NLP for Law and Social Science

ETH Zurich, Spring 2022

1. Introduction

Hybrid-lecture norms for those joining online:
(1) Turn on video and set audio to mute
(2) In Participants panel, set zoom name to full name
(3) say "hi" in the chat

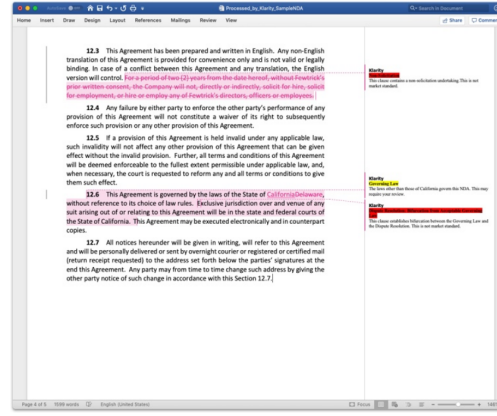
Klarity reviews NDAs under commercial market standard.

👌 Klarity highlights standard language in green.

👀 Language that requires your attention is in yellow.

👿 Non-market standard language and red-flags are in red.

Language that is not marked is boilerplate and doesn't deserve your attention.



The World's First Robot Lawyer

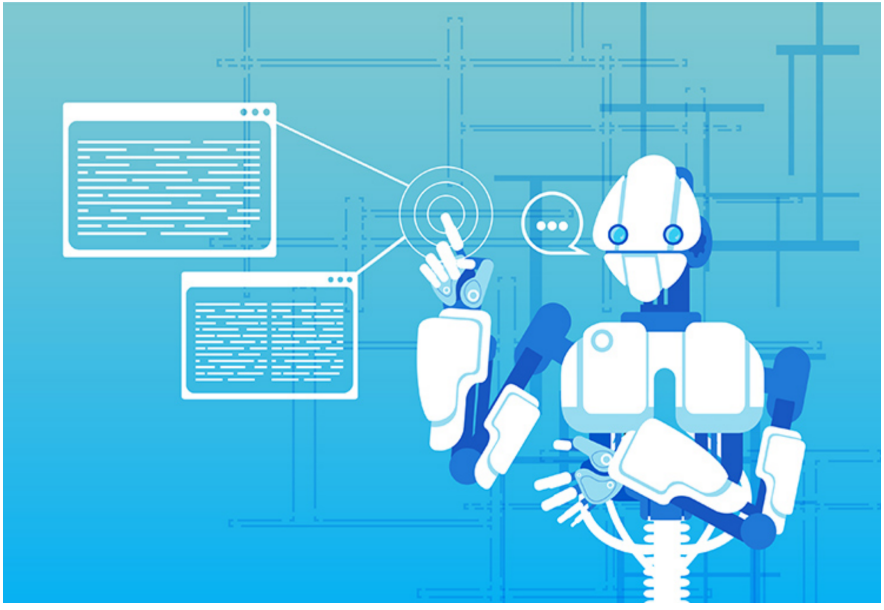
The DoNotPay app is the home of the world's first robot lawyer. Fight corporations, beat bureaucracy and sue anyone at the press of a button.

Sign Up/Login

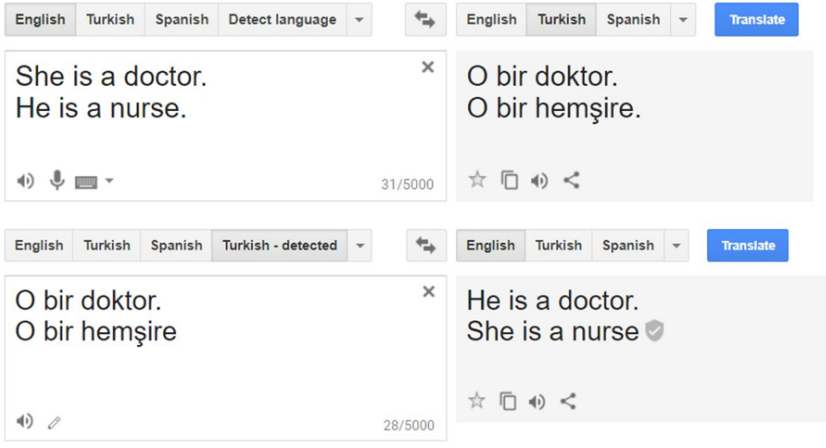
THINGS YOU CAN DO WITH DONOTPAY

- Fight Corporations
- Beat Bureaucracy
- Find Hidden Money
- Sue Anyone
- Automatically Cancel Your Free Trials

Your Court-Appointed Chatbot – Is Artificial Intelligence Threatening the Legal Profession?



Language Models can be Biased



Source: fastai NLP course.

Howard, co-founder of Fast.AI agrees. “I’ve been trying to warn people about this for a while,” he says. “We have the technology to totally fill Twitter, email, and the web up with reasonable-sounding, context-appropriate prose, which would drown out all other speech and be impossible to filter.”

`https://transformer.huggingface.co/doc/distil-gpt2`

Welcome

- This course focuses on applications of **natural language processing** in **law** and **social science**.

Welcome

- This course focuses on applications of **natural language processing** in **law** and **social science**.
- Engineering goals:
  - Develop skills in applied natural language processing
  - Apply to machine interpretation and generation of natural language texts – e.g. legal and political documents.

Welcome

- This course focuses on applications of **natural language processing** in **law** and **social science**.
- Engineering goals:
  - Develop skills in applied natural language processing
  - Apply to machine interpretation and generation of natural language texts – e.g. legal and political documents.
- Scientific goals:
  - Relate text data to metadata to understand social forces.

