

# Sequencing Legal DNA

NLP for Law and Political Economy  
ETH Zurich, Spring 2021

**Welcome to the course!**

## **Instructions before we begin:**

- (1) Turn on video and set audio to mute
- (2) In Participants panel, set zoom name to “Full Name, School, Dept/Major”  
(ex: “Leon Smith, ETH Computer Science”)
  - (3) Say “hi” in the chat

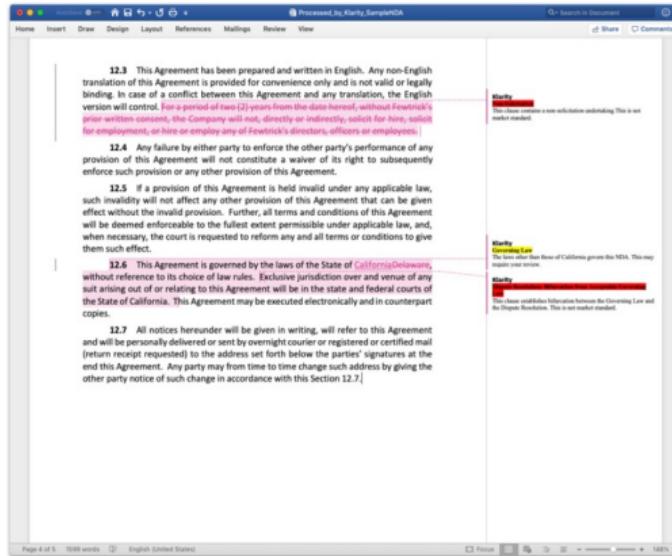
# Klarity reviews NDAs under commercial market standard.

👉 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.



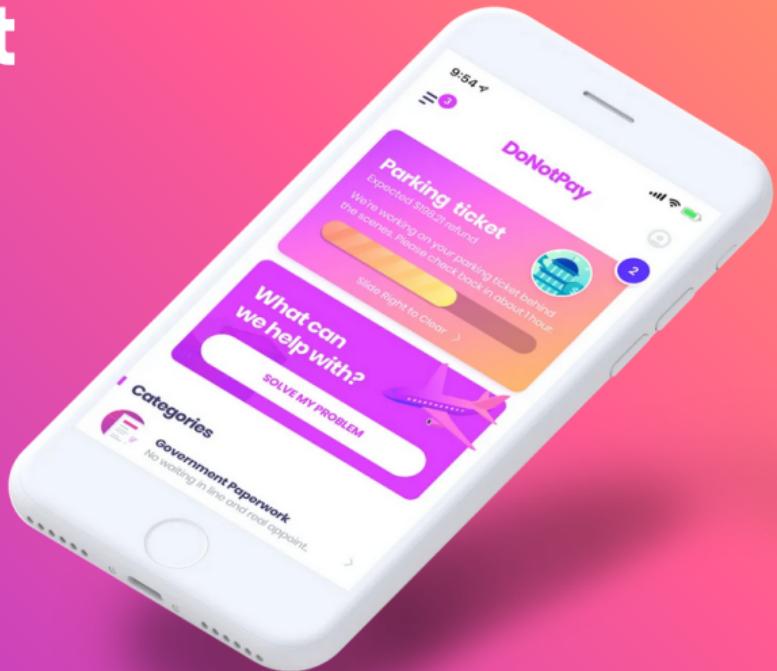
# The World's First Robot Lawyer

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.

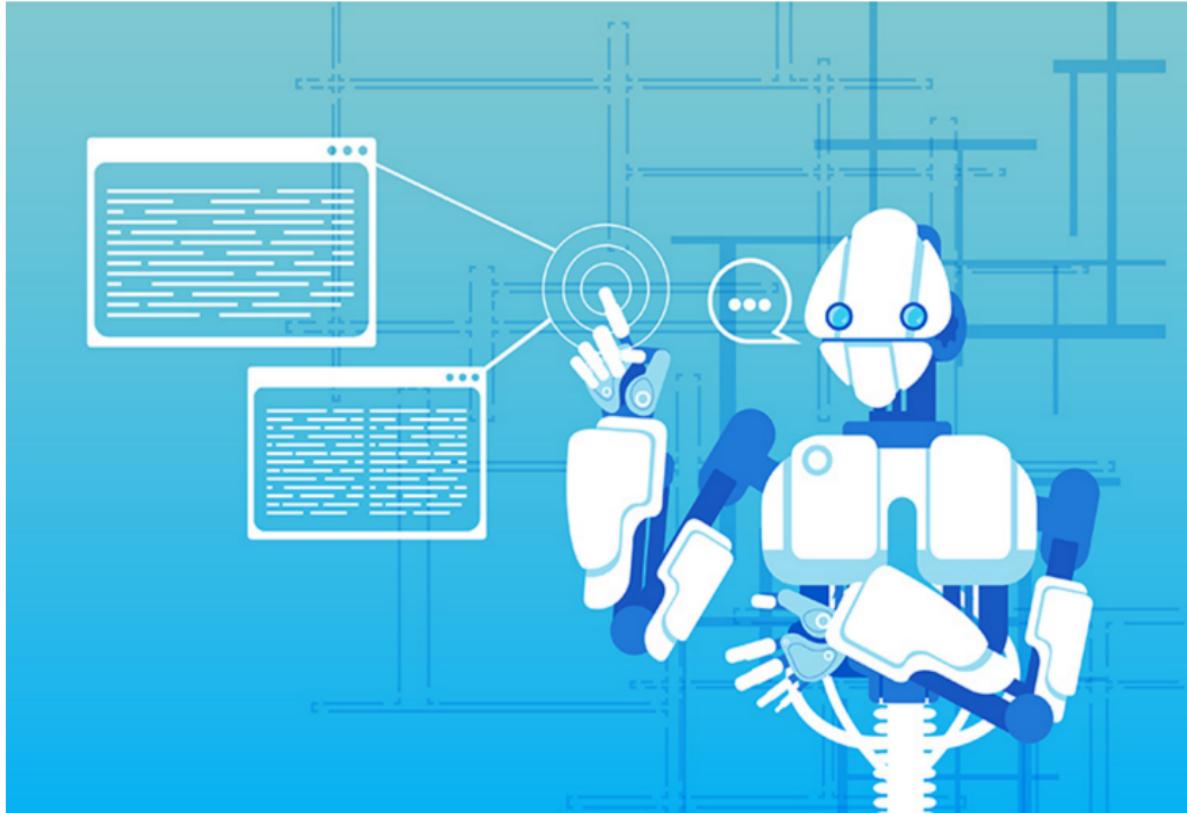
[Sign Up/Login](#)

## 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

The image shows a user interface for translating English sentences into Turkish. It features two main sections, each with a "Translate" button and language selection menus (English, Turkish, Spanish, Detect language).

