Automated Text Analysis for Political Science

...and Measurement With Machine Learning

Andy Halterman and Aidan Milliff

MIT Political Methodology Lab Workshop, 2 April 2021

Who has a text problem?

Three phrases of quantitative (measurement) work

Research design, method design, and method engineering

Research design, method design, and method engineering

1. research design
2. “method design”
3. “method engineering”

Research Design

Research Design

The research design phase is a standard component of all empirical work:

- questions
- theories
- hypotheses
- observable implications of hypotheses (“likelihood functions”)
- operationalization and measurement
- identification strategies

Research Design in Text

How does the text come to be generated and what purpose does it serve in the research design?

Research Design in Text

How does the text come to be generated and what purpose does it serve in the research design?

- Ideology $\longrightarrow$ speech in Congress (Quinn et al. 2010)
- Political strategy $\longrightarrow$ Senate press releases (Grimmer 2010)
- Threat perception $\longrightarrow$ speech in the NSC (Landau-Wells)
- Human rights record $\longrightarrow$ newspaper reporting (Nielsen 2013)
- Events predictive of conflict $\longrightarrow$ newspaper reporting (Mueller and Rauh 2017)

“Method” Design

“Method” Design

Conceptually, how do you represent or transform your text [data] to produce an output that’s useful for your question and research design?

“Method” Design

Conceptually, how do you represent or transform your text [data] to produce an output that’s useful for your question and research design?

1. supervised or unsupervised? (Grimmer and Stewart 2013)
2. what do you map your data onto, or, what form does the output take?
3. how do you represent your data, or, how do you get your data into a matrix?

(a) Supervised vs. unsupervised

(a) Supervised vs. unsupervised

Supervised: Teach a machine to recover known labels well*, potentially by approximating a true DGP.

Quantities of interest:

- $\widehat{y}$, the predicted values for new observations
- the parameters of the model ($\widehat{\beta}$ or $\widehat{\theta}$).

(a) Supervised vs. unsupervised

Supervised: Teach a machine to recover known labels well*, potentially by approximating a true DGP.

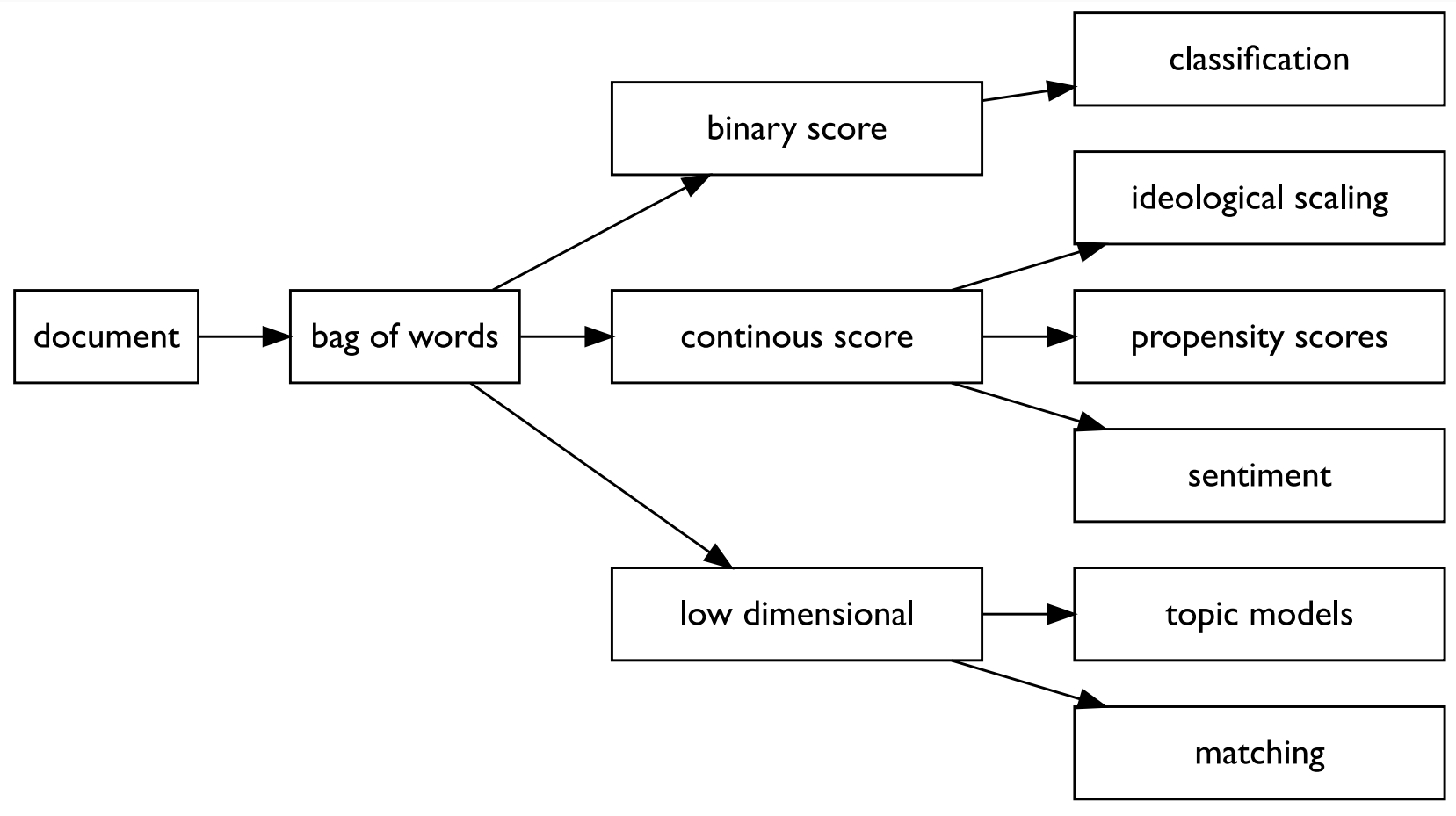
Quantities of interest:

- $\widehat{y}$, the predicted values for new observations
- the parameters of the model ($\widehat{\beta}$ or $\widehat{\theta}$).

Unsupervised: usefully categorize or summarize your data. ("Find a low-dimensional representation"). This often takes the form of data being put into clusters indicated by z.

(b) What do you want to map your data onto?

(b) What do you want to map your data onto?



(c) How do you get your data into a matrix?

(c) How do you get your data into a matrix?

Traditional approaches:

- bag of words (word $\longrightarrow$ vocab number $\longrightarrow$ one hot vector $\longrightarrow$ elementwise sum)
- bigrams, ngrams
- same thing, but scaling words by importance (tf-idf)
- stemming, stopword removal, etc. (Denny and Spirling 2018)

Bag-of-words boardwork

“my representative supports the estate tax”

“no taxation without representation”

Bag-of-words boardwork

“my representative supports the estate tax”

“no taxation without representation”

```
represent=1
support=2
estate=3   [0 0 1 0 0 0]
tax=4
no=5
without=6
```

Bag-of-words boardwork

“my representative supports the estate tax”

“no taxation without representation”

```
represent=1
support=2
estate=3   [0 0 1 0 0 0]
tax=4
no=5
without=6

 1  2  3  4  5
[1  1  1  1  0 ]
[1  0  0  1  0 ]
```

Beyond bag of words

Newer stuff:

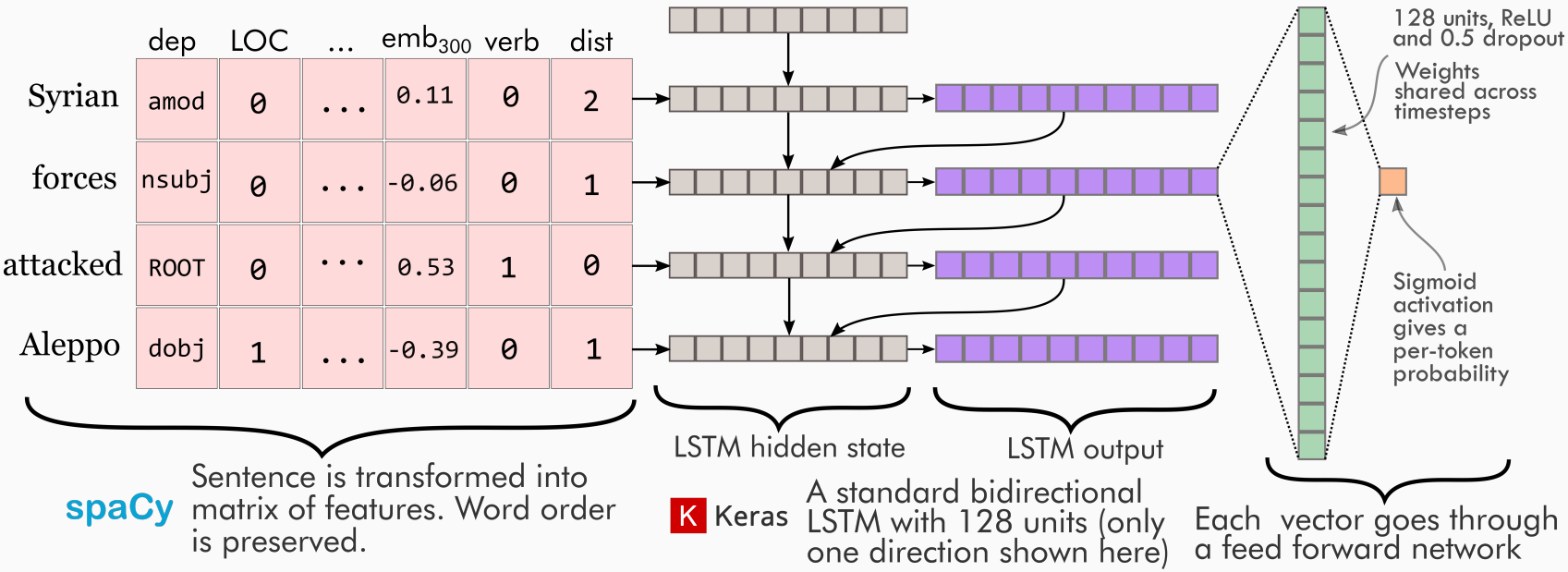
- embeddings
- grammatical role in the sentence, part-of-speech
- small pieces of the document (e.g. named entities)
- raw words in order
- raw characters in order

