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

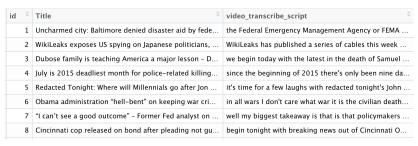
- ▶ Topic modeling is an unsupervised method of classification of text documents
- Useful for data that is neither classified nor labeled (no variables to guide the classification algorithm)

Topic Modeling

- Topic modeling is an unsupervised method of classification of text documents
- Useful for data that is neither classified nor 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



- 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
                 the
## 2
             federal
## 3
       1
           emergency
## 4
         management
## 5
       1
              agency
## 6
                   or
## 7
                fema
## 8
                 has
              denied
##
## 10
            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 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

- ► Text Mining with R
- ► An Introduction to Topic Modeling