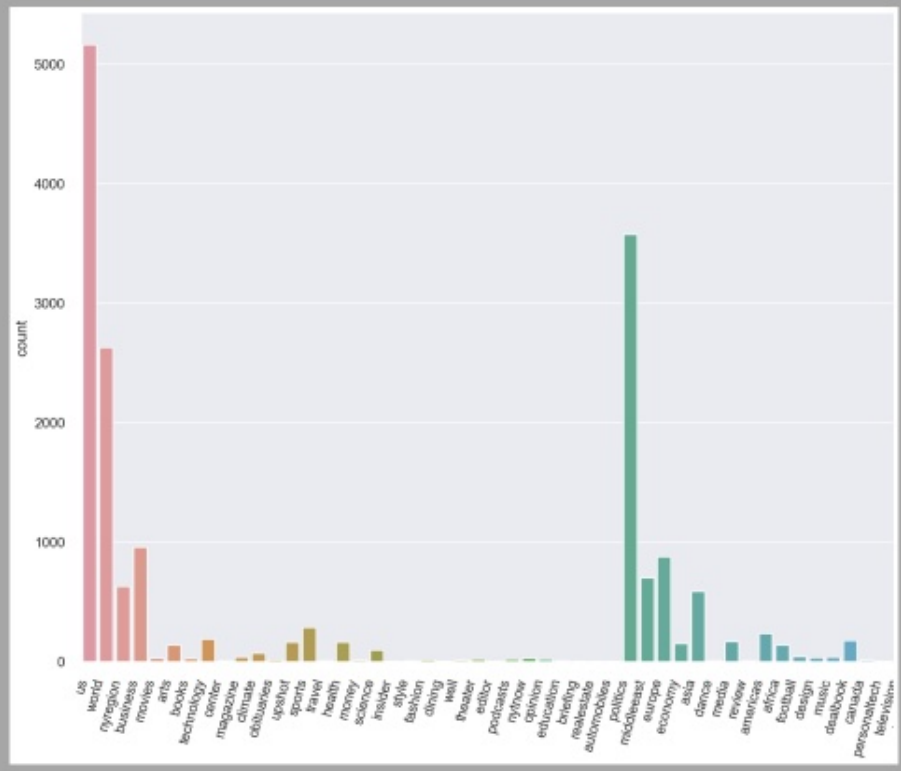

Reclassifying the Times
July 2019
All the news that's fit...
What was in the data set and why we decided it was worth reading
spaCy for NLP
The challenges, advantages and drawbacks of using the most powerful NLP tools in data science . . . for a 5-day project
The bottom line
Our findings after five days of wrestling with industrial-strength NLP packages
Processing the press
Wrangling the news using spaCy, scikitlearn, pandas, and gensim for natural language processing


The Data
• 10,700 articles taken as snapshots from the landing page of the NYT website by archive.org from 2014-2018
• Uncategorized but tagged with persistent URLs
• Can we predict the article's section of the paper?
• Are the categories meaningful, or are there other, latent groupings?
Methods
Tools

Methods

- Latent Dirichlet Allocation
- Dimensionality reduction with LSA
- Random forest

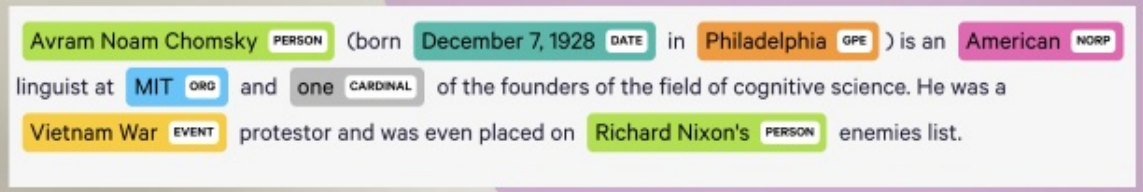



Libraries, Tools, and Models
• spaCy
• Industry standard for NLP
• Optimized for high volume
• gensim
• Customizable models
• Compatible with spaCy
• scikit-learn
• Standard statistical library
• Evaluation metrics


Reclassifying the Times
July 2019
All the news that's fit...
What was in the data set and why we decided it was worth reading
spaCy for NLP
The challenges, advantages and drawbacks of using the most powerful NLP tools in data science . . . for a 5-day project
The bottom line
Our findings after five days of wrestling with industrial-strength NLP packages
Processing the press
Wrangling the news using spaCy, scikitlearn, pandas, and gensim for natural language processing

