Text Data in Economics

Warwick QAPEC Summer School

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.

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.
- Economics:
  - Relate text data to metadata to understand economic 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.

What we will do

1. **Read text documents as data**.

1. **Read text documents as data**.
2. **Unsupervised learning techniques for interpreting corpora**.

1. **Read text documents as data**.
2. **Unsupervised learning techniques for interpreting corpora**.
3. **Supervised learning for regression and classification**.

1. **Read text documents as data**.
2. **Unsupervised learning techniques for interpreting corpora**.
3. **Supervised learning for regression and classification**.
4. **Word/document embedding for identifying dimensions of language**.

1. **Read text documents as data**.
2. **Unsupervised learning techniques for interpreting corpora**.
3. **Supervised learning for regression and classification**.
4. **Word/document embedding for identifying dimensions of language**.
5. **Syntactic/semantic parsing – identifying actors, actions, and attributes.**

Logistics

Learning Materials

Course Content Overview

Schedule

- See syllabus for the schedule.
- all sessions are recorded.
- 10 lectures:
  - up to 40 minutes of lecturing
  - ~20 minutes for student presentations/discussions of papers
- 4 TA sessions
  - going over the example notebooks and problem set solutions

Course Learning Materials

- Course Syllabus (see chat).
- Course Repo:
  - `https://github.com/elliottash/text_econ_2022`

Teaching Assistant

- Sandra Stampi-Bombelli (alessandra.stampi@gess.ethz.ch)
  - has taken the advanced version of this course
  - thesis research using some new word embedding tools
- TA Sessions:
  - see schedule on syllabus
  - recorded part: will go over code notebooks and homeworks
  - non-recorded part: office hours to answer questions

Course Communication

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

Assignments

- 3 problem sets (based on the coding material)
- 1 group presentation on one of the course readings
- 1 referee report on one of the course readings, written individually.

Assignments

- 3 problem sets (based on the coding material)
- 1 group presentation on one of the course readings
- 1 referee report on one of the course readings, written individually.
- optional (limited spots)
  - a course project in groups of 2 students
  - to be workshopped in-person in September

Composition of Class (Survey)

...