Welcome

- This course focuses on applications of **natural language processing** in **law** and **social science**.
- Engineering goals:
  - Develop skills in applied natural language processing
  - Apply to machine interpretation and generation of natural language texts – e.g. legal and political documents.
- Scientific goals:
  - Relate text data to metadata to understand social forces.
  - Understand the motivations and decisions of judges and public officials through their writings and speeches.

Welcome

- This course focuses on applications of **natural language processing** in **law** and **social science**.
- Engineering goals:
  - Develop skills in applied natural language processing
  - Apply to machine interpretation and generation of natural language texts – e.g. legal and political documents.
- Scientific goals:
  - Relate text data to metadata to understand social forces.
  - Understand the motivations and decisions of judges and 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**.

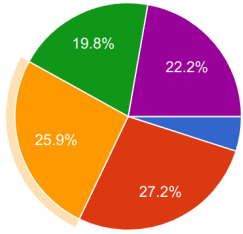
1. **Read text documents as data**.
2. **Unsupervised learning techniques for interpreting corpora**.
3. **Supervised learning for regression and classification**.
4. **Word/document embedding for identifying dimensions of language**.
5. **Discourse analytics – summarization, question answering.**

Lecture Times

- Mondays, 1415h-16h
  - ~10 minute break, usually 15h-1510h

Which of the following best matches your preferences about the course format?

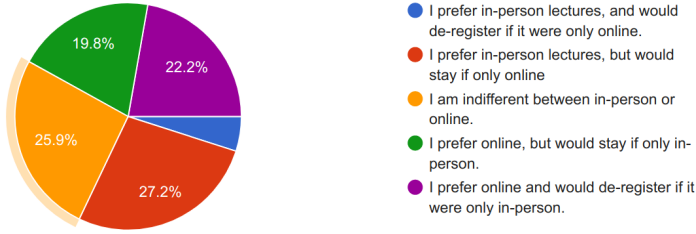
81 responses



- I prefer in-person lectures, and would de-register if it were only online.
- I prefer in-person lectures, but would stay if only online
- I am indifferent between in-person or online.
- I prefer online, but would stay if only in-person.
- I prefer online and would de-register if it were only in-person.

Which of the following best matches your preferences about the course format?

81 responses

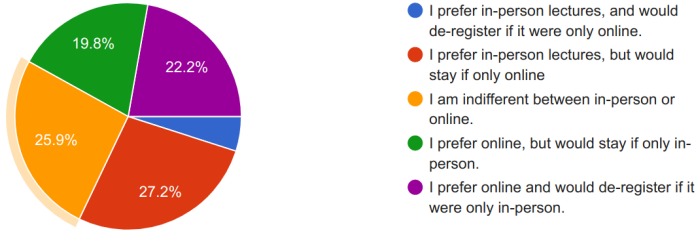


Let's make the most of hybrid learning!

- Lectures will be recorded, but live attendance is required and absences require permission of instructor.
  - To make questions or comments from online, use the "raise hand" function.

Which of the following best matches your preferences about the course format?

81 responses



Let's make the most of hybrid learning!

- Lectures will be recorded, but live attendance is required and absences require permission of instructor.
  - To make questions or comments from online, use the "raise hand" function.
- We will keep track of course participation through in-class activities (e.g. zoom polls, group work).

Course Learning Materials

- Course Syllabus (see chat).
- Bibliography of References:
  - `https://bit.ly/NLP-reading-list`
- Course Repo:
  - `https://github.com/elliottash/nlp_lss_2022`

Teaching Assistant

- Afra Amini (afra.amini@infk.ethz.ch)
- TA Sessions:
  - Fridays, 10am-1145am on zoom (and recorded)
  - recorded part: will go over code notebooks and homeworks
  - non-recorded part: office hours to answer questions

Course Communication

- Course announcements will be done on Moodle (and sent by email).
- Post questions/concerns on the weekly course Q&A moodle thread.
  - if you want to post anonymously, send to Afra beforehand and she can post it.

Course Workload

3 ECTS credits $\approx$ 90 hours of work

Course Workload

3 ECTS credits $\approx$ 90 hours of work

- 12 lectures, 1.75 hours each $=$ 21 hours
- 11 NLP programming homework assignments, ~1.5 hours each $\approx$ 17 hours
- Required readings (three papers) $\approx$ 6 hours
- 2 response essays, ~6 hours each $\approx$ 12 hours
- Final assignment / take-home test, 4 hours
- $\approx$ **60 required hours.**

Course Workload

3 ECTS credits ≈ 90 hours of work

- 12 lectures, 1.75 hours each = 21 hours
- 11 NLP programming homework assignments, ~1.5 hours each ≈ 17 hours
- Required readings (three papers) ≈ 6 hours
- 2 response essays, ~6 hours each ≈ 12 hours
- Final assignment / take-home test, 4 hours
- ≈ **60 required hours.**
- ≈ 30 hours at student discretion:
  - 12 optional TA sessions, 1 hour each ≈ 12 hours
  - leaves ~14 hours for additional study time

Course Projects

2 additional ECTS credits $\approx$ 60 additional hours of work

- About twice as much out-of-class work expected
  - previous course projects have turned into conference/journal publications.
  - two projects turned into funded Innoisuisse startups.

Course Projects

2 additional ECTS credits $\approx$ 60 additional hours of work

- About twice as much out-of-class work expected
  - previous course projects have turned into conference/journal publications.
  - two projects turned into funded Innoisuisse startups.
- Can be done individually or in small groups (up to 4).