Advantages of spaCy

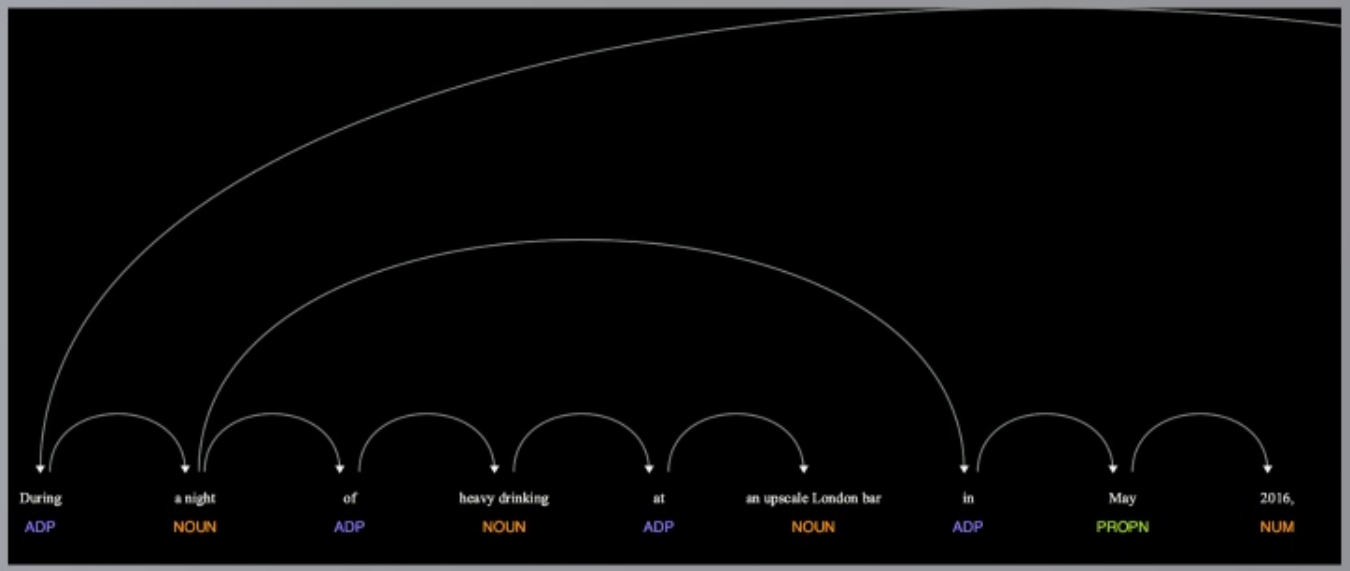
- Optimized for large-scale information extraction, Cython-based
- Most accurate linguistic tagging and parsing
- Extremely flexible and customizable



Process

How we used spaCy

- Used Spacy's stop word list plus various titles ("Mr."), "Washington"
- Displacy dependency visualizer.
- Topic model run: Latent Dirichlet Allocation

