Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It

Matthew J. Denny[1] Arthur Spirling

Penn State University New York University

October 15, 20016

[1]Work supported by NSF Grant: DGE-1144860

Text-As-Data Research

1. Awesome Research Design!
2. Collect Awesome Text Data!
3. ...
4. Perform Awesome Analysis!
5. Publish Awesome Paper!

Raw Text

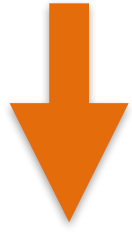
103d CONGRESS
1st Session

H. R. 3

[Report No. 103-375, Part I]

To amend the Federal Election Campaign Act of 1971 to provide for a voluntary system of spending limits and benefits for congressional election campaigns, and for other purposes.

Preprocessing



Document-Term Matrix

amend	federal	section	spending	…
56	34	20	75	…
24	13	41	0	…
…	…	…	…	…

Common Preprocessing Decisions

P – **Punctuation Removal**

N – **Number Removal**

L – **Lowercasing**

S – **Stemming**

W – **Stopword Removal**

I – **Infrequent Term Removal**

'3' – **n-gram Inclusion**

7 binary choices $\longrightarrow 2^7 = 128$ specifications.

Supervised Learning

Unsupervised Learning

What Could Possibly Go Wrong?

Motivating Example

- UK Manifestos Corpus (1918–2001)
- Labour, Liberal, Conservative Parties
- Wordfish
  - Place documents in ideological space
- Process:
  1. Select preprocessing specification
  2. Run Wordfish

1983 Labour Manifesto

Mass unemployment costs the country £15 billion, £16 billion, £17 billion a year, astronomic figures never conceived possible before, and they move higher still every month.

Mass unemployment is the main reason why most families in Britain, all but the very rich, are paying more in taxes today than they did four years ago when the Conservatives promised to cut them for everybody.

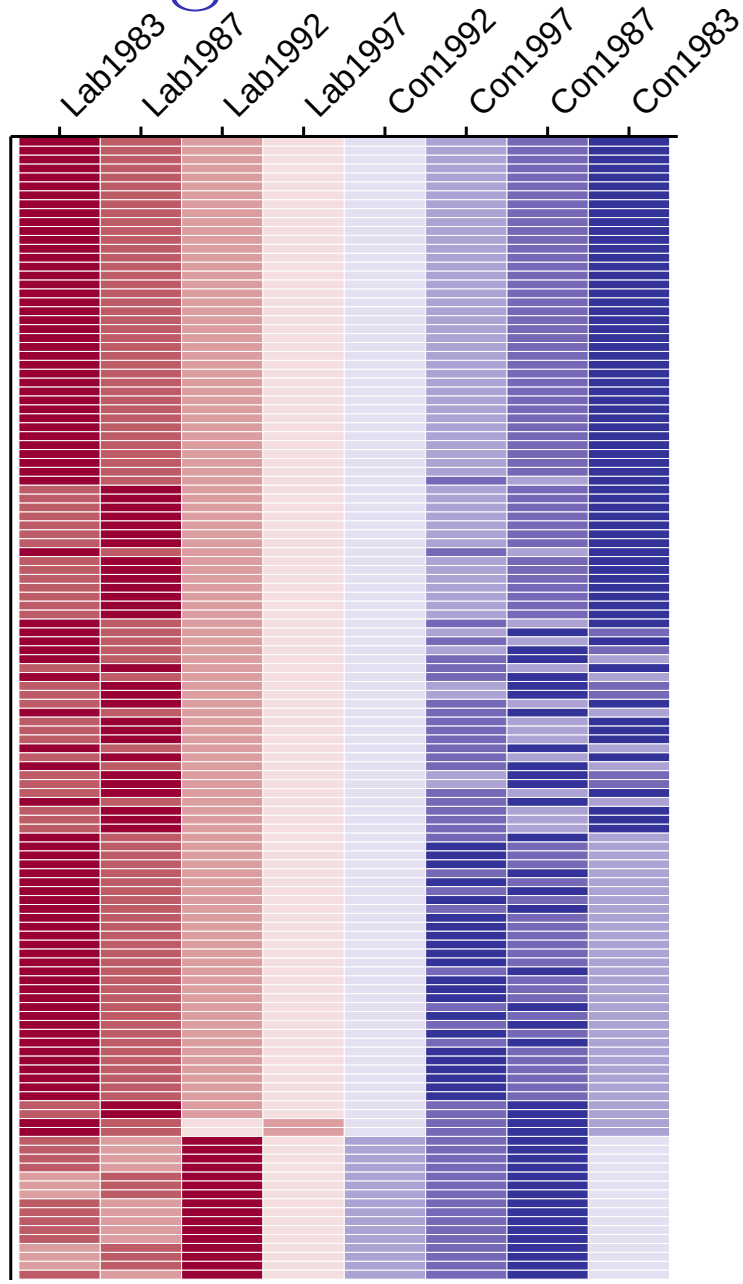
Mass unemployment is the main reason why we are wasting our precious North Sea oil riches. Since 1979 Mrs. Thatcher's government has had the benefit of £20 billion in tax revenues from the North Sea. It has all been swallowed by the huge, mounting cost of mass unemployment. And the oil won't last for ever, although, according to Mrs. Thatcher's economics, the unemployment will.

A-Priori Rankings

- Focus on 8 Manifestos.
  1. Four general elections (1983–1997)
  2. Labour and Conservative parties
- Lab $_{1983}$: "longest suicide note in history", extremely left–wing.

$$\text{Lab}_{1983} < \text{Lab}_{1987} < \text{Lab}_{1992} < \text{Lab}_{1997} <$$
$$\text{Con}_{1992} < \text{Con}_{1997} < \text{Con}_{1987} < \text{Con}_{1983}$$

Wordfish Rankings



Forking Paths

- 12 unique document rankings
- Substantially different conclusions.

