Text Data in Business and Economics

Benjamin W. Arold, University of Cambridge

1. Overview

Welcome

- This course focuses on applications of **natural language processing** in **applied economics**

Welcome

- This course focuses on applications of **natural language processing** in **applied economics**
- Methods:
  - Develop skills in applied natural language processing
  - Convert natural language texts – e.g. legal and political documents – to data
  - Use same/similar methods to analyze image data and audio data

Welcome

- This course focuses on applications of **natural language processing** in **applied economics**
- Methods:
  - Develop skills in applied natural language processing
  - Convert natural language texts – e.g. legal and political documents – to data
  - Use same/similar methods to analyze image data and audio data
- Economics:
  - Relate text data to metadata to understand economic/political/social forces
  - e.g., analyze the motivations and decisions of public officials through their writings and speeches
  - Assess the real-world impacts of language on government and the economy

Logistics

Learning Materials

Course Content Overview

Schedule

- 10 lectures (5 meetings, with 2 lectures each):
- Course Syllabus:
  - https://docs.google.com/document/d/1FDIEEsO_v2Lwm_Kkkf_T17lj8jsQr_FohZBIxtTyt9Q/edit?tab=t.0
- Course Repo:
  - https://github.com/BenjaminArold/Course_Text_Data_2024

Course Communication

- Course announcements will be done via email (if you have not been getting emails from me already, let me know)

Logistics

Learning Materials

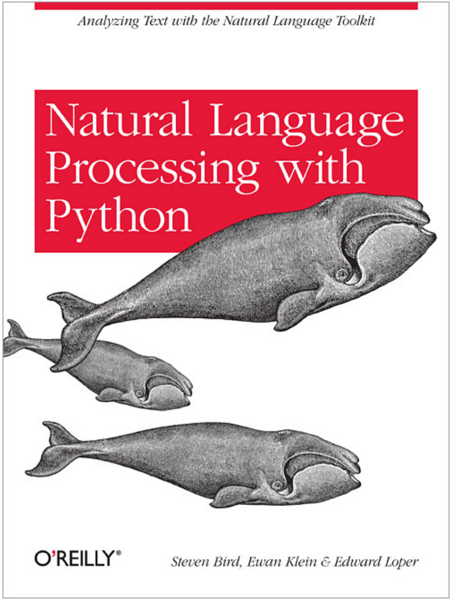
Course Content Overview

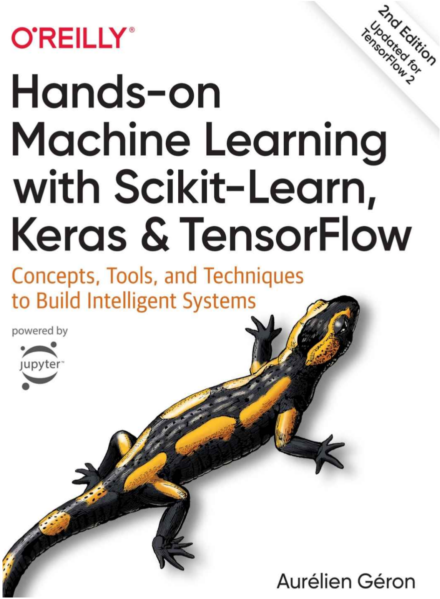
Course Bibliographies

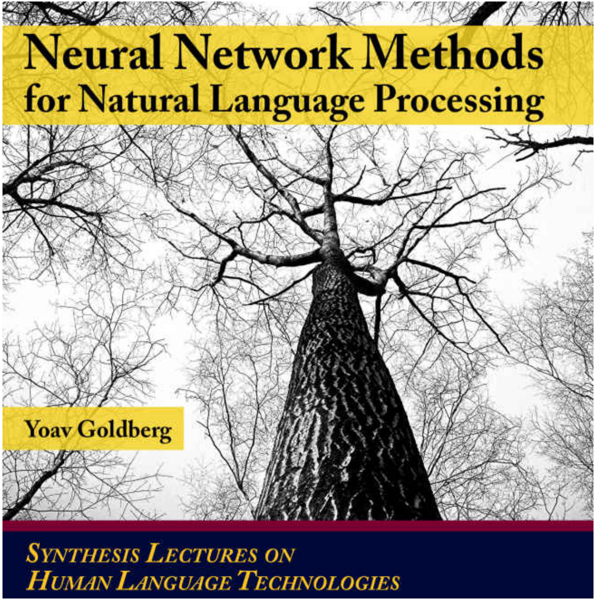
- Bibliography of references:
  - reference readings on tools/methods
  - not required, but useful to complement the slides

