



Tracking Theoretical Shifts over Time

A Natural Language Processing Analysis of

Cass R. Sunstein

Omar CHERIF, Dalila LADLI, Maxime MÉLOUX

Supervised by prof. Samuel FEREY

June 28, 2022





Table of Contents

Introduction

- ▶ Introduction
- ▶ Corpus acquisition and processing
- ▶ NLP for validating hypotheses
- ▶ NLP tools as a source of new hypotheses
- ▶ Conclusion



Related Work

Introduction

Applying NLP to law and economics:

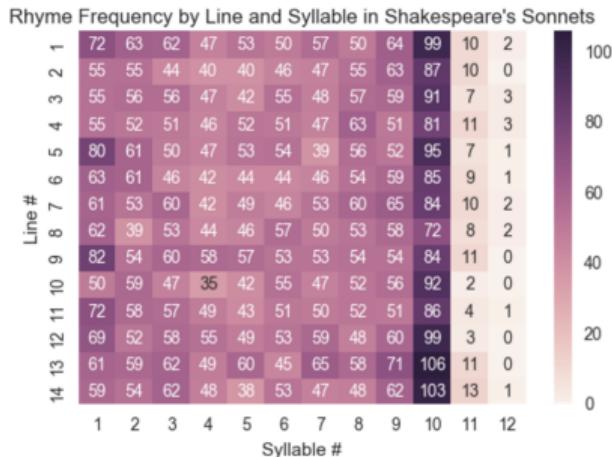
- *Five reasons for the use of network analysis in the history of economics* [Catherine and Doehe, 2018]
- *What topic modeling could reveal about the evolution of economics* [Ambrosino et al., 2018]
- *Forty Years of Behavioral Economics* [Truc, 2021]



A Unique Approach

Introduction

- Ideological analysis over time
- Single-author corpus?
 - Literature only [Culpeper and Archer, 2020]





Application - 1/2

Introduction

- Chicago school of economics (neoclassical economics)
- *Homo economicus* → perfect rationality
- Psychologists (Amos Tversky, Daniel Kahneman) in the 1960s
- Behavioral economics: bounded rationality, heuristics, biases
- Prospect theory (losses \neq gains)



Application - 2/2

Introduction

- Cass R. Sunstein (born 1954)
- University of Chicago Law School
- Republicanism (not the Republican party)
- Libertarian paternalism, *Nudge* in the 2000s



Working Hypotheses

Introduction

In Sunstein's articles:

- Can we detect a transition from republicanism to behavioral economics?
When?
- How to represent subtopics of republicanism (citizenship, civic virtue...) over time?
- How to model the interaction between global topics (poverty, autonomy...)?



Methodology

Introduction

Three main phases:

- Corpus acquisition and cleaning
- Frequentist, statistical analysis (mostly supervised)
- Machine learning analysis (mostly unsupervised)

Perspectives

- Semi-supervised, "objective" text interpretation
- Field orthogonal to ours, with an expert in economics



Table of Contents

Corpus acquisition and processing

- ▶ Introduction
- ▶ Corpus acquisition and processing
- ▶ NLP for validating hypotheses
- ▶ NLP tools as a source of new hypotheses
- ▶ Conclusion



Selection and retrieval

Corpus acquisition and processing

- Two sources: JSTOR¹ and Chicago Unbound²
- 1987-2005: $215 + 415 = 620$ articles
- Duplicates/quality filtering: working versions?
- Final list: 231 PDF files

¹<https://www.jstor.org/>

²<https://chicagounbound.uchicago.edu/>



Text extraction

Corpus acquisition and processing

- Three main OCR tools
- LAPDFText > CERMine > Tesseract/OCRmyPDF (for our use)



Manual cleaning

Corpus acquisition and processing

Duration: 40-60 hours.

In this respect the distribution of powers was consonant with Madison's own hostility to rapid change in government, captured in his antipathy to "turbulence," but disso-

930 BRIGHAM YOUNG UNIVERSITY LAW
REVIEW [1987] nant with Jefferson's belief that turbulence is healthy for a republic.

In this respect the distribution of powers was consonant with Madison's own hostility to rapid change in government, captured in his antipathy to "turbulence," but dissonant with Jefferson's belief that turbulence is healthy for a republic.

Figure 1: Cleaning of a page/line break (green), page footer (yellow) and header (blue).

Not pictured: missing pages (usually a few), structured data (figures, tables).



Automatic cleaning

Corpus acquisition and processing

Duration: 15-20 hours.

Judicial control of regulatory behavior has come in the form of the "hardlook" doctrine, initially developed by the United States Court of Appeals for the District of Columbia CIRCUIT,² and subsequently endorsed by the Supreme Court.³ The hard-look doctrine has both procedural and substantive elements. Procedurally, it requires regulatory agencies to generate detailed explanations for their decisions—to consider reasonable alternatives."⁴

Judicial control of regulatory behavior has come in the form of the "hard-look" doctrine, initially developed by the United States Court of Appeals for the District of Columbia Circuit, and subsequently endorsed by the Supreme Court. The hard-look doctrine has both procedural and substantive elements. Procedurally, it requires regulatory agencies to generate detailed explanations for their decisions — to consider reasonable alternatives.

Figure 2: Cleaning of footnote numbers (yellow), typographical/OCR errors (green), bad de-hyphenation (blue) and dash-hyphen confusion (red).