Specification	Most Left	Most Right
P-N-S-W-I-3	Lab 1983	Cons 1983
N-S-W-3	Lab 1987	Cons 1987
N-L-3	Lab 1992	Cons 1987
N-L-S	Lab 1983	Cons 1992

Another Example: Topic Models

- Senate Press Releases (Grimmer, 2010)
- Sample of 1,000 documents
  - 100×10 Senators.
- Note: no n-grams (computational cost).
- Procedure:
  1. Find optimal number of topics for each specification (perplexity).
  2. Run topic model (LDA)

Sen. Sanders, April 1, 2008

"It is no secret to anyone in Vermont that the American economy today is in pretty serious trouble: that the middle class is shrinking, poverty is increasing and the gap between the very rich and everyone else is growing wider. It is also true that despite all the rhetoric about family values, the American worker now works the longest hours of anyone in a major country, and that many of our families are stressed out and exhausted.

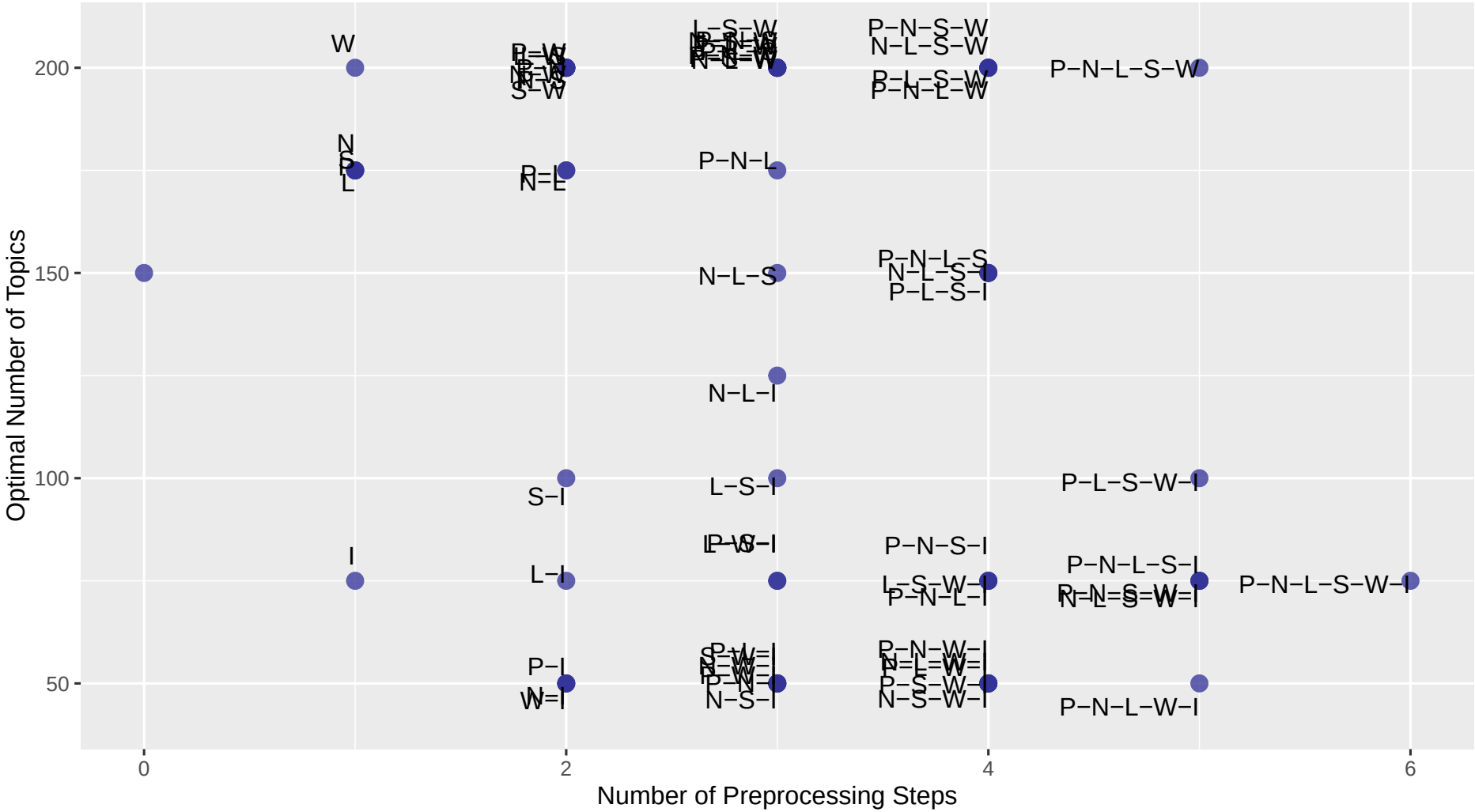
"It is no secret that our health care system is disintegrating, that 47 million Americans have no health insurance and, despite that, we spend twice as much per capita on health care as any other nation.

"It is no secret that the way we treat our children is nothing less than shameful; that we have the highest rate of childhood poverty in the industrialized world; our childcare system is totally inadequate; that too many of our kids drop out of school and that the cost of college is increasingly unaffordable. And, in my view, one of the results of how we neglect many of our children is that we end up with more people behind bars, in jail, than any other country on earth. There is a correlation between the highest rate of childhood poverty and the highest rate of incarceration."

