Sequencing Legal DNA

NLP for Law and Political Economy

1. Course Overview and Introduction

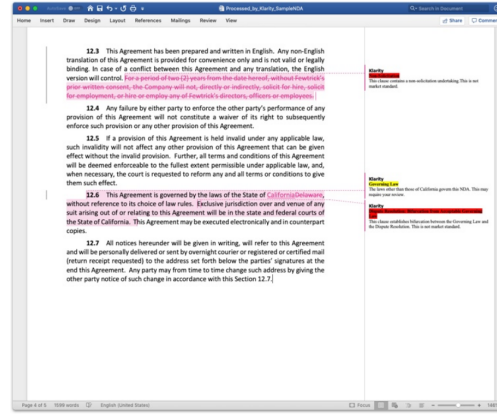
Klarity reviews NDAs under commercial market standard.

👌 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

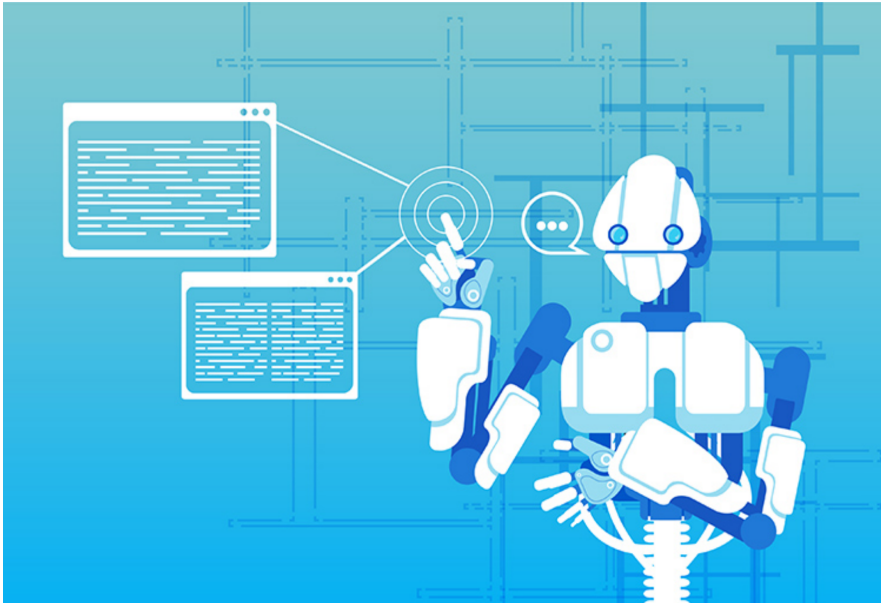
The DoNotPay app is the home of the world's first robot lawyer. Fight corporations, beat bureaucracy and sue anyone at the press of a button.

Sign Up/Login

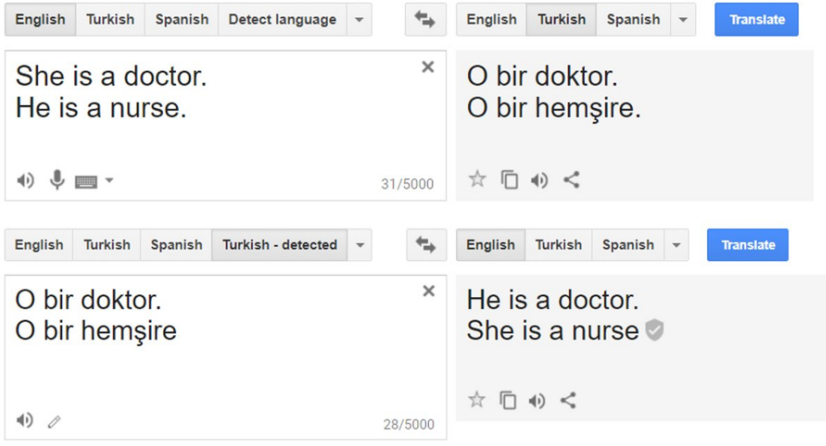
THINGS YOU CAN DO WITH DONOTPAY

- 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 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.
   - Automated legal annotation using relation extraction and knowledge graphs.

1. **Read text documents as data:**
   - Convert texts to features – words to phrases to embeddings.
   - Automated legal annotation using relation extraction and knowledge graphs.
2. **Unsupervised learning techniques for interpreting corpora:**
   - Matrix factorization, topic models, and clustering.
   - Analyze citation networks to understand flows of ideas.

1. **Read text documents as data:**
   - Convert texts to features – words to phrases to embeddings.
   - Automated legal annotation using relation extraction and knowledge graphs.
2. **Unsupervised learning techniques for interpreting corpora:**
   - Matrix factorization, topic models, and clustering.
   - Analyze citation networks to understand flows of ideas.
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.
   - Automated legal annotation using relation extraction and knowledge graphs.
2. **Unsupervised learning techniques for interpreting corpora:**
   - Matrix factorization, topic models, and clustering.
   - Analyze citation networks to understand flows of ideas.
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 isolating 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.
   - Automated legal annotation using relation extraction and knowledge graphs.
2. **Unsupervised learning techniques for interpreting corpora:**
   - Matrix factorization, topic models, and clustering.
   - Analyze citation networks to understand flows of ideas.
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 isolating dimensions of language:**
   - Analyze moral/legal values, ideology, and prejudice.
   - Can legal language be "debiased" using embedding models?
5. **Legal text generation**
   - What about a chatbot to provide legal advice?
   - Or a robot judge – taking facts, making a decision, and writing a legal opinion explaining it?

Outline

Lecture Times

- Mondays, 1:15pm-3pm
  - LFW C5
- ~10 minute break, 2pm-2:10pm

Online Course Materials

- Course Syllabus:
  - `https://bit.ly/2sNwBS7`
- Course Repo:
  - `https://github.com/elliottash/legal_dna_2020`

Teaching Assistants

- Claudia Marangon (`claudia.marangon@gess.ethz.ch`)
  - will assist with grading homework assignments.
  - can answer questions about lectures and notebooks.
- Selina Lorusso (`selina.lorusso@gess.ethz.ch`)
  - will assist with administrative side.
  - can help with logistical issues such as due dates, submitting assignments, and using the message forum.

Course Message Forum

- Course communication will be done through the course message forum (as feasible):

  `https://robot-judge.com/forums/forum/sequencing-legal-dna/`

- Students should register for the forum and provide the username in the course survey.
  - If you have not taken the survey (including auditors), it is available here: `https://forms.gle/p8inFY6bnhwtzQTE8`.

Course Message Forum

- Course communication will be done through the course message forum (as feasible):

  `https://robot-judge.com/forums/forum/sequencing-legal-dna/`

- Students should register for the forum and provide the username in the course survey.
  - If you have not taken the survey (including auditors), it is available here: `https://forms.gle/p8inFY6bnhwtzQTE8`.
- In particular:
  - Course announcements will be posted as threads on the Announcements Forum – all students should subscribe there.

Course Message Forum

- Course communication will be done through the course message forum (as feasible):

  `https://robot-judge.com/forums/forum/sequencing-legal-dna/`

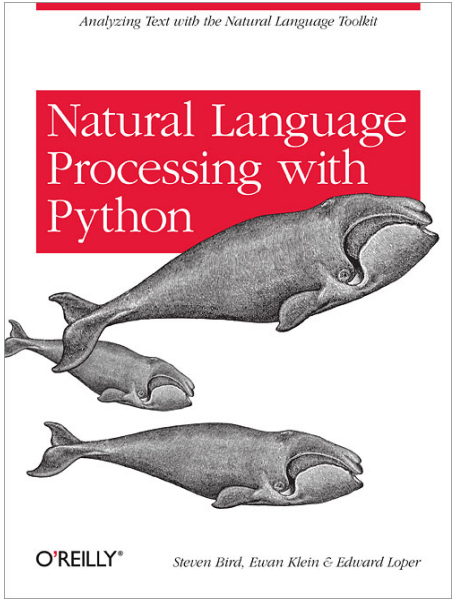
- Students should register for the forum and provide the username in the course survey.
  - If you have not taken the survey (including auditors), it is available here: `https://forms.gle/p8inFY6bnhwtzQTE8`.
- In particular:
  - Course announcements will be posted as threads on the Announcements Forum – all students should subscribe there.
  - Forum should be used for questions about course content or technical issues.

