

D001 Economic Analysis of Non-Standard Data

Benjamin W. Arold

1. Overview

Welcome

- ▶ This course covers methods for processing and analysing non-standard data in economics
- ▶ Focus on text-as-data, image-as-data, and audio-as-data

Welcome

- ▶ This course covers methods for processing and analysing non-standard data in economics
- ▶ Focus on text-as-data, image-as-data, and audio-as-data
- ▶ Methods:
 - ▶ Develop skills in applied natural language processing
 - ▶ Convert natural language texts – e.g. economic, legal, and political documents – to data
 - ▶ Use same/similar methods to analyze image data and audio data

Welcome

- ▶ This course covers methods for processing and analysing non-standard data in economics
- ▶ Focus on text-as-data, image-as-data, and audio-as-data
- ▶ Methods:
 - ▶ Develop skills in applied natural language processing
 - ▶ Convert natural language texts – e.g. economic, 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

- ▶ 9 lectures, 2 hours each:
- ▶ Course Syllabus:
 - ▶ <https://docs.google.com/document/d/1RpfXwZl0PM3WyM5cz7cbLD6EVTr9kJNjHyGuApVmEWM/edit?tab=t.0>
- ▶ FAQ Sheet:
 - ▶ <https://docs.google.com/document/d/1xX3FGpIFA2ChVCY5p3gKNcWN4WJU12VMs75GyweP5U/edit?tab=t.0>
- ▶ Course Repo:
 - ▶ <https://www.vle.cam.ac.uk/course/view.php?id=254153>

Logistics

Learning Materials

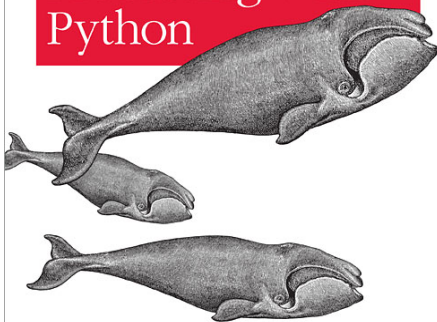
Course Content Overview

Course Bibliographies

- ▶ Materials referenced in syllabus:
 - ▶ Lecture slides
 - ▶ Papers (required and suggested reading)
 - ▶ Jupyter Notebooks
 - ▶ Additional Material

Analyzing Text with the Natural Language Toolkit

Natural Language Processing with Python



O'REILLY®

Steven Bird, Ewan Klein & Edward Loper

O'REILLY®

Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow

Concepts, Tools, and Techniques
to Build Intelligent Systems

powered by



Aurélien Geron

2nd Edition
Updated for
TensorFlow 2



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

ISBN 978-0-521-87600-1
ISBN 978-0-521-87601-8

Main Python packages for NLP

- ▶ Python 3 is ideal for text data / natural language processing, image analysis, and audio analysis
 - ▶ 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

Coding Examples on Moodle (discussed in class):

<https://www.vle.cam.ac.uk/course/view.php?id=254153>

Discussant Presentations

- ▶ At the end of most lectures, we will have one presentation on one of the economics articles listed in the syllabus

- ▶ Please sign up here:

https://docs.google.com/spreadsheets/d/1n_ZtNSmweG1iThvRetm9KQFacjzZVXmpDoxNCfUEWwo/edit?gid=0#gid=0

- ▶ Critical presentations are up to 15 minutes, should present and critique:
 - ▶ research question
 - ▶ main focus: text/image/audio data and methods
 - ▶ empirical methods
 - ▶ results
 - ▶ contribution

Logistics

Learning Materials

Course Content Overview

Text, Image, and Audio Data is high-dimensional

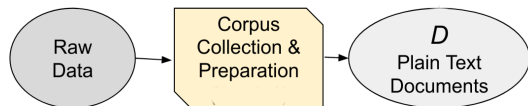
- ▶ Let's start with text analysis
- ▶ Sample of documents, each n_L words long, drawn from vocabulary of n_V words
- ▶ The unique representation of each document has dimension $n_V^{n_L}$
 - ▶ e.g., a sample of 30-word Twitter messages using only the one thousand most common words in the English language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

“Text as Data”, GKT 2017

Summarize analysis in three steps:

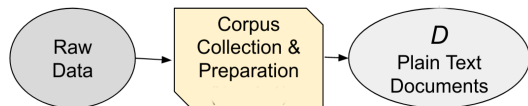
- ▶ convert raw text D to numerical array \mathbf{C}
 - ▶ The elements of \mathbf{C} are counts over tokens (words or phrases)
- ▶ map \mathbf{C} to predicted values $\hat{\mathbf{V}}$ of unknown outcomes \mathbf{V}
 - ▶ Learn $\hat{\mathbf{V}}(\mathbf{C})$ using machine learning
 - ▶ e.g. supervised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

Corpora



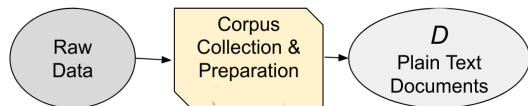
- ▶ Text data is a sequence of characters called documents.
- ▶ The set of documents is the corpus, which we will call D .

Corpora



- ▶ Text data is **unstructured**:
 - ▶ the information we want is mixed together with (lots of) information we don't
- ▶ 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 **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
- ▶ Text data is a sequence of characters called documents.
- ▶ The set of documents is the corpus, which we will call D .

This course is about relating documents to metadata

- ▶ This course is on **analysis of text, image, and audio** data for **economics**:
 - ▶ start with text analysis (natural language processing NLP), then transfer methods to image and audio
 - ▶ text documents are not that meaningful by themselves
 - ▶ we want to relate **text** data to **metadata**

This course is about relating documents to metadata

- ▶ This course is on **analysis of text, image, and audio** data for **economics**:
 - ▶ start with text analysis (natural language processing NLP), then transfer methods to image and audio
 - ▶ text documents are not that meaningful by themselves
 - ▶ we want to relate **text** data to **metadata**
- ▶ 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 **analysis of text, image, and audio** data for **economics**:
 - ▶ start with text analysis (natural language processing NLP), then transfer methods to image and audio
 - ▶ text documents are not that meaningful by themselves
 - ▶ we want to relate **text** data to **metadata**
- ▶ 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

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

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