Perplexity to Select Number of Topics

- 10-fold cross validation.
- Split data into train/test sets (80/20).
- Find minimum *perplexity* over num. topics.
- topics = $\{25, 50, 75, 100, 125, 150, 175, 200\}$

Optimal Number of Topics

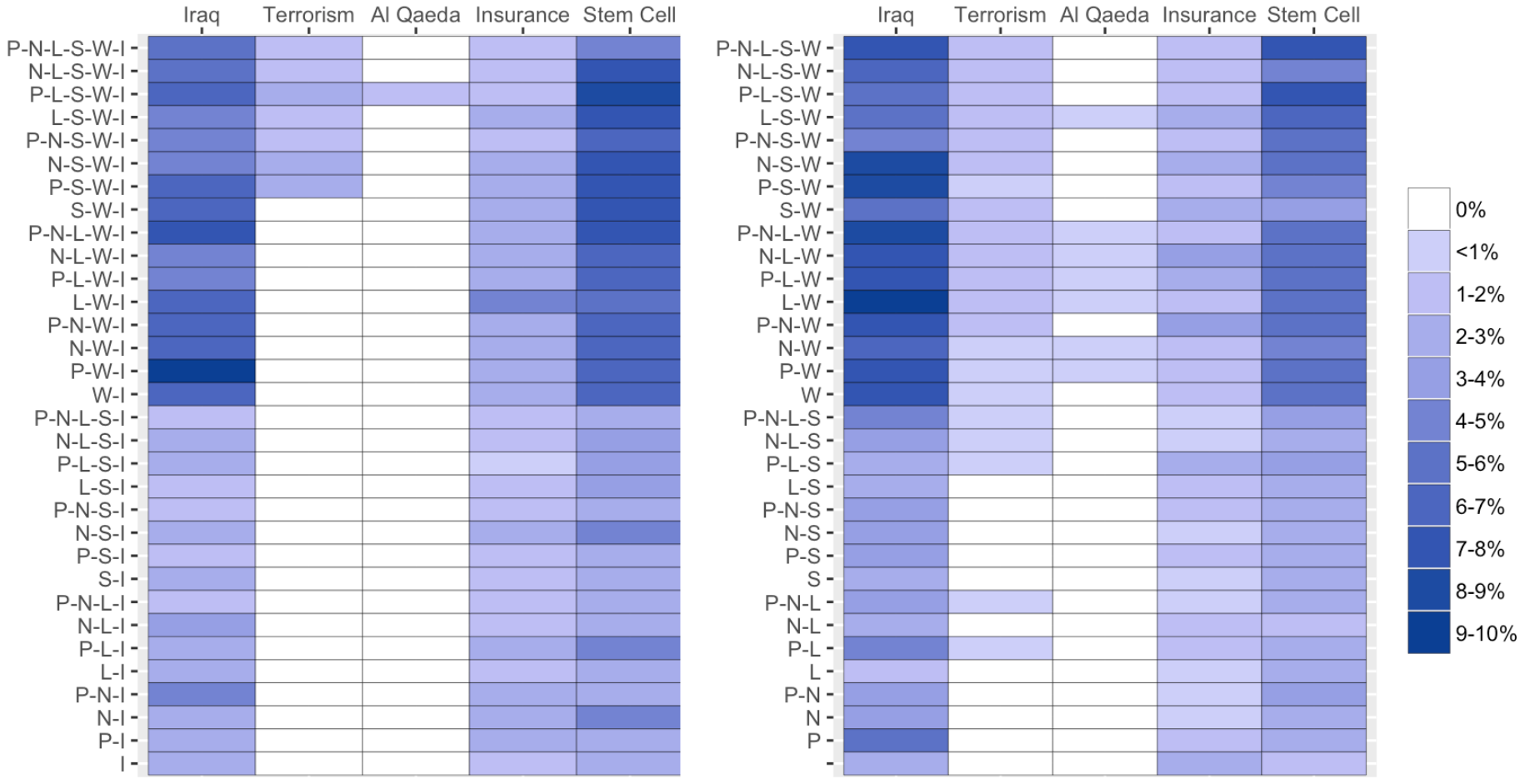


Key Terms Example

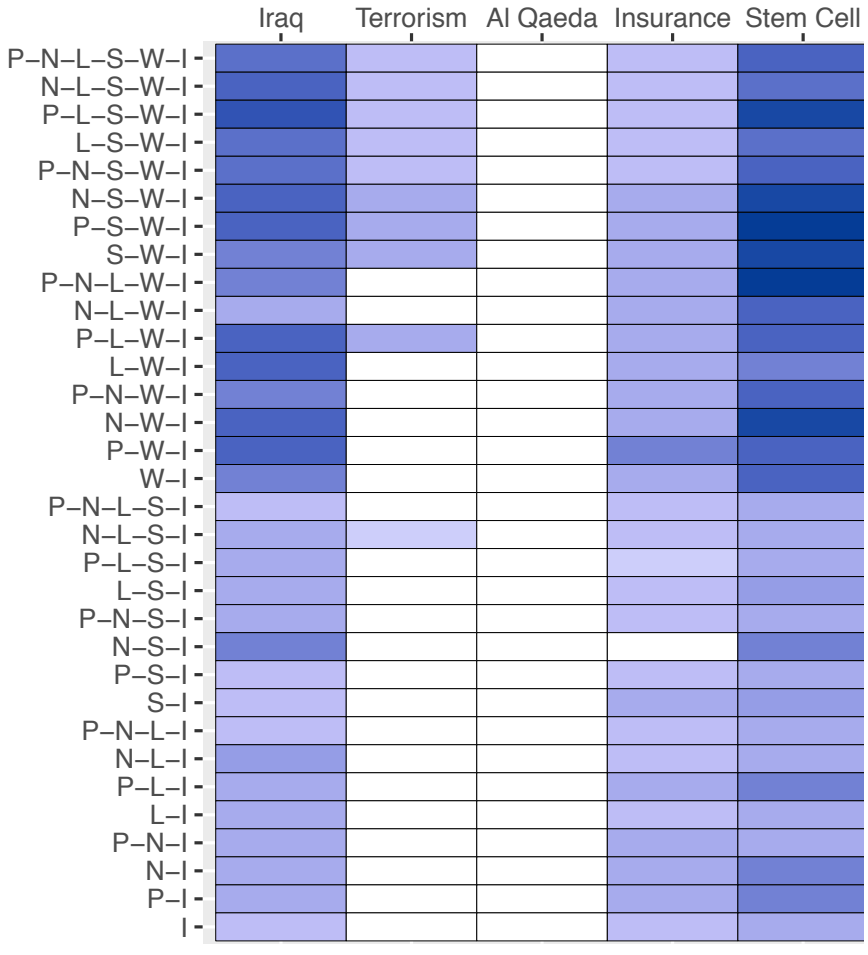
- Select five "key terms".
- How many topic top-terms are they in?