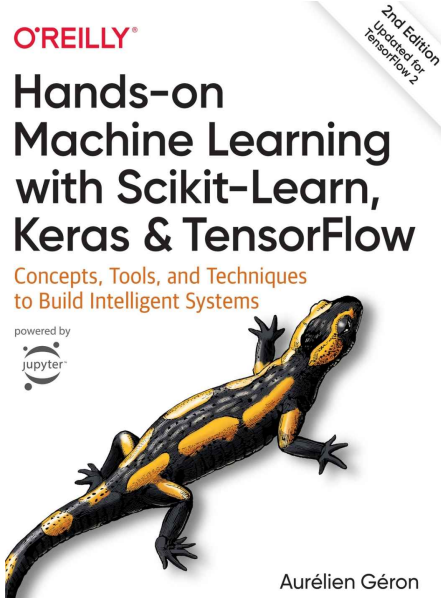
Course Participation Requirement

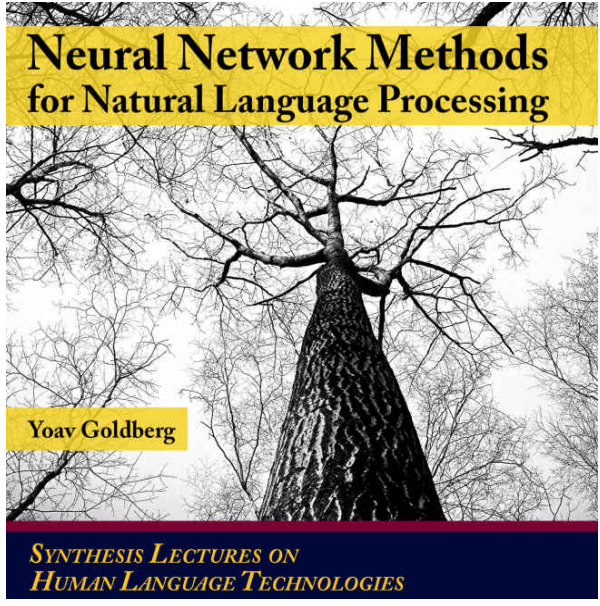
- Some participation on the course forum is required. Recommended:
  - log in to read some posts at least once a week (preferably the morning of lecture days)
  - up-vote or down-vote at least two posts per week
  - Comment on at least one post per week.

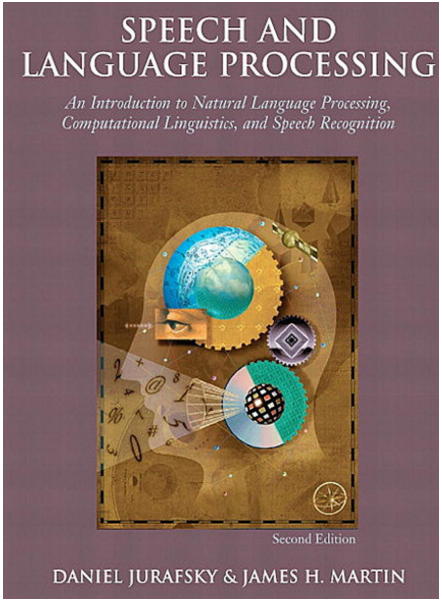
Syllabus has a long readings list

- None of the readings are mandatory
  - (except for the purposes of the response essays, to be discussed shortly).
- Material in the slides is based on the readings so they can be used as an additional reference.


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

Example Code is in Python

- Python 3.7 is ideal for text data and machine learning.
  - You can use Anaconda or download the packages we need to a pip environment.
- See the syllabus for lists of packages.
- If you prefer to use a different programming language, let me know.

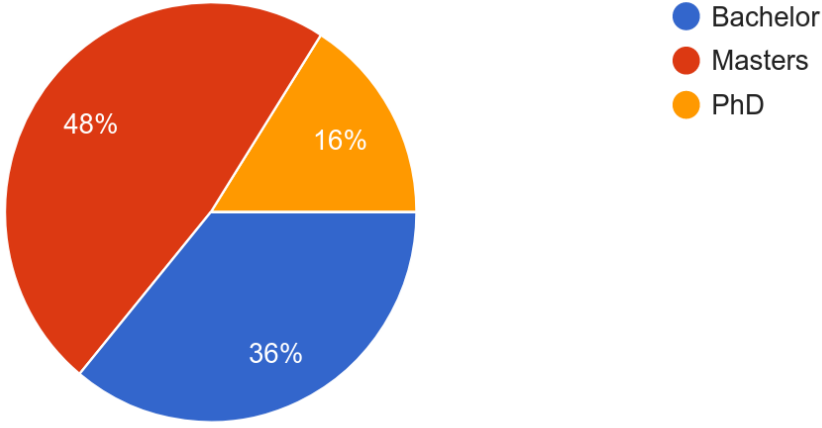
Homework Assignments

- There are three homework assignments
  - upload to assignment dropbox (see syllabus), first one is due March 11th
- More details next week.

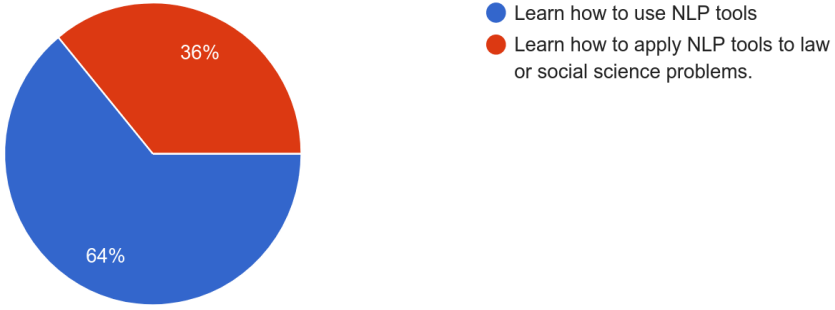
Office Hours Etc

- I will be available to meet after lectures.
- Can also set up appointments by email: `ashe@ethz.ch`.
- We will have two rounds of meetings during the term to discuss the course projects.

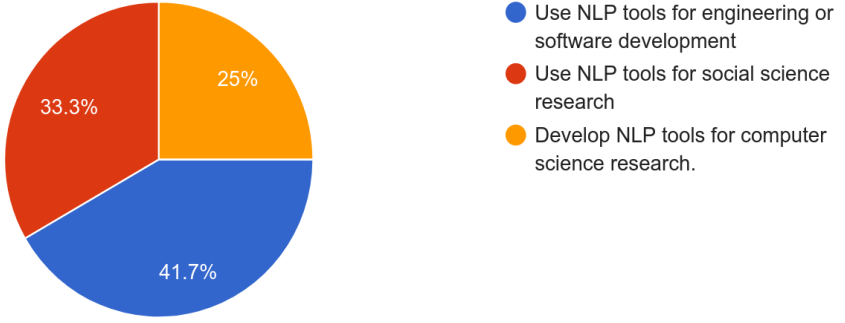
Class Survey Results



Class Survey Results



Class Survey Results



Topics Timeline

Week 01 Feb. 17	Course Overview and Text Data Essentials
Week 02 Feb. 24	Tokens and N-Grams
Week 03 March 2	Document Distance and Topic Models
Week 04 March 9	Supervised Learning with Text
Week 05 March 16	Neural Nets and Word Embeddings
Week 06 March 23	Word Embeddings and “Bias” in Language
Week 07 March 30	Syntactic Parsers and Semantic Role Labeling
Week 08 April 6	Convolutions, Recurrence, and Attention
April 13 & 20	(Spring Break)
Week 09: April 27	Document Embeddings
Week 10: May 4	Text Generators
Week 11: May 11	Causal Inference with Text Data
Week 12: May 18	Information Extraction and Knowledge Graphs

Course Projects

- The main course product is a research paper.
  - Can be done individually or in small groups (preferably 2, up to 4 with good reason).
  - Do an original app or analysis using methods learned in the course.