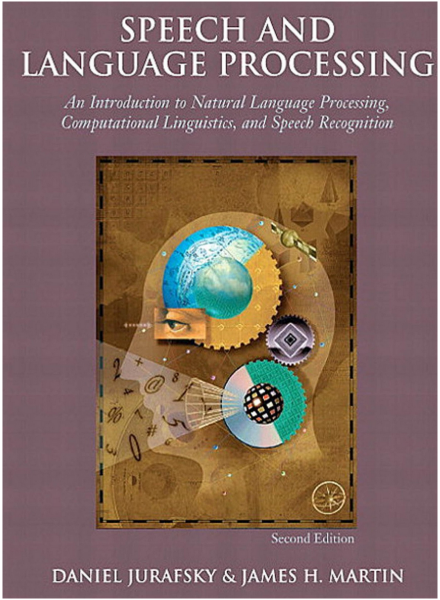
Course Bibliographies

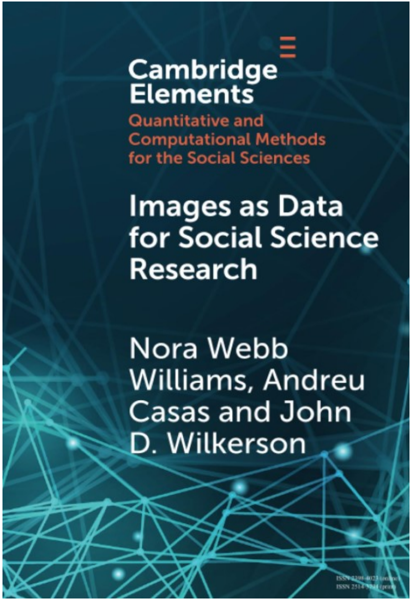
- Bibliography of references:
  - reference readings on tools/methods
  - not required, but useful to complement the slides
- Bibliography of applications:
  - economics application papers, for class presentations


Analyzing Text with the Natural Language Toolkit
Natural Language Processing with Python
O'REILLY®
Steven Bird, Ewan Klein & Edward Loper


O'REILLY
2nd Edition
Updated for TensorFlow 2
Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow
Concepts, Tools, and Techniques to Build Intelligent Systems
powered by
jupyter
Aurélien Géron


Neural Network Methods
for Natural Language Processing
Yoav Goldberg
SYNTHESIS LECTURES ON
HUMAN LANGUAGE TECHNOLOGIES


SPEECH AND
LANGUAGE PROCESSING
An Introduction to Natural Language Processing,
Computational Linguistics, and Speech Recognition
Second Edition
DANIEL JURAFSKY & JAMES H. MARTIN


Cambridge
Elements
Quantitative and
Computational Methods
for the Social Sciences
Images as Data
for Social Science
Research
Nora Webb
Williams, Andreu
Casas and John
D. Wilkerson

Main Python packages for NLP

- Python 3 is ideal for text data and natural language processing
  - Can use Anaconda or download the packages we need to a pip environment
    - nltk – broad collection of pre-neural-nets NLP tools
    - scikit-learn – ML package with nice text vectorizers, clustering, and supervised learning
    - xgboost – gradient-boosted machines for supervised learning
    - gensim – topic models and embeddings
    - spaCy – tokenization, NER, parsing, pre-trained vectors
    - huggingface – pre-trained transformer models
    - tensorflow / keras – deep learning-based text/image/audio analysis
    - librosa - library for audio analysis

Coding Practice and Assignments

Main Coding Examples on GitHub (discussed in class):`https://github.com/BenjaminArold/Course_Text_Data_2024/tree/main/notebooks`

Additional Assignments on GitHub (not discussed in class): `https://github.com/BenjaminArold/Course_Text_Data_2024/tree/main/assignments`

Discussant Presentations

- At the end of most lectures, we will have one discussant presentations on one of the economics articles listed in the syllabus
- Please sign up here: https://docs.google.com/spreadsheets/d/1ASbOxEPEwhZeDo6JefnZZGUxBGdYbMjRwdnESTCZUvM/edit#gid=0
- Critical presentations are up to 15 minutes, should present and critique:
  - research question
  - text-image-audio-analysis methods
  - empirical methods
  - results
  - contribution

Topics Covered

Text-as-Data:

- Dictionaries, Tokenization, and Document Distance
- Topic Models and ML with text, Word Embeddings and Linguistic Parsing
- Transformers, LLMs