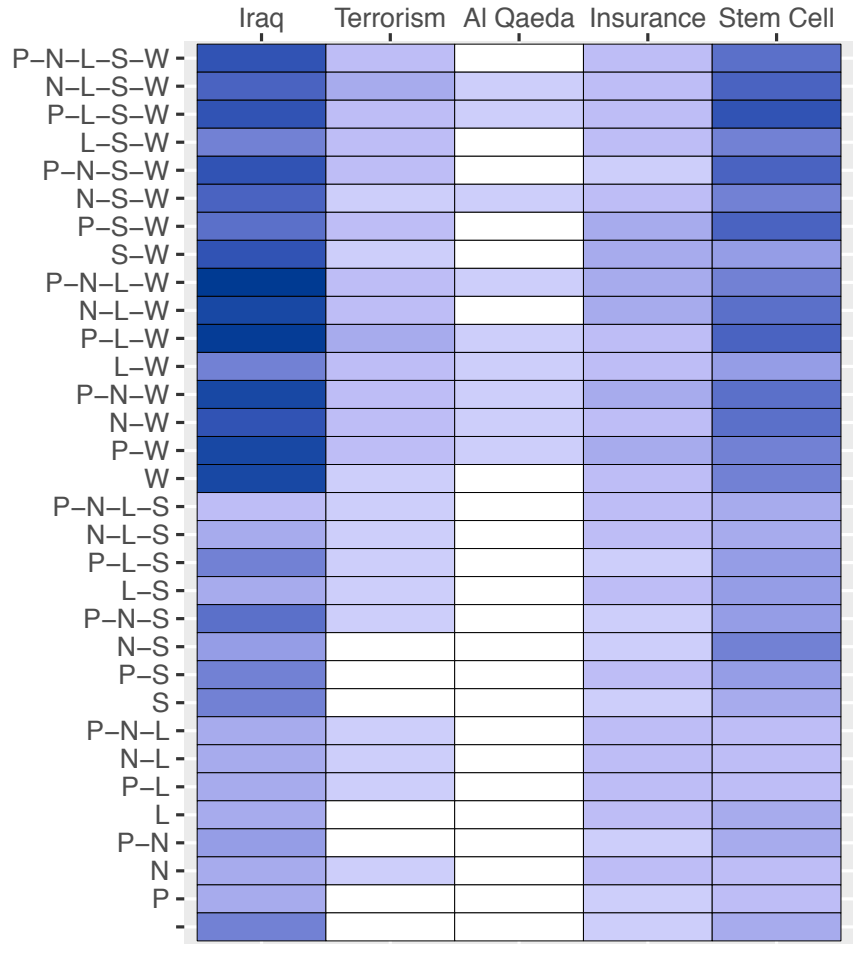
iraq
terror(ism)
(al) **qaeda**
insur(ance)
stem (cell)

Key Terms in Topic Top-Terms



Key Terms: Average of 40 Initializations





Forking Paths

- Different preprocessing $\longrightarrow$ different conclusions.
- Are we **doomed**?