Course Projects

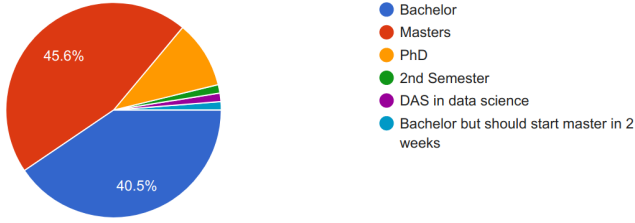
2 additional ECTS credits $\approx$ 60 additional hours of work

- About twice as much out-of-class work expected
  - previous course projects have turned into conference/journal publications.
  - two projects turned into funded Innoisuisse startups.
- Can be done individually or in small groups (up to 4).
- Information session after Week 2 lecture (~10 minutes)
  - we have a list of potential topic ideas.

Composition of Class (Survey)

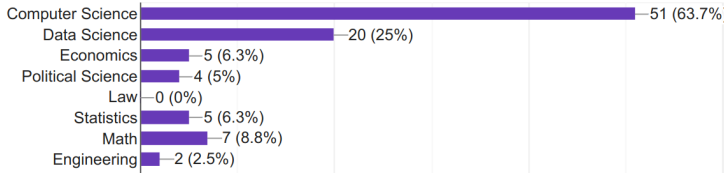
What is your current degree status?

79 responses



What is your major or concentration? (check all that apply, including previous degrees)

80 responses



Course Bibliographies

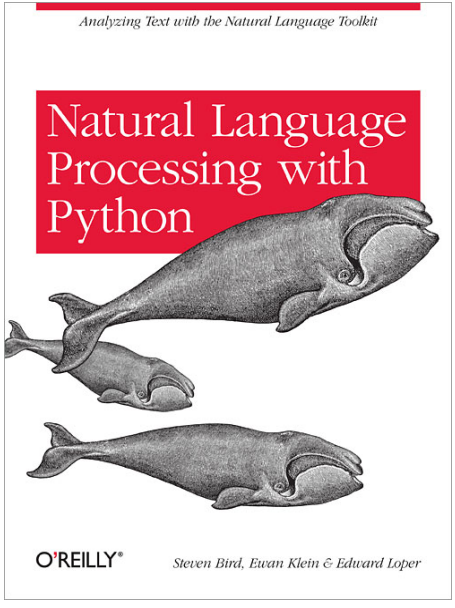
- Required readings for in-class discussions indicated in syllabus.

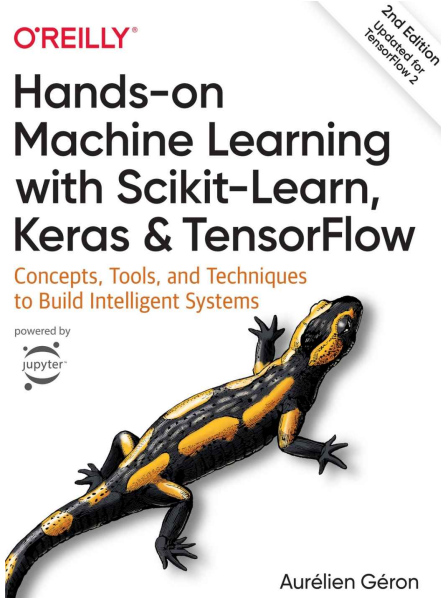
Course Bibliographies

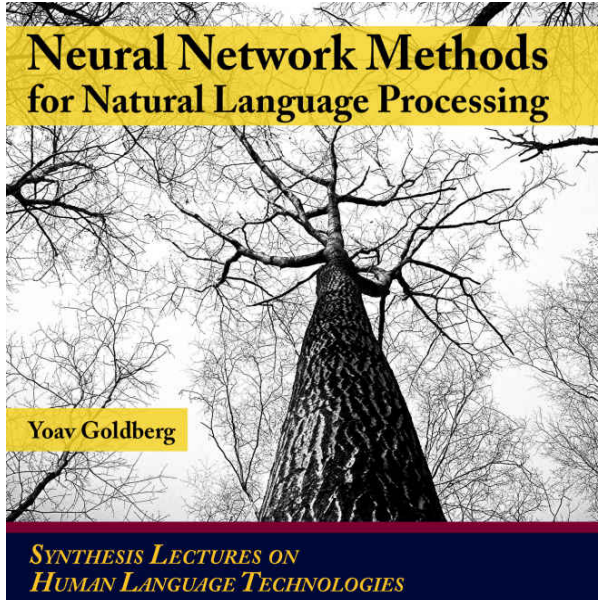
- Required readings for in-class discussions indicated in syllabus.
- Bibliography of references:
  - reference readings on tools/methods
  - not required, but useful to complement the slides

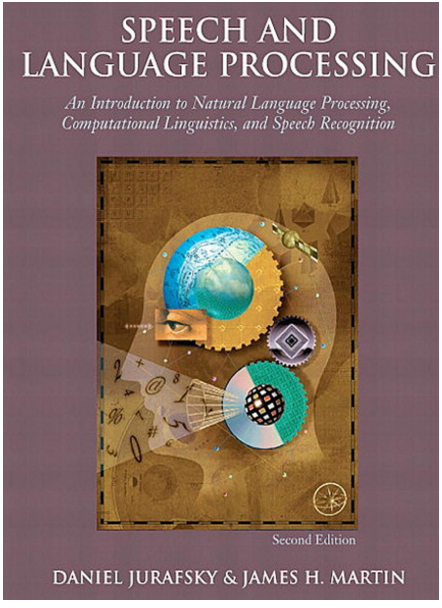
Course Bibliographies

- Required readings for in-class discussions indicated in syllabus.
- Bibliography of references:
  - reference readings on tools/methods
  - not required, but useful to complement the slides
- Bibliography of applications:
  - social science papers for response essays (more next week)