Custom embeddings trained on human rights text

```
model.most_similar("torture")
('illtreatment', 0.6942635178565979),
('beatings', 0.668760180473327 6),
('extrajudicial', 0.6468784809112549),
('mistreatment', 0.6361342668533325),
('maltreatment', 0.633286714553833), ...

model.most_similar("putin")
[('yushchenko', 0.879678130149 8413),
('bouteflika', 0.872167527675 6287),
('yeltsin', 0.845860242843 6279),
('saakashvili', 0.843777596950531),
('rakhmonov', 0.840139210224 1516), ...
```

Beyond bag-of-words (example)



3. “Method Engineering”

“Method Engineering”

This phase involves making decisions like

- logit or probit?
- SVM or neural net?
- EM or MCMC?
- recurrent neural net or convolutional neural net?

Back to text

Description	Representation	Model	Citation
Estimating the ideological position of political parties from manifesto text	continuous score	EM Poisson of words	Slapin and Proksch (2008)
Predicting legislators' votes from bill text	low dimensional	ideal point LDA	Gerrish and Blei (2011)
Learning conservative and liberal terms from a classification model on political speeches	binary (feature weights of interest)	SVM	Diermeier et al. (2012)
Estimate topics and correlation within a corpus of Islamic cleric's text	low dimensional	structural topic model (extension of LDA)	Nielsen (2017)
Estimate document ideology from coded sentences	continuous	IRT	Benoit et al. (2016)
(Proportion of blogs with) positive or negative sentiment toward presidential candidates	continuous	SVM	Hopkins and King (2010)
Classifying news articles for signals of impending mass killing	binary	SVM	Halterman, Ulfelder, and Valentino (2016)

More tools and ecosystem:

R tools:

- `tm`: tools for tokenizing, stemming, stopword removal, etc.
- `stm`: the go-to topic modeling package in political science
- `tidytext`: a package and paradigm for working with text in a "tidy" way

