Named entity recognition using topic models

Traveler's Lab
Wesleyan University
September 25, 2018

Topic models

- A type of “soft” clustering used for content discovery and classification
- Topics are collections of words
  - E.g. *weather*: snow, rain, wind, hot, cold, humid, thunder
- Topic names are picked by an analyst - “beauty is in the eyes of the beholder”
  - E.g. what’s the difference between *weather* and *climate*?

Example

- Five documents, two topics − restaurants and loans

"Due to bad loans, the bank agreed to pay the fines"
"If you are late to pay off your loans to the bank, you will face fines"
"How will you pay off the loans you will need for the restaurant you want opened?"
"There is a new restaurant that just opened on Warwick street"
"A new restaurant opened in downtown"

Soft clustering

- ‘pay’ and ‘loans’ belong to two topics

	Terms						
Docs	bank	fines	loans	pay	new	opened	restaurant
d_1	1	1	1	1	0	0	0
d_2	1	1	1	1	0	0	0
d_3	0	0	1	1	0	1	1
d_4	0	0	0	0	1	1	1
d_5	0	0	0	0	1	1	1

Soft clustering, ctd.

- Regular clustering is ‘hard’ − would assign the word only to one topic
- Soft clustering returns a probability of an item belonging to cluster
- In topic models, probability of word in topic, or topic in document

Topic model for 2 topics

document	`1`	`2`
<chr>	<dbl>	<dbl>
d_1	0.679	0.321
d_2	0.679	0.321
d_3	0.321	0.679
d_4	0.391	0.609
d_5	0.391	0.609

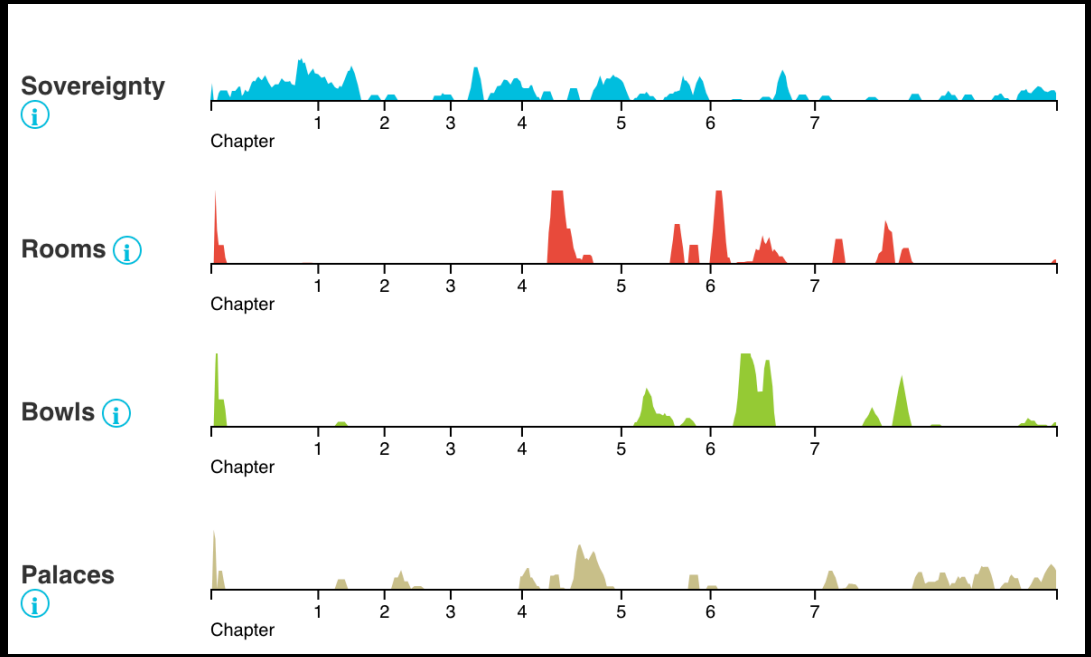
	term	`1`	`2`
	<chr>	<dbl>	<dbl>
1	bank	0.216	0.0103
2	fines	0.216	0.0103
3	loans	0.0103	0.320
4	new	0.216	0.0103
5	opened	0.0103	0.320
6	pay	0.320	0.0103
7	restaurant	0.0103	0.320

Widely used by DH

- “Distant reading” in digital humanities
- Off-the-shelf solutions: Mallet, JSTOR Text Analyzer
- JSTOR Labs https://labs.jstor.org/projects/
- JSTOR Topicgraph https://labs.jstor.org/topicgraph/

JSTOR Topicgraph

- JSTOR topics found in *Imperial Matter: Ancient Persia and the Archeology of Empires*
- "Document" is a chunk of text (500 or 1000 words)
- Topic names picked by JSTOR staff



Named entity recognition

- Entity − upper-case word within a sentence
- Use the word context of an entity
- Take n-words from left and right side of entity
  - E.g. “bishop of ___”, “son of ___”
  - But consider “went to ___” - ambiguous

Decisions

- Finding a pattern for entity
  - Single or multi-word, are dots allowed? *Korykos* vs. *St. Thomas*
- Size of the window: 1, 2, or 3 words
  - Larger window means fewer ‘documents’
- Suffix to show the side (left or right)
  - Reduces term frequencies
- How to handle punctuation
  - Commas limit the size of a context window

Regular expression patterns

- Pattern for the entity

```
"((St[.]? )?[A-Z][a-z]+ )"
```

- Pattern for the entity and the context

```
m = gregexpr("( [a-z]+){2} ((St[.]? )?[A-Z][a-z]+ )([a-z]+ ){2}",
             theo_text, perl=F)
```

- Symbols: ? – occurs 0 or 1 time, {2} occurs 2 times, + - occurs 1 or more times

Model fitting

```
mod = LDA(x=dtm, k=2, method="Gibbs",
          control=list(alpha=1.1, delta=0.1, seed=12345,
                       burnin=1000,
                       iter=10000, thin=1))
```

- Parameter ‘alpha’ affects the proportions of topics.
- alpha > 1 – proportions closer to 50/50
- Run time depends on dtm size, number of iterations

Results

- Table with probabilities of topics in documents
- Each document – context of one entity
- Topic numbering is arbitrary

	document	1	2
1	Aaron	0.93859649	0.06140351
2	Abandanes	0.66129032	0.33870968
3	Abas	0.71978022	0.28021978
4	Abasgia	0.17741935	0.82258065
5	Abasgians	0.40990991	0.59009009
6	Abdela	0.66129032	0.33870968
7	Abdelas	0.89083558	0.10916442
8	Abderachman	0.50000000	0.50000000
9	Abimelech	0.83175355	0.16824645
10	Aboubacharos	0.85211268	0.14788732
11	Aboulabas	0.50000000	0.50000000