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

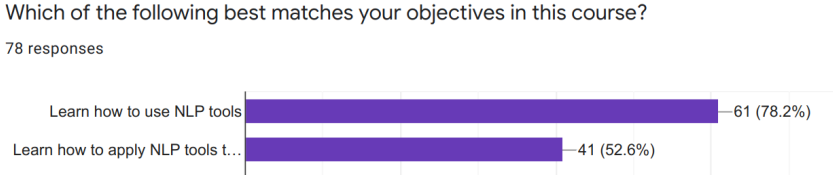
Python is a Course Pre-Requisite

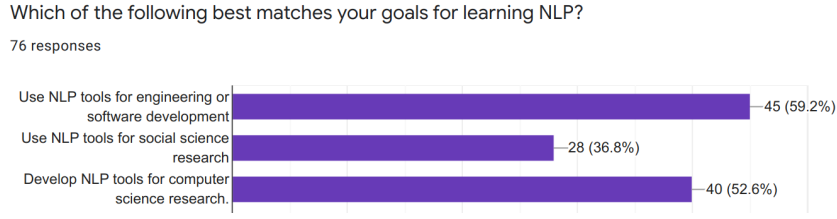
- Example Code Notebooks:
  `https://github.com/elliottash/nlp_lss_2022/tree/master/notebooks`
- Python 3 is ideal for text data and natural language processing.
  - Can use Anaconda or download the packages we need to a pip environment.
  - See the syllabus for list of packages we will use.
- In first TA session this Friday, can try to help with setup questions.

Main Python packages for NLP

- nltk – broad collection of pre-neural-nets NLP tools
- gensim – topic models and embeddings
- spaCy – tokenization, NER, parsing, pre-trained vectors
- huggingface – source for pre-trained transformer models.

Course Objectives (Student Self-Reports)





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

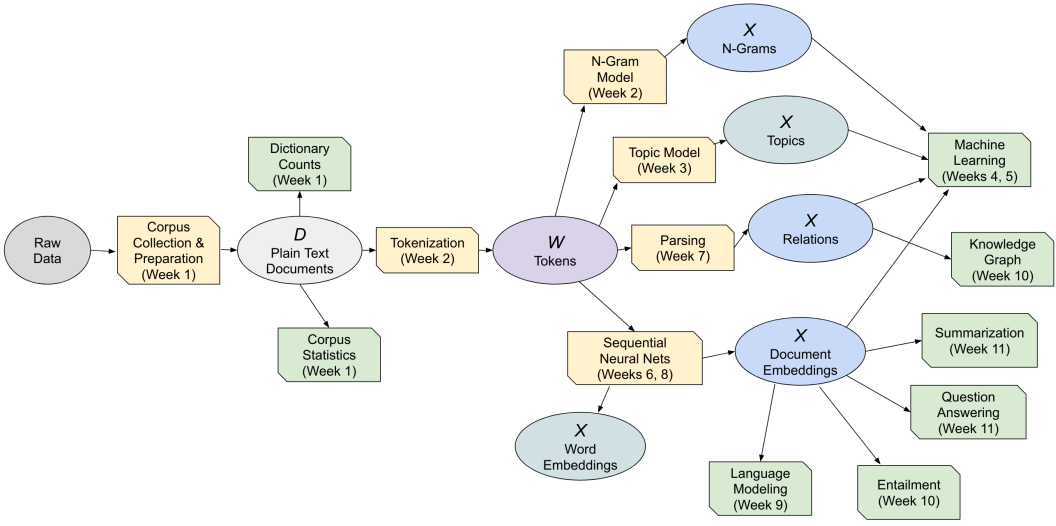
- Massive increase in availability of unstructured text datasets:
  - new social structures (the internet, email)
  - digitization efforts (govt documents, Google)
- Parallel increase in computational resources:
  - cheap disk space
  - efficient database solutions
  - compute: CPUs $\rightarrow$ GPUs $\rightarrow$ 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
- These trends are especially salient in **law** and **political science.**

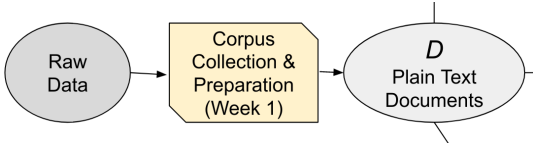
Big Data, Big Analytics

- 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
- These trends are especially salient in **law** and **political science.**
  - The social phenomena of interest – **legal and political institutions** – are composed of thousands, potentially millions, of lines of **unstructured text**.
  - 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)

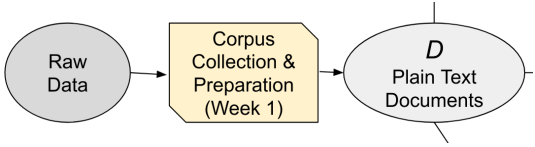
Any logistical questions about the course?

Corpora



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

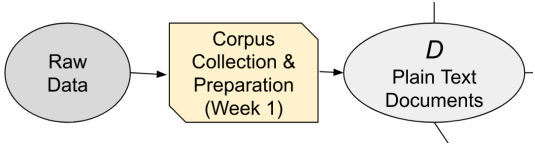
Corpora



- 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.

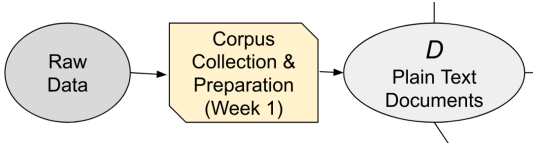
Corpora



- 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.

