Cypher Workshop - NLP and Web Scraping

DSC at W&M, April 2021

What is web scraping?

- Programmatically collecting data from websites
- How does it differ from an API?
- When would we use it? Any concerns?
- Python libraries: Beautiful Soup, requests, and Newspaper3k

Image source: *parsehub.com*

Intro to NLP

- Drawing quantitative results from text data
- Examples of text data sources
- Methods: machine learning & deep learning, lexicon-based methods, statistical methods
- Word embeddings

The NLP Pipeline

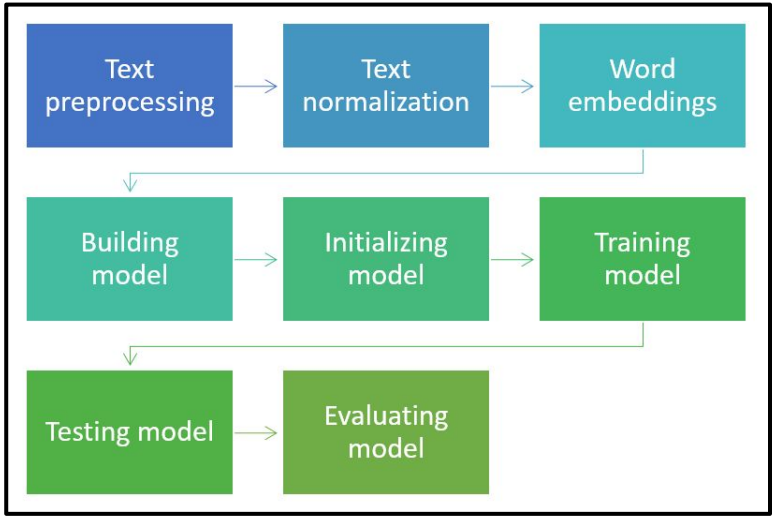

Image source: *towardsdatascience.com*

Common NLP tasks

- Named entity recognition
- Sentiment analysis and document classification
- Topic modeling
- Semantic similarity

Image source: *towardsdatascience.com*

NLP applications

- All of Google's search engine capabilities
- Autocorrect and word/phrase suggestions
- Geolocation using text
- Applied research

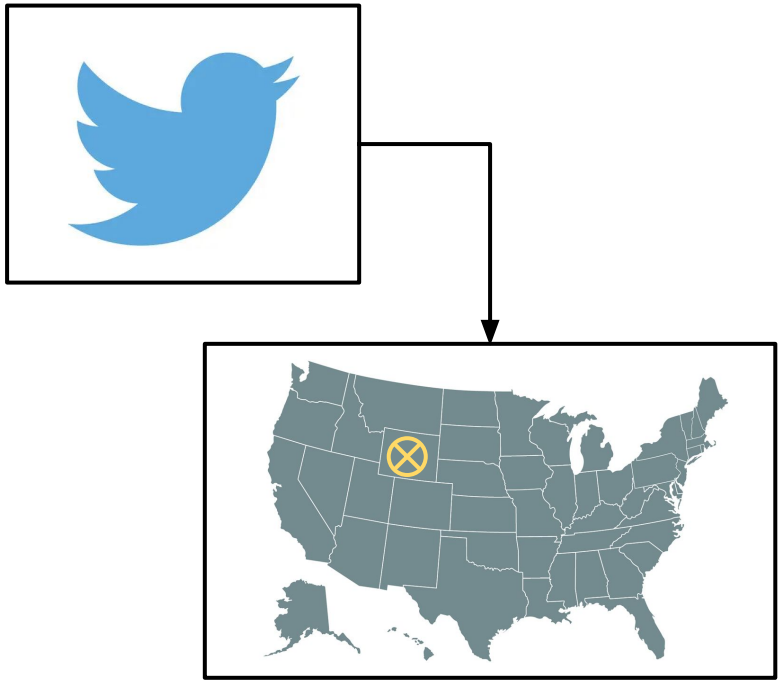

Image source: *hrc.org*, *twitter.com*

NLP tools in python

- Natural language toolkit (NLTK)
- spaCy
- Stanza
- Tensorflow
- Gensim

Image sources: *stanfordnlp.github.io/stanza*
tensorflow.org
spacy.io
clay-atlas.com

Further resources

- Our general website and GitHub site, for more workshops
- A hackathon project that we worked on - basic text processing in the wild!
- Links to NLP resources are included in the Kaggle notebook