What's Next

Taylor Arnold[1] Lauren Tilton[2]

04 July 2017

[1] Assistant Professor of Statistics
University of Richmond
@statsmaths

[2] Assistant Professor of Digital Humanities
University of Richmond
@nolauren

Summary of today

- Preliminaries
- Topic I: Tokenising text
- Topic II: The NLP Pipeline
- Topic III: Modelling Textual Data
- Conclusions

Things we did not cover (but wanted to)

- pre-processing data (i.e., HTML tags)
- sentiment analysis
- coreferences
- word embeddings
- non-western languages
- reconstructing text
- many more examples of using various
  - parts of speech codes
  - dependency types
  - entity types

Other references

The **cleanNLP** paper:

> *Taylor Arnold (2017). A Tidy Data Model for Natural Language Processing using cleanNLP. The R Journal, 9(2), 1-20.*

Our book:

> *Taylor Arnold and Lauren Tilton (2015). "Humanities Data in R: Exploring Networks, Geospatial Data, Images, and Text." Quantitative Methods in the Humanities and Social Sciences Springer, New York, NY, USA.*

Pedagogical paper:

> *Taylor Arnold and Lauren Tilton , "Basic Text Processing in R", Programming Historian, (2017-03-27).*

And the talk:

> *Wednesday at 14:06 in Plenary 4.0 as part of the Text Mining session.*