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.
- The tools from Weeks 2 (Tokenization) and 3 (Dimension Reduction) are focused on this step:
  - transforming an unstructured corpus D to a usable matrix X.

This course is about relating documents to metadata

- This course is on **applied** NLP:
  - the documents are not that meaningful by themselves.
  - we want to relate **text** data to **meta**data.

This course is about relating documents to metadata

- This course is on **applied** NLP:
  - the documents are not that meaningful by themselves.
  - we want to relate **text** data to **meta**data.
- e.g., measuring positive-negative sentiment Y in judicial opinions.
  - not that meaningful by itself.

This course is about relating documents to metadata

- This course is on **applied** NLP:
  - the documents are not that meaningful by themselves.
  - we want to relate **text** data to **meta**data.
- e.g., measuring positive-negative sentiment Y in judicial opinions.
  - not that meaningful by itself.
- but how about sentiment Y_{ijt} in opinion i by judge j at time t:
  - how does sentiment vary over time t?
  - does judge from party p_j express more negative sentiment toward defendants from group g_i?

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.

What should we use as the document in these contexts? (discuss in pairs)

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
3. measuring whether newspapers use higher or lower sentiment toward different groups.

Handling Corpora

- There is already a vast amount of data out there that has already been compiled (e.g. CourtListener, Twitter, New York Times, Google Trends, Wikipedia).

Handling Corpora

- There is already a vast amount of data out there that has already been compiled (e.g. CourtListener, Twitter, New York Times, Google Trends, Wikipedia).
- This won't be on an assignment but everyone in this class should learn how to:
  1. query REST API's
  2. run a web scraper in selenium
  3. do pre-processing on corpora, e.g. to remove HTML markup, fix errors associated with OCR.
- I also recommend everyone to become familiar with hugginface datasets (`https://huggingface.co/docs/datasets/`)

Handling Corpora

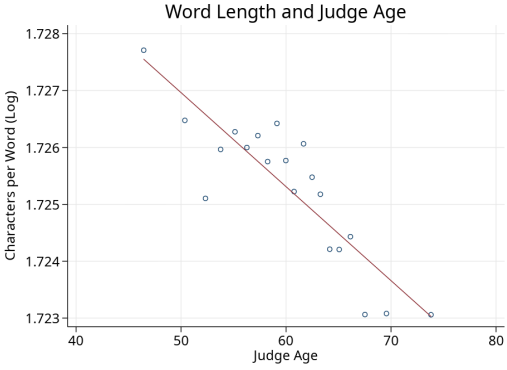
- There is already a vast amount of data out there that has already been compiled (e.g. CourtListener, Twitter, New York Times, Google Trends, Wikipedia).
- This won't be on an assignment but everyone in this class should learn how to:
  1. query REST API's
  2. run a web scraper in selenium
  3. do pre-processing on corpora, e.g. to remove HTML markup, fix errors associated with OCR.
- I also recommend everyone to become familiar with hugginface datasets (`https://huggingface.co/docs/datasets/`)
- All of the tools that we discuss in this class are available in many languages, and machine translation is now quite good and automatable (e.g. `huggingface.co/docs/transformers/master/en/model_doc/marian`).

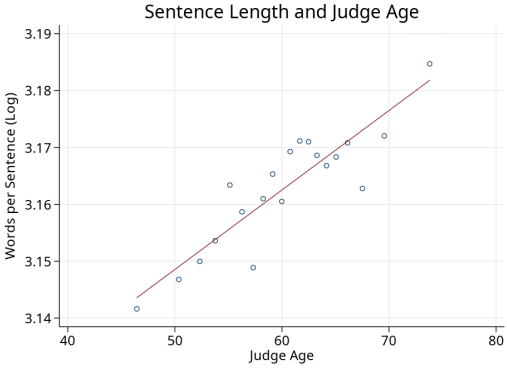
Judge Age and Writing Style

Ash, Goessmann, and MacLeod (2022)

Judge Age and Writing Style

Ash, Goessmann, and MacLeod (2022)





Optimal Legal Complexity (Katz and Bommarito 2014)

- More legal detail is needed to properly specify rules and target incentives to activities and groups.
  - but there are costs to understanding/following/maintaining complex laws, so there is a trade off.

Optimal Legal Complexity (Katz and Bommarito 2014)

- More legal detail is needed to properly specify rules and target incentives to activities and groups.
  - but there are costs to understanding/following/maintaining complex laws, so there is a trade off.
- Katz and Bommarito measure complexity/detail from the text – number of words for code title, and also *word entropy* $\approx$ diversity of the vocabulary.

Optimal Legal Complexity (Katz and Bommarito 2014)

- More legal detail is needed to properly specify rules and target incentives to activities and groups.
  - but there are costs to understanding/following/maintaining complex laws, so there is a trade off.
- Katz and Bommarito measure complexity/detail from the text – number of words for code title, and also *word entropy* $\approx$ diversity of the vocabulary.

Five largest and smallest titles by token count

Title	Tokens	Tokens per section
Public Health and Welfare (Title 42)	2,732,251	369.22
Internal Revenue Code (Title 26)	1,016,995	487.07
Conservation (Title 16)	947,467	200.48
Commerce and Trade (Title 15)	773,819	336.88
Agriculture (Title 7)	751,579	274.00
President (Title 3)	7,564	120.06
Intoxicating Liquors (Title 27)	6,515	144.78
Flag and Seal, Seat of Govt. and the States (Title 4)	5,598	119.11
General Provisions (Title 1)	3,143	80.59
Arbitration (Title 9)	2,489	80.29

Optimal Legal Complexity (Katz and Bommarito 2014)

- More legal detail is needed to properly specify rules and target incentives to activities and groups.
  - but there are costs to understanding/following/maintaining complex laws, so there is a trade off.
- Katz and Bommarito measure complexity/detail from the text – number of words for code title, and also *word entropy* $\approx$ diversity of the vocabulary.

