Introduction to Topic Modeling

Evgenia (Zhenya) Olimpieva

2/08/2021

Topic Modeling

- Topic modeling is an **unsupervised** method of classification of text documents

Topic Modeling

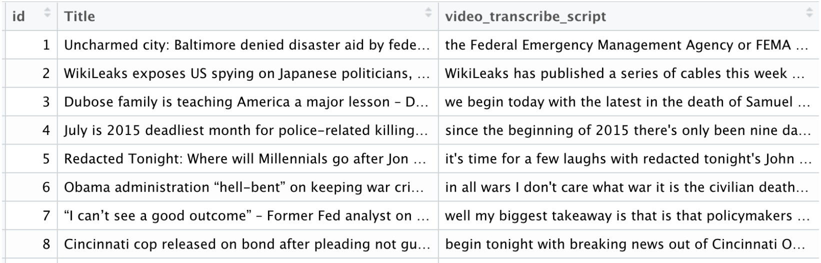
- Topic modeling is an **unsupervised** method of classification of text documents
- Useful for data that is neither classified or labeled (no variables to guide the classification algorithm)

Topic Modeling

- Topic modeling is an **unsupervised** method of classification of text documents
- Useful for data that is neither classified or labeled (no variables to guide the classification algorithm)
- Topic Modeling helps **discover** underlying groups in textual data.

Data

- Scraped RT America YouTube videos (2015:2017)
- N = 3569



id	Title	video_transcribe_script
1	Uncharmed city: Baltimore denied disaster aid by fede…	the Federal Emergency Management Agency or FEMA …
2	WikiLeaks exposes US spying on Japanese politicians, …	WikiLeaks has published a series of cables this week …
3	Dubose family is teaching America a major lesson – D…	we begin today with the latest in the death of Samuel …
4	July is 2015 deadliest month for police-related killing…	since the beginning of 2015 there's only been nine da…
5	Redacted Tonight: Where will Millennials go after Jon …	it's time for a few laughs with redacted tonight's John …
6	Obama administration “hell-bent” on keeping war cri…	in all wars I don't care what war it is the civilian death…
7	“I can’t see a good outcome” – Former Fed analyst on …	well my biggest takeaway is that is that policymakers …
8	Cincinnati cop released on bond after pleading not gu…	begin tonight with breaking news out of Cincinnati O…

Steps

1. Convert data to Tidy Text
2. Convert Tidy Text to DTM (Document Term Matrix)
3. Run Topic Model (LDA or Latent Dirichlet allocation)
4. Convert output of LDA to Tidy Data
5. Classify Documents into Topics

Tidy Text

Tidy text is a one-token-per-row dataset.

```
RT_tidy[1:10,-2]
```

```
##    id       word
## 1   1        the
## 2   1    federal
## 3   1  emergency
## 4   1 management
## 5   1     agency
## 6   1         or
## 7   1       fema
## 8   1        has
## 9   1     denied
## 10  1   maryland
```

Document Term Matrix (DTM)

- A document-term matrix describes the frequency of terms that occur in each document.
- In a document-term matrix, rows correspond to documents and columns correspond to terms.
- Will have many many many columns. . . . but we never have to look at it!
- It is important because the topicmodels package requires DTM and not Tidy Text.

Latent Dirichlet allocation (LDA)

Latent Dirichlet allocation is one of the most common algorithms for topic modeling. It does two things. It does some Bayesian magin and as the result:

- Models every document as a **mixture** of topics.
- Models every topic as a **mixture** of words.

LDA is a mixed-membership model. Every document has an association with each topic with a certain probabiliy.

Latent Dirichlet allocation (LDA)

- Topic Model doesnt tell you how many topics might be in your corpus.
- You have to tell the model that using human intuition!
- That’s the parameter “k” that you have to decide upon and there are NO RULES on how you will decide.
- Topic model also won’t name the topics for you - you will have to do it yourself based on the words associated with each topic.

Acknowledgments