**Top Section:**

- Input: She is a doctor.  
He is a nurse.
- Output: O bir doktor.  
O bir hemşire.
- Details: 31/5000 tokens used.

**Bottom Section:**

- Input: O bir doktor.  
O bir hemşire
- Output: He is a doctor.  
She is a nurse ✓
- Details: 28/5000 tokens used.

The interface includes standard translation controls like copy, save, and share, as well as audio playback icons.

Source: fastai NLP course.

# OPENAI'S NEW MULTITALENTED AI WRITES, TRANSLATES, AND SLANDERS

*A step forward in AI text-generation that also spells trouble*

By James Vincent | Feb 14, 2019, 12:00pm EST

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 to ***Sequencing Legal DNA***

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

# Welcome to ***Sequencing Legal DNA***

- ▶ This course focuses on applications of **natural language processing** in **law** and **social science**.
- ▶ Scientific goals:
  - ▶ Understand the motivations and decisions of judges and public officials through their writings and speeches.

# Welcome to ***Sequencing Legal DNA***

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

# Welcome to ***Sequencing Legal DNA***

- ▶ This course focuses on applications of **natural language processing** in **law** and **social science**.
- ▶ Scientific goals:
  - ▶ 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.
- ▶ Engineering goals:
  - ▶ Develop tools for “sequencing legal DNA” – machine interpretation and generation of legal documents.

*What we will do*

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

- Convert texts to features – words to phrases to embeddings.
- Syntactic / semantic parsing, relation extraction

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

- Convert texts to features – words to phrases to embeddings.
- Syntactic / semantic parsing, relation extraction

**2. Unsupervised learning techniques for interpreting corpora:**

- Document distance, topic models, and clustering.

- 1. Read text documents as data:**
  - Convert texts to features – words to phrases to embeddings.
  - Syntactic / semantic parsing, relation extraction
- 2. Unsupervised learning techniques for interpreting corpora:**
  - Document distance, topic models, and clustering.
- 3. Supervised learning for regression and classification:**
  - From linear models to ensembles to DNNs.
  - Model explanation methods to show what is going on inside the black box – and to better understand judicial explanations.

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

- Convert texts to features – words to phrases to embeddings.
- Syntactic / semantic parsing, relation extraction

**2. Unsupervised learning techniques for interpreting corpora:**

- Document distance, topic models, and clustering.

**3. Supervised learning for regression and classification:**

- From linear models to ensembles to DNNs.
- Model explanation methods to show what is going on inside the black box – and to better understand judicial explanations.

**4. Word embedding for identifying dimensions of language:**

- Analyze moral/legal values, ideology, and prejudice.
- Can legal language be “debiased” using embedding models?

- 1. Read text documents as data:**
  - Convert texts to features – words to phrases to embeddings.
  - Syntactic / semantic parsing, relation extraction
- 2. Unsupervised learning techniques for interpreting corpora:**
  - Document distance, topic models, and clustering.
- 3. Supervised learning for regression and classification:**
  - From linear models to ensembles to DNNs.
  - Model explanation methods to show what is going on inside the black box – and to better understand judicial explanations.
- 4. Word embedding for identifying dimensions of language:**
  - Analyze moral/legal values, ideology, and prejudice.
  - Can legal language be “debiased” using embedding models?
- 5. Legal discourse analytics**
  - Summarization / information extraction
  - Question answering / claim checking

Why did you sign up for this course?

**Zoom Poll 1.1**

# Logistics

Course Overview

Corpora

Quantity of Text as Data

Dictionary-Based Methods

Wrapping Up

## Lecture Times

- ▶ Mondays, 1215h-14h
  - ▶ Zoom
- ▶ ~10 minute break, 13h-1310h

## Online Lecture Norms

### **Let's make the most of online learning!**

- ▶ Lectures will be recorded, but live attendance is required and absences require permission of instructor.
- ▶ Keep video on if connection allows.
- ▶ Stay muted when not talking.
- ▶ To make questions or comments, use the “raise hand” function.

# Online Lecture Norms

## **Let's make the most of online learning!**

- ▶ Lectures will be recorded, but live attendance is required and absences require permission of instructor.
- ▶ Keep video on if connection allows.
- ▶ Stay muted when not talking.
- ▶ To make questions or comments, use the “raise hand” function.
- ▶ We will keep track of course participation through in-class activities (e.g. zoom polls, group work).

## Online Course Materials

- ▶ Course Syllabus:
  - ▶ <http://bit.ly/NLP-syll>
- ▶ Course Repo:
  - ▶ [https://github.com/elliottash/legal\\_dna\\_2021](https://github.com/elliottash/legal_dna_2021)

## Teaching Assistants

- ▶ Dominik Stammbach (`dominik.stammbach@gess.ethz.ch`)
- ▶ Claudia Marangon (`claudia.marangon@gess.ethz.ch`)
- ▶ TA Sessions: Fridays, 2pm-3pm
  - ▶ will go over code notebooks, homeworks, and some additional material
  - ▶ not mandatory – attend if you are new to the tools.
  - ▶ we will also post recordings.
- ▶ TA Office Hours: Mondays, 4pm-5pm
  - ▶ can answer questions about lectures and notebooks.

## Course Communication

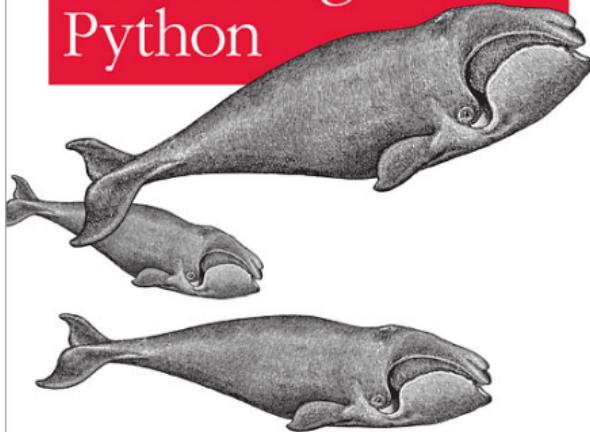
- ▶ Course communication will be done through eDoz.
- ▶ Questions welcome via email, to me or to the TA's.
- ▶ I will be available in the zoom 5 minutes early, during the mid-lecture break, and until 15 minutes after the end of lecture.
- ▶ Will schedule meetings with students doing projects.

## Syllabus has a long readings list

- ▶ There are only a handful of required readings (highlighted).
- ▶ Other readings can be used as reference:
  - ▶ to complement the slides
  - ▶ to be used for reading response essays (more next week)
  - ▶ for projects

*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



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

## Example Code is in Python

- ▶ Python 3.7 is ideal for text data and natural language processing.
  - ▶ You can use Anaconda or download the packages we need to a pip environment.
  - ▶ See the syllabus for lists of packages.
- ▶ First TA session this Friday will go over setting up python and installing needed packages.