Automatic cleaning - The Monster

Corpus acquisition and processing

```
[43805/68296] ... rence signals a fundamental departure from the more narrow right-duty relations [characteris-] tic of the common law. In the face of that departure, courts ought to be close...  
Unknown word: [characteris], possible split: [character is], spaced out: [characteris] / [characteris]  
[1] character [2] character [3] characterize [4] characterizes [5] characteristic [6] characterized [7] characterless [8] charcuterie [9] charters [10] chanceries  
[Add] [R]eplace [S]plit [H]yphenize [C]ontext Merge [L]eft [M]erge right [O]ne-time [F]ast replace [N]umber removal [D]etach words [E]tach right [Ne(w)] rule [I]gnore  
Choose action: a  
87% [REDACTED] | 46748/68309 [00:45<00:12, 1696.86it/s]  
  
[43805/68295] ... rence signals a fundamental departure from the more narrow right-duty relations [characteris-tic] of the common law. In the face of that departure, courts ought to be closely at...  
Unknown word: [characteris-tic], possible split: [character is tic], spaced out: [characteris - tic] / [characteris - tic]  
[1] characteristic [2] characteristics [3] uncharacteristic [4] characterization [5] characteristically [6] characterizations [7] characterize [8] characterized [9] characterizes [10] characterizing  
[Add] [R]eplace [S]plit [H]yphenize [C]ontext Merge [L]eft [M]erge right [O]ne-time [F]ast replace [N]umber removal [D]etach words [E]tach right [Ne(w)] rule [I]gnore  
Choose action: i  
85% [REDACTED] | 46748/68309 [00:45<00:12, 1696.86it/s]  
  
[46864/68204] ... ny, while at the same time subjecting aggressive enforcement action to judicial [review.68] This result is hardly likely to fit with Congress' goals in enacting regulatory...  
Unknown word: [review.68], possible split: [review 68], spaced out: [review. 68] / [review. 68]  
[1] review [2] reviewal [3] reviews [4] reviewed [5] reviewer [6] reviewers [7] reviewing [8] reviews [9] preview [10] previewed  
[Add] [R]eplace [S]plit [H]yphenize [C]ontext Merge [L]eft [M]erge right [O]ne-time [F]ast replace [N]umber removal [D]etach words [E]tach right [Ne(w)] rule [I]gnore  
Choose action: n  
70% [REDACTED] | 48153/68309 [00:46<00:12, 1681.86it/s]  
  
[48181/68292] ... a case based solely on the Constitution, without a statute creating a cause of [action.70] But statutory cases present a different issue.  
The argument for judicial refusal...  
Unknown word: [action.70], possible split: [action 70], spaced out: [action. 70] / [action. 70]  
[1] action [2] actions [3] actin [4] acting [5] actinide [6] actinium [7] actionable [8] auction [9] auctioned [10] auctioneer  
[Add] [R]eplace [S]plit [H]yphenize [C]ontext Merge [L]eft [M]erge right [O]ne-time [F]ast replace [N]umber removal [D]etach words [E]tach right [Ne(w)] rule [I]gnore  
Choose action: n  
74% [REDACTED] | 50298/68309 [00:48<00:08, 2082.46it/s]  
  
[50405/68290] ... I think that these considerations would justify several conclusions. First, W [arth] itself may well have been rightly decided. Because the plaintiffs' claim was co...  
Unknown word: [arth], possible split: [arth], spaced out: [arth] / [arth]  
[1] arch [2] argh [3] art [4] arty [5] arty [6] auth [7] earth [8] earth [9] earth [10] 11th  
[Add] [R]eplace [S]plit [H]yphenize [C]ontext Merge [L]eft [M]erge right [O]ne-time [F]ast replace [N]umber removal [D]etach words [E]tach right [Ne(w)] rule [I]gnore  
Choose action: i  
83% [REDACTED] | 50931/68309 [00:54<00:04, 2761.02it/s]  
  
[50958/68289] ... he virtue of helping to correct the fundamental flaw of Data Processing, and of [reestablishing] that it is both desirable and inevitable for courts to focus on legislative ins...  
Unknown word: [reestablishing], possible split: [reestablishing] / [reestablishing]  
[1] establishing [2] disestablishing [3] reestablish [4] reestablished [5] reestablishes [6] destabilizing [7] establish [8] established [9] establishes [10] establishment  
[Add] [R]eplace [S]plit [H]yphenize [C]ontext Merge [L]eft [M]erge right [O]ne-time [F]ast replace [N]umber removal [D]etach words [E]tach right [Ne(w)] rule [I]gnore  
Choose action: a  
80% [REDACTED] | 58864/68309 [00:58<00:07, 1285.42it/s]  
  
[59032/68289] ... lification in note 2 supra.  
Compare The End on Animals v Espy, BIA Supp 142 [(DOC] 1993), granting standing to people interested in studying bison, in the context...  
Unknown word: [DOC], possible split: [D_O_C], spaced out: [DOC] / [DOC]  
[1] d.c. [2] ddt [3] dec [4] dfc [5] doc [6] dsc [7] c [8] 2d [9] 3d [10] abc  
[Add] [R]eplace [S]plit [H]yphenize [C]ontext Merge [L]eft [M]erge right [O]ne-time [F]ast replace [N]umber removal [D]etach words [E]tach right [Ne(w)] rule [I]gnore  
Choose action: a  
87% [REDACTED] | 50385/68309 [01:00<00:21, 423.58it/s]  
  
[59207/68289] ... sfenders of Wildlife, 112 F.2d 2138 (1902).  
The issue arose in Public Citizen v [USTR,] 822 F Supp 21 (DOC 1993), in which the district court found standing without cl...  
Unknown word: [USTR], possible split: [US TIR], spaced out: [USTR] / [USTR]
```



Automatic cleaning - The Monster

Corpus acquisition and processing

```
[43865/68295] ... rence signals a fundamental departure from the more narrow right-duty relations [characteris-tic]
Unknown word: [characteris-tic], possible split: [character is tic], spaced out: [characteris - tic] / [characteris - tic]
[1] characteristic   [2] characteristics   [3] uncharacteristic   [4] characterization   [5] characteristically
[A]dd      [R]eplace    [S]plit     [H]yphenize    [C]ontext     Merge [L]eft     [M]erge right    [O]ne-time
Choose action: 1
```



Automatic cleaning - The Monster

Corpus acquisition and processing

Type	Number of items learned
Unknown words	11,648
Word substitution rules	2,078

Table 1: Statistics obtained using the Monster.



Resulting Corpus

Corpus acquisition and processing

- 231 text files
- 43 thousand paragraphs
- 2.8 million words
- 17.5 million characters



Table of Contents

NLP for validating hypotheses

- ▶ Introduction
- ▶ Corpus acquisition and processing
- ▶ NLP for validating hypotheses
- ▶ NLP tools as a source of new hypotheses
- ▶ Conclusion



TF-IDF

NLP for validating hypotheses

Term Frequency-Inverse Document Frequency (TF-IDF)

- Basic statistical method, extracts words specific to each document
- Counts word frequencies in each document (year) and over the whole corpus
- Applies the following formula for each year

$$\text{tfidf}(t, d, D) = \frac{f_d(t)}{\sum_{w \in D} f_d(w)} \log \left(\frac{|D|}{|\{d \in D : t \in d\}|} \right)$$

- The bigger the TF-IDF at time t , the more important the word



Results

NLP for validating hypotheses

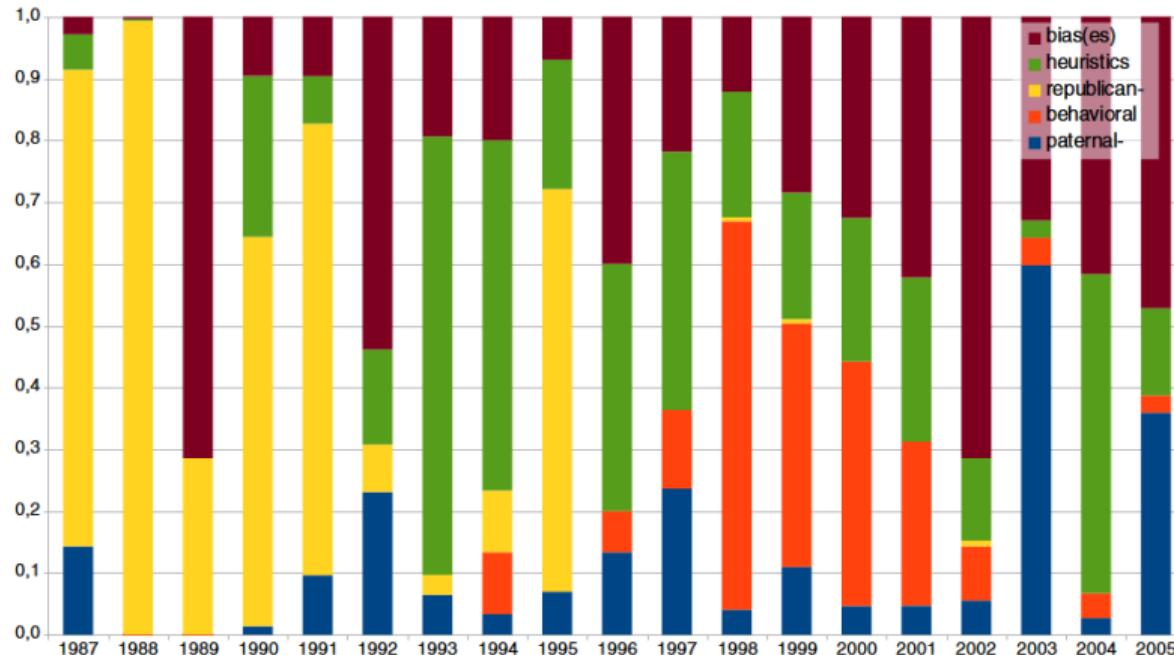


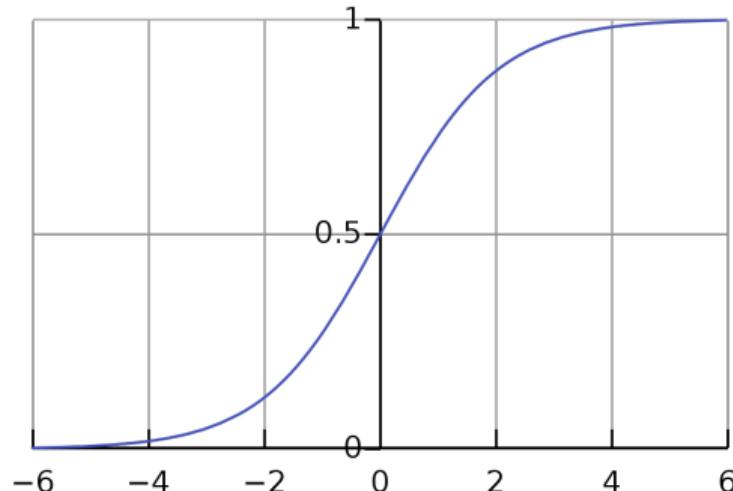
Figure 3: The normalized TF-IDF of selected words over time.