Logistics

Learning Materials

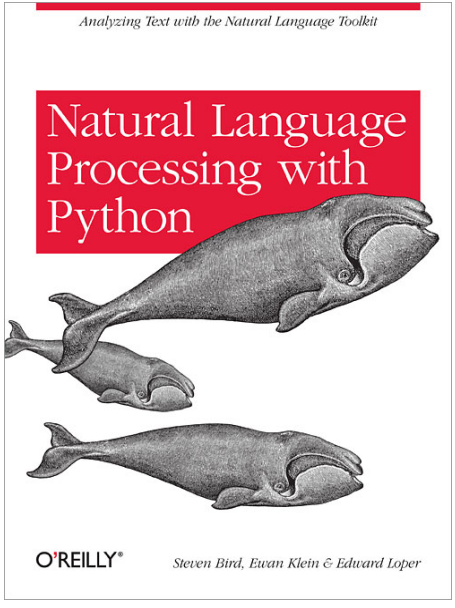
Course Content Overview

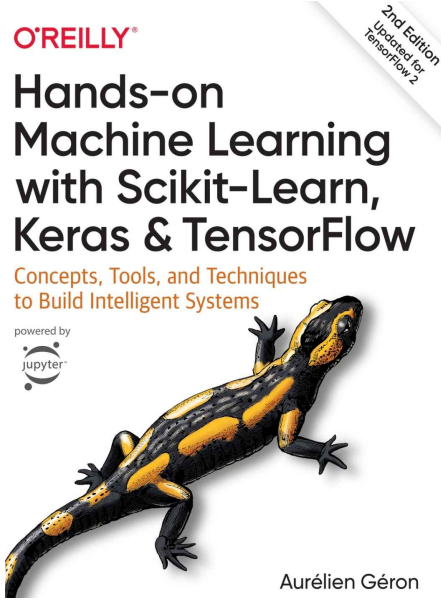
Course Bibliographies

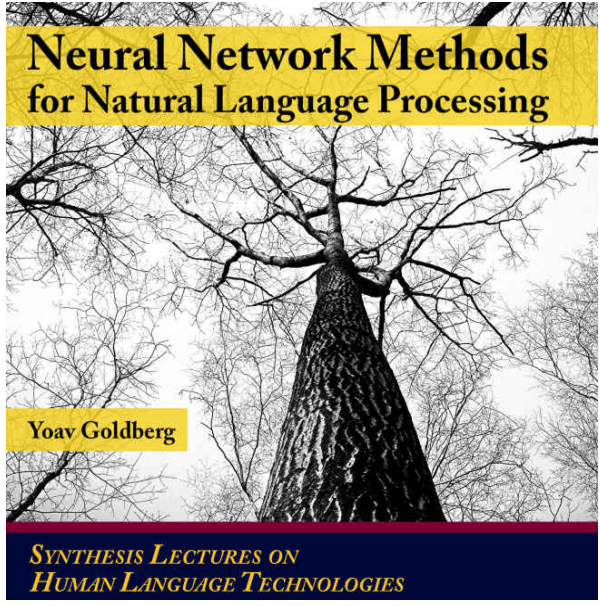
- Bibliography of references:
  - reference readings on tools/methods
  - not required, but useful to complement the slides

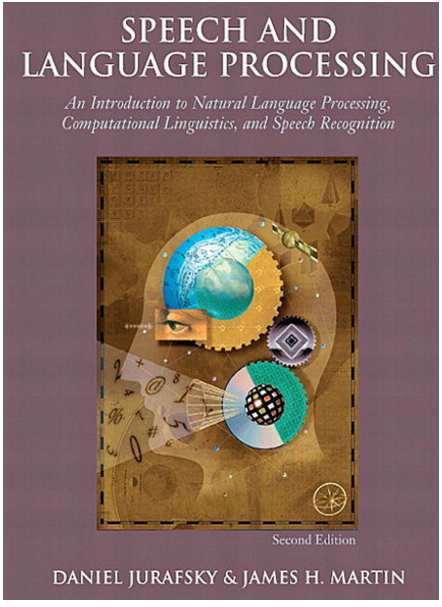
Course Bibliographies

- 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

Python is a Course Pre-Requisite

- Example Code Notebooks: `https://github.com/elliottash/text_econ_2022/tree/master/notebooks`
- Python 3 is ideal for text data and natural language processing.
  - Can use Anaconda or download the packages we need to a pip environment.
  - See the syllabus for list of packages we will use.
- if you want to try to use Stata or R instead, let me know.
  - i could use your help in translating the course materials

Main Python packages for NLP

- 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 – source for pre-trained transformer models.

Coding Practice and Homework Assignments

Coding Examples on GitHub:

`https://github.com/elliottash/nlp_lss_2022/tree/master/notebooks`

Homework Assignments on GitHub:

`https://github.com/elliottash/nlp_lss_2022/tree/master/homework`

- Timeline for code material:
  - Coding notebook for Week t will be reviewed in TA Session on Friday of week t.
  - Homework for Week t:
    - due Thursday night in Week $t+1$, uploaded on EduFlow.
    - Homeworks will be checked in the TA session on Friday of Week $t+1$
- E.g.:
  - notebook 1 will be reviewed this Friday Feb 25th (Week 1 TA Session)
  - homework 1 will be due next Thursday (March 3rd) and reviewed next friday (Week 2 TA session, March 4th)
  - notebook 2 will be reviewed on Week 2 TA session on March 4th
  - and so on.