## 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
- ▶ 10 NLP programming homework assignments,  $\sim$ 1.5 hours each  $\approx$  15 hours
- ▶ Required readings (two papers and a few short articles/snippets)  $\approx$  6 hours
- ▶ 3 response essays,  $\sim$ 6 hours each  $\approx$  18 hours
- ▶ Take-home exam, 2 hours
- ▶  **$\approx$  62 required hours.**

## Course Workload

**3 ECTS credits  $\approx$  90 hours of work**

- ▶ 12 lectures, 1.75 hours each = 21 hours
- ▶ 10 NLP programming homework assignments,  $\sim$ 1.5 hours each  $\approx$  15 hours
- ▶ Required readings (two papers and a few short articles/snippets)  $\approx$  6 hours
- ▶ 3 response essays,  $\sim$ 6 hours each  $\approx$  18 hours
- ▶ Take-home exam, 2 hours
- ▶  **$\approx$  62 required hours.**
- ▶  $\approx$  28 hours at student discretion:
  - ▶ 12 optional TA sessions, 1 hour each  $\approx$  12 hours
  - ▶ leaves  $\sim$ 16 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

## Course Projects

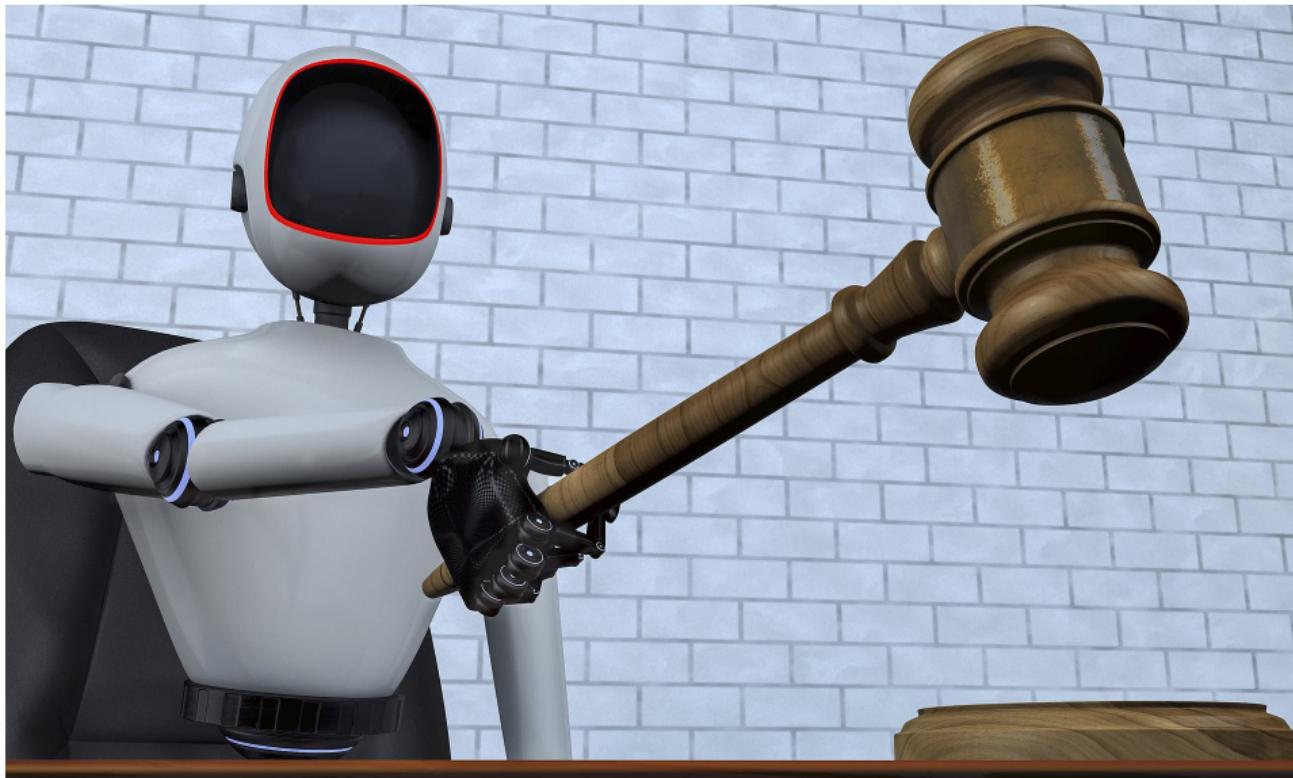
**2 additional ECTS credits  $\approx$  60 additional hours of work**

- ▶ About twice as much out-of-class work expected
  - ▶ previous course projects have turned into conference/journal publications.
  - ▶ two projects turned into funded Innouisse startups.
- ▶ Can be done individually or in small groups (preferably 2, up to 4 with good reason).

## 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 Innouisse startups.
- ▶ Can be done individually or in small groups (preferably 2, up to 4 with good reason).
- ▶ Information session after Week 2 lecture (~10 minutes)
  - ▶ we have a list of potential topics and postdoc advisors.