Topics Covered

Text-as-Data:

- Dictionaries, Tokenization, and Document Distance
- Topic Models and ML with text, Word Embeddings and Linguistic Parsing
- Transformers, LLMs

Image-as-Data:

- Classical ML, Convolutional Neural Nets, More Deep Learning

Topics Covered

Text-as-Data:

- Dictionaries, Tokenization, and Document Distance
- Topic Models and ML with text, Word Embeddings and Linguistic Parsing
- Transformers, LLMs

Image-as-Data:

- Classical ML, Convolutional Neural Nets, More Deep Learning

Audio-as-Data:

- Classical ML, Recurrent Neural Nets, Ethical Considerations

Text is high-dimensional

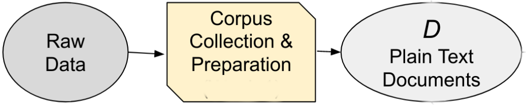
- sample of documents, each n_L words long, drawn from vocabulary of n_V words
- The unique representation of each document has dimension $n_V^{n_L}$
  - e.g., a sample of 30-word Twitter messages using only the one thousand most common words in the English language
    - $\rightarrow$ dimensionality $= 1000^{30} = 10^{32}$

"Text as Data", GKT 2017

Summarize analysis in three steps:

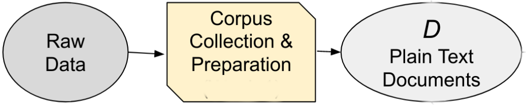
- convert raw text D to numerical array $\boldsymbol{C}$
    - The elements of $\boldsymbol{C}$ are counts over tokens (words or phrases)
- map $\boldsymbol{C}$ to predicted values $\hat{\boldsymbol{V}}$ of unknown outcomes $\boldsymbol{V}$
    - Learn $\hat{\boldsymbol{V}}(\boldsymbol{C})$ using machine learning
    - e.g. supervised learning for some labeled $\boldsymbol{C}_i$ and V_i
    - or unsupervised learning of topics/dimensions just from $\boldsymbol{C}$
- use $\hat{\boldsymbol{V}}$ for subsequent descriptive or causal analysis

Corpora



- Text data is a sequence of characters called documents.
- The set of documents is the corpus, which we will call D.

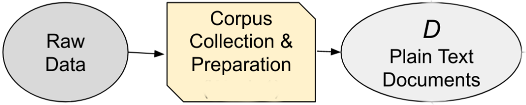
Corpora



- Text data is a sequence of characters called documents.
- The set of documents is the corpus, which we will call D.

- Text data is **unstructured**:
  - the information we want is mixed together with (lots of) information we don't

Corpora



- Text data is a sequence of characters called documents.
- The set of documents is the corpus, which we will call D.

- Text data is **unstructured**:
  - the information we want is mixed together with (lots of) information we don't
- All text data approaches will throw away some information:
  - The trick is figuring out how to retain valuable information

This course is about relating documents to metadata

- This course is on NLP for **economics**:
  - the documents are not that meaningful by themselves
  - we want to relate **text** data to **metadata**
  - we want to see how these methods can be applied for image and audio analysis

This course is about relating documents to metadata

- This course is on NLP for **economics**:
  - the documents are not that meaningful by themselves
  - we want to relate **text** data to **metadata**
  - we want to see how these methods can be applied for image and audio analysis
- e.g., measuring positive-negative sentiment Y in political speeches
  - not that meaningful by itself

This course is about relating documents to metadata

- This course is on NLP for **economics**:
  - the documents are not that meaningful by themselves
  - we want to relate **text** data to **metadata**
  - we want to see how these methods can be applied for image and audio analysis
- e.g., measuring positive-negative sentiment Y in political speeches
  - not that meaningful by itself
- but how about sentiment Y_{ijkt} in speech i by politician j on topic k at time t:
  - how does sentiment vary over time t?
  - does politician from party p_j express more negative sentiment toward topic k?

What counts as a document?

The unit of analysis (the "document") will vary depending on your question

- needs to be fine enough to fit the relevant metadata variation
- should not be finer – would make dataset more high-dimensional without relevant empirical variation

What counts as a document?

The unit of analysis (the “document”) will vary depending on your question

- needs to be fine enough to fit the relevant metadata variation
- should not be finer – would make dataset more high-dimensional without relevant empirical variation

E.g., what should we use as the document in these contexts?

1. predicting whether a judge is right-wing or left-wing in partisan ideology, from their written opinions
2. predicting whether parliamentary speeches become more emotive in the run-up to an election