Transition Modeling

NLP for validating hypotheses

$$h(x) = \frac{1}{1 + e^{-x}}$$





Why the sigmoid?

NLP for validating hypotheses

$$f(L, k, x, x_0, b) = \frac{L}{1 + e^{-k(x-x_0)}} + b$$

- Organic approximation
- Helps us filter words
- Visualizes intensity of change and length



Word selection

NLP for validating hypotheses

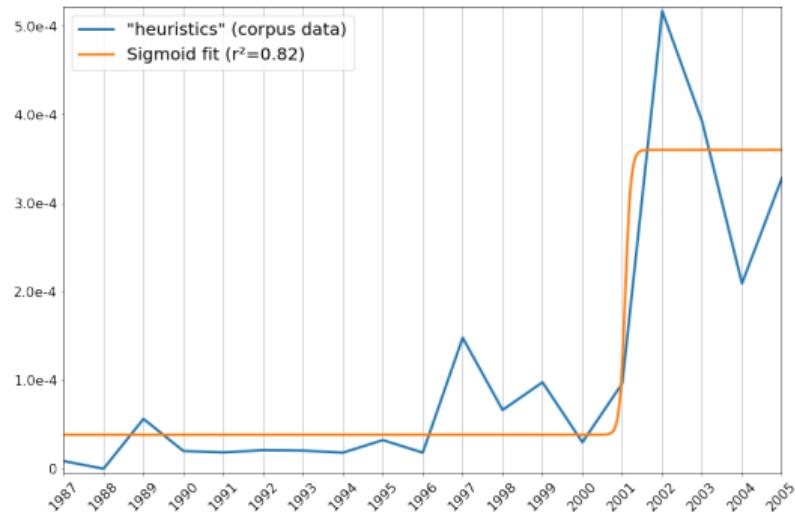
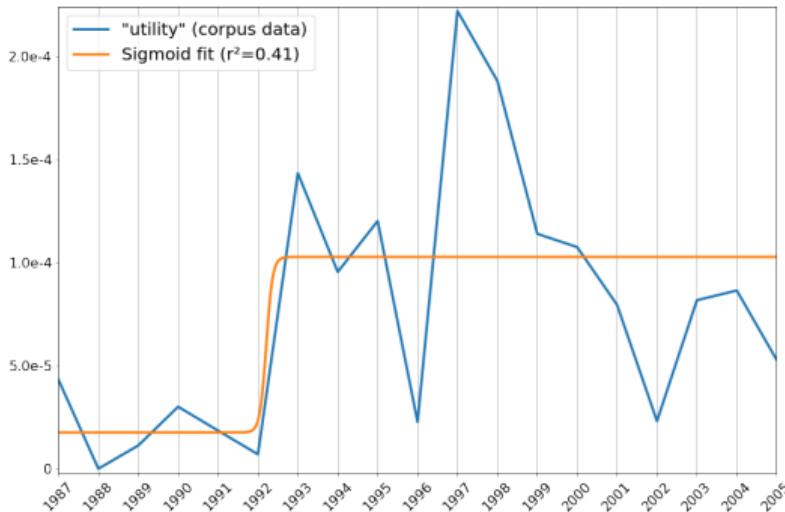
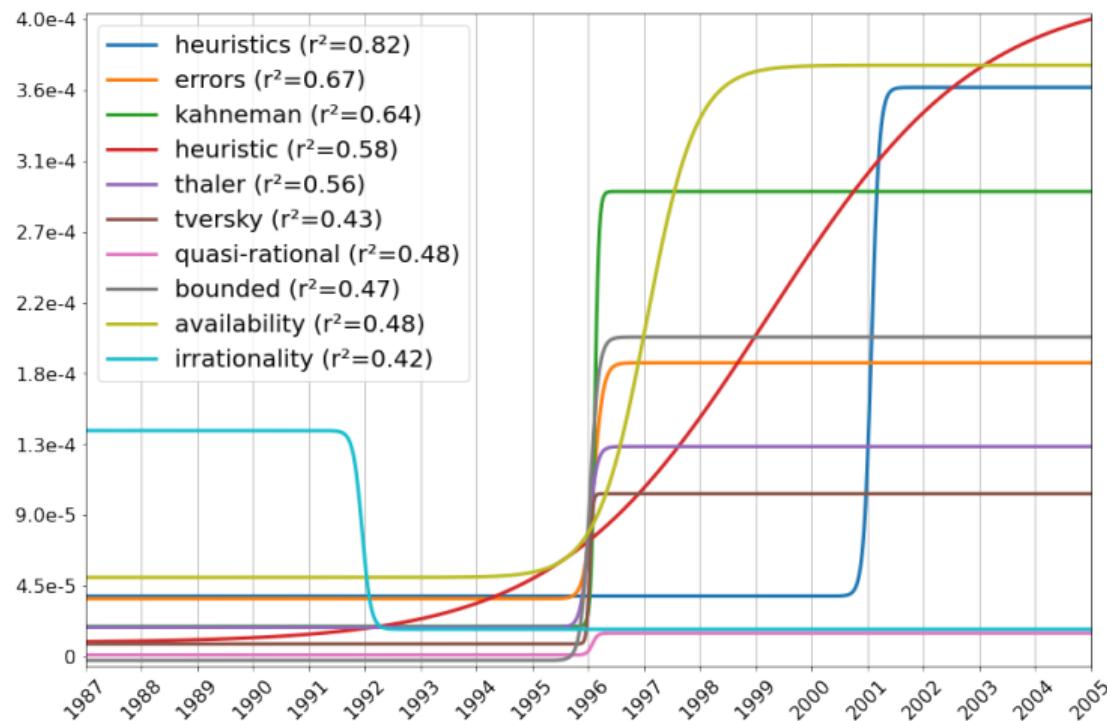


Figure 4: Observed versus fitted data for a low (left) and a high (right) value of r^2 .



