Improving Topic Document Modeling

RoboTop

Sources and methods

Topic modeling with LDA

LDA on 15000 news articles from 15 publications

Add multiple copies of Named Entities

How good are topic modeling results?

Topics mostly look reasonable ...

But not always

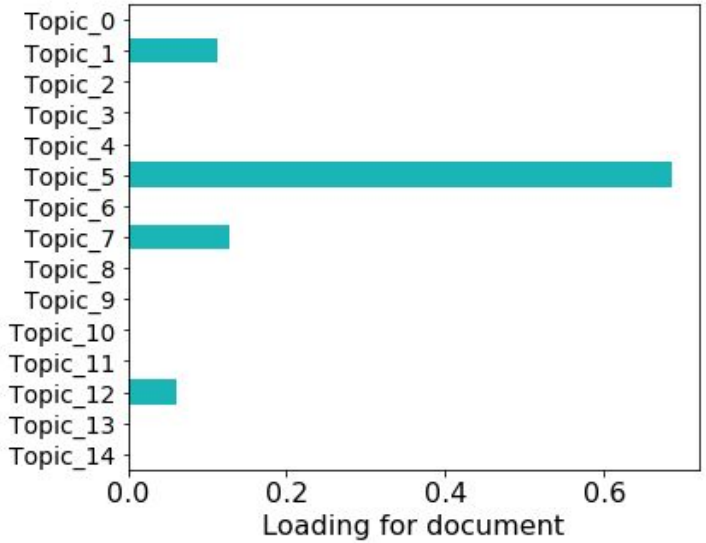
In addition, document match to topics is not completely reliable

What would a human do with this task?

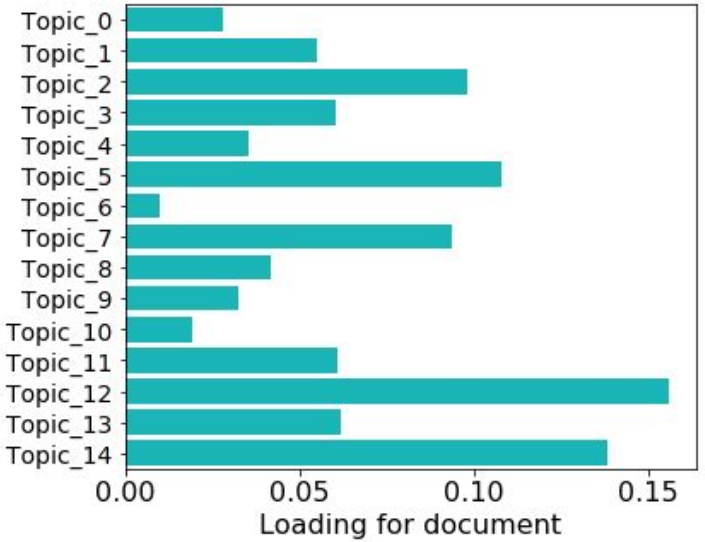
Maybe leave many of them out, and only assign a topic to a subset - 15 topics plus “other”

Can we get a machine to do this?

Which documents are problematic?



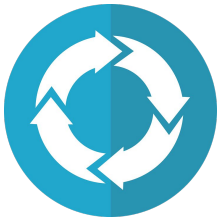
Low entropy



High entropy

Model, discard, model again

- Perform LDA to get topics
- Look at each document's distribution over topics
- Calculate the entropy - high entropy means evenly distributed over many topics
- Throw out documents with entropy in the top 5%
- Repeat
- Stopping criterion: mean entropy drops below target



How well does this strategy work?