Logistics

Learning Materials

Course Content Overview

Course Objectives (Student Self-Reports)

xxx

Big Data, Big Analytics

Big Data, Big Analytics

- Massive increase in availability of unstructured text datasets:
  - new social structures (the internet, email)
  - digitization efforts (govt documents, Google)

Big Data, Big Analytics

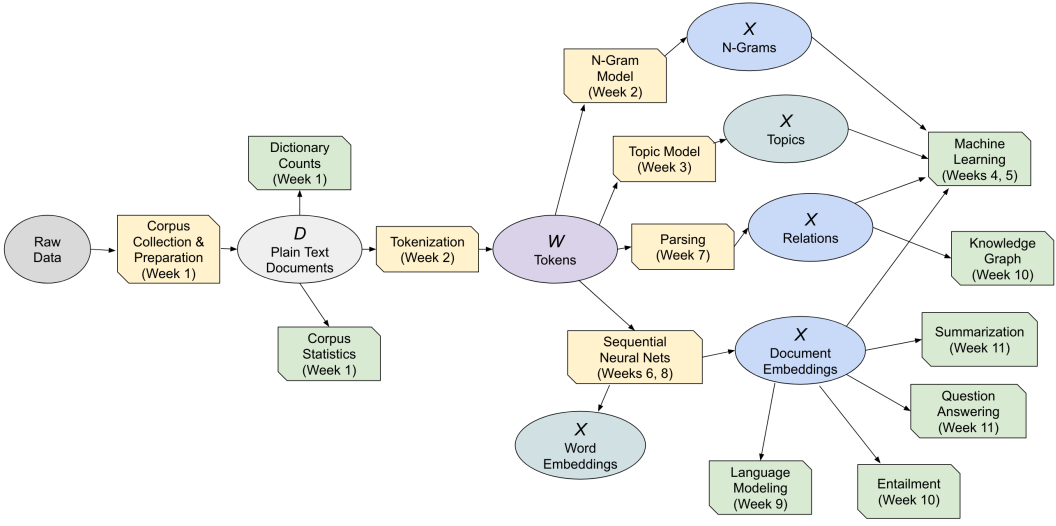
- Massive increase in availability of unstructured text datasets:
  - new social structures (the internet, email)
  - digitization efforts (govt documents, Google)
- Parallel increase in computational resources:
  - cheap disk space
  - efficient database solutions
  - compute: CPUs $\rightarrow$ GPUs $\rightarrow$ TPUs

Big Data, Big Analytics

- Massive increase in availability of unstructured text datasets:
  - new social structures (the internet, email)
  - digitization efforts (govt documents, Google)
- Parallel increase in computational resources:
  - cheap disk space
  - efficient database solutions
  - compute: CPUs $\rightarrow$ GPUs $\rightarrow$ TPUs
- Parallel development of tools for natural language analysis
  - text by itself is not very useful
  - machine learning, natural language processing, causal inference

Big Data, Big Analytics

- Massive increase in availability of unstructured text datasets:
  - new social structures (the internet, email)
  - digitization efforts (govt documents, Google)
- Parallel increase in computational resources:
  - cheap disk space
  - efficient database solutions
  - compute: CPUs $\rightarrow$ GPUs $\rightarrow$ TPUs
- Parallel development of tools for natural language analysis
  - text by itself is not very useful
  - machine learning, natural language processing, causal inference
- many interesting economic variables are in text!
  - e.g. economic news, political speeches, laws
  - We cannot read them – somehow we must teach the computers to read them for us.


Raw Data
Corpus Collection & Preparation (Week 1)
D Plain Text Documents
Dictionary Counts (Week 1)
Corpus Statistics (Week 1)
Tokenization (Week 2)
W Tokens
N-Gram Model (Week 2)
X N-Grams
Topic Model (Week 3)
X Topics
Parsing (Week 7)
X Relations
Sequential Neural Nets (Weeks 6, 8)
X Word Embeddings
X Document Embeddings
Language Modeling (Week 9)
Entailment (Week 10)
Machine Learning (Weeks 4, 5)
Knowledge Graph (Week 10)
Summarization (Week 11)
Question Answering (Week 11)

Any logistical questions about the course?