Results 1/2

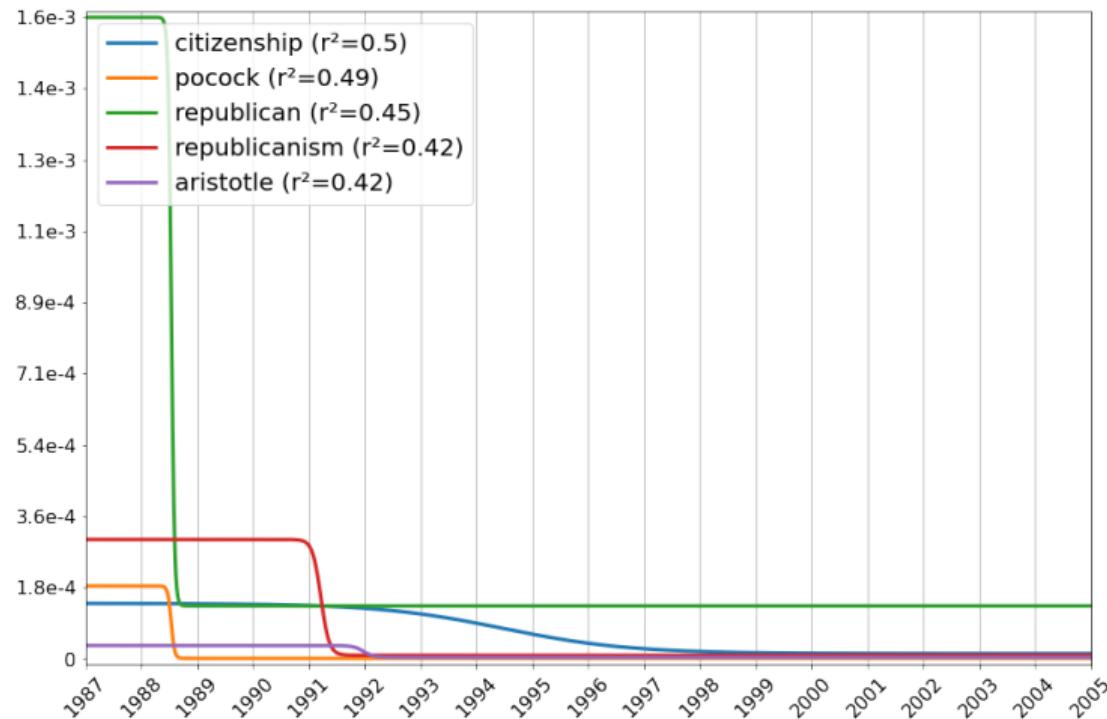
NLP for validating hypotheses





Results 2/2

NLP for validating hypotheses





Semantic Analysis with Tropes

NLP for validating hypotheses



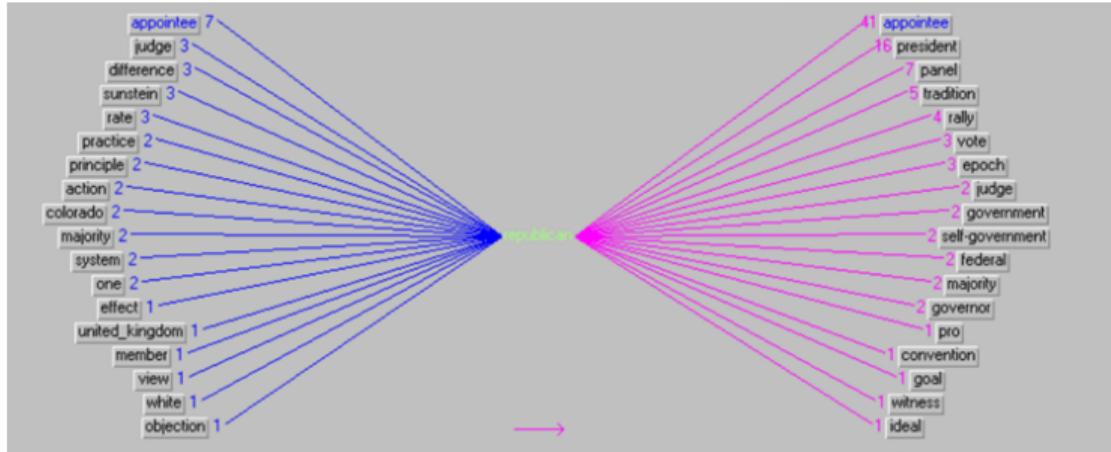
Tropes

<http://www.tropes.fr>



Semantic Analysis with Tropes

NLP for validating hypotheses





Results

NLP for validating hypotheses



Figure 5: "Heuristic" before and after 1994.



Results

NLP for validating hypotheses

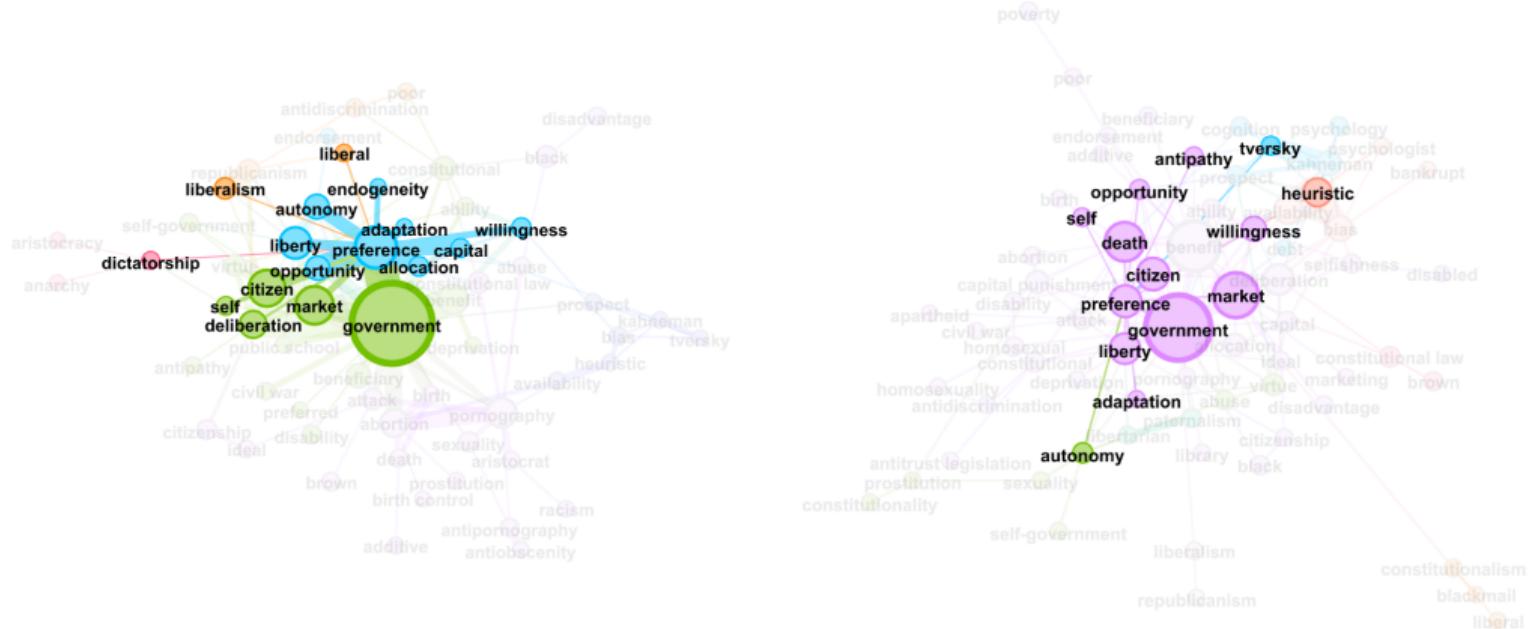


Figure 6: "Preferences" before and after 1994.



Results

NLP for validating hypotheses

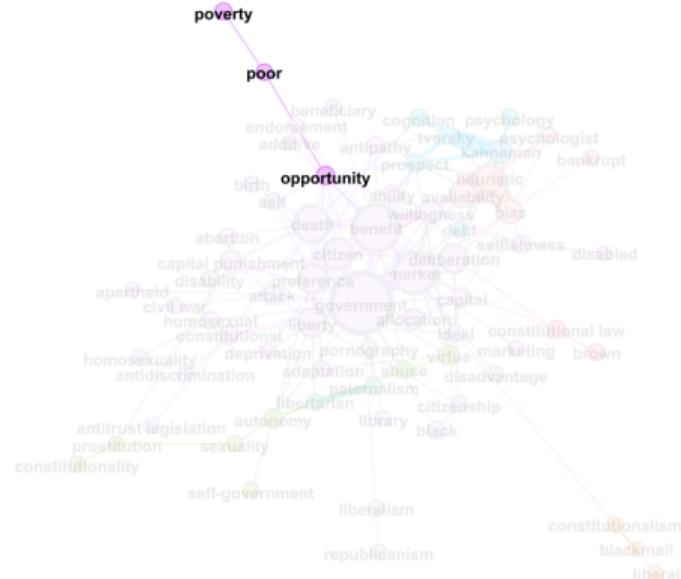
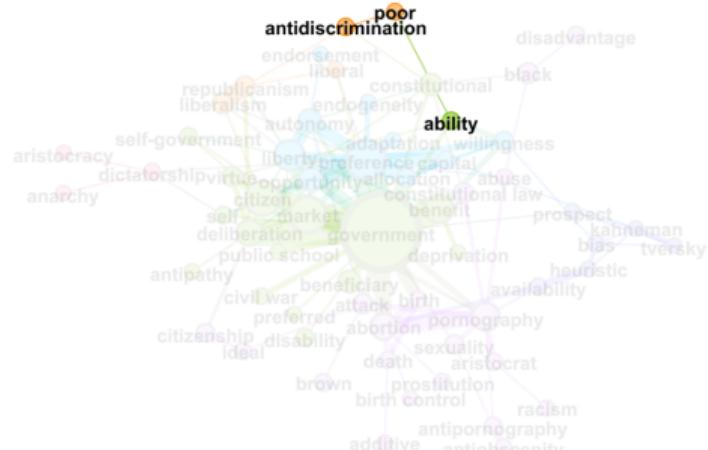


Figure 7: "Poor" before and after 1994.