Metrics - improved coherence and perplexity scores

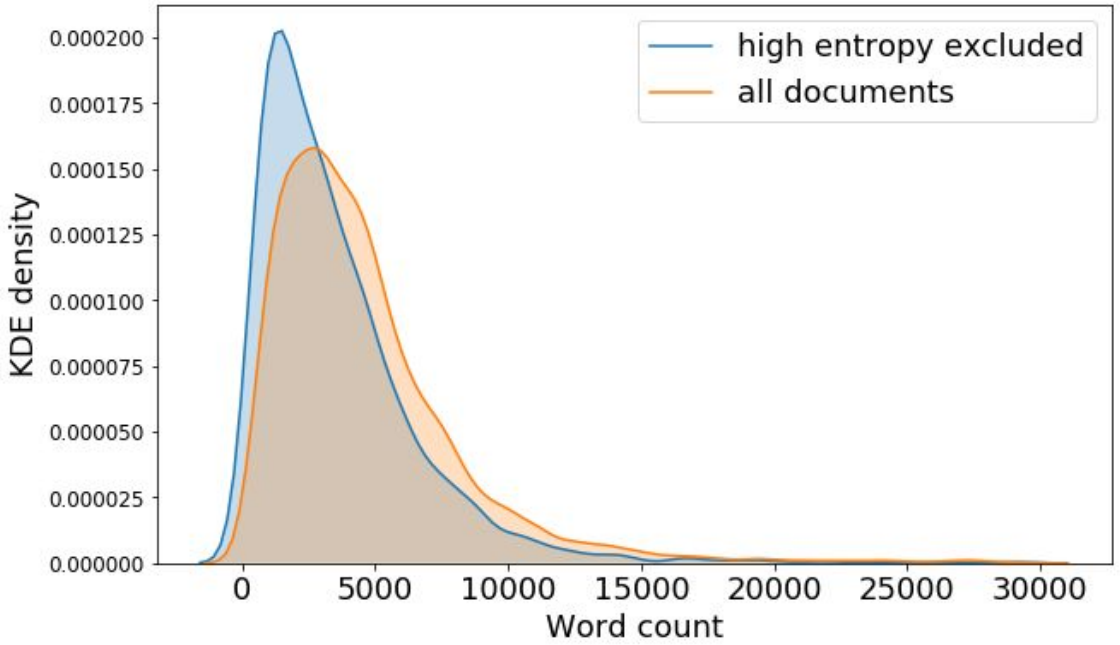
	Coherence	Log(Perplexity)
All documents model	0.49	-9.7
Low entropy only model	0.62	-6.9

Manual inspection - all topics appear coherent

Top 4 documents for each category also look fine

What did it do?

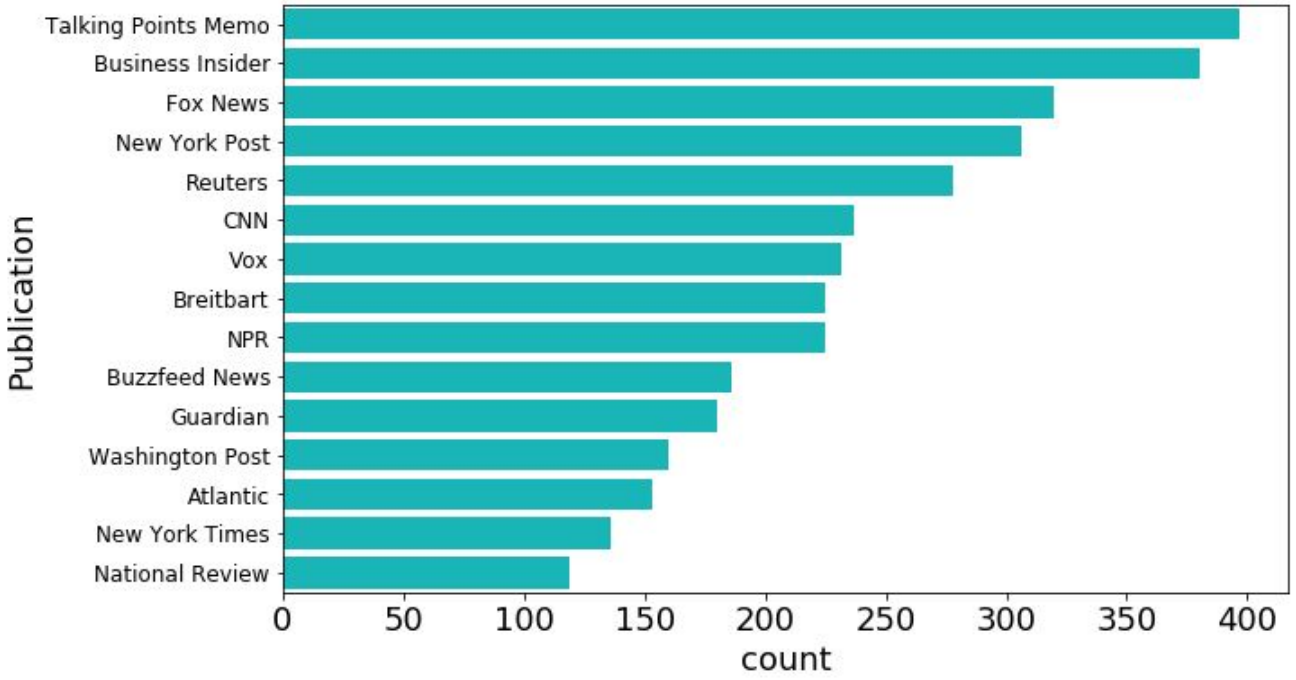
Documents not eliminated evenly - shift in distribution of content length to shorter articles



What did it do?