Course Projects

- The main course product is a research paper.
  - Can be done individually or in small groups (preferably 2, up to 4 with good reason).
  - Do an original app or analysis using methods learned in the course.
- First deliverable: 1/2 page description of topic (March 30th)
- More info in Week 3.

Course Projects

- The main course product is a research paper.
  - Can be done individually or in small groups (preferably 2, up to 4 with good reason).
  - Do an original app or analysis using methods learned in the course.
- First deliverable: 1/2 page description of topic (March 30th)
- More info in Week 3.
- 2 extra credits available for course project (5 credits total):
  - about twice as much work expected.
  - should be aimed at publication in a conference/journal, or development of a useful piece of software.

Outline

Big Data, Big Analytics

- Massive increase in unstructured text due to:
  - new social structures (the internet, email)
  - new/improved data collection
  - digitization efforts (govt documents, Google)

Big Data, Big Analytics

- Massive increase in unstructured text due to:
  - new social structures (the internet, email)
  - new/improved data collection
  - digitization efforts (govt documents, Google)
- Tools to analyze text advancing in parallel
  - text by itself is not very useful
  - machine learning, natural language processing, causal inference

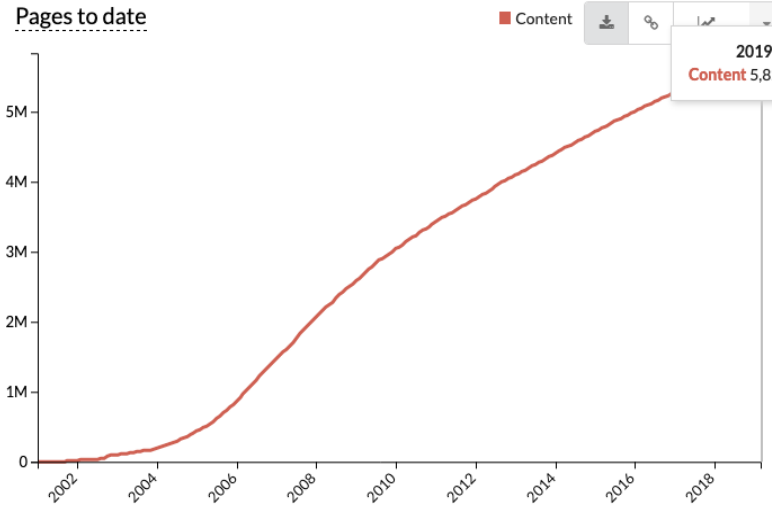
Big Data, Big Analytics

- Massive increase in unstructured text due to:
  - new social structures (the internet, email)
  - new/improved data collection
  - digitization efforts (govt documents, Google)
- Tools to analyze text advancing in parallel
  - text by itself is not very useful
  - machine learning, natural language processing, causal inference
- Where are these trends most salient?
  - **law** and **political economy**

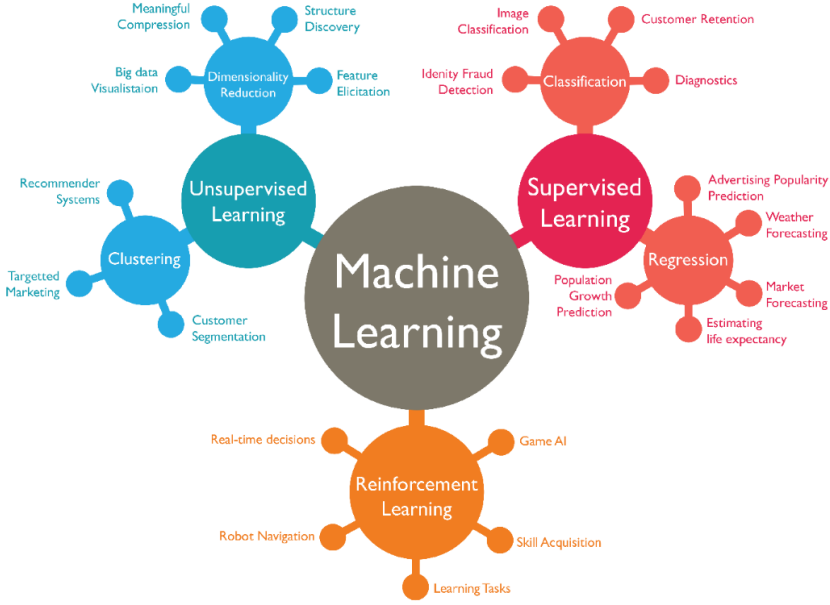
Big Data, Big Analytics

- Massive increase in unstructured text due to:
  - new social structures (the internet, email)
  - new/improved data collection
  - digitization efforts (govt documents, Google)
- Tools to analyze text advancing in parallel
  - 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.

of Wikipedia Pages, 2001-2019



The Machine Learning Landscape



Diversification of Text Data Methods

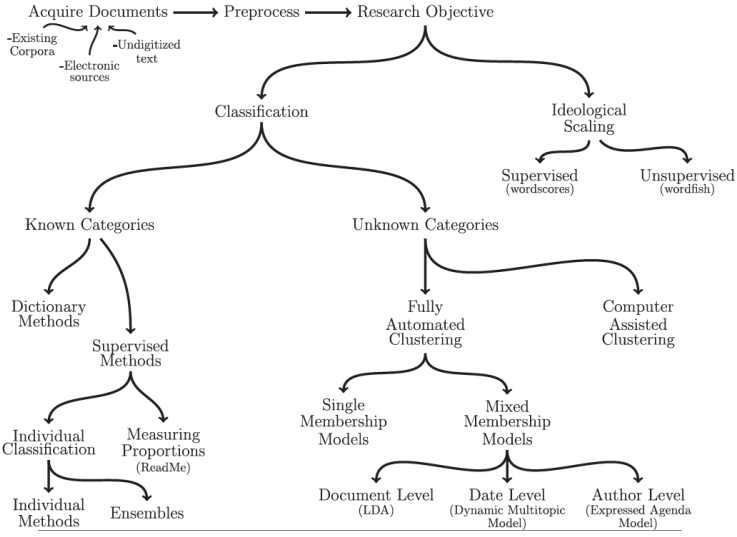


Fig. 1 An overview of text as data methods.

Source: Stewart and Grimmer (2013).

All Quantitative Models of Language are Wrong — Some are Useful

Grimmer & Stewart 2013

- Data generation process for text is unknown.
- Language is complex:
  - "Time flies like an arrow; fruit flies like a banana."

All Quantitative Models of Language are Wrong — Some are Useful

Grimmer & Stewart 2013

- Data generation process for text is unknown.
- Language is complex:
  - "Time flies like an arrow; fruit flies like a banana."
- Models necessarily fail to capture language, but may be useful for specific tasks
  - there is no globally best method – depends on question and context.