Results

NLP for validating hypotheses

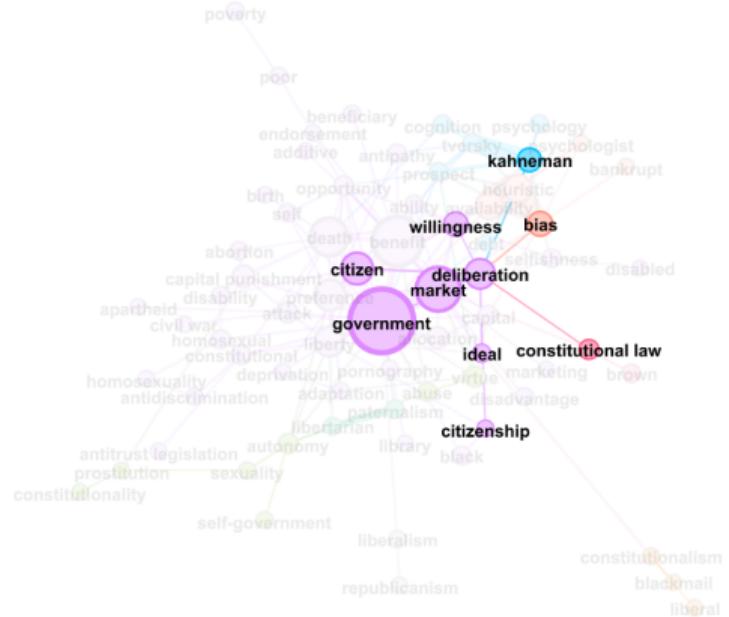
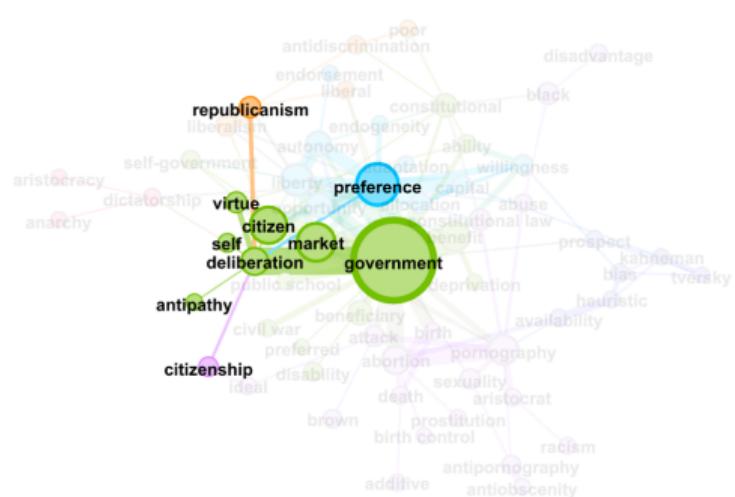


Figure 8: "Deliberation" before and after 1994.



Global results: 1987-1993 / 1994-2005

NLP for validating hypotheses

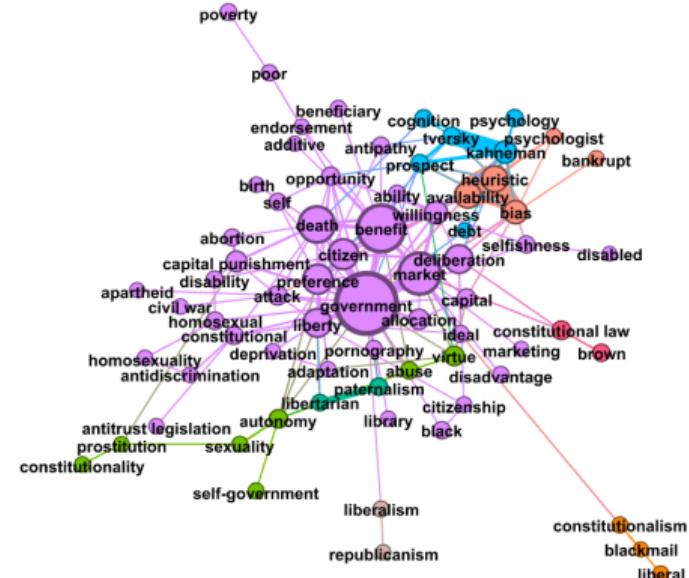
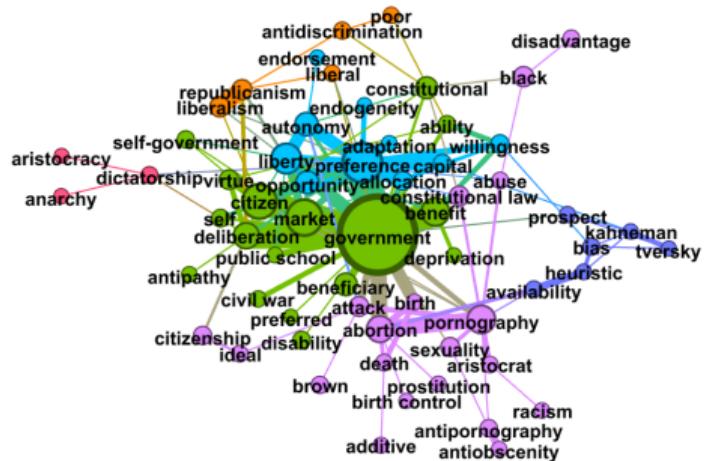


Figure 9: The entire graph before and after 1994.



Table of Contents

NLP tools as a source of new hypotheses

- ▶ Introduction
- ▶ Corpus acquisition and processing
- ▶ NLP for validating hypotheses
- ▶ NLP tools as a source of new hypotheses
- ▶ Conclusion



What is topic modeling?

NLP tools as a source of new hypotheses

- Extraction of main themes
- Topics
- Weights



Latent Dirichlet Allocation

NLP tools as a source of new hypotheses

LDA [Blei et al., 2003] solves this problem:

- One of the first unsupervised methods
- Words that appear in the same context are part of the same topic
- We can select the number n of topics



Results

NLP tools as a source of new hypotheses

1987 (topics 1 and 2)	
.011*power	.017*state
.008*administrative	.014*court
.008*judicial	.012*statute
.008*understanding	.010*regulatory
.008*contract	.010*private
.006*state	.009*democracy
.006*agency	.009*protection
.006*justice	.009*provide
.006*lochner	.009*discrimination
.006*system	.009*agency

1995 (topics 1 and 2)	
.020*right	.009*lawyer
.011*people	.007*theory
.009*rule	.007*social
.008*category	.006*people
.008*constitutional	.006*constitutional
.007*political	.006*political
.007*particular	.006*general
.006*might	.005*right
.006*think	.005*rule
.006*change	.005*agreement

2005 (topics 1 and 2)	
.009*would	.014*precautionary
.008*people	.013*principle
.008*paper	.010*discount
.007*activity	.008*risk
.006*agency	.007*regulation
.006*perhaps	.007*supra
.006*engage	.006*benefit
.006*relatively	.020*sunstein
.006*procedure	.011*market
.006*likely	.011*posner



Cons

NLP tools as a source of new hypotheses

This method still has flaws:

- Still relies on expert input
- Not consistent
- Depends on the n value



Dynamic and Embedding Topic Modeling

NLP tools as a source of new hypotheses

- [Blei and Lafferty, 2006] introduces Dynamic Topic Modeling (DTM)
- [Dieng et al., 2020] also upgrades topic modeling to Embedding Topic Modeling (ETM)

While we did not use either, it led us to a great discovery...