Five largest and smallest titles by token count

Title	Tokens	Tokens per section
Public Health and Welfare (Title 42)	2,732,251	369.22
Internal Revenue Code (Title 26)	1,016,995	487.07
Conservation (Title 16)	947,467	200.48
Commerce and Trade (Title 15)	773,819	336.88
Agriculture (Title 7)	751,579	274.00
President (Title 3)	7,564	120.06
Intoxicating Liquors (Title 27)	6,515	144.78
Flag and Seal, Seat of Govt. and the States (Title 4)	5,598	119.11
General Provisions (Title 1)	3,143	80.59
Arbitration (Title 9)	2,489	80.29

Five highest and lowest titles by word entropy

Title	Word entropy
Commerce and Trade (Title 15)	10.80
Public Health and Welfare (Title 42)	10.79
Conservation (Title 16)	10.75
Navigation and Navigable Waters (Title 33)	10.67
Foreign Relations and Intercourse (Title 22)	10.67
Intoxicating Liquors (Title 27)	9.01
President (Title 3)	8.89
National Guard (Title 32)	8.50
General Provisions (Title 1)	8.49
Arbitration (Title 9)	8.24

Overview of Dictionary-Based Methods

- Dictionary-based text methods use a pre-selected list of words or phrases to analyze a corpus.
  - use regular expressions for this task (see notebook)

Overview of Dictionary-Based Methods

- Dictionary-based text methods use a pre-selected list of words or phrases to analyze a corpus.
  - use regular expressions for this task (see notebook)
- Corpus-specific: counting sets of words or phrases across documents
  - (e.g., number of times a judge says “justice” vs “efficiency”)

Overview of Dictionary-Based Methods

- Dictionary-based text methods use a pre-selected list of words or phrases to analyze a corpus.
  - use regular expressions for this task (see notebook)
- Corpus-specific: counting sets of words or phrases across documents
  - (e.g., number of times a judge says “justice” vs “efficiency”)
- General dictionaries: WordNet, LIWC, MFD, etc.

Measuring uncertainty in macroeconomy

Baker, Bloom, and Davis (QJE 2016)

Measuring uncertainty in macroeconomy

Baker, Bloom, and Davis (QJE 2016)

For each newspaper on each day since 1985, submit the following query:

1. Article contains "uncertain" OR "uncertainty", AND
2. Article contains "economic" OR "economy", AND
3. Article contains "congress" OR "deficit" OR "federal reserve" OR "legislation" OR "regulation" OR "white house"

Normalize resulting article counts by total newspaper articles that month.

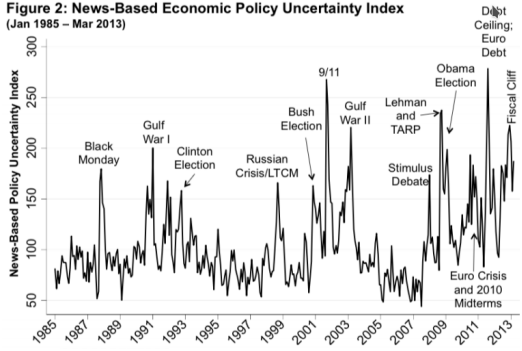
Measuring uncertainty in macroeconomy

Baker, Bloom, and Davis (QJE 2016)

For each newspaper on each day since 1985, submit the following query:

1. Article contains "uncertain" OR "uncertainty", AND
2. Article contains "economic" OR "economy", AND
3. Article contains "congress" OR "deficit" OR "federal reserve" OR "legislation" OR "regulation" OR "white house"

Normalize resulting article counts by total newspaper articles that month.



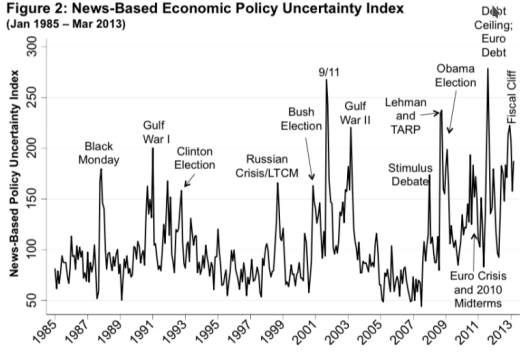
Measuring uncertainty in macroeconomy

Baker, Bloom, and Davis (QJE 2016)

For each newspaper on each day since 1985, submit the following query:

1. Article contains "uncertain" OR "uncertainty", AND
2. Article contains "economic" OR "economy", AND
3. Article contains "congress" OR "deficit" OR "federal reserve" OR "legislation" OR "regulation" OR "white house"

Normalize resulting article counts by total newspaper articles that month.



- but see Keith et al (2020), showing some problems with this measure (`https://arxiv.org/abs/2010.04706`).

WordNet

- English word database: 118K nouns, 12K verbs, 22K adjectives, 5K adverbs

The noun "bass" has 8 senses in WordNet.
1. bass[1] - (the lowest part of the musical range)
2. bass[2], bass part[1] - (the lowest part in polyphonic music)
3. bass[3], basso[1] - (an adult male singer with the lowest voice)
4. sea bass[1], bass[4] - (the lean flesh of a saltwater fish of the family Serranidae)
5. freshwater bass[1], bass[5] - (any of various North American freshwater fish with lean flesh (especially of the genus Micropterus))
6. bass[6], bass voice[1], basso[2] - (the lowest adult male singing voice)
7. bass[7] - (the member with the lowest range of a family of musical instruments)
8. bass[8] - (nontechnical name for any of numerous edible marine and freshwater spiny-finned fishes)

Figure 19.1 A portion of the WordNet 3.0 entry for the noun *bass*.