All Quantitative Models of Language are Wrong — Some are Useful

Grimmer & Stewart 2013

- Data generation process for text is unknown.
- Language is complex:
  - "Time flies like an arrow; fruit flies like a banana."
- Models necessarily fail to capture language, but may be useful for specific tasks
  - there is no globally best method – depends on question and context.
- Few theories to help – have to *validate* that a method works.

Outline

Overview

- Text data is a sequence of characters called **documents**.
- The set of documents is the **corpus**.

Overview

- Text data is a sequence of characters called **documents**.
- The set of documents is the **corpus**.
- Text data is **unstructured**:
  - the information we want is mixed together with (lots of) information we don't.
  - How to separate the two?

Overview

- Text data is a sequence of characters called **documents**.
- The set of documents is the **corpus**.
- Text data is **unstructured**:
  - the information we want is mixed together with (lots of) information we don't.
  - How to separate the two?
- All text data approaches will throw away some information:
  - The trick is figuring out how to retain valuable information.

Documents and metadata

- For small corpora, you might have the text and metadata together in a spreadsheet.
- For larger corpora, you might have:
  - A document is a text file (or an item in a relational database).
  - A corpus is a folder of text files.
  - The filenames for the text files should contain an identifier for linking to metadata.

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.
- Looking at how judges decide different types of cases $\rightarrow$ a case would be a document.

What counts as a document?

- The unit of analysis (the "document") will vary depending on your question.
- Looking at how judges decide different types of cases $\rightarrow$ a case would be a document.
- Looking at how judges differ within a court $\rightarrow$ you might aggregate all of a judge's cases as a document.

What counts as a document?

- The unit of analysis (the "document") will vary depending on your question.
- Looking at how judges decide different types of cases $\rightarrow$ a case would be a document.
- Looking at how judges differ within a court $\rightarrow$ you might aggregate all of a judge's cases as a document.
- Looking for factual vs legal statements within a case $\rightarrow$ a document might be a section, paragraph, or sentence.

Outline

Publicly Available 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).
- Chris Bail curates a list of these corpora:
  - `https://docs.google.com/spreadsheets/d/1I7cvuCBQxosQK2evTcdL3qtglaEPc0WFEs6rZMx-xiE/edit`

Publicly Available 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).
- Chris Bail curates a list of these corpora:
  - `https://docs.google.com/spreadsheets/d/1I7cvuCBQxosQK2evTcdL3qtglaEPc0WFEs6rZMx-xiE/edit`
- Many proprietary corpora are becoming available for research:
  - Lexis
  - Web of Science

Web Scraping

- A web scraper is a computer program that:
  - loads/reads in a web page
  - finds some information on it
  - grabs the information
  - stores it in a dataset

Web Scraping

- A web scraper is a computer program that:
  - loads/reads in a web page
  - finds some information on it
  - grabs the information
  - stores it in a dataset
- Once upon a time you could collect virtually any piece of information from the internet by screen scraping.
  - But now web sites make it difficult with restrictive terms of use, bot-blockers, javascript, etc.
  - Still, a little creativity (and selenium) goes a long way.

What a web site looks like to us

What a web site looks like to a computer

```
<!DOCTYPE html>
<html lang="en" dir="ltr" class="client-nojs">
<head>
<meta charset="UTF-8" />
<title>World Health Organization ranking of health systems in 2000 - Wikipedia, the free encyclopedia</title>
<meta name="generator" content="MediaWiki 1.26wmf10" />
<link rel="alternate" href="android-app://org.wikipedia/http/en.m.wikipedia.org/wiki/World_Health_Organization_ranking_of_health_systems_in_2000" />
<link rel="alternate" type="application/x-wiki" title="Edit this page" href="/w/index.php?title=World_Health_Organization_ranking_of_health_systems_in_2000&amp;action=edit" />
<link rel="edit" title="Edit this page" href="/w/index.php?title=World_Health_Organization_ranking_of_health_systems_in_2000&amp;action=edit" />
<link rel="apple-touch-icon" href="/static/apple-touch/wikipedia.png" />
<link rel="shortcut icon" href="/static/favicon/wikipedia.ico" />
<link rel="search" type="application/opensearchdescription+xml" href="/w/opensearch_desc.php" title="Wikipedia (en)" />
<link rel="EditURI" type="application/rsd+xml" href="//en.wikipedia.org/w/api.php?action=rsd" />
<link rel="alternate" hreflang="x-default" href="/wiki/World_Health_Organization_ranking_of_health_systems_in_2000" />
<link rel="copyright" href="//creativecommons.org/licenses/by-sa/3.0/" />
<link rel="alternate" type="application/atom+xml" title="Wikipedia Atom feed" href="/w/index.php?title=Special:RecentChanges&amp;feed=atom" />
<link rel="canonical" href="https://en.wikipedia.org/wiki/World_Health_Organization_ranking_of_health_systems_in_2000" />
<link rel="stylesheet" href="//en.wikipedia.org/w/load.php?debug=false&amp;lang=en&amp;modules=ext.uls.nojs%7Cext.visualEditor.viewPageTarget.noscript%7Cext.wikihiero%7Cmediawiki.legacy.commonPrint%2Cshared%7Cmediawiki.sectionAnchor%7Cmediawiki.skinning.interface%7Cmediawiki.ui.button%7Cskins.vector.styles%7Cwikibase.client.init&amp;only=styles&amp;skin=vector&amp;*" />
<meta name="ResourceLoaderDynamicStyles" content="" />
<link rel="stylesheet" href="//en.wikipedia.org/w/load.php?debug=false&amp;lang=en&amp;modules=site&amp;only=styles&amp;skin=vector&amp;*" />
https://en.wikipedia.org/w/index.php?title=World_Health_Organization a:lang(mzn),a:lang(ps),a:lang(ur){text-decoration:none}
```

Browser Automation

- Many web sites are designed to be difficult to scrape.
- Python has nice solutions for simulating a human browser:
  - selenium (chromedriver, phantomjs)

API's

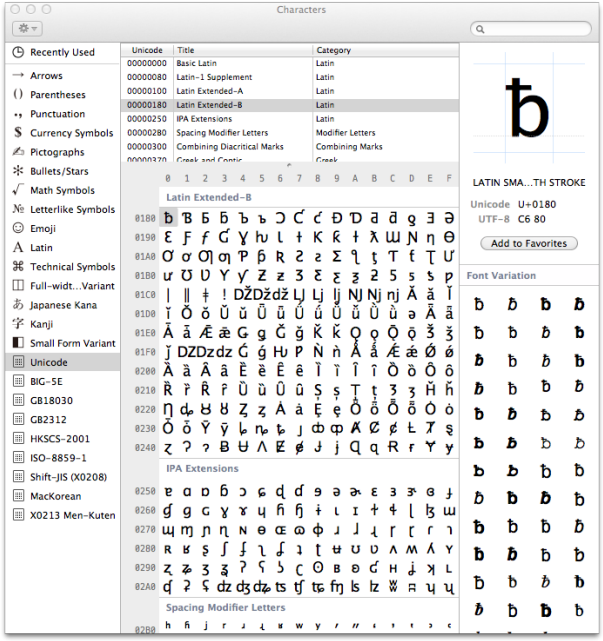
- API = Application Programming Interface
  - These are developer-oriented tools that provide access to cleaner data.
- Chris Bail's list of API's that could be interesting for research:
  - `https://docs.google.com/spreadsheets/d/1ZEr3okdlb0zctmX0MZKo-gZKPsq5WGn1nJOxPV7al-Q/edit`

Other Languages

- All of the tools that we discuss in this class are available in many languages.
- spaCy has full functionality in English, German, Spanish, Portuguese, French, Italian, and Dutch.
  - beta functionality in dozens of other languages including Chinese and Arabic
  - See `https://spacy.io/usage/models`.
- Can also translate (e.g., googletrans gives link to google API).
- The machine learning models are language-independent.