Bertopic

NLP tools as a source of new hypotheses

BERTopic [Grootendorst, 2022]: topic modeling algorithm using word embeddings (through BERT) to represent the evolution of topics.





What is BERT?

NLP tools as a source of new hypotheses

Bidimensional Encoder Representations from Transformers (BERT)

- Series of large transformer models
- Shown to perform very well in a large array of NLP tasks
- Internally relies on word embeddings



Better results!

NLP tools as a source of new hypotheses

Topic #	Words
0	punitive, awards, damages, jury, juries, dollar, judgments, punishment, punitive damages, outrage
1	speech, amendment, free, first amendment, pornography, first, government, free speech, expression, viewpoint
2	id, see id, see, id id, id 38, 38, 28 id, id 535, id 67, 66 id
3	discrimination, affirmative, affirmative action, action, racial, race, caste, equality, black, principle
4	president, executive, power, congress, presidential, framers, powers, control, authority, independent
5	us, 1976, see, 1973, usc, united, cir, 424, eg, 410
6	epa, air, pollution, clean, environmental, act, standards, clean air, emissions, air act
7	risk, risks, people, probability, slovic, availability, heuristic, fear, availability heuristic, information
8	interpretation, statutory, interpretive, statutes, meaning, principles, courts, norms, legislative, text
9	standing, injury, plaintiff, article, injury fact, suit, court, congress, iii, article iii



Groups of topics

NLP tools as a source of new hypotheses

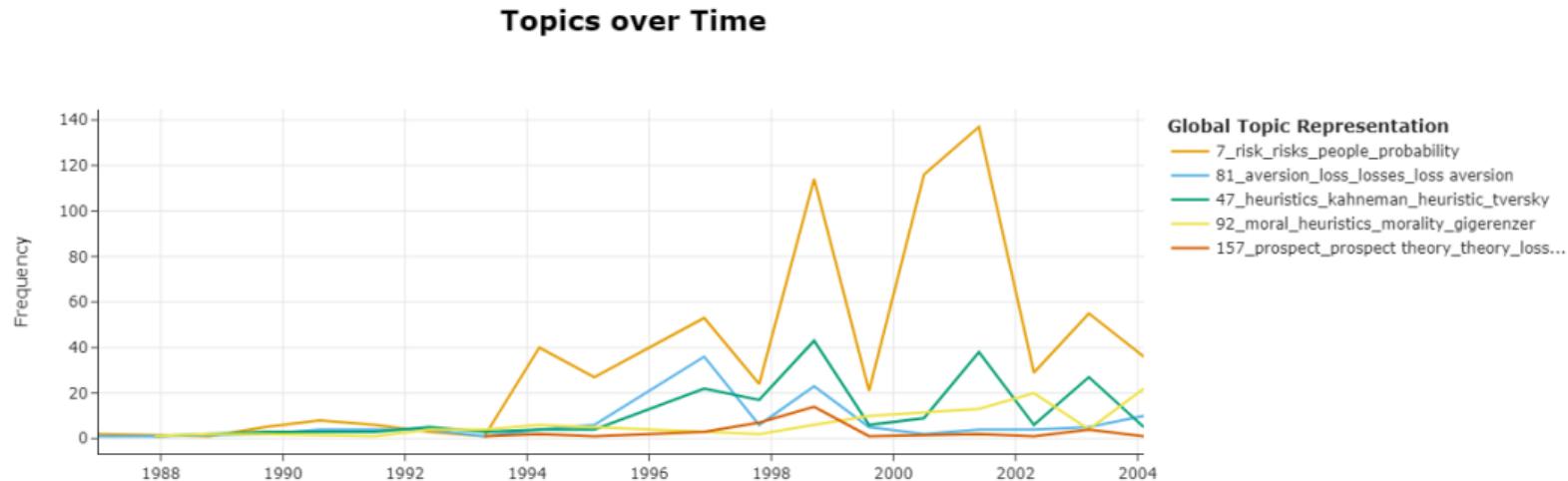
Category	Topics
Political philosophy	36, 40, 44, 51, 75, 130
Economic approaches	24, 146
Ethics	125, 134
Judicial interpretations	19, 33, 39, 50, 155
Libertarian paternalism	13, 74
Moral philosophy	3, 13, 18, 35, 61, 74, 121
Behavioral economics	7, 47, 81, 92, 157
Endowment effect	65
Miscellaneous	12, 55

Table 2: Groups of topics extracted from BERTopic, labeled by prof. Ferey.



Example of frequency analysis

NLP tools as a source of new hypotheses





Example of frequency analysis

NLP tools as a source of new hypotheses

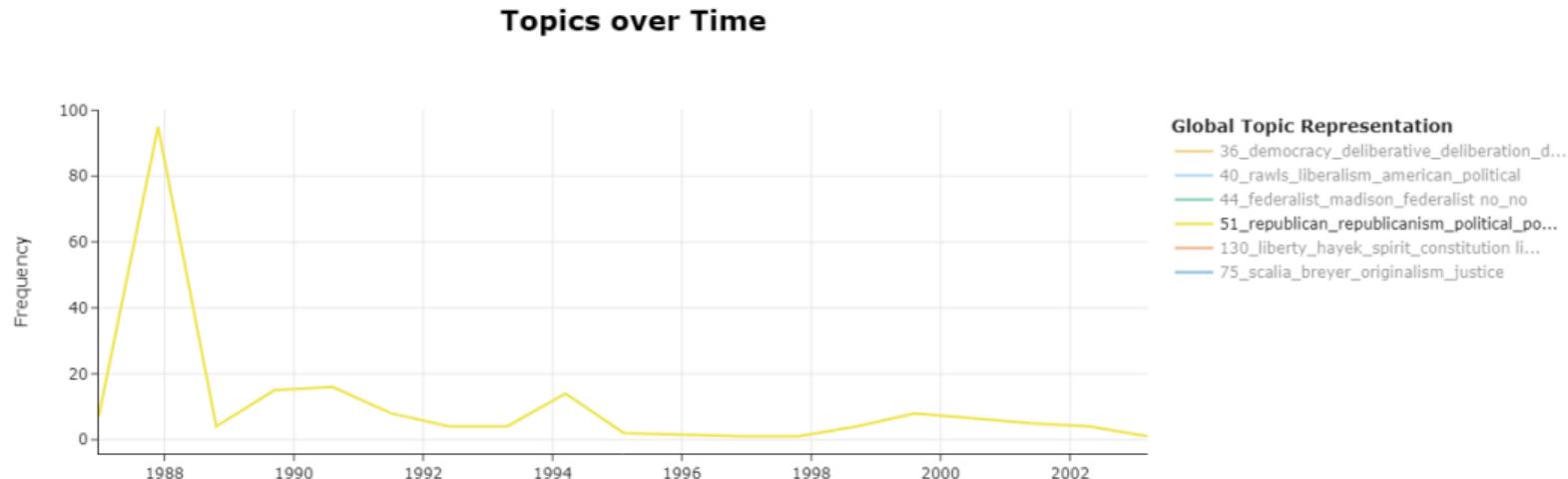




Table of Contents

Conclusion

- ▶ Introduction
- ▶ Corpus acquisition and processing
- ▶ NLP for validating hypotheses
- ▶ NLP tools as a source of new hypotheses
- ▶ Conclusion



Results and interpretation

Conclusion

- The hypothesis seems true, through multiple methods
- Republican → Behavioral economics
- Sub-topics: poverty, deliberation, preferences
- Objective text interpretation?



Takeaways?