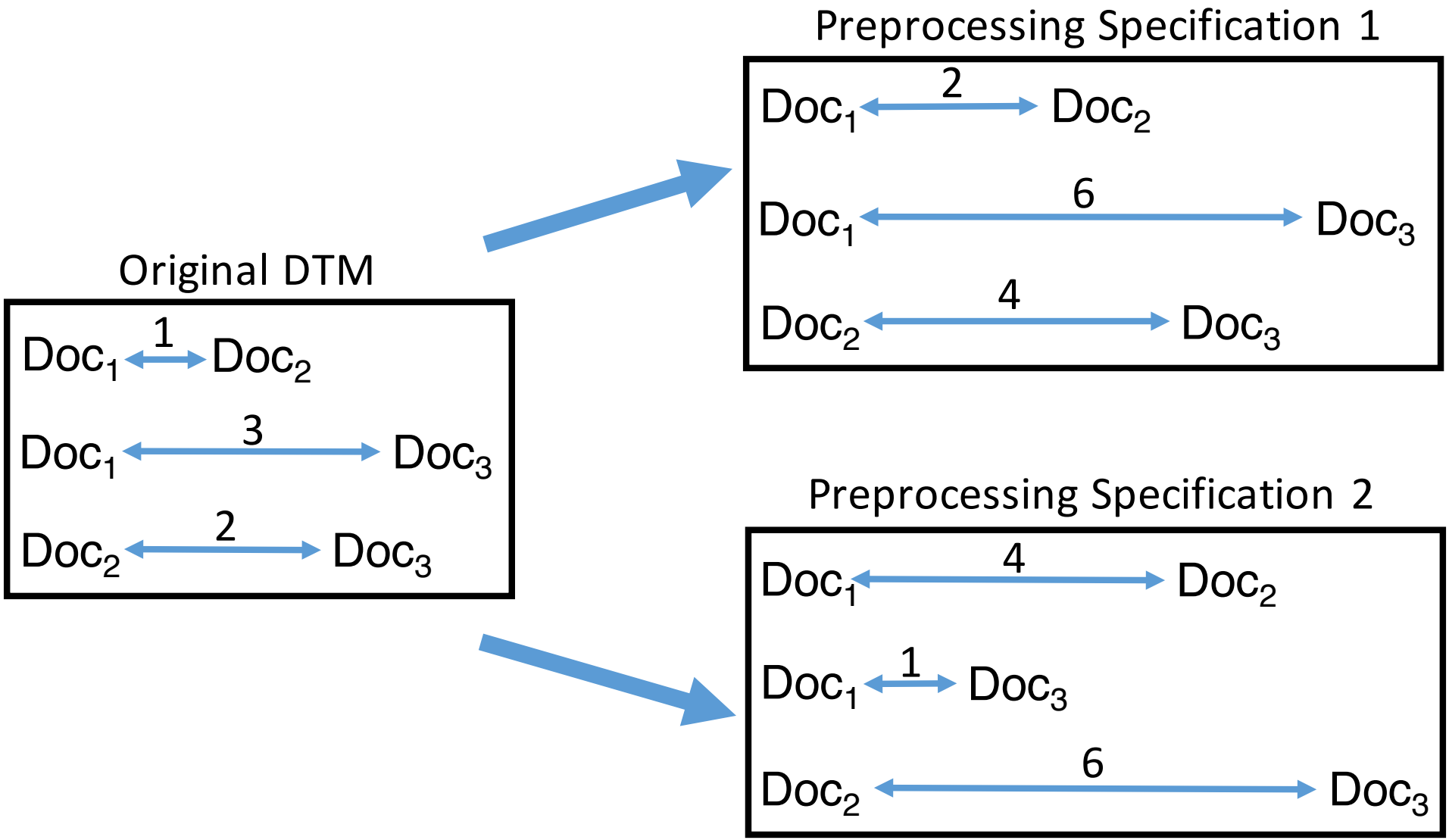
Our Solution: `preText`

- Assess consequences of preprocessing choices.
- Characterize a number of corpora.
- Easy to use `R` package!

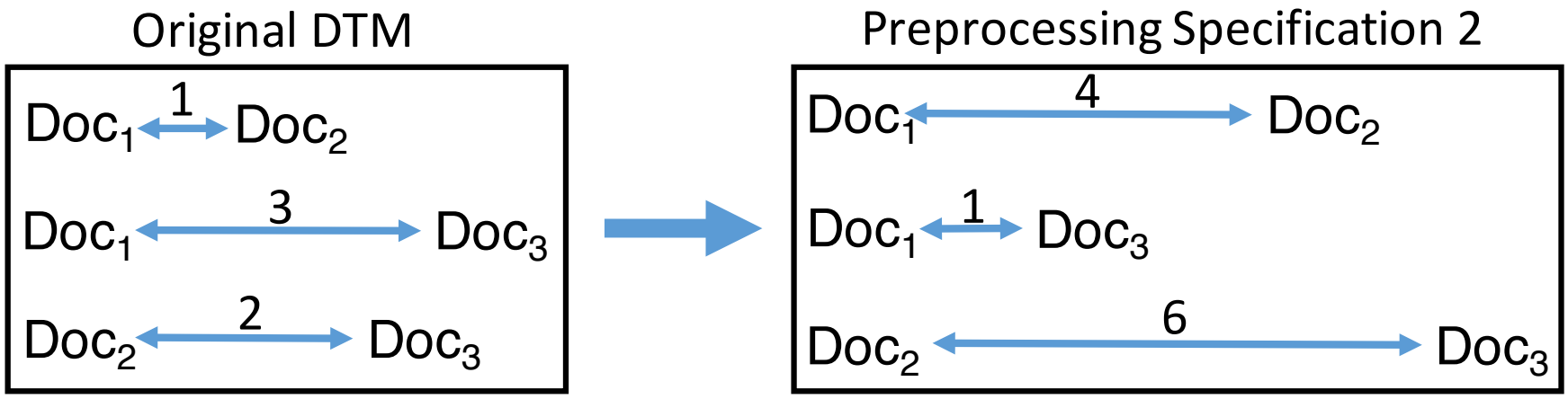
Overview: Movements in Pairwise Document Distances

- No preprocessing as base case.
- Compare how **pairwise document distances** change with preprocessing.
- Measure how unusual these changes are.

Example With Three Documents



Ranking Distance Changes



Original DTM	Preproc. Spec. 2	Abs. Difference
d(1,3) = 3	d(2,3) = 6	Δ d(1,3) = 2
d(2,3) = 2	d(1,2) = 4	Δ d(2,3) = 1
d(1,2) = 1	d(1,3) = 1	Δ d(1,2) = 1

Comparing Preprocessing Specifications

- Each specification will have a **largest mover**.
- Rank in other specifications $(M_1, ..., M_{127})$?

$$\mathbf{v_{M_1}} = (2_{M_2}, 14_{M_3}, 2_{M_4}, 3_{M_5}, \ldots, 15_{M_{127}}).$$

- Average of $\mathbf{v_{M_i}} \longrightarrow$ how **unusual**.