Shift in the distribution of publications

Less complex documents and shorter content correspond to fewer topics.



Thank you

Appendix

Future work

Evaluate coherence and perplexity metrics more thoroughly. Do they correspond well with manual evaluations of topics and documents?

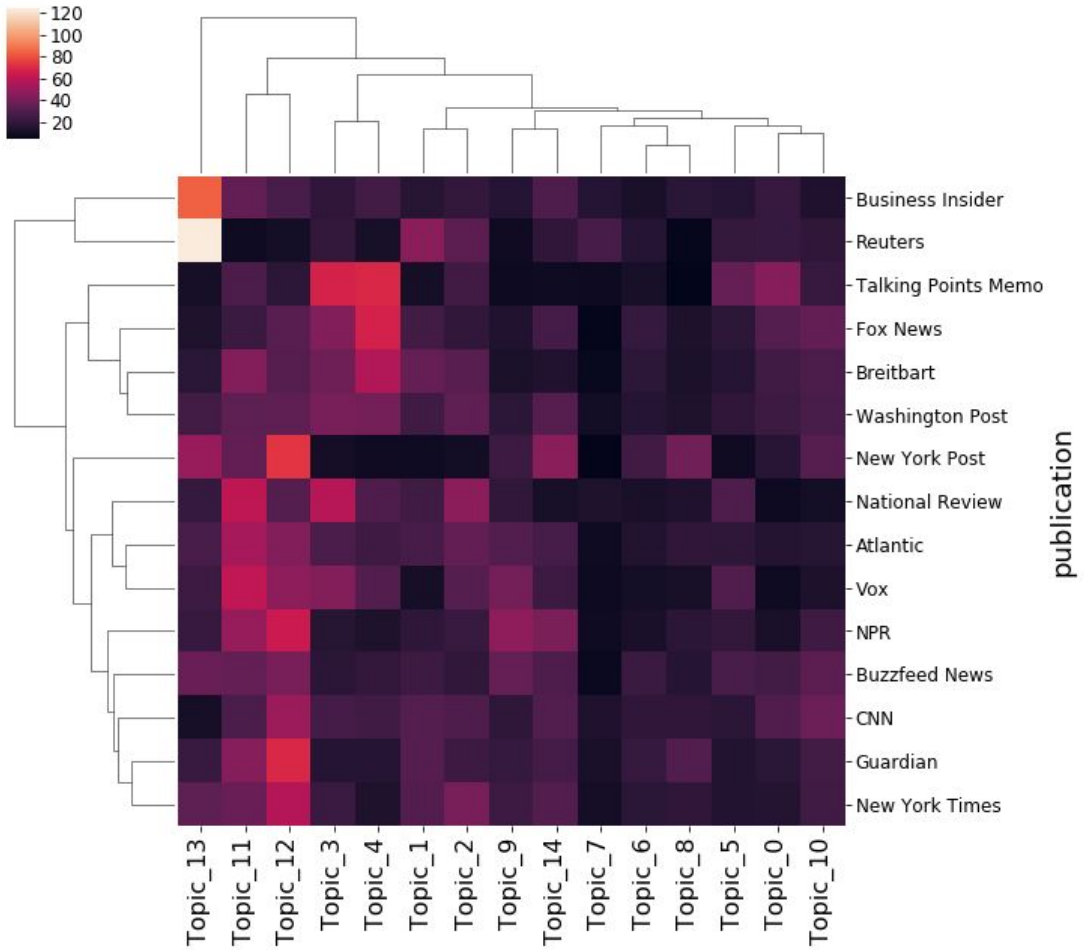
Compare filtering by entropy with other filtering approaches

Supervised or semi-supervised approaches may perform better

Another direction: compare publications and topics

Reuters and Business Insider appear to be outliers, due to their heavy emphasis on business stories.

The Guardian is not a clear outlier, in spite of the fact that it is British, while the other publications are American.

