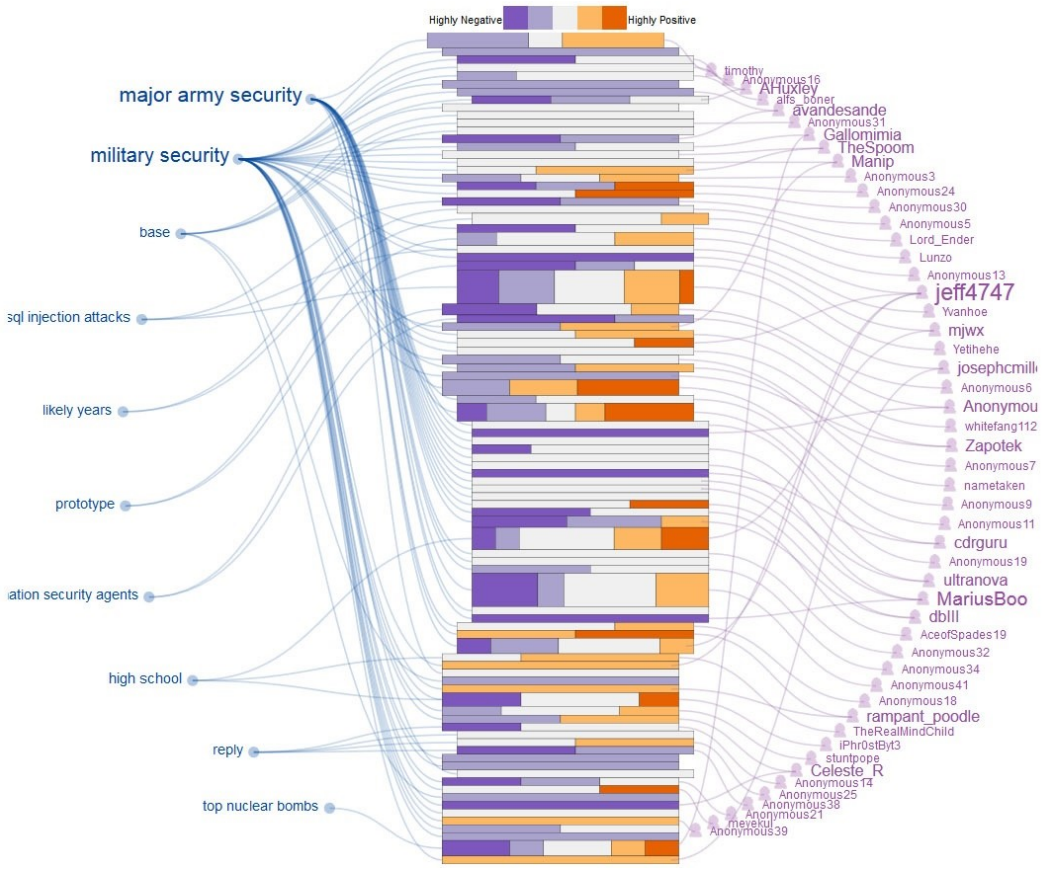


Image: http://www.cs.ubc.ca/cs-research/lci/research-groups/natural-language-processing/image/convis/3.jpg

fluquid

1) Content Extraction

- Challenge:
  Given a document,
  extract the main text information
  as plaintext
- Libraries
  - html-text
  - boilerpipe (java)
  - dragnet
  - apache tika (java; supports many formats)
- Example - Readability

Image: http://webdata-scraping.com/media/2016/04/web_scraping_spider.png

fluquid

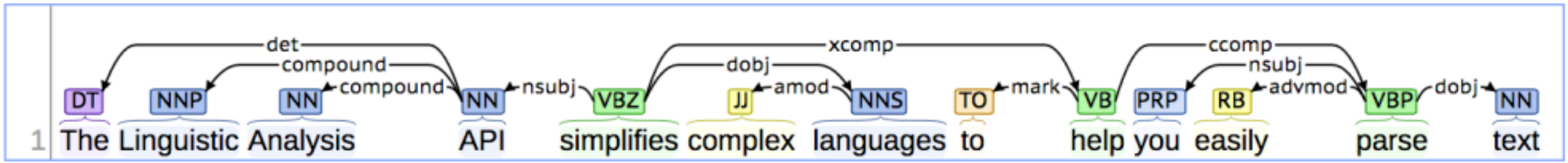
2) Parsing

- Spacy 2 is awesome!
  - Sentence segmentation
  - Word segmentation
  - Lemmatization/stemming
  - Parsing
  - POS (part of speech)
  - Word vectors
  - Word/sentence similarity
  - etc.

