Fundamentals of Natural Language Processing

A PRACTITIONER'S PERSPECTIVE

Andrew Therriault

JUNE 26, 2019

HARVARD UNIVERSITY

Let’s Get Started…

- My Background:
  - Academic: NYU Politics PhD (2011), Vanderbilt Post-Doc (2012)
  - Professional: Data Science Consultant (2012-2014); Director of Data Science, Democratic National Committee (2014-2016); Chief Data Officer, City of Boston (2016-2018); Infrastructure Data Science Manager, Facebook (2018-)
- What we’re doing today:
  - Discussing the basics of NLP and exploring its applications
  - Doing a deep dive into what NLP looks like in practice for unsupervised and supervised learning applications
  - Trying out NLP on your own data

An Introduction to NLP and its Applications

Some Very Basic Definitions

- **Natural Language Processing (NLP):** The use of computers to process linguistic information into structured data and extract useful insights
- **Document:** A single unit of observation used in NLP, which could be as short as a word or as long as a book
- **Corpus:** A collection of documents from a shared context that are processed and analyzed using NLP

Conceptual Framework

- Text *is* structured data, but lacks the simple structure needed for quantitative analysis and ML
- Extensive preprocessing and parsing is needed to convert plain text into structured quantities
- How to parse a given document depends upon the specific application and the corpus being used
- Once parsed, documents can be analyzed like other quantitative datasets or used in more complex applications

Common NLP Tools for Python Users

- Natural Language Toolkit, https://www.nltk.org
- scikit-learn, https://scikit-learn.org
- CoreNLP, https://stanfordnlp.github.io/CoreNLP/
- Gensim, https://radimrehurek.com/gensim/
- spaCy, https://spacy.io/
- AllenNLP, https://allennlp.org/

Basics of Text Preprocessing & Parsing

- Data ingestion from files, scraping, APIs
- Tokenizing documents into characters, words, n-grams, and sentences
- Word stemming and lemmatization
- Removing stop words
- Creating a “bag of words” model
- Parsing regular expressions and metadata

Advanced Document Parsing

- Text matching
- Word embeddings
- Named entity recognition
- Part-of-Speech Tagging
- Syntax parsing

Unsupervised Learning Applications

- Unsupervised learning aims to extract information from documents to derive insights about individual documents and the corpus as a whole
- Common use cases:
  - Document clustering
  - Topic modeling and extraction
  - Document and word similarity
- Unsupervised learning can be used on its own or as a precursor to supervised learning

Supervised Learning Applications

- Supervised learning uses parsed text as an input to an ML model for classification or regression
- Some very common examples:
  - Spam filtering
  - Topic classification
  - Sentiment analysis
- The model's predictions may tell us either about the documents themselves or about something related to the document

More Complex Applications of NLP

- Search Engines
- Document Summarization
- Machine Translation
- Chatbots
- Information Retrieval

Some Real Examples From My Work

- Politics:
  - Coding issues discussed by political candidates
  - Measuring voter knowledge in survey responses
- Government:
  - Predicting opioid overdoses from EMS case reports
  - Recommending 311 service request types
- Tech:
  - Triaging help desk support emails
  - Flagging sensitive information in a database
  - Preventing vandalism in open-sourced maps

Demo: NLP for Machine Learning Applications

Developing Your Own NLP Projects

Trying It For Yourself

- For the rest of the workshop, you'll work through some of these methods yourself
- Pick a dataset of your own to work with
  - Don't know where to start? Try http://http://archive.ics.uci.edu/ml/
- Given the limited time available, focus on getting something that works rather than making it perfect

Areas to Explore

- Processing and Descriptive Analysis
  - Does the dataset present unique challenges in loading and processing? What are the pros and cons of various processing choices on this data? What are the most common words / n-grams, and how are they distributed across the corpus?
- Unsupervised Learning
  - Can the documents be grouped into clusters that make sense? Does topic modeling reveal interesting themes?
- Supervised Learning
  - Can we build a model from this data to categorize documents or predict some outcome form them? How well does that model generalize outside of the training sample?

Next Steps

- Continue developing your own projects
- Read more on NLP motivations, theory, methods, and examples
- Try out more advanced topics
  - Using word embeddings to enhance your data
  - Build more flexible models using neural networks
  - Incorporate NLP into real-time applications

Thank you!

- Twitter: @therriaultphd
- github.com/therriault
- andrew.therriault@gmail.com