Conclusion

- Applying many methods from the master's
- Customer - consultant relationship
- Good preparation for interdisciplinary work - common in NLP



Q&A

*Thank you for listening!
Your feedback will be highly appreciated.*



References I

Conclusion

-  Ahmed, A. and Xing, E. (2010).
Staying Informed: Supervised and Semi-Supervised Multi-View Topical Analysis of Ideological Perspective.
In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1140–1150, Cambridge, MA. Association for Computational Linguistics.
-  Ambrosino, A., Cedrini, M., Davis, J., Fiori, S., Guerzoni, M., and Nuccio, M. (2018).
What topic modeling could reveal about the evolution of economics.
Journal of Economic Methodology, 25:1–20.



References II

Conclusion

-  Blei, D. M. and Lafferty, J. D. (2006).
Dynamic topic models.
In *Proceedings of the 23rd international conference on Machine learning*, ICML '06, pages 113–120, New York, NY, USA. Association for Computing Machinery.
-  Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003).
Latent dirichlet allocation.
The Journal of Machine Learning Research, 3(null):993–1022.
-  Catherine, H. and Doehne, M. (2018).
Five reasons for the use of network analysis in the history of economics.
Journal of Economic Methodology, 25(4):311–328.



References III

Conclusion

-  Culpeper, J. and Archer, D. (2020).
Shakespeare's language: Styles and meanings via the computer.
Language and Literature: International Journal of Stylistics,
29:096394702094943.
-  Dieng, A. B., Ruiz, F. J. R., and Blei, D. M. (2020).
Topic Modeling in Embedding Spaces.
Transactions of the Association for Computational Linguistics, 8:439–453.
-  Grootendorst, M. (2022).
Bertopic: Neural topic modeling with a class-based tf-idf procedure.
arXiv preprint arXiv:2203.05794.



References IV

Conclusion

-  Jelveh, Z., Kogut, B., and Naidu, S. (2018).
Political Language in Economics.
SSRN Scholarly Paper ID 2535453, Social Science Research Network,
Rochester, NY.
-  Truc, A. (2021).
Forty Years of Behavioral Economics.
SSRN Scholarly Paper ID 3762621, Social Science Research Network,
Rochester, NY.



What would we change?

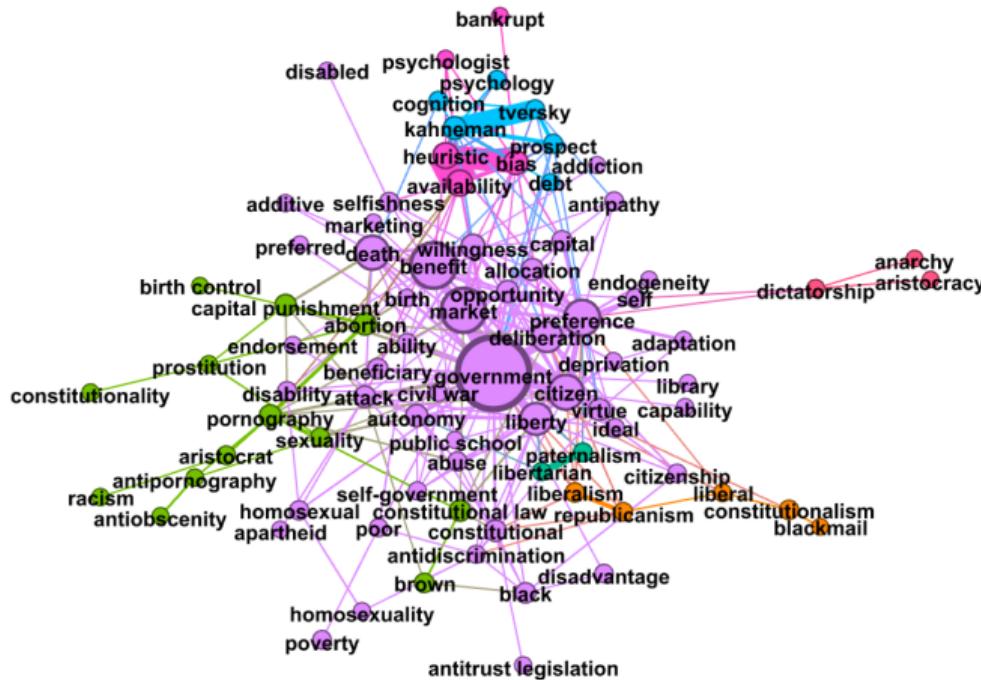
Conclusion

- Rely more on existing work (literature)
- Corpus cleaning: start automation earlier
- Give ourselves more deadlines



Global results: 1987-2005

Conclusion





Text extraction

Conclusion

- Question of footnotes, de-hyphenation?
- Mostly LAPDFText (80%), manual footnote insertion
- CERMine for the remaining articles (more footnote work)
- 4 bad articles: OCRmyPDF + manual de-hyphenation



Text extraction

Conclusion

Neglect CASSR.SUNSTEIN csunstei@uchicago.edu
KarNl.LlewellDyinstS.servicPer,ofofJurisprudLeanwcSechooDl,epartmoefnPtoliticSac
lience, UniversoiftCyhicagLoawSchoo11,111East60thStreeCth,icagIoL,60637U, SA
Whenstronegmotioanrseinvolvепde,oplteendtofocuosnthebadnesosftheoutcomrea,thet
rhanonthe
probabiltihtyattheoutcomweilolccurTheresulti"npgrobabilnietgylecht"elptsoexpl
aienxcessirveactions tolow-probabriilsitkoysfcatastropTheer.rorisshtoswa
workinkgnowledogfeprobabilnitezylecptr,oducing
publifcearthamtighgtreatelyxeetdhedisountheadrmA.sa
resulotfprobabilnitezylecpte,oploefteanre
farmorceoncernaebdouttheriskosfterroritshmanaboustatisticlaalrlygerrisktshat
thecyonfroinotrdinary
lifeIntheconteoxftteroriasnmdanalogoruisskts,helegaslystefmrequenrtelsypontd
osprobabilnitezylect,
resultiinnregulatitohnamtighbteunjustifoireedvencounterproduBcuttipvueb.lifc
earisitselafcosta,nd itisassociatwedithmanoythecrosts,nt hefor
mof"rippeleffectpsr"oducedbyfearA.sa normatmivaetter,
governmsehnnotulrdeduceevenunjustiffieeadrif,thethebenefoitftsheresponcsaenbeshow
tnooutweithhecots. JELClassificatioKno:D,8 Terroristsshowa workingknowledgeof
threenoteworthypointsaboutfear.Of these, thefirsttwoare well-knownT.he

Figure 10: An article requiring the use of OCRmyPDF.