Python tools

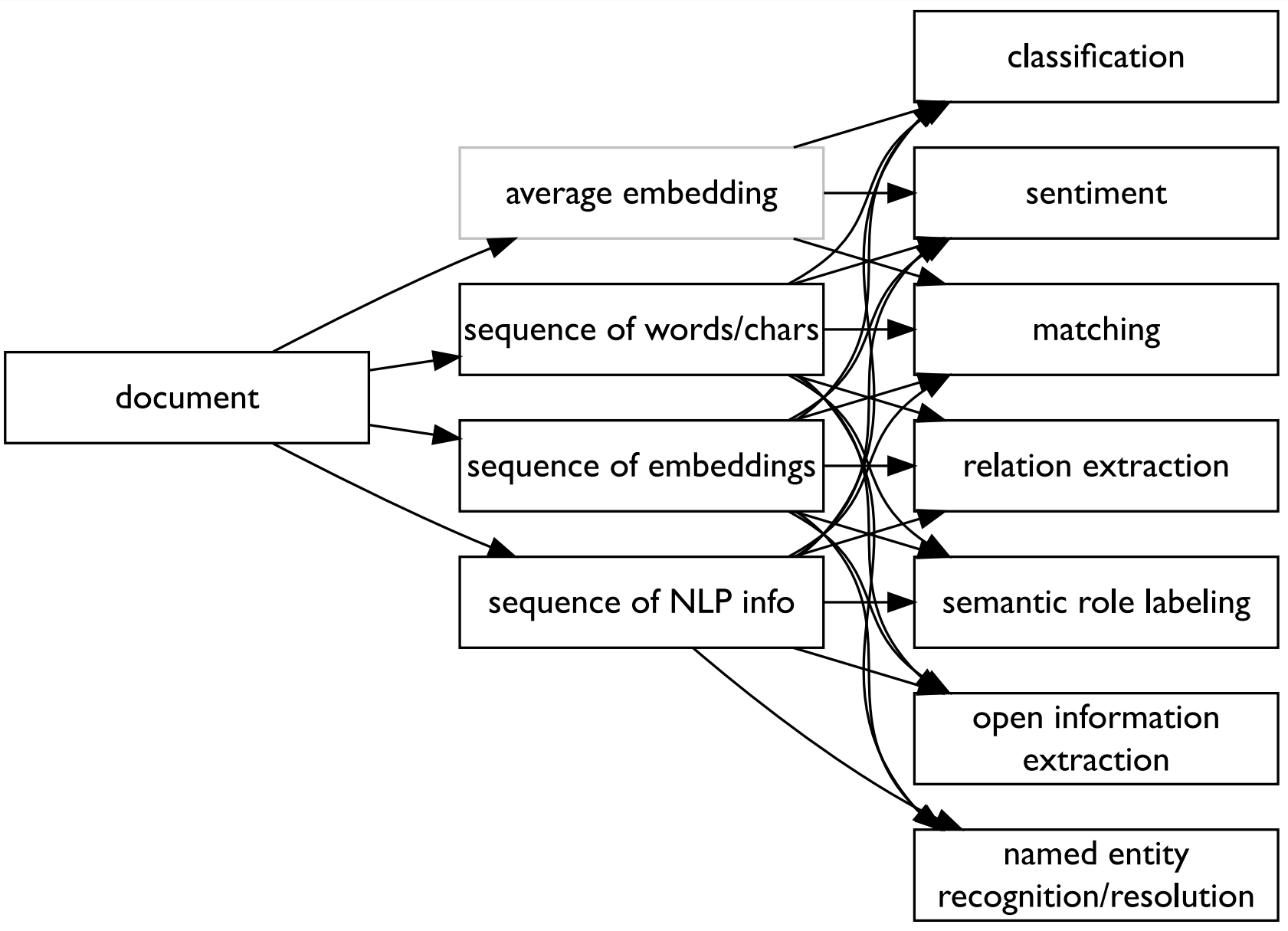
- `requests` and `BeautifulSoup` for scraping
- `spaCy`: natural language processing with a nice API
- `gensim`: "topic modeling for humans" with corpus management
- `scikit-learn`: machine learning library
- `keras`, `tensorflow`, and `PyTorch`: neural network libraries
- `Prodigy`: annotate data for text classification, NER, etc.

Summary

1. design your research well
2. what is the text doing for you?
3. what do you need to produce from each document?
4. how are you representing your document?
5. which model/algorithm/package will you use?

Extra Slides

3rd generation representations



Word Embeddings

Word embeddings a.k.a. word vectors a.k.a. distributed/dense/low-dimensional representations

Skipgram with negative sampling (Mikolov et al. 2013)

$$p(c|w,\theta) = \frac{\exp(v_c \cdot v_w)}{\sum_{c' \in C} \exp(v'_c \cdot v_w)}$$

- find v_c, v_w, vector representations of the word and context to maximize this $p(c|w)$. - implicitly, these vectors turn out to be a matrix decomposition on the pointwise mutual information matrix (Levy and Goldberg 2014)

References

Denny, Matthew J, and Arthur Spirling. 2018. "Text Preprocessing for Unsupervised Learning: Why It Matters, When It Misleads, and What to Do About It." *Political Analysis*, 1–22. https://doi.org/10.1017/pan.2017.44.

Grimmer, Justin. 2010. "A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases." *Political Analysis* 18 (1): 1–35. https://doi.org/10.1093/pan/mpp034.

Grimmer, Justin, and Brandon M Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21 (3): 267–97. https://doi.org/10.1093/pan/mps028.

Levy, Omer, and Yoav Goldberg. 2014. "Neural Word Embedding as Implicit Matrix Factorization." In *Advances in Neural Information Processing Systems*, 2177–85.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. "Distributed Representations of Words and Phrases and Their Compositionality." In *Advances in Neural Information Processing Systems*, 3111–9.

Mueller, Hannes, and Christopher Rauh. 2017. "Reading Between the Lines: Prediction of Political Violence Using Newspaper Text." *American Political Science Review*, 1–18.

Nielsen, Richard A. 2013. "Rewarding Human Rights? Selective Aid Sanctions Against Repressive States." *International Studies Quarterly* 57 (4): 791–803.

Quinn, Kevin M, Burt L Monroe, Michael Colaresi, Michael H Crespin, and Dragomir R Radev. 2010. "How to Analyze Political Attention with Minimal Assumptions and Costs." *American Journal of Political Science* 54 (1): 209–28.