- Textacy
  - Extends spacy functionality
- syntaxnet
  - Parser and language understanding engine developed by Google
  - For more advanced use cases

Enhanced Dependencies:

Image:https://stanfordnlp.github.io/CoreNLP/images/Cate-Blanchett.png



3) Named Entity Extraction

- Entities:
  persons, organizations, locations, date, time, money, email, social media, postal address, etc.
- NER, Disambiguation
  - spacy - basic entity extraction
  - stanbol - pretty good for "production use"
  - dbpedia spotlight - between stanbol and AIDA
  - AIDA - very good, but slow
- Normalization
  - cleanco - companies
  - probablepeople - person names
  - python-phonenumbers - international phone numbers
  - libpostal - postal addresses
- webstruct - train your own NER with annotated training data

Image: https://pbs.twimg.com/media/Ct_oP9AXYAExsNq.jpg

fluquid

4) Topic Modelling

- Goal: Dimensionality Reduction from 50k+-dimensional token space to "topic" manifold
- Assumption: Every document covers several different "topics"
- A topic is comprised of words that often co-occur
- Approach: Analyze which words co-occur more frequently with each other than with other words
- Can be used as a basis for clustering, similarity, etc.
- Libraries
  - gensim LDA
  - sklearn NMF
- Demo

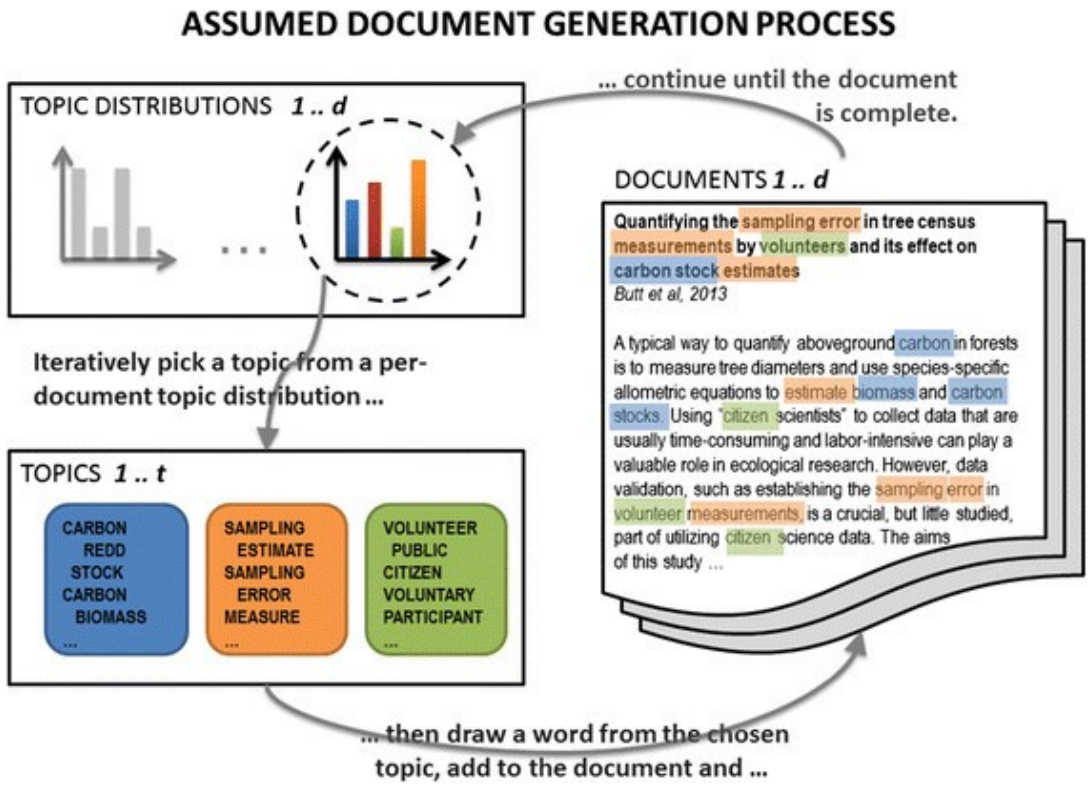


Image: http://bit.ly/2A0hbcA

fluquid

5) Sentiment Analysis

- Identify what sentiment an expression carries
  - Polarity, Subjectivity
  - Paragraph, Sentence, Entity
- Challenges:
  - Generally messy and often does not produce great results
  - Sarcasm, Irony, Context
  - Mixed sentiments in any single statement
- Libraries
  - vaderSentiment
  - twitter-sent-dnn
- Examples
  - cryptocurrencies
  - twitter "performance review" tweets

Image: https://thumbs.dreamstime.com/t/reaction-smileys-vector-clip-art-30534441.jpg

fluquid

Metadata

- Use pre-structured information from web data where available
- Formats
  - Metadata (schema.org)
  - Microdata (vcard)
  - json-ld
  - OpenGraph
  - Twitter Card
- Libraries
  - Extruct
  - Apache Any23 (java)

Image: https://i2.wp.com/blog.parse.ly/wp-content/uploads/2015/08/Metadata-Tags-Use.jpg

fluquid

Miscellaneous

- Language Detection
  - cld2-cffi
- Find many possible terms in text
  - pyahocorasick
- Structured Data Extraction
  - Pydepta
  - Demo
- Unicode Normalization
  - unidecode

Image: http://windows.ischool.syr.edu/wp-content/uploads/2009/06/visit-with-clare-gail-008.jpg

fluquid

Questions?

- Content Extraction in R
  - boilerpipeR
- Wordpress Plugin Scanner
  - sorry, it's not open-source yet; but I will open-source it soon at github.com/fluquid
- Extract Bibliography from Academic Papers
  - grobid (GeneRation Of BIbliographic Data)
  - pdfextract
  - CERMINE
- Find similar skills, capabilities
  - gensim word2vec
  - spacy even comes with semantic sentence similarity ;)