## Questions

Logistics

Course Overview

Corpora

Quantity of Text as Data

Dictionary-Based Methods

Wrapping Up

# 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 → GPUs → TPUs

# Big Data, Big Analytics

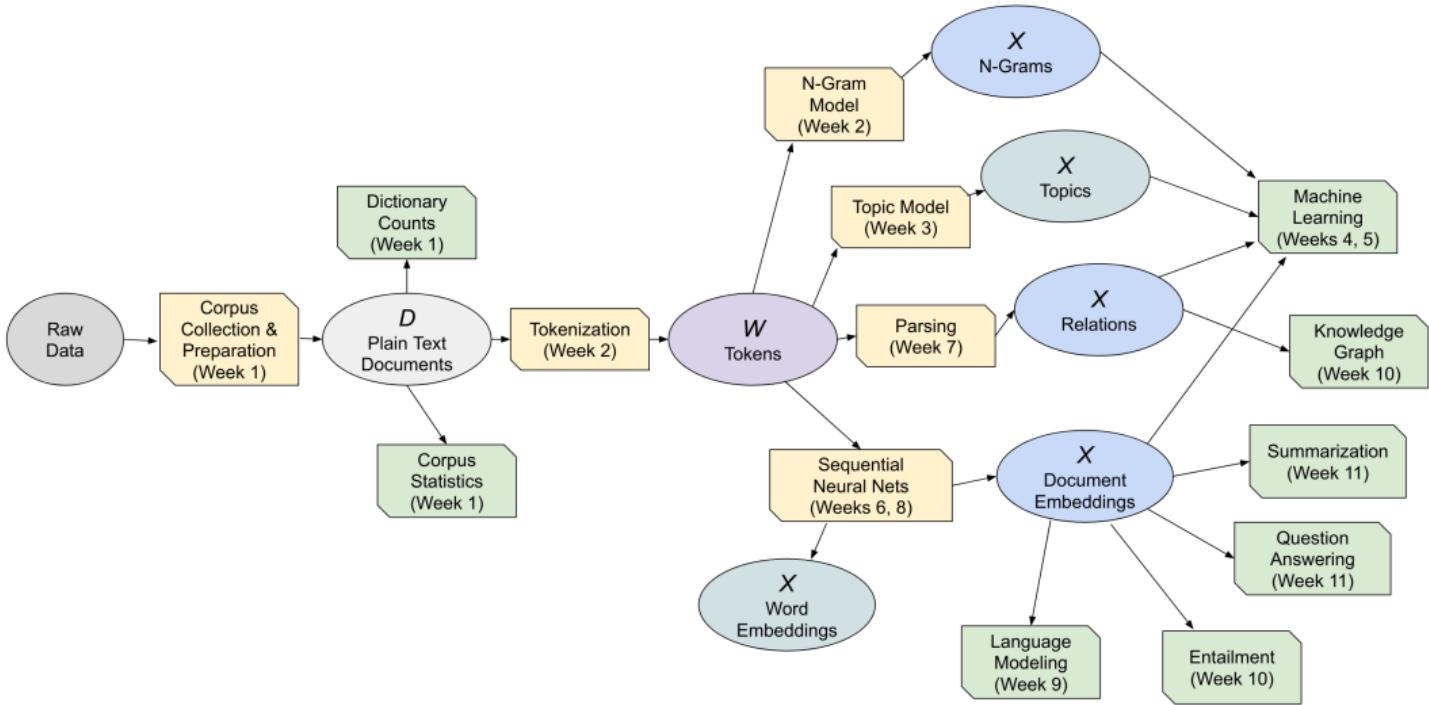
- ▶ Massive increase in availability of unstructured text datasets:
  - ▶ new social structures (the internet, email)
  - ▶ digitization efforts (govt documents, Google)
- ▶ Parallel increase in computational resources:
  - ▶ cheap disk space
  - ▶ efficient database solutions
  - ▶ compute: CPUs → GPUs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

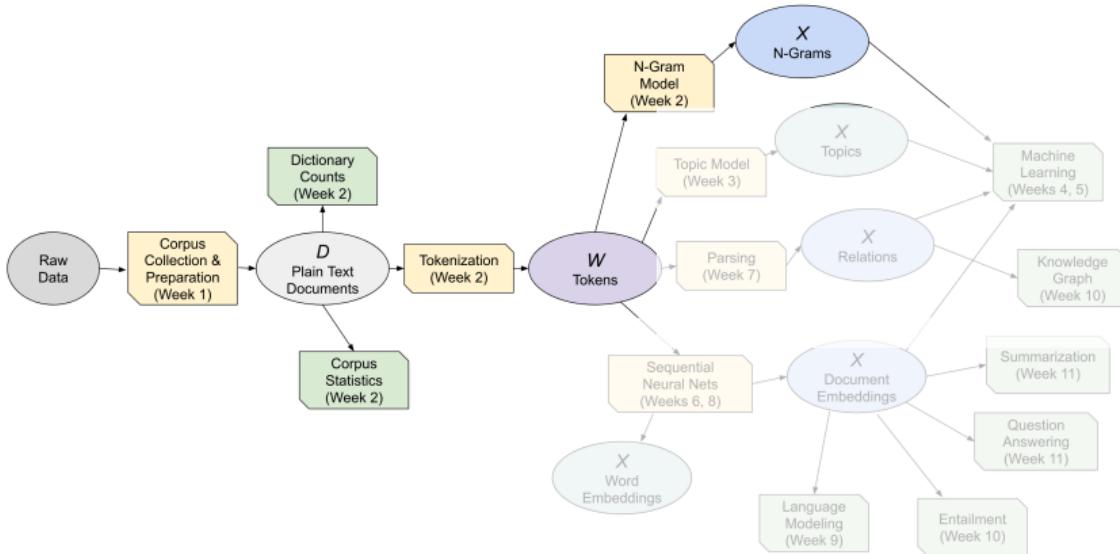
# Big Data, Big Analytics

- ▶ 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 → GPUs → TPUs
- ▶ Parallel development of tools for natural language analysis
  - ▶ text by itself is not very useful
  - ▶ machine learning, natural language processing, causal inference
- ▶ Where are these trends most salient?
  - ▶ **law and political economy**

# Big Data, Big Analytics

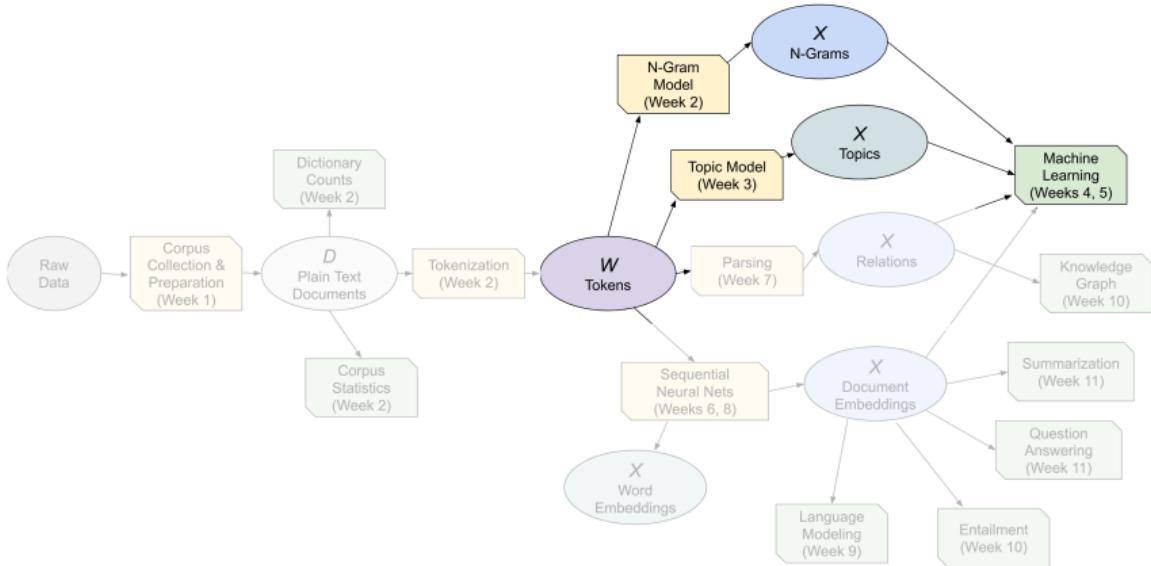
- ▶ Massive increase in availability of unstructured text datasets:
  - ▶ new social structures (the internet, email)
  - ▶ digitization efforts (govt documents, Google)
- ▶ Parallel increase in computational resources:
  - ▶ cheap disk space
  - ▶ efficient database solutions
  - ▶ compute: CPUs → GPUs → TPUs
- ▶ Parallel development of tools for natural language analysis
  - ▶ text by itself is not very useful
  - ▶ machine learning, natural language processing, causal inference
- ▶ Where are these trends most salient?
  - ▶ **law and political economy**
  - ▶ 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.