`preText` Scores

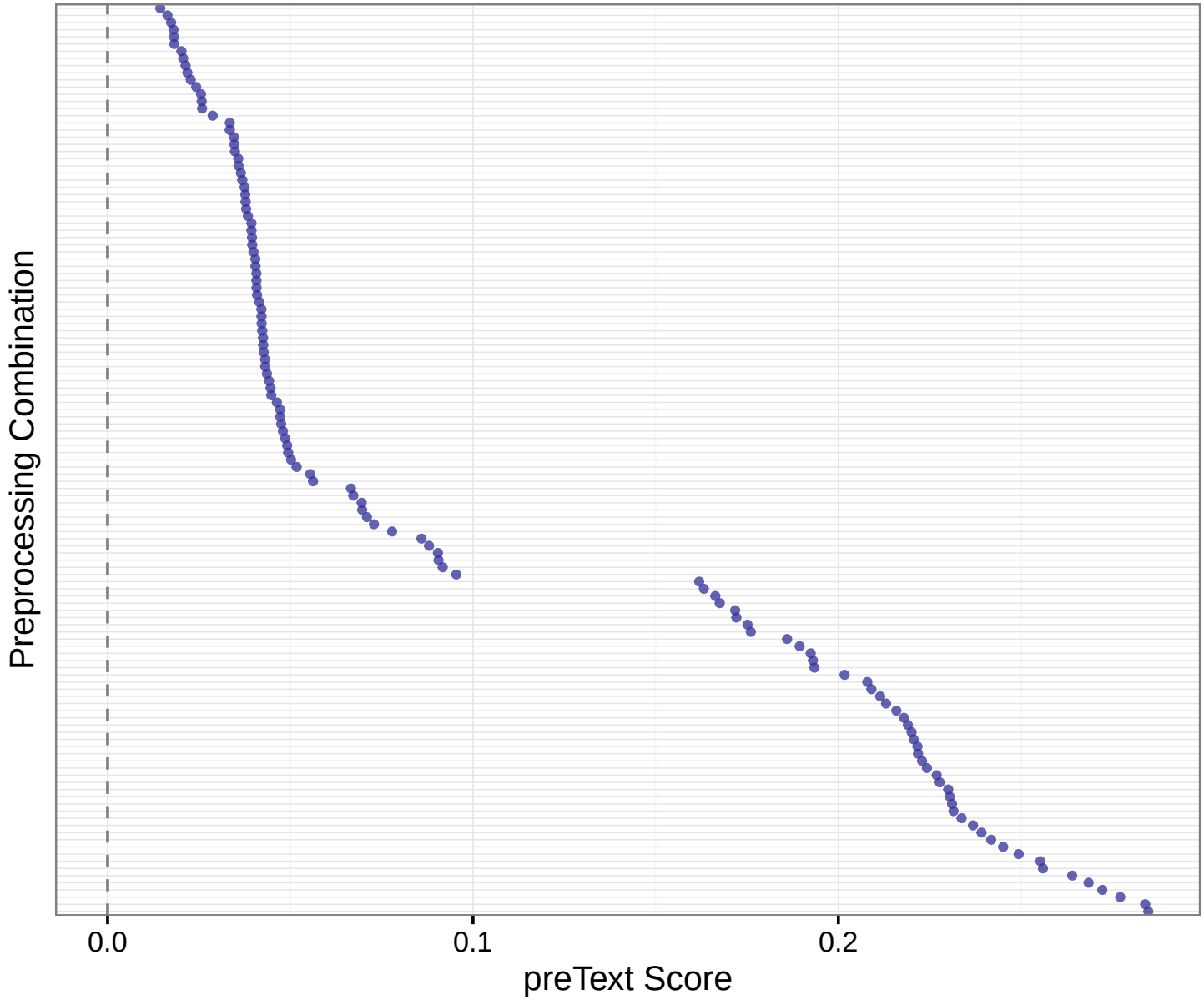
- Consider top k largest moving doc pairs.
- Average across $\mathbf{v_{M_i}} \longrightarrow \mathbf{v_{M_i}}^{(k)}$
- Normalize by $\frac{n(n-1)}{2}$ (n = num docs)

$$\texttt{preText score}_i = \frac{2\mathbf{v_{M_i}}^{(k)}}{n(n-1)}$$

Interpreting `preText` Scores

- `preText` scores range between 0 and 1.
- **Lower** score $\longrightarrow$ **"typical"** changes in document distances.
- **Higher** score $\longrightarrow$ **"atypical"** changes in document distances.

