Assessing the Consequences of Text Preprocessing Decisions

Matthew J. Denny\textsuperscript{1}  Arthur Spirling

Penn State University  New York University

October 15, 20016

\textsuperscript{1}Work supported by NSF Grant: DGE-1144860
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 \( \rightarrow 2^7 = 128 \) 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
A-Priori Rankings

- Focus on 8 Manifestos.

2. Labour and Conservative parties


<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;</td>
<td>&lt;</td>
<td>&lt;</td>
<td>&lt;</td>
<td>&lt;</td>
<td>&lt;</td>
<td>&lt;</td>
<td>&lt;</td>
</tr>
</tbody>
</table>
Wordfish Rankings
Forking Paths

- 12 unique document rankings

- Substantially different conclusions.

<table>
<thead>
<tr>
<th>Specification</th>
<th>Most Left</th>
<th>Most Right</th>
</tr>
</thead>
<tbody>
<tr>
<td>P-N-S-W-I-3</td>
<td>Lab 1983</td>
<td>Cons 1983</td>
</tr>
<tr>
<td>N-S-W-3</td>
<td>Lab 1987</td>
<td>Cons 1987</td>
</tr>
<tr>
<td>N-L-3</td>
<td>Lab 1992</td>
<td>Cons 1987</td>
</tr>
<tr>
<td>N-L-S</td>
<td>Lab 1983</td>
<td>Cons 1992</td>
</tr>
</tbody>
</table>
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)
Perplexity to Select Number of Topics

- Split data into train/test sets (80/20).

- Find minimum perplexity over num. topics.

- topics = \{25, 50, 75, 100, 125, 150, 175, 200\}

- 10-fold cross validation.
Optimal Number of Topics

Number of Preprocessing Steps

Optimal Number of Topics
Key Terms Example

- Select five “key terms”.

- How many topic top-terms are they in?

<table>
<thead>
<tr>
<th>iraq</th>
</tr>
</thead>
<tbody>
<tr>
<td>terror (ism)</td>
</tr>
<tr>
<td>(al) qaeda</td>
</tr>
<tr>
<td>insur (ance)</td>
</tr>
<tr>
<td>stem (cell)</td>
</tr>
</tbody>
</table>
Key Terms in Topic Top-Terms
Key Terms: Average of 40 Initializations
Forking Paths

- Different preprocessing $\rightarrow$ 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

Original DTM

- **Doc1** → **Doc2** (1)
- **Doc1** ↔ **Doc3** (3)
- **Doc2** → **Doc3** (2)

Preprocessing Specification 1

- **Doc1** → **Doc2** (2)
- **Doc1** ↔ **Doc3** (6)
- **Doc2** → **Doc3** (4)

Preprocessing Specification 2

- **Doc1** → **Doc2** (4)
- **Doc1** → **Doc3** (1)
- **Doc2** ↔ **Doc3** (6)
Ranking Distance Changes

Original DTM

Doc1 <-> Doc2
Doc1 <-> Doc3
Doc2 <-> Doc3

Preprocessing Specification 2

Doc1 <-> Doc3
Doc2 <-> Doc3

<table>
<thead>
<tr>
<th>Original DTM</th>
<th>Preproc. Spec. 2</th>
<th>Abs. Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>d(1,3) = 3</td>
<td>d(2,3) = 6</td>
<td>Δ d(1,3) = 2</td>
</tr>
<tr>
<td>d(2,3) = 2</td>
<td>d(1,2) = 4</td>
<td>Δ d(2,3) = 1</td>
</tr>
<tr>
<td>d(1,2) = 1</td>
<td>d(1,3) = 1</td>
<td>Δ d(1,2) = 1</td>
</tr>
</tbody>
</table>
Comparing Preprocessing Specifications

- Each specification will have a largest mover.

- Rank in other specifications $(M_1, \ldots, M_{127})$?

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

- Average of $\mathbf{v}_{M_i} \rightarrow$ how unusual.
preText Scores

▷ Consider top $k$ largest moving doc pairs.

▷ Average across $\mathbf{v}_{M_i} \rightarrow \mathbf{v}_{M_i}^{(k)}$

▷ Normalize by $\frac{n(n-1)}{2}$ ($n = \text{num docs}$)

$$\text{preText score}_i = \frac{2\mathbf{v}_{M_i}^{(k)}}{n(n-1)}$$
Interpreting preText Scores

- preText scores range between 0 and 1.
- Lower score $\rightarrow$ “typical” changes in document distances.
- Higher score $\rightarrow$ “atypical” changes in document distances.
preText Scores for Press Releases

![Graph showing the relationship between preText Score and Preprocessing Combination. The graph displays a downward trend as the preText Score increases.]
Which Steps Matter?

Regression Coefficient

UK Manifestos
- Use Ngrams
- Stemming
- Remove Stopwords
- Remove Punctuation
- Remove Numbers
- Remove Infrequent Terms
- Lowercase

State Of The Union Speeches
- Regression Coefficient
- Top 100 Pairs

Indian Treaties
- Lowercase
- Remove Infrequent Terms
- Remove Numbers
- Remove Punctuation
- Remove Stopwords
- Use Ngrams

Death Row Statements
- Regression Coefficient
- Top 100 Pairs

Press Releases
- Lowercase
- Remove Infrequent Terms
- Remove Numbers
- Remove Punctuation
- Remove Stopwords
- Use Ngrams
Common Trends? (Danger!)

- Stopping, punctuation: highly variable.
- Stemming, numbers, lowercasing: not much effect.
- Including n-grams: potentially “good”.
- Infrequent terms: potentially “bad”.
Summary

- Preprocessing matters.
- Forking paths of inference.
- Our solution: preText.

- General Advice:
  - Some steps seem innocuous.
  - Always check – tell reader!
Happy Sloths Love R Packages!

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

Regression Coefficient