Outline

Character Encodings



Corpus cleaning

- What we usually do:
  - remove HTML markup, extra white space, and unicode

Corpus cleaning

- What we usually do:
  - remove HTML markup, extra white space, and unicode
- But HTML markup is often valuable:
  - HTML markup for section header names.
  - Legal database web sites often have HTML tags for citations to other cases.

Corpus cleaning

- What we usually do:
  - remove HTML markup, extra white space, and unicode
- But HTML markup is often valuable:
  - HTML markup for section header names.
  - Legal database web sites often have HTML tags for citations to other cases.
- Other cleaning steps:
  - page numbers
  - hyphenations at line breaks
  - table of contents, indexes, etc.
- These are all corpus-specific, so inspect ahead of time.

Regular Expressions

- Regular Expressions, implemented in the Python package **re**, provide a powerful string matching tool.
  - A systematic string matching protocol – can match arbitrary string patterns
  - e.g., use `'utilit\w+'` to match utility, utilities, utilitarian, ...
  - Important for identifying speaker names (in political documents) section headers (in statutes), citations (in judicial opinions), etc.

Regular Expressions

- Regular Expressions, implemented in the Python package **re**, provide a powerful string matching tool.
  - A systematic string matching protocol – can match arbitrary string patterns
  - e.g., use `'utilit\w+'` to match utility, utilities, utilitarian, ...
  - Important for identifying speaker names (in political documents) section headers (in statutes), citations (in judicial opinions), etc.
- See NLTK book Chapter 3.4-3.5 for an introduction, and nice online lessons at RegexOne (linked in syllabus).

OCR (Optical Character Recognition)

- Your data might be in PDF's or images. Needs to be converted to text
- The best solution (that I know of) is ABBYY FineReader, which is expensive but might be available at the library.
- My colleague Joe Sutherland at Columbia has a nice open-source package for OCR:
  - `https://github.com/jlsutherland/doc2text`
- And I just saw this last month:
  - `https://github.com/faustomorales/keras-ocr`

Should you run a spell checker?

- The short answer is no:
  - Most corpora have important specialized vocabulary that would be flagged by standard spell-checkers.

Should you run a spell checker?

- The short answer is no:
  - Most corpora have important specialized vocabulary that would be flagged by standard spell-checkers.
  - In most empirical contexts, it's safe to assume that spelling errors (especially OCR errors) are uncorrelated with treatment assignment.

Should you run a spell checker?

- The short answer is no:
  - Most corpora have important specialized vocabulary that would be flagged by standard spell-checkers.
  - In most empirical contexts, it's safe to assume that spelling errors (especially OCR errors) are uncorrelated with treatment assignment.
- Better solutions:
  - drop short words (one or two letters) and long words (over 12 letters).
  - get doc frequencies for each word and filter out rare words
    - or use word embeddings and trust that misspellings will be nearby the true word.

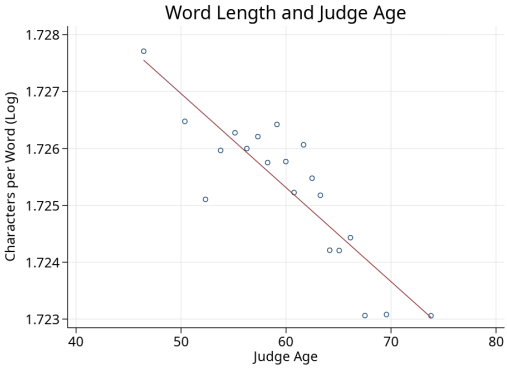
Should you run a spell checker?

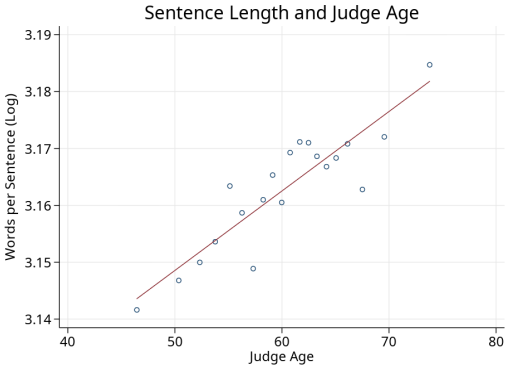
- The short answer is no:
  - Most corpora have important specialized vocabulary that would be flagged by standard spell-checkers.
  - In most empirical contexts, it's safe to assume that spelling errors (especially OCR errors) are uncorrelated with treatment assignment.
- Better solutions:
  - drop short words (one or two letters) and long words (over 12 letters).
  - get doc frequencies for each word and filter out rare words
    - or use word embeddings and trust that misspellings will be nearby the true word.
- **But**:
  - There are cases where spelling errors could be correlated with treatment (for example, increasing legislator salaries might change both policy priorities and spelling error rates). Run a spell-checker and see if there is a treatment effect on misspelling rates.

Outline

Judge Age and Writing Style

Ash and MacLeod (2020)





Optimal legal complexity

Katz and Bommarito (2014)

"Everything should be made as simple as possible, but no simpler."

--not Einstein

Optimal legal complexity

Katz and Bommarito (2014)

"Everything should be made as simple as possible, but no simpler."

--not Einstein

- 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)

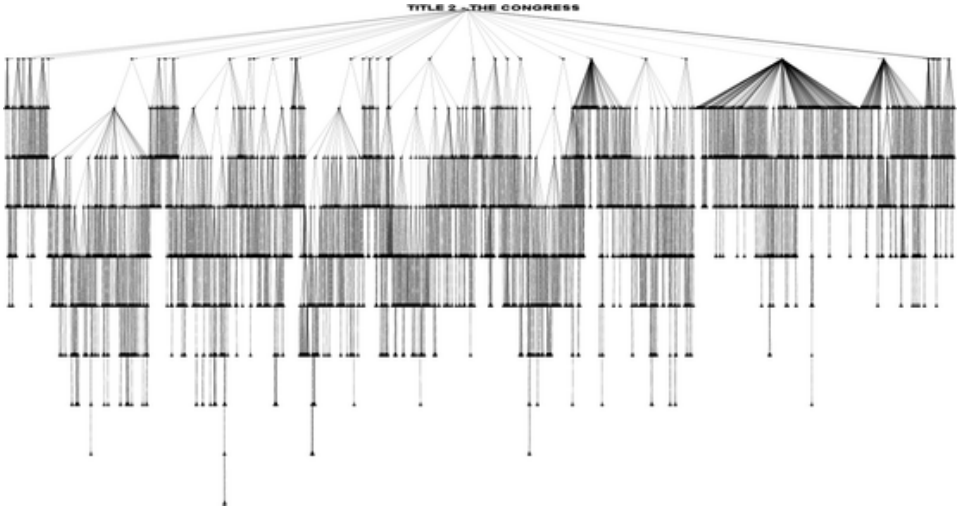
"Everything should be made as simple as possible, but no simpler."

--not Einstein

- 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.

The U.S. Code

Katz and Bommarito (2014)



- The U.S. Code consists of 49 titles, which can be further subdivided into subtitle, chapter, subchapter, part, subpart, section, subsection, paragraph, subparagraph, clause, and subclause.