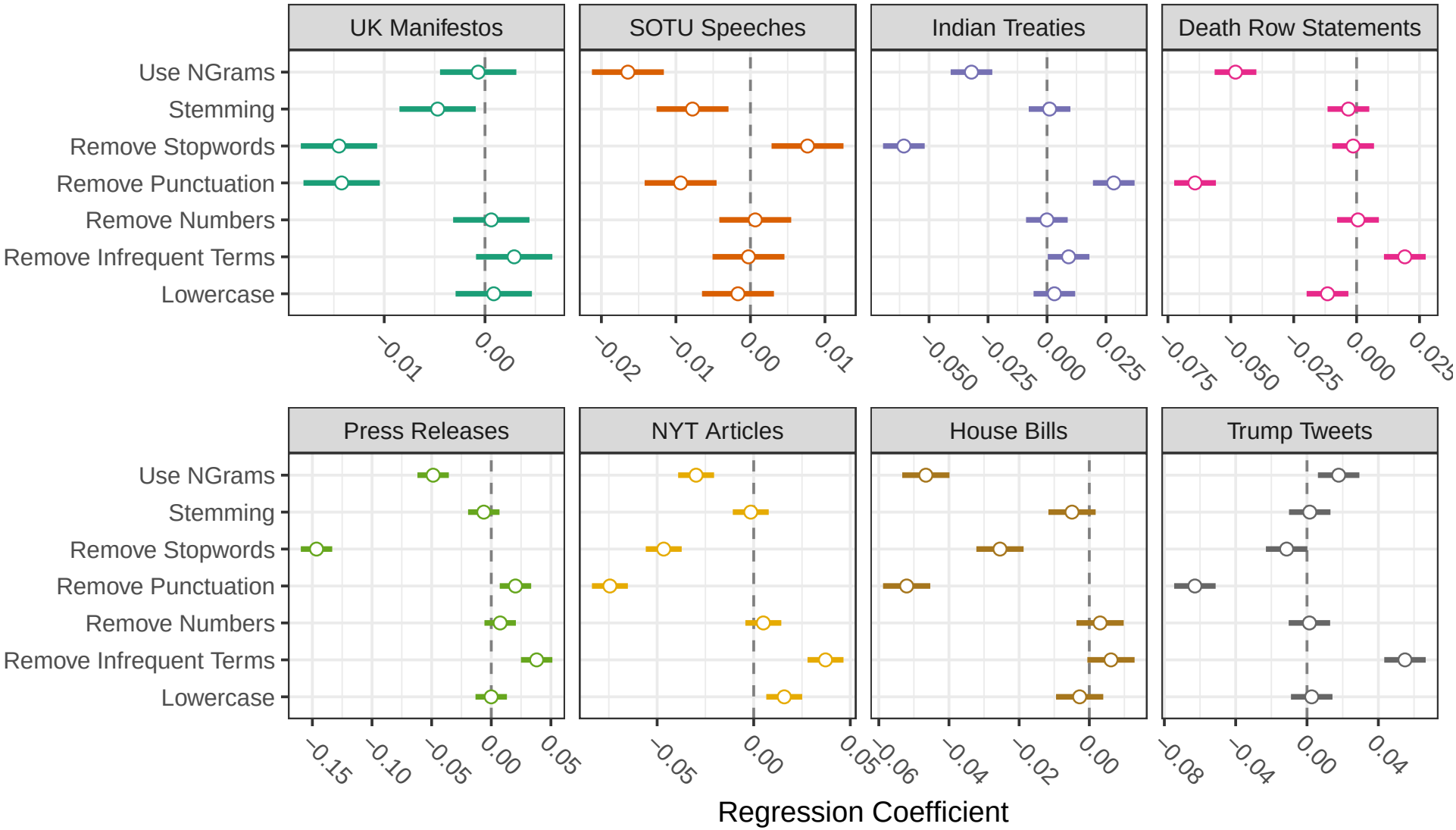
preText Scores for Press Releases



Regression Analysis

$$\texttt{preText}\ \text{score}_i = \beta_0 + \beta_1 \text{Punctuation}_i + \beta_2 \text{Numbers}_i + \beta_3 \text{Lowercase}_i + \beta_4 \text{Stem}_i + \beta_5 \text{Stop Words}_i + \beta_6 \text{N-Grams}_i + \beta_7 \text{Infrequent Terms}_i + \varepsilon_i$$

Regression Analysis Results



Different preprocessing steps "matter" for different corpora

What To Do About It

1. Significant parameter estimates serve as an "early warning".

2. Conservative approach: average results over all specifications.

3. Depends on how good your "theory" is.

4. *A priori* reasons for selecting a particular specification.

Three Cases

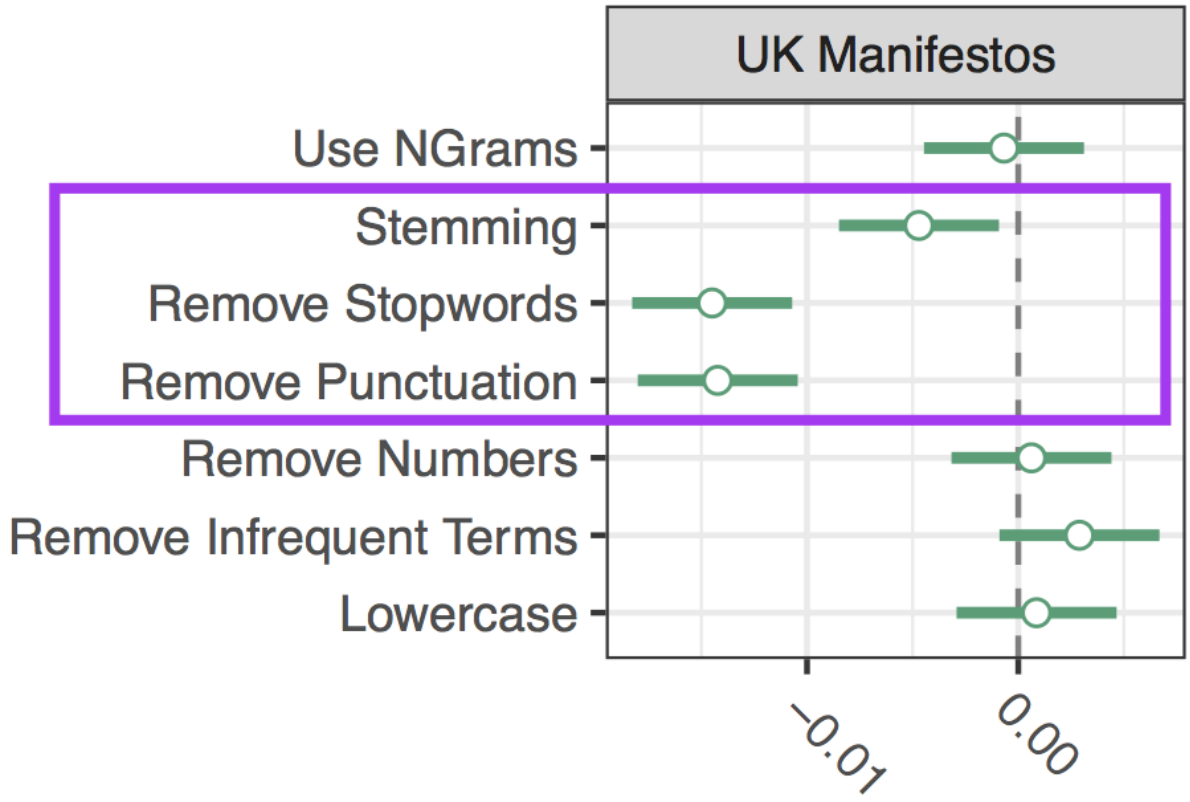
1. All Parameter Estimates Are Not Significantly Different From Zero.

