Natural Language Processing

Data Skills for Empirical Research

Winter, 2020

NORTHWESTERN UNIVERSITY

Kellogg School of Management

Dealing with Unstructured Data

Vast quantities of information are encoded as unstructured data, in the form of natural language text.

But it can be hard to make this information available for computational analysis at scale.

Information Extraction

As we learned in the regular expression workshop, regex are great tackling lots of information extraction problems.

Information Extraction

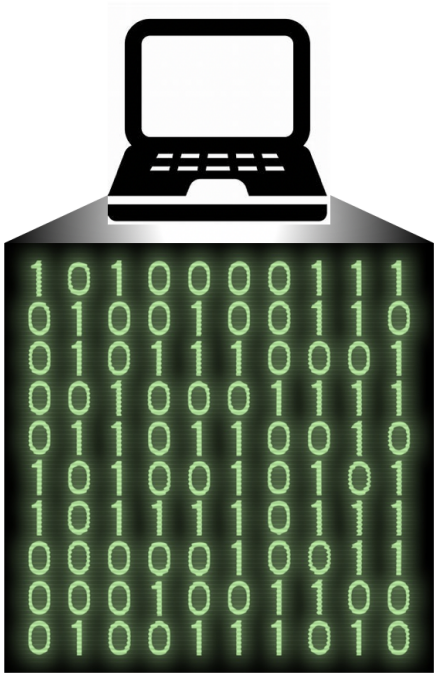
But...

Information Extraction

Google just made another giant move in its Silicon Valley land grab. The interr

NLP to the Rescue

A field concerned with the ability of a computer to understand, analyze, manipulate, and potentially generate human language



NLP to the Rescue

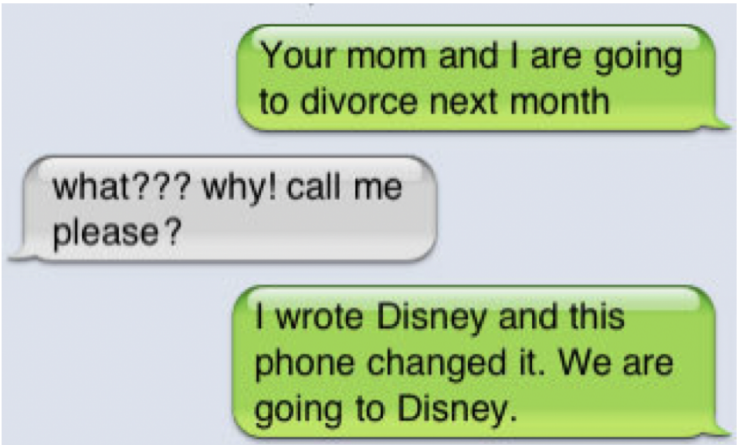
Autocomplete

natural language processing is

natural language processing issues
natural language processing is **fun**
natural language processing is
natural language processing is **divided into**
natural language processing is **machine learning**
natural language processing israel
natural language processing is **used for what**
natural language processing is **a field of quizlet**

Spam Filter

Autocorrect



NLP Application Types

Supervised Methods

- Information Extraction
  - Named Entities
  - Relations
  - Temporal Expressions
  - Events
- Entity Linking
- Text Classification, including Sentiment Analysis

2

Unsupervised Methods

- Text Feature Analysis
- Topic Modeling
- Language Models

NLP Tasks

Lower-Level Tasks

- Sentence splitting
- Tokenization
- Lemmatizing / stemming
- Removing “stop” words, punctuation
- Vectorizing documents
- Word embeddings

NLP Tasks

Higher-Level Tasks

- Part of Speech Tagging
- Word Sense Disambiguation
- Coreference Resolution
- Semantic Role Labeling
- Syntactic Parsing
- Dialogue Modeling

Resources

- Introductory textbook: https://web.stanford.edu/~jurafsky/slp3
- Python resources:
  - spaCy: https://spacy.io/
  - scikit-learn: https://scikit-learn.org/stable/tutorial/text_analytics/working_with_text_data.html
- R resources:
  - https://cran.r-project.org/web/views/NaturalLanguageProcessing.html
  - tidytext: https://cran.r-project.org/web/packages/tidytext/index.html
  - tm: https://cran.r-project.org/web/packages/tm/index.html