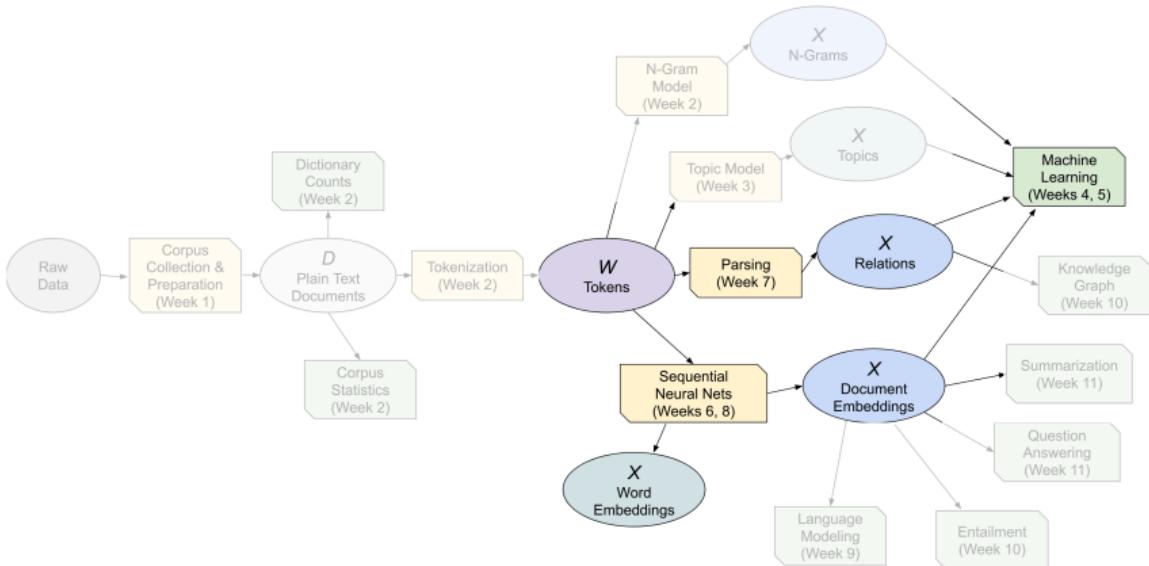
- ▶ Week 01 Introduction
  - ▶ Intro to Corpora
  - ▶ Text Complexity Methods
  - ▶ Dictionary Methods

- ▶ Week 02 Tokenization
  - ▶ N-grams
  - ▶ Word pieces
  - ▶ Parts of Speech
  - ▶ Named Entity Recognition

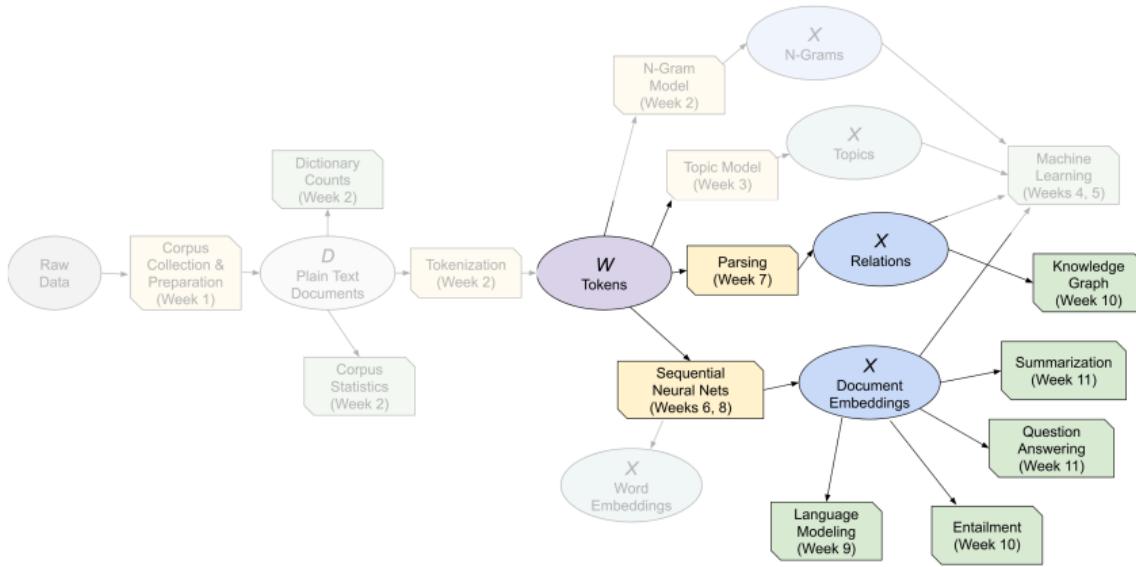


- ▶ Week 03 March 8 Dimensionality and Distance
  - ▶ SVD
  - ▶ NMF
  - ▶ Topic Models
  - ▶ Text-based ideal points