2. Strong Theory, Some Parameter Estimates Are Significantly Different From Zero.

3. Weak Theory, Some Parameter Estimates Are Significantly Different From Zero.

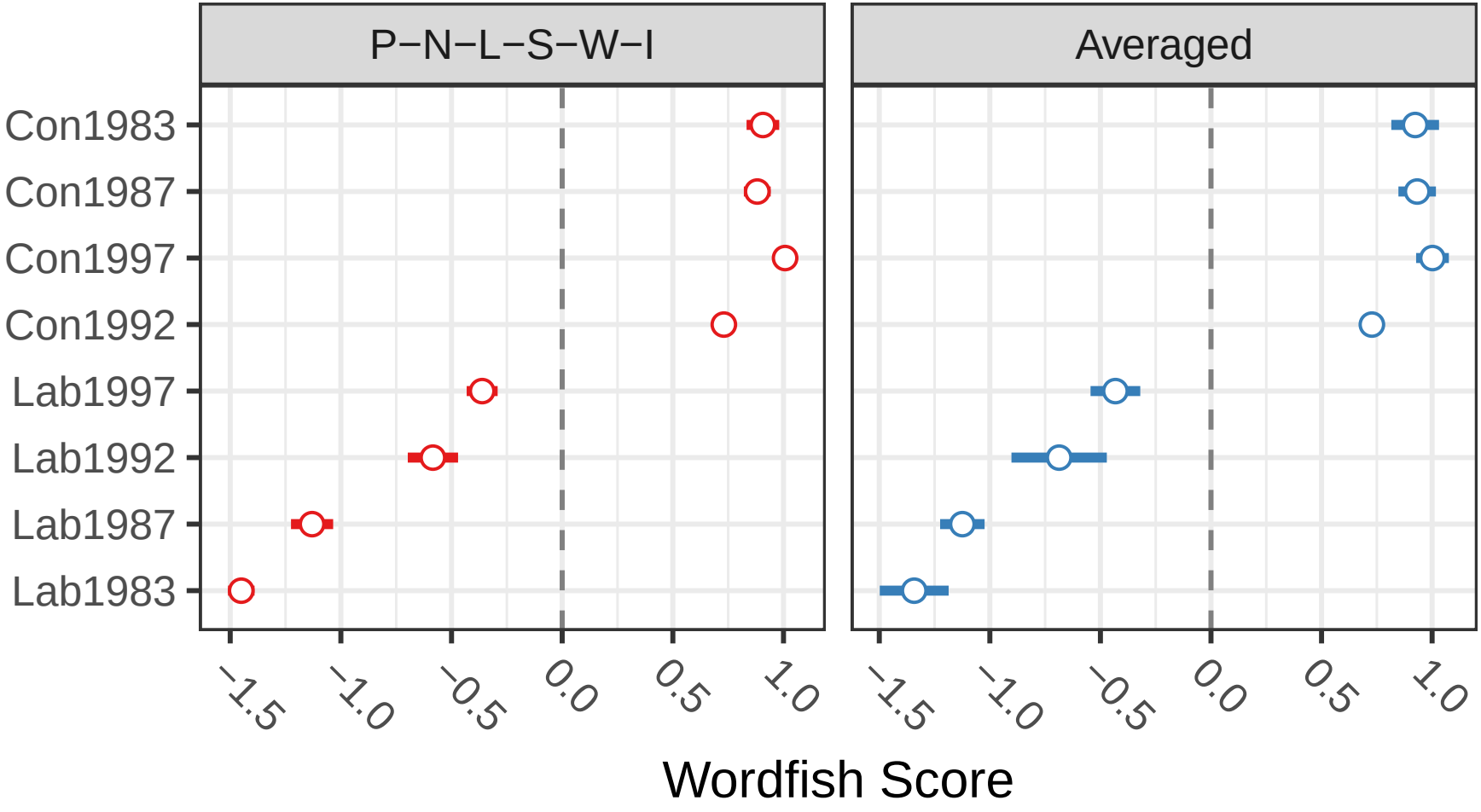
Returning To The UK Wordfish Example

- Weak "theory" $\longrightarrow$ P-N-L-S-W-I



- $2^3 = 8$ combinations of choices to average over.

Model Averaging



- **Theoretical Specification: "Wrong"!**
- **Averaged: Less "Wrong"!**

Summary

- **Preprocessing matters.**
- **Forking paths** of inference.
- Our solution: `preText`.
- General Advice:
  - Represent uncertainty.
  - **Always check – tell reader!**

Software and Paper

- `install.packages("preText")`
- ssrn.com/abstract=2849145
- github.com/matthewjdenny/preText