Number of Clauses $\approx$ Number of Words

Measuring Complexity (Katz and Bommarito 2014)

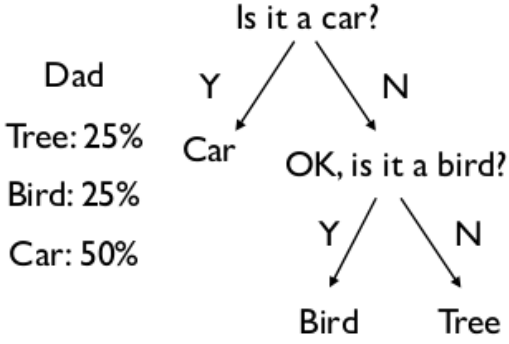
Five largest and smallest titles by structural size

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

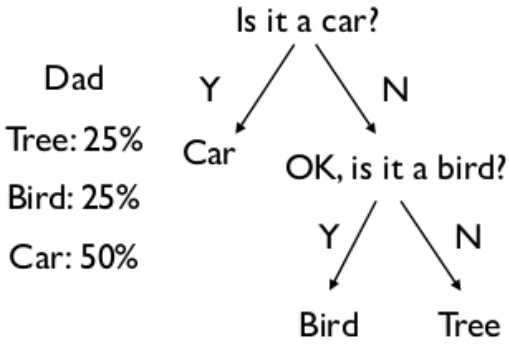
Title	Tokens	Tokens per section
Public Health and Welfare (Title 42)	2,732,251	369.22
Internal Revenue Code (Title 26)	1,016,995	487.07
Conservation (Title 16)	947,467	200.48
Commerce and Trade (Title 15)	773,819	336.88
Agriculture (Title 7)	751,579	274.00
President (Title 3)	7,564	120.06
Intoxicating 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

Digression: Twenty Questions (DeDeo 2018)



- The optimal set of questions minimizes the expected number of turns until you guess the answer.

Digression: Twenty Questions (DeDeo 2018)



- The optimal set of questions minimizes the expected number of turns until you guess the answer.
  - The length of the optimal script is closely approximated by the **entropy**

$$H(X) = -\sum_{i=1}^{N} \Pr(x_i) \log_2 \Pr(x_i)$$

where one guesses from words $\{x_1, x_2, ..., x_N\}$.

Axioms for Shannon Entropy

Shannon (1948) wanted a function $H(\vec{p})$ for a vector of probabilities $\vec{p}$, that would satisfy four axioms:

1. continuity (small changes in $p_i \rightarrow$ small changes in H)
2. symmetry (invariance to re-ordering of $\vec{p}$
3. Condition of maximum information (H maximized when all p_i are equal
4. [next slide]

4. Coarse Graining

- Consider a set $X = \{p_a, p_b, p_c\}$
  - Consider a subset $X' = \{p_a, p_{bc}\}$ where $p_{bc} = p_b + p_c$ (we don't distinguish b and c)

4. Coarse Graining

- Consider a set $X = \{p_a, p_b, p_c\}$
  - Consider a subset $X' = \{p_a, p_{bc}\}$ where $p_{bc} = p_b + p_c$ (we don't distinguish b and c)
- Coarse graining requires

$$H(X) = H(X') + p_{bc} H(G)$$

where $G = \{\frac{p_b}{p_{bc}}, \frac{p_c}{p_{bc}}\}$ the distribution for making the finer-grained distinction between b and c.

4. Coarse Graining

- Consider a set $X = \{p_a, p_b, p_c\}$
  - Consider a subset $X' = \{p_a, p_{bc}\}$ where $p_{bc} = p_b + p_c$ (we don't distinguish b and c)
- Coarse graining requires

$$H(X) = H(X') + p_{bc} H(G)$$

  where $G = \{\frac{p_b}{p_{bc}}, \frac{p_c}{p_{bc}}\}$ the distribution for making the finer-grained distinction between b and c.
- The unique function satisfying these four assumptions is

$$H(X) = -\sum_{x \in X} \Pr(x) \log \Pr(x)$$

Continuous Distributions and Cross Entropy

- For continuous distribution with pdf $p(x)$:

$$H(p(x)) = -\int_x p(x)\log(p(x))dx$$

  - broad distributions have higher entropy.

Continuous Distributions and Cross Entropy

- For continuous distribution with pdf $p(x)$:

$$H(p(x)) = -\int_x p(x)\log(p(x))dx$$

  - broad distributions have higher entropy.

- The cross entropy for distributions $p(x)$ and $q(x)$ is a measure of similarity between distributions:

$$H(p,q) = -\int_x p(x)\log((q(x))dx$$

  - In machine learning, can be used as a loss function, where p is true and q is the model prediction.
  - It is the standard loss function for logistic regression.

Entropy in Text

- In text data, the probabilities p_i can be interpreted as the probability (frequency) of observing particular tokens: characters, words, n-grams, etc.
  - In general, documents with more diverse vocabularies/topics will have higher entropy.
  - an alternative: run compression algorithms on text and measure bytes of compressed data.

Word Length and Word Entropy

Measuring Complexity (Katz and Bommarito 2014)

Ten titles with highest average word size

Title	Avg. word size
Domestic Security (Title 6)	6.90
War and National Defense (Title 50)	6.83
Public Printing and Documents (Title 44)	6.74
Foreign Relations and Intercourse (Title 22)	6.74
Public Contracts (Title 41)	6.73
Crimes and Criminal Procedure (Title 18)	6.16
Intoxicating Liquors (Title 27)	6.15
Internal Revenue Code (Title 26)	6.10
Flag and Seal, Seat of Govt. and the States (Title 4)	6.10
Bankruptcy (Title 11)	6.07

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

Outline

Overview of Dictionary-Based Methods

- Dictionary-based text methods use a pre-selected list of words or phrases to analyze a corpus.

Overview of Dictionary-Based Methods

- Dictionary-based text methods use a pre-selected list of words or phrases to analyze a corpus.
- Corpus-specific (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.
- Corpus-specific (e.g., number of times a judge says “justice” vs “efficiency”)
- General dictionaries: WordNet, LIWC, MFD, etc.

Corpus-specific words

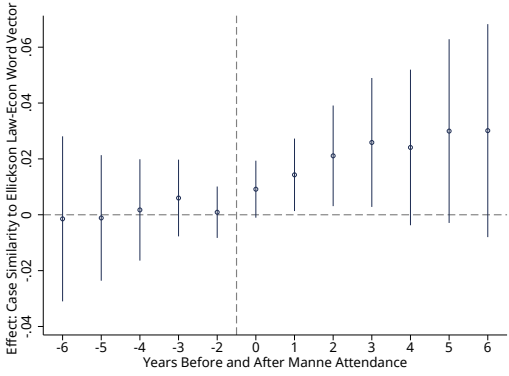
- Sometimes counting sets of words or phrases across documents can provide useful evidence.

Corpus-specific words

- Sometimes counting sets of words or phrases across documents can provide useful evidence.
- Ash, Chen, and Naidu (2019):
  - We analyze the use of economics reasoning in the judiciary.
  - For example, use of the word "efficiency" or "deterrence" after attending a two-week intensive summer course in economics.