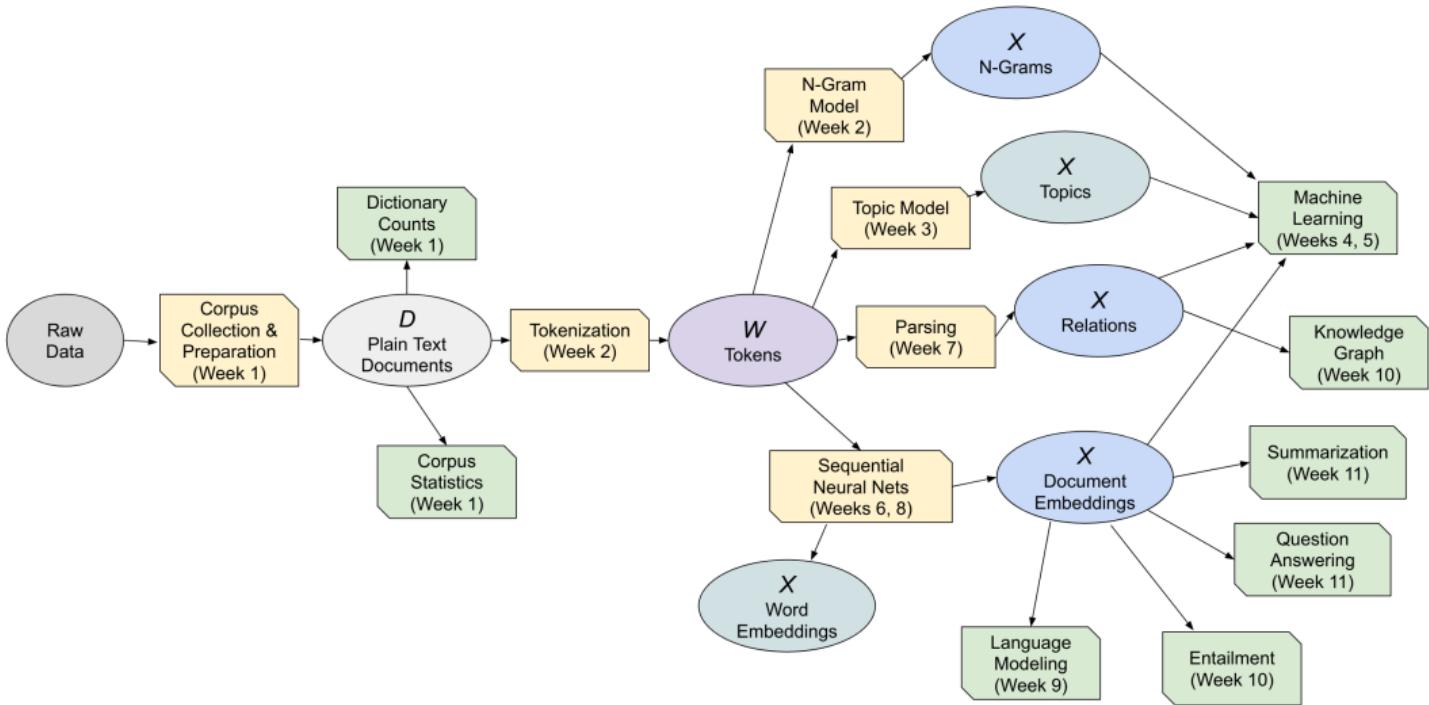
- ▶ Week 04 March 15 Machine Learning for NLP
  - ▶ text regression
  - ▶ text classification
  - ▶ elastic net, logistic regression, xgboost
- ▶ Week 05 March 22 Deep Learning for NLP
  - ▶ multilayer perceptrons
  - ▶ recurrent neural nets



- ▶ Week 06 March 29 Word Embeddings
  - ▶ word2vec, glove
- ▶ Week 07 April 12th Parsing and Relation Extraction
  - ▶ syntactic / semantic parsing
  - ▶ relation extraction
- ▶ Week 08 April 26th Transformer Embeddings
  - ▶ birectional transformer architectures (e.g. BERT, RoBERTa)
  - ▶ document embeddings



- ▶ Week 09: May 3rd Language Modeling
  - ▶ autoregressive transformers (e.g. GPT)
  - ▶ transformer autoencoders (e.g. BART)
  - ▶ conditioned generation
- ▶ Week 10: May 10th Local Semantics
  - ▶ entailment
  - ▶ stance detection
  - ▶ knowledge graphs
- ▶ Week 11: May 17th Global Semantics
  - ▶ summarization
  - ▶ question answering
  - ▶ claim checking
- ▶ Week 12: May 31st Causal Inference with Text



Logistics

Course Overview

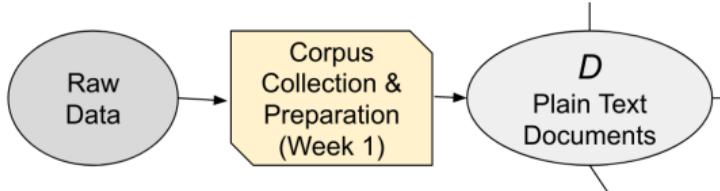
Corpora

Quantity of Text as Data

Dictionary-Based Methods

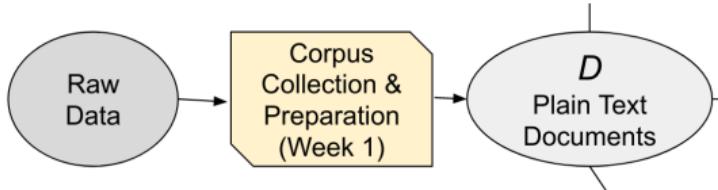
Wrapping Up

# Corpora



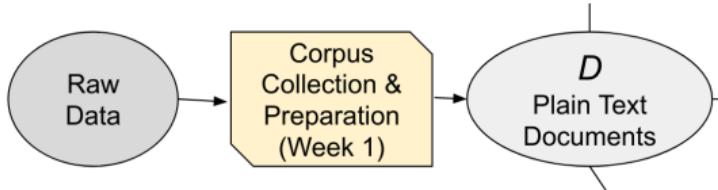
- ▶ Text data is a sequence of characters called documents.
- ▶ The set of documents is the corpus, which we will call  $D$ .

# Corpora



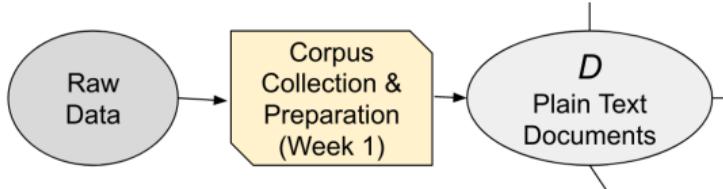
- ▶ Text data is **unstructured**:
  - ▶ the information we want is mixed together with (lots of) information we don't.
- ▶ Text data is a sequence of characters called documents.
- ▶ The set of documents is the corpus, which we will call  $D$ .

# 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  $D$ .

# 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.
- ▶ The tools from Weeks 2 (Tokenization) and 3 (Dimension Reduction) are focused on this step:
  - ▶ transforming an unstructured corpus  $D$  to a usable matrix  $X$ .
- ▶ Text data is a sequence of characters called documents.
- ▶ The set of documents is the corpus, which we will call  $D$ .

## What counts as a document?

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

## What counts as a document?

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

### **Zoom Poll 1.2**

## Handling Corpora

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

## Handling Corpora

- ▶ There is already a vast amount of data out there that has already been compiled (e.g. CourtListener, Twitter, New York Times, Reuters, Google, 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.
- ▶ All of the tools that we discuss in this class are available in many languages, and machine translation is now extremely good and automatable.