Corpus evaluation

Conclusion

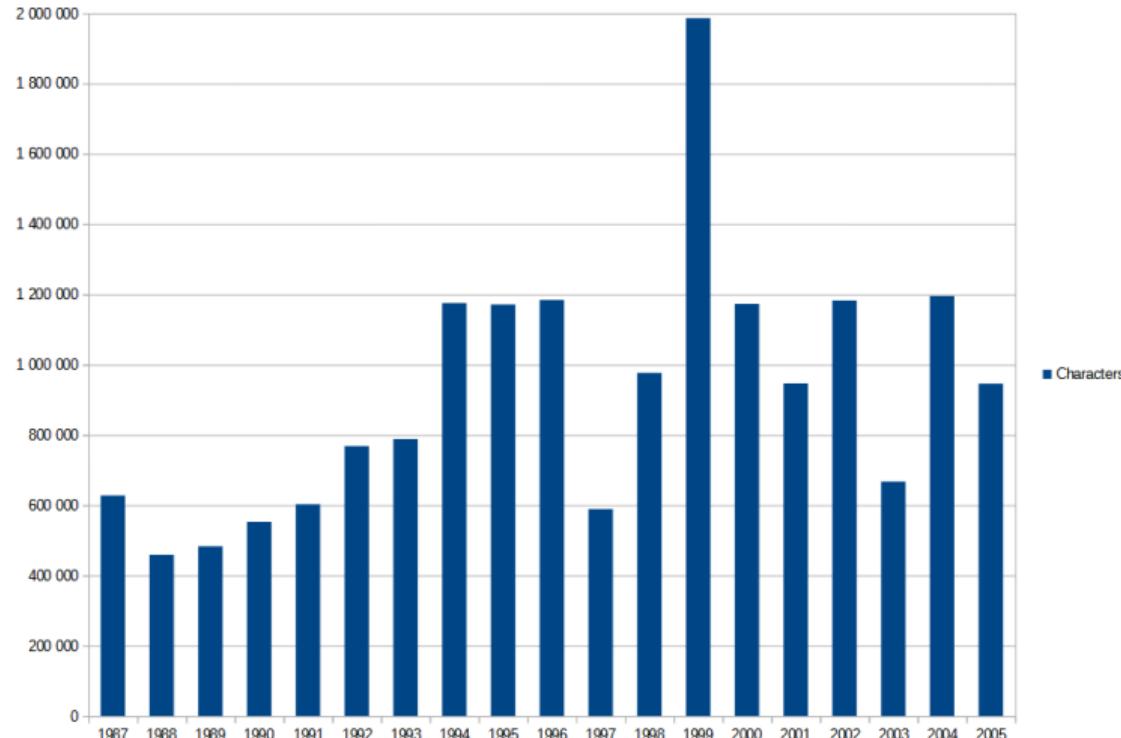
Corpus	Minimum	Maximum	Average	Std. deviation
CERMine	11.618	269.665	22.057	20.778
LAPDFTText	11.018	39.360	18.418	3.315
Manual cleaning	10.995	23.275	17.085	2.268
Manual + auto. cleaning	10.495	22.130	16.705	2.151

Table 3: Statistics about the average sentence perplexity computed over all articles.



Resulting corpus

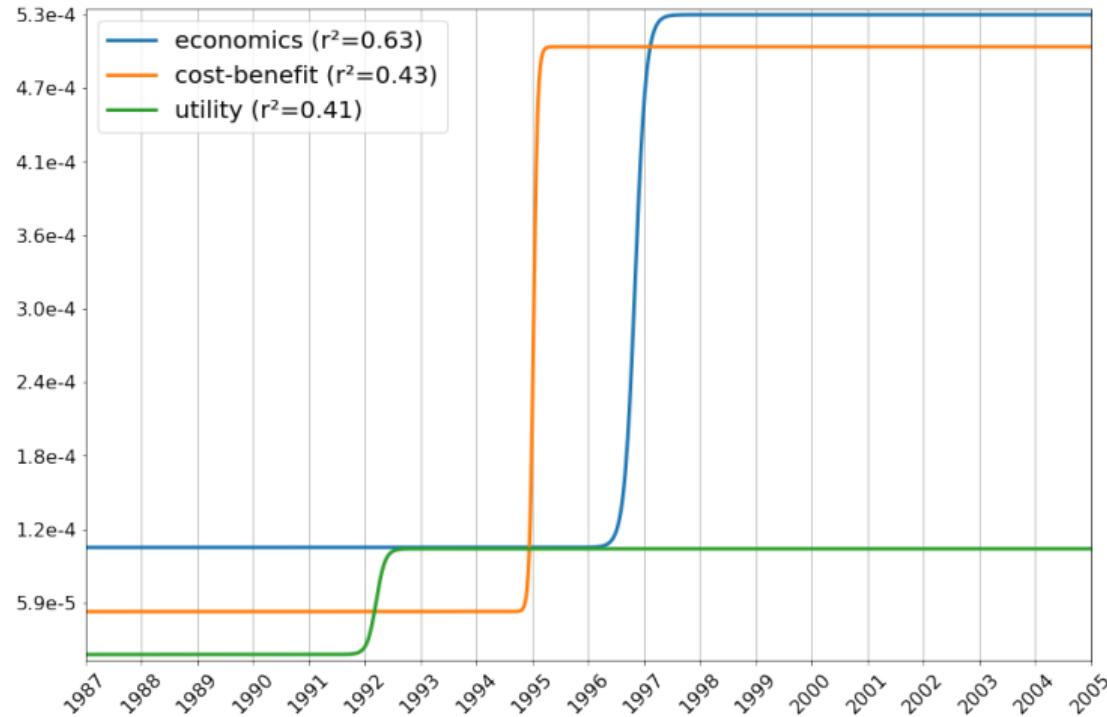
Conclusion





Results 3/3

Conclusion





Supervised approaches

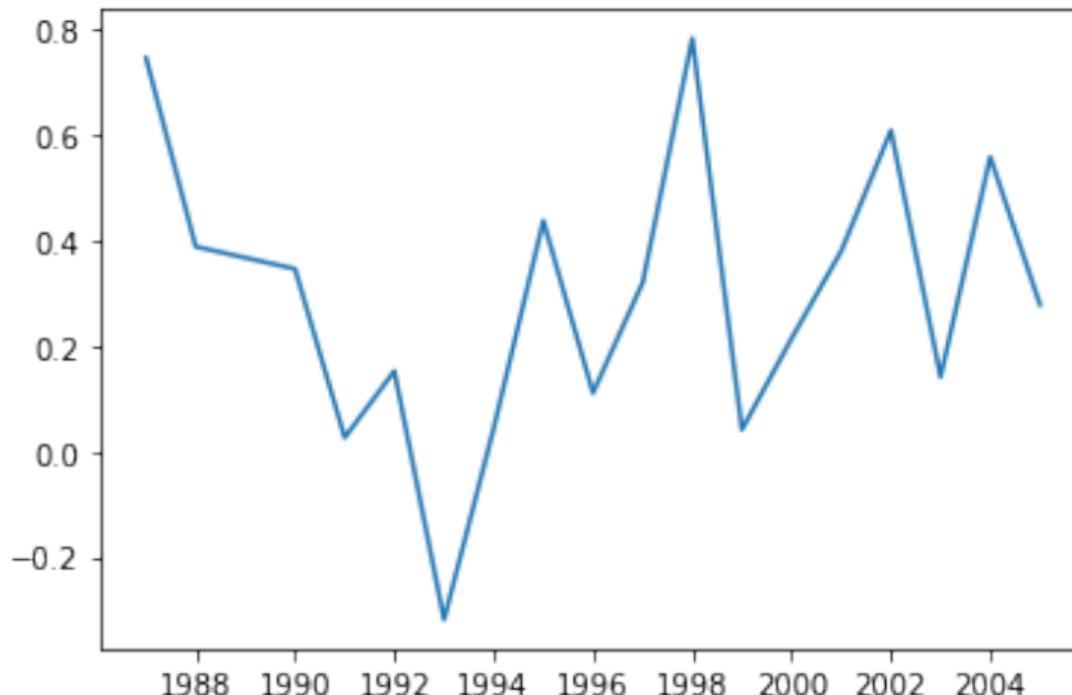
Conclusion

companies cut john families kids class american governor nuclear give fight gore ago back jim americans history fund oil didnt year country 1 budget cuts job jobs al 000 laden bin agree national lost kerry ill **years** presidents rights today bush health **president** parents middle number united choice social children schools left college debt countries day **america** insurance drug **security** big bring general things theyve **plan** school percent weapons program support benefits forces **question** means care put bill respect states theyre war vice **world** fact **tax** thing ive pay problem talk military iraq great trillion **im** life medicare billion million **good** public safe congress prescription education time kind **people** difference terrorists **dont** wrong long 2 made **make** hussein change important saddam hes clear drugs senate administration law **money** working doesnt man spending mr peace making part lead leadership nation high intelligence policy **troops** **government** move programs coming destruction child find threat business lot side **weve** called issue interest **youre** voted small state seniors energy hard lets afghanistan **strong** decision qaida thought deal **work** end local sense set vote marriage terror problems wont protect gun understand **federal** hope reform **system** increase nations matter senator talks continue record texas place lives east folks **taxes** freedom decisions washington citizens free opponent relief youve



Sentiment Analysis with NLTK

Conclusion





Future work

Conclusion

- Posner-Becker blog
- [Jelveh et al., 2018]
- [Ahmed and Xing, 2010]