- Synonym sets (synsets) are a group of near-synonyms, plus a gloss (definition).
  - also contains information on antonyms (opposites), holonyms/meronyms (part-whole).
- Nouns are organized in categorical hierarchy (hence "WordNet")
  - "hypernym" – the higher category that a word is a member of.
  - "hyponyms" – members of the category identified by a word.

WordNet Supersenses (Word Categories)

Category	Example	Category	Example	Category	Example
ACT	*service*	GROUP	*place*	PLANT	*tree*
ANIMAL	*dog*	LOCATION	*area*	POSSESSION	*price*
ARTIFACT	*car*	MOTIVE	*reason*	PROCESS	*process*
ATTRIBUTE	*quality*	NATURAL EVENT	*experience*	QUANTITY	*amount*
BODY	*hair*	NATURAL OBJECT	*flower*	RELATION	*portion*
COGNITION	*way*	OTHER	*stuff*	SHAPE	*square*
COMMUNICATION	*review*	PERSON	*people*	STATE	*pain*
FEELING	*discomfort*	PHENOMENON	*result*	SUBSTANCE	*oil*
FOOD	*food*			TIME	*day*

Figure 19.2 Supersenses: 26 lexicographic categories for nouns in WordNet.

WordNet Supersenses (Word Categories)

Category	Example	Category	Example	Category	Example
ACT	*service*	GROUP	*place*	PLANT	*tree*
ANIMAL	*dog*	LOCATION	*area*	POSSESSION	*price*
ARTIFACT	*car*	MOTIVE	*reason*	PROCESS	*process*
ATTRIBUTE	*quality*	NATURAL EVENT	*experience*	QUANTITY	*amount*
BODY	*hair*	NATURAL OBJECT	*flower*	RELATION	*portion*
COGNITION	*way*	OTHER	*stuff*	SHAPE	*square*
COMMUNICATION	*review*	PERSON	*people*	STATE	*pain*
FEELING	*discomfort*	PHENOMENON	*result*	SUBSTANCE	*oil*
FOOD	*food*			TIME	*day*

Figure 19.2 Supersenses: 26 lexicographic categories for nouns in WordNet.

Supersense	Verbs denoting . . .
body	grooming, dressing and bodily care
change	size, temperature change, intensifying
cognition	thinking, judging, analyzing, doubting
communication	telling, asking, ordering, singing
competition	fighting, athletic activities
consumption	eating and drinking
contact	touching, hitting, tying, digging
creation	sewing, baking, painting, performing
emotion	feeling
motion	walking, flying, swimming
perception	seeing, hearing, feeling
possession	buying, selling, owning
social	political and social activities and events
stative	being, having, spatial relations
weather	raining, snowing, thawing, thundering

General Dictionaries

- Function words (e.g. *for*, *rather*, *than*)
  - also called stopwords
  - can be used to get at non-topical dimensions, identify authors.

General Dictionaries

- Function words (e.g. *for*, *rather*, *than*)
  - also called stopwords
  - can be used to get at non-topical dimensions, identify authors.
- LIWC (pronounced "Luke"): Linguistic Inquiry and Word Counts
  - 2300 words 70 lists of category-relevant words, e.g. "emotion", "cognition", "work", "family", "positive", "negative" etc.

General Dictionaries

- Function words (e.g. *for*, *rather*, *than*)
  - also called stopwords
  - can be used to get at non-topical dimensions, identify authors.
- LIWC (pronounced "Luke"): Linguistic Inquiry and Word Counts
  - 2300 words 70 lists of category-relevant words, e.g. "emotion", "cognition", "work", "family", "positive", "negative" etc.
- Mohammad and Turney (2011):
  - code 10,000 words along four emotional dimensions: joy–sadness, anger-fear, trust-disgust, anticipation-surprise
- Warriner et al (2013):
  - code 14,000 words along three emotional dimensions: valence, arousal, dominance.

Sentiment Analysis

Extract a "tone" dimension – positive, negative, neutral

- standard approach is lexicon-based, but they fail easily: e.g., "good" versus "not good" versus "not very good"

Sentiment Analysis

Extract a "tone" dimension – positive, negative, neutral

- standard approach is lexicon-based, but they fail easily: e.g., "good" versus "not good" versus "not very good"
- huggingface model hub has a number of transformer-based sentiment models

Sentiment Analysis

Extract a "tone" dimension – positive, negative, neutral

- standard approach is lexicon-based, but they fail easily: e.g., "good" versus "not good" versus "not very good"
- huggingface model hub has a number of transformer-based sentiment models
- Off-the-shelf scores may be trained on biased corpora, eg online writing – may not work for legal text, for example.
  - Hamilton et al (2016) and Zorn and Rice (2019) show how to make domain-specific sentiment lexicons using word embeddings (more on this later).

Problems with Sentiment Analyzers: NLP System Bias

```
text_to_sentiment("Let's go get Italian food")
2.0429166109
text_to_sentiment("Let's go get Chinese food")
1.4094033658
text_to_sentiment("Let's go get Mexican food")
0.3880198556
```

```
text_to_sentiment("My name is Emily")
2.2286179365
text_to_sentiment("My name is Heather")
1.3976291151
text_to_sentiment("My name is Yvette")
0.9846380213
text_to_sentiment("My name is Shaniqua")
-0.4704813178
```

Is this sentiment model racist?