Logistics

Course Overview

Corpora

Quantity of Text as Data

Dictionary-Based Methods

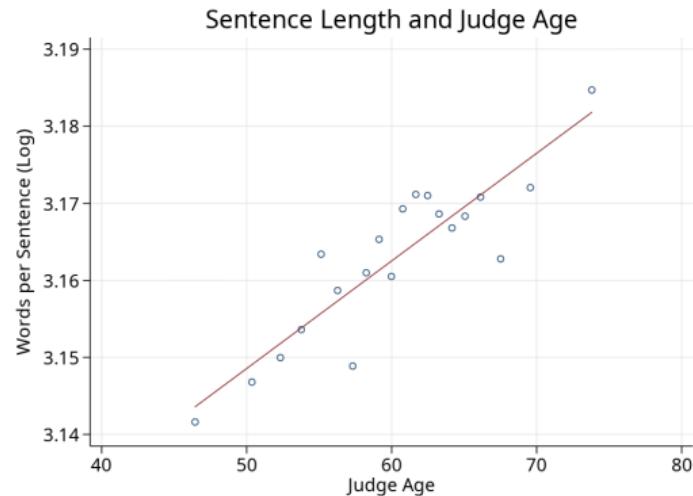
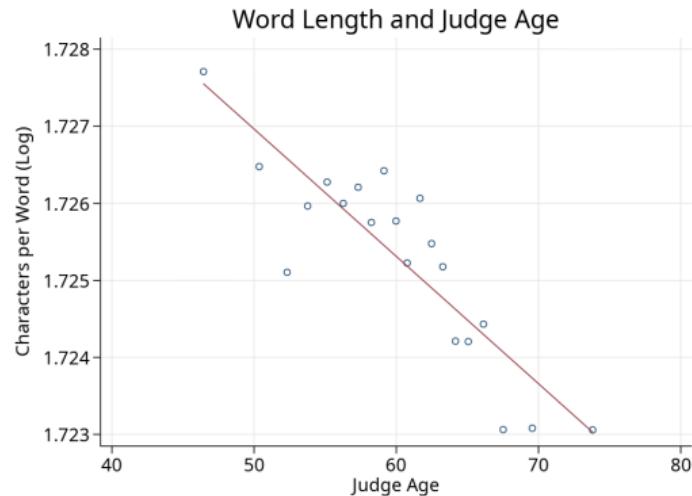
Wrapping Up

# Judge Age and Writing Style

Ash and MacLeod (2021)

# Judge Age and Writing Style

Ash and MacLeod (2021)



# Optimal legal complexity

Katz and Bommarito (2014)

## Optimal legal complexity

Katz and Bommarito (2014)

- ▶ More detail is needed in law to properly target incentives to activities and groups.
  - ▶ but there are costs to understanding/following complex laws, so there is a trade off.

# Optimal legal complexity

Katz and Bommarito (2014)

- ▶ More detail is needed in law to properly target incentives to activities and groups.
  - ▶ but there are costs to understanding/following complex laws, so there is a trade off.
- ▶ Analyzing this issue empirically requires a measure of complexity/detail.
  - ▶ e.g., compressing a text file with gzip, and then taking the ratio of the compressed file size to decompressed filesize, gives a measure of entropy or structure in the document.

# Number of Clauses ≈ Number of Words

Measuring Complexity (Katz and Bommarito 2014)

Five largest and smallest titles by structural size

Title	V
Public Health and Welfare (Title 42)	110,605
Internal Revenue Code (Title 26)	51,553
Conservation (Title 16)	33,062
Agriculture (Title 7)	29,191
Education (Title 20)	28,096
Arbitration (Title 9)	68
General Provisions (Title 1)	84
Flag and Seal, Seat of Government, and the States (Title 4)	221
Intoxicating Liquors (Title 27)	224
Census (Title 13)	272

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

Logistics

Course Overview

Corpora

Quantity of Text as Data

Dictionary-Based Methods

Wrapping Up

## 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)

- ▶ Baker, Bloom, and Davis measure economic policy uncertainty using Boolean search of newspaper articles.

# Measuring uncertainty in macroeconomy

Baker, Bloom, and Davis (QJE 2016)

- ▶ Baker, Bloom, and Davis measure economic policy uncertainty using Boolean search of newspaper articles.
- ▶ 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”

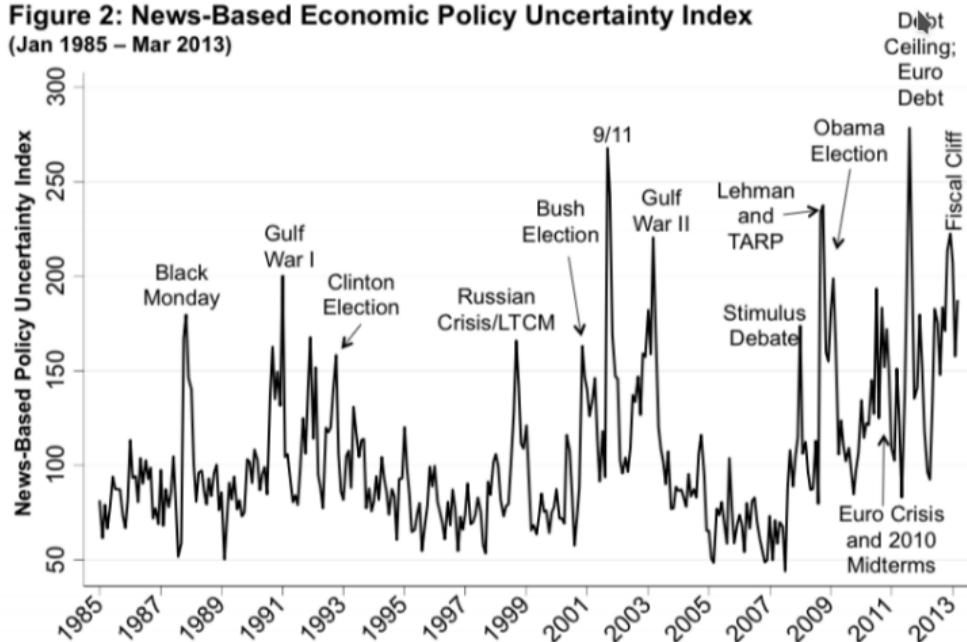
# Measuring uncertainty in macroeconomy

Baker, Bloom, and Davis (QJE 2016)

- ▶ Baker, Bloom, and Davis measure economic policy uncertainty using Boolean search of newspaper articles.
- ▶ 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

**Figure 2: News-Based Economic Policy Uncertainty Index**  
(Jan 1985 – Mar 2013)



## WordNet

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

# WordNet

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

The noun “bass” has 8 senses in WordNet.