Impact of Economics Training on Economics Language

Ash, Chen, and Naidu (2019)



After attendance, Economics Trained Judges increase use of a selection of terms related to law and economics

Measuring uncertainty in macroeconomy

Baker, Bloom, and Davis (2016)

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

Measuring uncertainty in macroeconomy

Baker, Bloom, and Davis (2016)

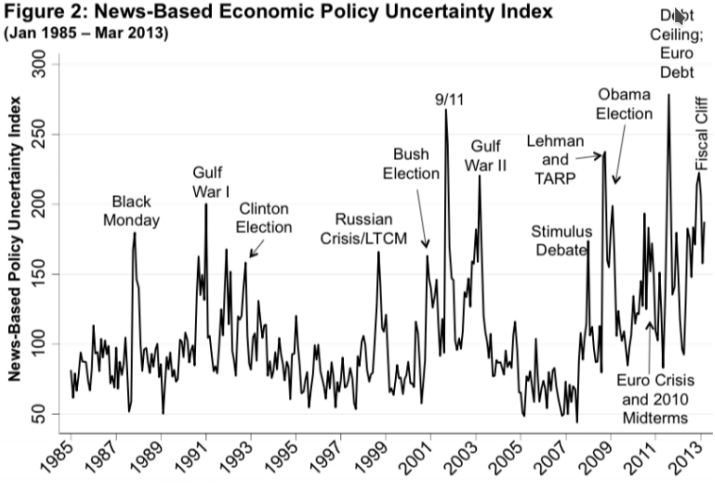
- Baker, Bloom, and Davis measure economic policy uncertainty using Boolean search of newspaper articles.
- For each newspaper on each day since 1985, submit the following query:
  - 1. Article contains "uncertain" OR "uncertainty", AND
  - 2. Article contains "economic" OR "economy", AND
  - 3. Article contains "congress" OR "deficit" OR "federal reserve" OR "legislation" OR "regulation" OR "white house"

Measuring uncertainty in macroeconomy

Baker, Bloom, and Davis (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



WordNet

- English word database:
  - 117,798 nouns, 11,529 verbs, 22,479 adjectives, and 4,481 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/whole-party).

WordNet Supersenses

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

WordNet Sense Relations

Relation	Also Called	Definition	Example
Hypernym	Superordinate	From concepts to superordinates	$breakfast^1 \rightarrow meal^1$
Hyponym	Subordinate	From concepts to subtypes	$meal^1 \rightarrow lunch^1$
Instance Hypernym	Instance	From instances to their concepts	$Austen^1 \rightarrow author^1$
Instance Hyponym	Has-Instance	From concepts to their instances	$composer^1 \rightarrow Bach^1$
Part Meronym	Has-Part	From wholes to parts	$table^2 \rightarrow leg^3$
Part Holonym	Part-Of	From parts to wholes	$course^7 \rightarrow meal^1$
Antonym		Semantic opposition between lemmas	$leader^1 \Longleftrightarrow follower^1$
Derivation		Lemmas w/same morphological root	$destruction^1 \Longleftrightarrow destroy^1$

Figure 19.3 Some of the noun relations in WordNet.

Relation	Definition	Example
Hypernym	From events to superordinate events	$fly^9 \rightarrow travel^5$
Troponym	From events to subordinate event	$walk^1 \rightarrow stroll^1$
Entails	From verbs (events) to the verbs (events) they entail	$snore^1 \rightarrow sleep^1$
Antonym	Semantic opposition between lemmas	$increase^1 \Longleftrightarrow decrease^1$

Figure 19.4 Some verb relations in WordNet.

WordNet Sense Relations

```
bass3, basso (an adult male singer with the lowest voice)
=> singer, vocalist, vocalizer, vocaliser
   => musician, instrumentalist, player
      => performer, performing artist
         => entertainer
            => person, individual, someone...
               => organism, being
                  => living thing, animate thing,
                     => whole, unit
                        => object, physical object
                           => physical entity
                              => entity
bass7 (member with the lowest range of a family of instruments)
=> musical instrument, instrument
   => device
      => instrumentality, instrumentation
         => artifact, artefact
            => whole, unit
               => object, physical object
                  => physical entity
                     => entity
```

Figure 19.5 Hyponymy chains for two separate senses of the lemma *bass*. Note that the chains are completely distinct, only converging at the very abstract level *whole, unit*.

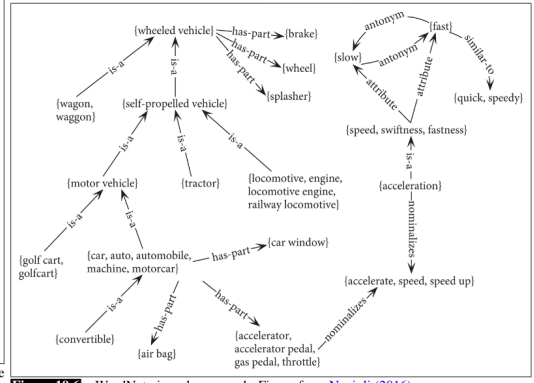


Figure 19.6 WordNet viewed as a graph. Figure from Navigli (2016).

WordNet can be used for shrinking vocabulary

- Wordnet provides synonym sets ("synsets"):
  - can replace words with the most frequent word from that synset.

WordNet can be used for shrinking vocabulary

- Wordnet provides synonym sets ("synsets"):
  - can replace words with the most frequent word from that synset.
- Wordnet provides hierachical relations between words:
  - ancestors of "dog" include "carnivore", "mammal", "vertebrate", "animal", and "physical entity"

WordNet can be used for shrinking vocabulary

- Wordnet provides synonym sets ("synsets"):
  - can replace words with the most frequent word from that synset.
- Wordnet provides hierachical relations between words:
  - ancestors of "dog" include "carnivore", "mammal", "vertebrate", "animal", and "physical entity"
  - Can replace words with a higher-level category to which it is member.

LIWC

- LIWC (pronounced “Luke”) stands for Linguistic Inquiry and Word Counts
  - 2300 words 70 lists of category-relevant words, e.g. “emotion”, “cognition”, “work”, “family”, “positive”, “negative” etc.
  - Info and publications at `liwc.net`
  - Invented in 1980s, now in third version

Emotion Lexicons

- 8 basic emotions, in four opposing pairs:
  - joy–sadness
  - anger–fear
  - trust–disgust
  - anticipation–surprise
- Mohammad and Turney (2011) code 10,000 words along the four dimensions (using Mturk)

Valence, Arousal, and Dominance

Warriner, Juperman, and Brysbaert (2013)

- Ratings for 14,000 words along three emotional dimensions:
  - valence – pleasantness of the stimulus
  - arousal – intensity of the emotion provoked by stimulus
  - dominance – degree of control exerted by the stimulus

Concreteness vs. Abstractness

- The degree to which the concept denoted by a word refers to a perceptible entity.
- Brysbaert, Warriner, and Kuperman (2014):
  - concreteness scores (1 to 5) for 37,058 words and 2896 two-word phrases.
  - concrete: “banana” (5)
  - abstract: “although” (1.07)

Function Words

- **Function words** are used for syntax rather than semantics (meaning):
  - e.g. *for*, *rather*, *than*

Function Words

- **Function words** are used for syntax rather than semantics (meaning):
  - e.g. *for*, *rather*, *than*
- There are many lists – e.g. LIWC has a list of 464 function words (the "funct" category).

Function Words

- **Function words** are used for syntax rather than semantics (meaning):
  - e.g. *for*, *rather*, *than*
- There are many lists – e.g. LIWC has a list of 464 function words (the "funct" category).
- Goldberg (2017) says you can take the top ~300 most-frequent words in a corpus as a list of function words.

Function Words: Applications

Hegolemeter: `http://textlab.econ.columbia.edu/hegel`

Function Words: Applications

Hegolemeter: `http://textlab.econ.columbia.edu/hegel`

- One of the earliest applications of text analysis used frequencies over function words to de-anonymze the *Federalist Papers* (Mostellar and Wallace 1964).

Sentiment Analysis

- There are many appraoches to sentiment analysis
  - e.g., in Python, nltk vader class provides positive, negative, and neutral scores for a document, and a composite score that combines all three.

Sentiment Analysis

- There are many appraoches to sentiment analysis
  - e.g., in Python, nltk vader class provides positive, negative, and neutral scores for a document, and a composite score that combines all three.
- Designed for online writing – hard to say how well it works on legal text, for example.
  - Hamilton-Clark-Leskovec-Jurafsky (2016) provide a method for making domain-specific sentiment lexicons using word embeddings (more on this later).