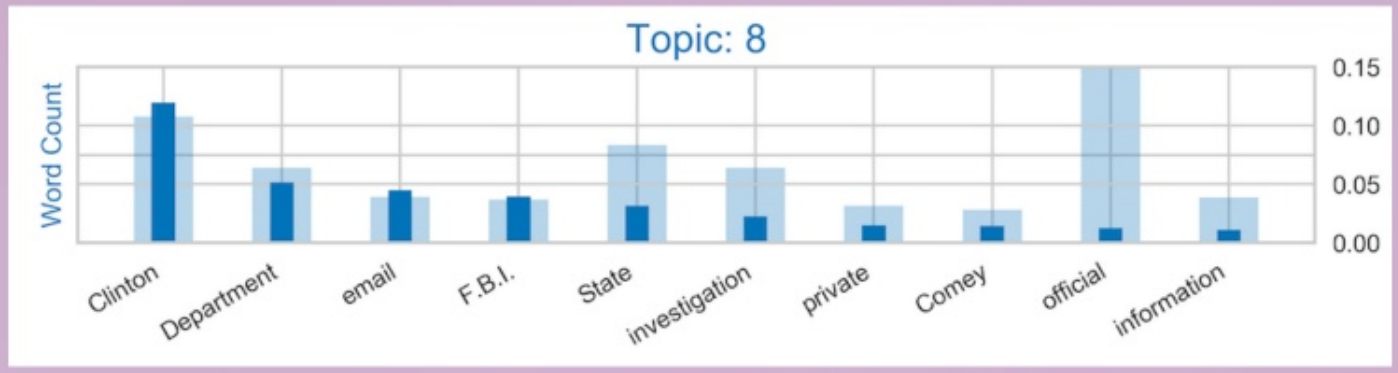


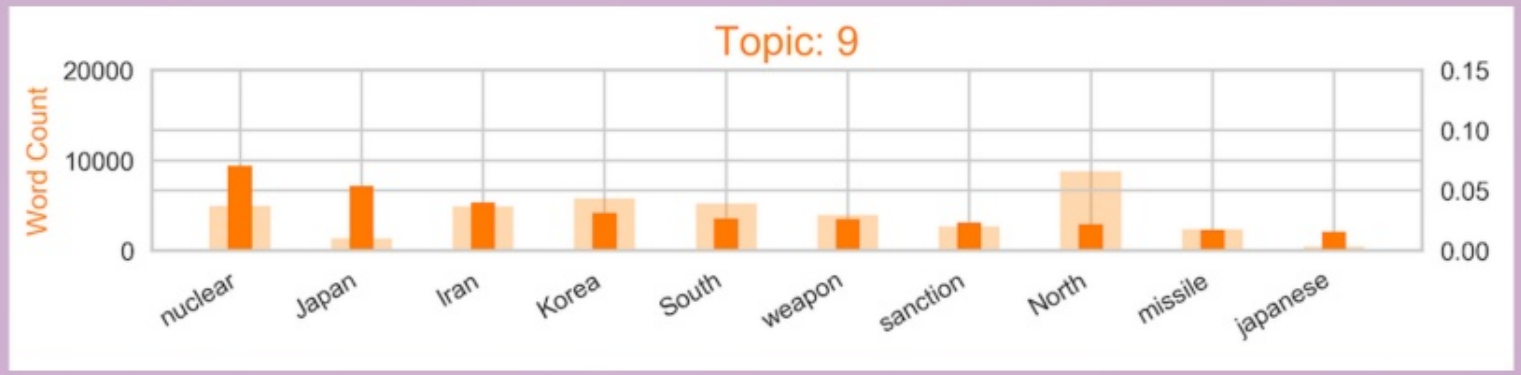

Reclassifying the Times
July 2019
All the news that's fit...
What was in the data set and why we decided it was worth reading
spaCy for NLP
The challenges, advantages and drawbacks of using the most powerful NLP tools in data science . . . for a 5-day project
The bottom line
Our findings after five days of wrestling with industrial-strength NLP packages
Processing the press
Wrangling the news using spaCy, scikitlearn, pandas, and gensim for natural language processing

LDA and results

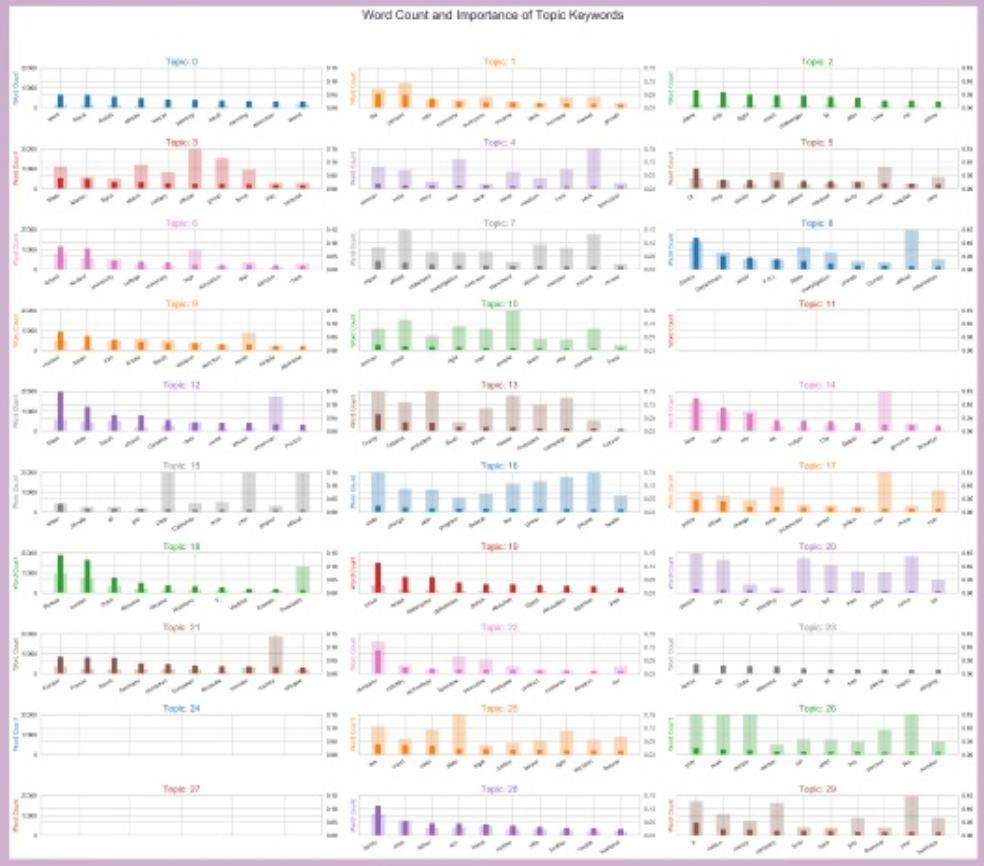
- Lemmatization gave the best results
- Unsupervised learning: measures of "accuracy" ~ coherency score
- Using sklearn iterative model, best = 0.58
- Top ten topic clusters matched to original articles side-by-side and adjudicated by humans