1. bass<sup>1</sup> - (the lowest part of the musical range)
2. bass<sup>2</sup>, bass part<sup>1</sup> - (the lowest part in polyphonic music)
3. bass<sup>3</sup>, basso<sup>1</sup> - (an adult male singer with the lowest voice)
4. sea bass<sup>1</sup>, bass<sup>4</sup> - (the lean flesh of a saltwater fish of the family Serranidae)
5. freshwater bass<sup>1</sup>, bass<sup>5</sup> - (any of various North American freshwater fish with lean flesh (especially of the genus Micropterus))
6. bass<sup>6</sup>, bass voice<sup>1</sup>, basso<sup>2</sup> - (the lowest adult male singing voice)
7. bass<sup>7</sup> - (the member with the lowest range of a family of musical instruments)
8. bass<sup>8</sup> - (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).

## WordNet Supersenses (Word Categories)

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

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

# WordNet Supersenses (Word Categories)

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

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

## WordNet can be used for shrinking vocabulary

e.g., for replacing rare words

- ▶ Replace rare words with more frequent word from same synset.

## WordNet can be used for shrinking vocabulary

e.g., for replacing rare words

- ▶ Replace rare words with more frequent word from same synset.
- ▶ Can use hierarchies and replace words with higher-level category:
  - ▶ e.g., ancestors of “dog” include “carnivore”, “mammal”, “vertebrate”, “animal”, and “physical entity”

## WordNet can be used for shrinking vocabulary

e.g., for replacing rare words

- ▶ Replace rare words with more frequent word from same synset.
- ▶ Can use hierarchies and replace words with higher-level category:
  - ▶ e.g., ancestors of “dog” include “carnivore”, “mammal”, “vertebrate”, “animal”, and “physical entity”
- ▶ For extreme dimension reduction, replace words with the supersense.

## General Dictionaries

- ▶ 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.
- ▶ Emotion lexicons:
  - ▶ Mohammad and Turney (2011) code 10,000 words along four dimensions: joy–sadness, anger–fear, trust–disgust, anticipation–surprise
  - ▶ Warriner et al (2013) code 14,000 words along three dimensions: valence, arousal, dominance.
- ▶ Function words (e.g. *for*, *rather*, *than*)
  - ▶ e.g. LIWC’s “funct” category.
  - ▶ can be used to get at non-topical dimensions, identify authors.

## Sentiment Analysis

- ▶ Extract a “tone” dimension – positive, negative neutral
  - ▶ lexicon-based solutions fail: e.g., “good” versus “not good” versus “not very good”

## Sentiment Analysis

- ▶ Extract a “tone” dimension – positive, negative neutral
  - ▶ lexicon-based solutions fail: e.g., “good” versus “not good” versus “not very good”
  - ▶ flair’s pre-trained sentiment model uses a context-sensitive neural net

## Sentiment Analysis

- ▶ Extract a “tone” dimension – positive, negative neutral
  - ▶ lexicon-based solutions fail: e.g., “good” versus “not good” versus “not very good”
  - ▶ flair’s pre-trained sentiment model uses a context-sensitive neural net
- ▶ Designed for online writing – hard to say how well it works on legal text, for example.
  - ▶ Hamilton et al (2016) provide a method for making domain-specific sentiment lexicons using word embeddings (more on this later).

Logistics

Course Overview

Corpora

Quantity of Text as Data

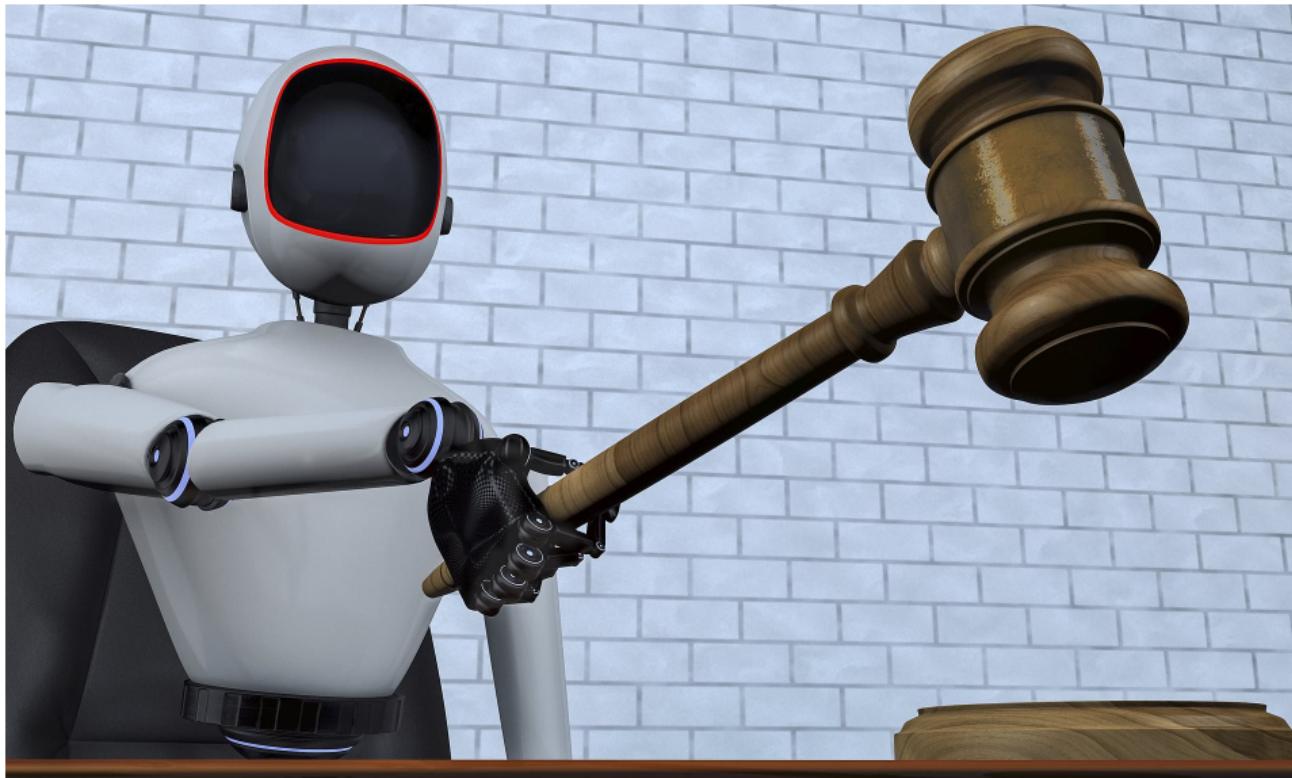
Dictionary-Based Methods

Wrapping Up

# First Homework Assignment

Homework Assignments Page: <http://bit.ly/NLP-HW>

- ▶ First homework: practice with some of the tools we already mentioned in today's lecture.
- ▶ to be uploaded via EduFlow on Moodle
- ▶ More info in the example code notebook, and in the first TA session on Friday.



**Meeting Adjourned!**