I'd hate to be the president

I'd hate to be the president

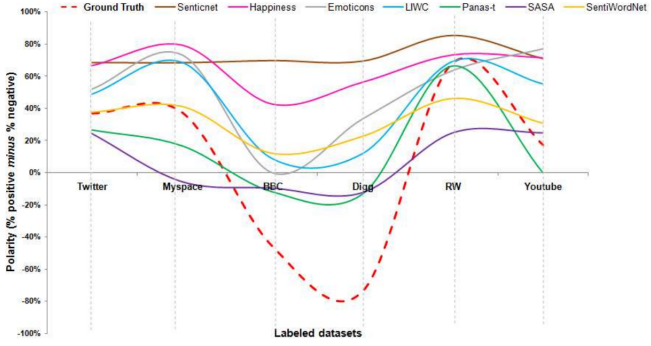


Figure 2: Polarity of the eight sentiment methods across the labeled datasets, indicating that existing methods vary widely in their agreement.

Folkore

Michalopolous and Xue 2019

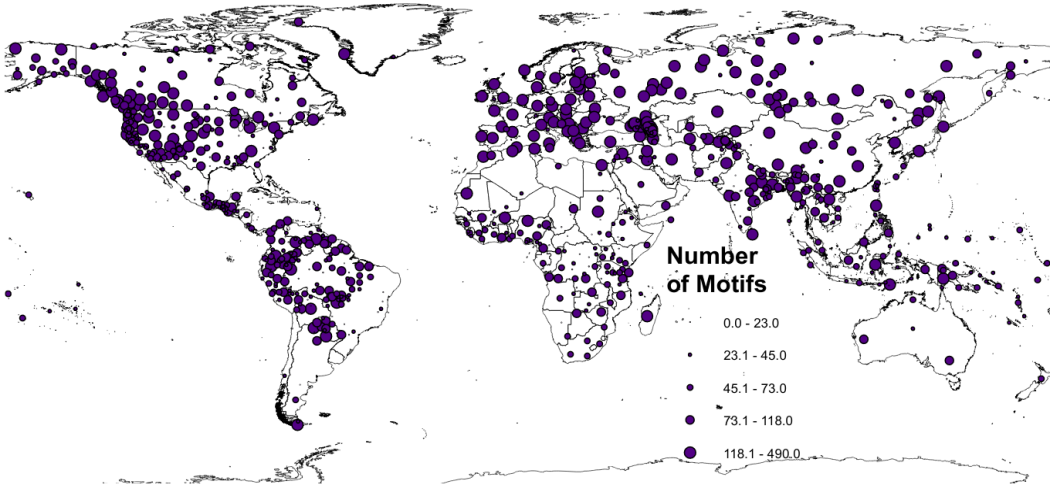


Figure 2: # of Motifs across Oral Traditions

Folkore

Michalopolous and Xue 2019

Table 1 - Panel C: Examples of Motifs in Each of the 12 categories of Berezkin's Catalogue

Motif Group	ID	Description
a	a3	The Moon is female or bisexual, the Sun is male.
b	a32	A figure or an imprint of some being or object are seen in the Moon.
c	b3a	Water is the original element, the dry earth appears later.
d	h28	Killed and destroyed (often burned) person or creature (usually ogre, fierce animal, powerful shaman) turns into a multitude of biting insects or into other small molesting creatures.
e	f9	For different reasons, sexual contact with a woman is deadly dangerous for a man.
f	g6	One of the trees is the principal, original one (emerged before all the other; ancestor of wild or cultivated plants; ocean or rivers inside it; world axis; higher than all the others; overshadows sky).
g	b82	Raven or other carrion-eating bird of dark color and a similar size was originally white.
h	l19b	A being with three or more heads is described in tales or represented in art.
i	m29b	In episodes related to deception, absurd, obscene or anti-social behavior the protagonist is fox, jackal or coyote.
j	k27n	Father or other kinsmen of hero's wife or bride try to kill or test him and/or suggest him difficult tasks.
k	m30	Person or creature who has no wings or is unable to fly on a long distance attempts to ascend to the sky or to fly far away but falls down or, deprived of his wings, remains in a place from which he is unable to return.
l	k38e	Loci or objects of three (rare – four) different materials are mentioned in such a way that all of them have positive connotations though unequal value (copper, silver and gold; silver, gold and diamonds, etc.)

Folkore

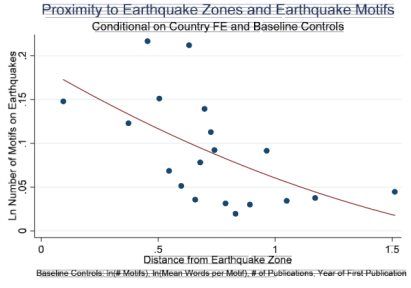


Figure 3a

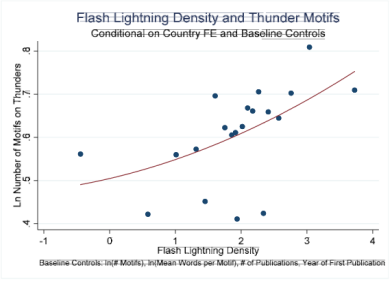


Figure 3b

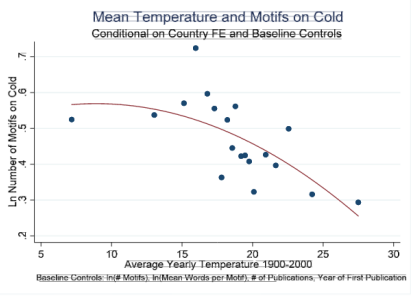


Figure 3c

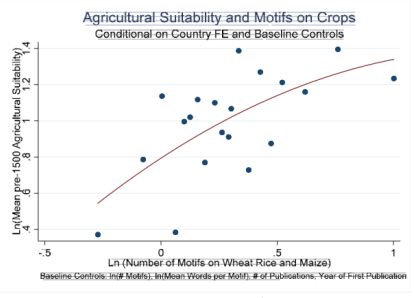


Figure 3d

Folkore

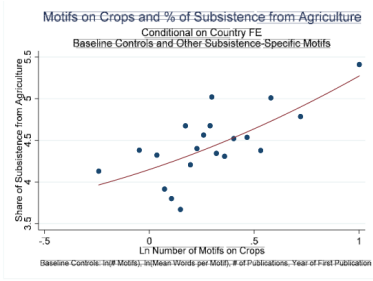


Figure 4a

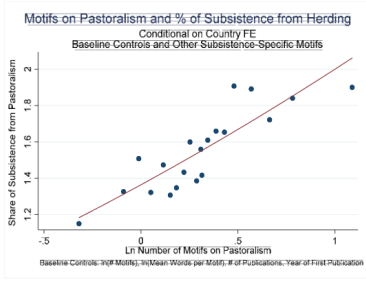


Figure 4b

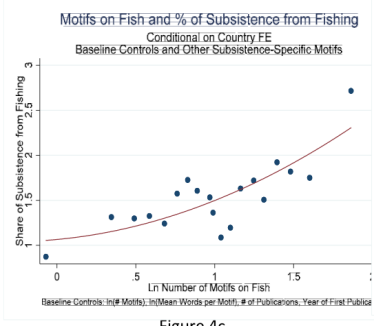


Figure 4c

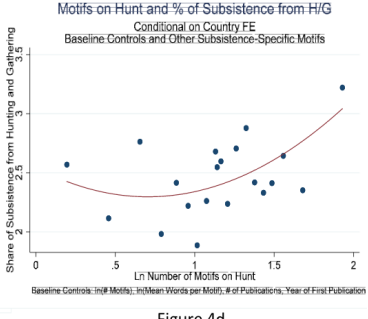


Figure 4d