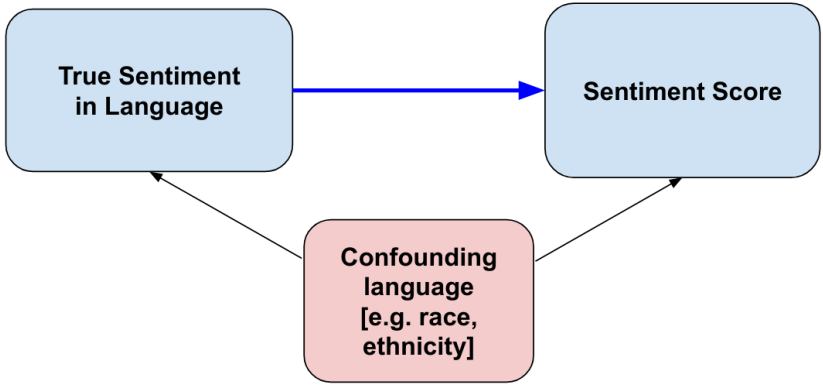
Source: Kareem Carr slides.

NLP "Bias" is statistical bias

- Sentiment scores that are trained on annotated datasets also learn from the correlated non-sentiment information.

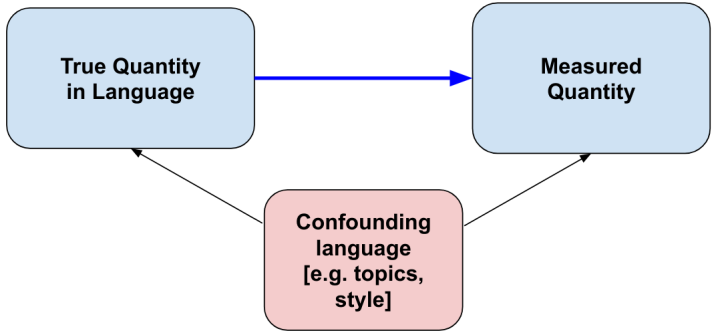
NLP "Bias" is statistical bias

- Sentiment scores that are trained on annotated datasets also learn from the correlated non-sentiment information.



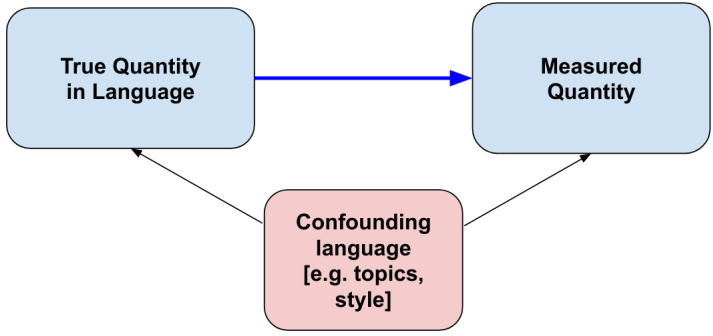
- Supervised sentiment models are confounded by correlated language factors.
  - e.g., in the training set maybe people complain about Mexican food more often than Italian food because Italian restaurants tend to be more upscale.

This is a universal problem



- supervised models (classifiers, regressors) learn features that are correlated with the label being annotated.
- unsupervised models (topic models, word embeddings) learn correlations between topics / contexts.

This is a universal problem



- supervised models (classifiers, regressors) learn features that are correlated with the label being annotated.
- unsupervised models (topic models, word embeddings) learn correlations between topics / contexts.
- **dictionary methods**, while having other limitations, mitigate this problem
  - the researcher intentionally "regularizes" out spurious confounders with the targeted language dimension.
  - helps explain why economists often still use dictionary methods.

Coding Practice and Homework Assignments

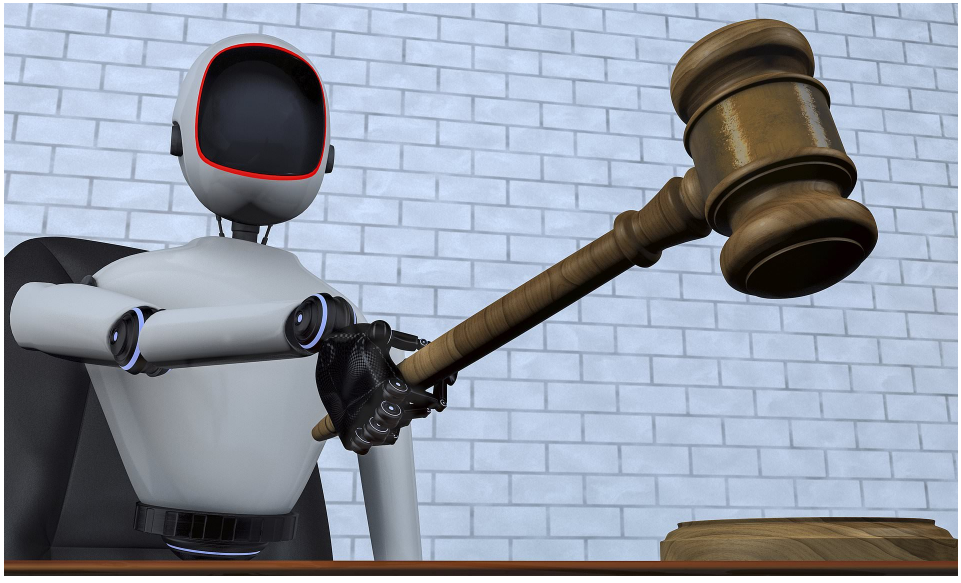
Coding Examples on GitHub:

`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.



Meeting Adjourned!