Sample topics

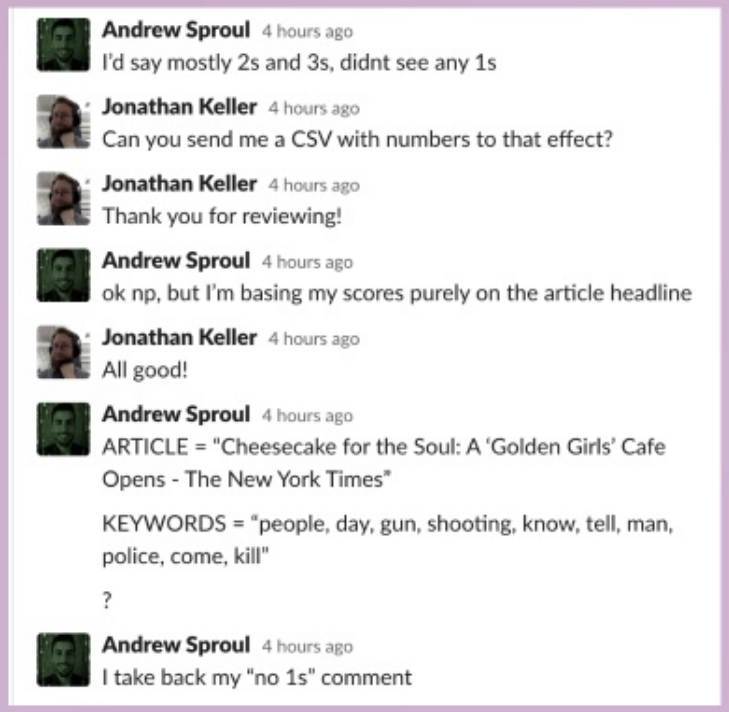




Word count weights



Survey says...


Andrew Sproul 4 hours ago
I'd say mostly 2s and 3s, didnt see any 1s
Jonathan Keller 4 hours ago
Can you send me a CSV with numbers to that effect?
Jonathan Keller 4 hours ago
Thank you for reviewing!
Andrew Sproul 4 hours ago
ok np, but I'm basing my scores purely on the article headline
Jonathan Keller 4 hours ago
All good!
Andrew Sproul 4 hours ago
ARTICLE = "Cheesecake for the Soul: A 'Golden Girls' Cafe Opens - The New York Times"
KEYWORDS = "people, day, gun, shooting, know, tell, man, police, come, kill"
?
Andrew Sproul 4 hours ago
I take back my "no 1s" comment


Reclassifying the Times
July 2019
All the news that's fit...
What was in the data set and why we decided it was worth reading
spaCy for NLP
The challenges, advantages and drawbacks of using the most powerful NLP tools in data science . . . for a 5-day project
The bottom line
Our findings after five days of wrestling with industrial-strength NLP packages
Processing the press
Wrangling the news using spaCy, scikitlearn, pandas, and gensim for natural language processing

Next steps

- Further improvements on LDA models, inc. hyperparameter refinement
- Random forest model
- Re-working spaCy workflow to incorporate native support for data type conversions via GoldParse
- Expand the bag of words model to a semantic graph of trigrams or n > 3


Reclassifying the Times
July 2019
All the news that's fit...
What was in the data set and why we decided it was worth reading
spaCy for NLP
The challenges, advantages and drawbacks of using the most powerful NLP tools in data science . . . for a 5-day project
The bottom line
Our findings after five days of wrestling with industrial-strength NLP packages
Processing the press
Wrangling the news using spaCy, scikitlearn, pandas, and gensim for natural language processing