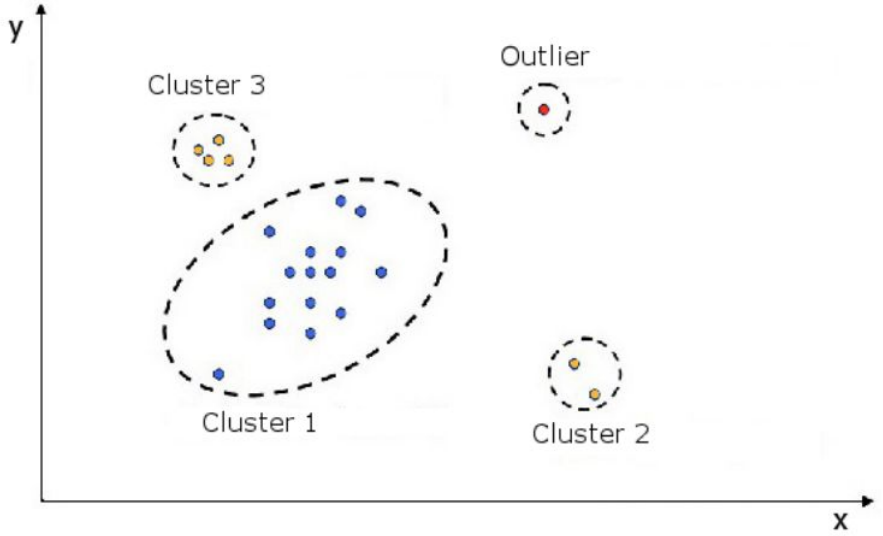


Why discard high-entropy documents?

“Outlier” documents may not clearly belong to any topic.

Two bad effects:

- Creation of incoherent topics
- Dilution of otherwise coherent topics



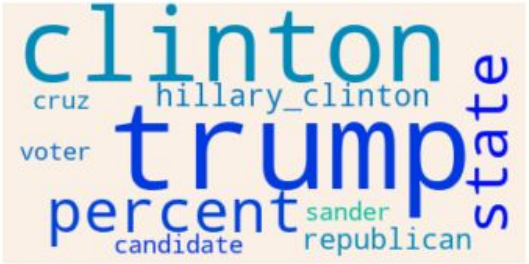
Coherence definition paper

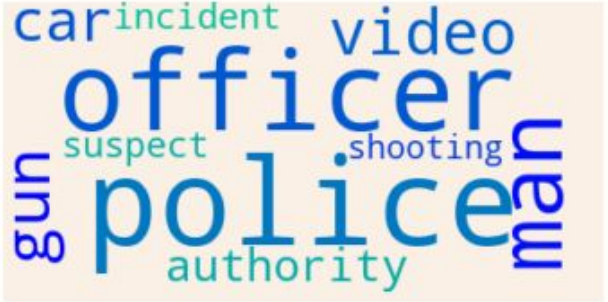
http://svn.aksw.org/papers/2015/WSDM_Topic_Evaluation/public.pdf

More topics


revenue
million
businesssale
investor
company
share
billion
year
analyst


research
doctor
year
study
brain
researcher
people
drug
patient
scientist


clinton
cruz
hillary_clinton
trump
voter
state
percent
sander
candidate
republican


car
incident
video
officer
suspect
shooting
gun
police
man
authority

Top articles for topics

Topic: Topic_0
Title: Salesforce still mulls bid for Twitter as shareholders resist: sources Pub: Reuters
Title: LinkedIn is crashing after weak earnings guidance Pub: Business Insider
Title: Yelp admits smaller businesses are fed up with its site Pub: New York Post
Title: Apple is suing one of its most important suppliers Pub: Business Insider

Topic: Topic_1
Title: What Doctors Learned From 42 Infants With Microcephaly Pub: NPR
Title: 1st Sexually Transmitted Zika Virus Case Confirmed in Dallas County Pub: Breitbart
Title: Studies Reinforce The Urgency Of Treating Pregnant Women With Malaria Pub: NPR
Title: Why Deep Breathing May Keep Us Calm - The New York Times Pub: New York Times

Topic: Topic_2
Title: How Hillary Clinton could win 270 electoral votes Pub: Vox
Title: Right now polls show Donald Trump losing every single swing state Pub: Vox
Title: Trump campaign to Cruz: GOP establishment is using you Pub: Fox News
Title: Caucuses vs Primaries Pub: Talking Points Memo