Automated Text Analysis in Political Science

Lecture 7: Topic models

May 14, 2019

dr. Martijn Schoonvelde

School of Politics and International Relations, UCD

Today's class

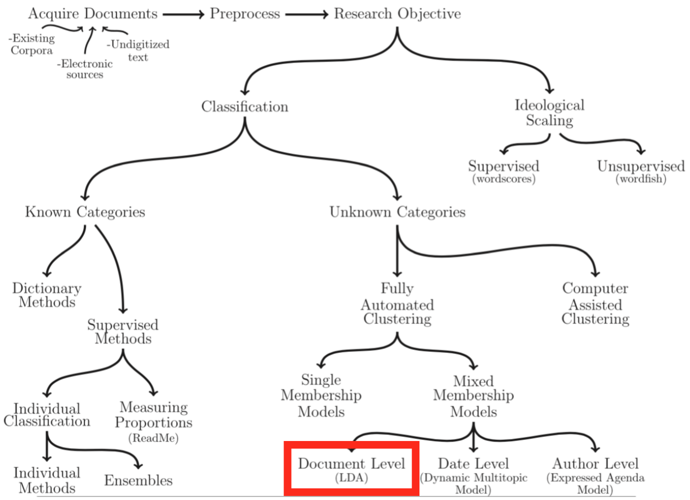
- Learn about and understand topic models (LDA in particular)
- Discover themes or topics in a corpus without fixing them in advance
- Practice LDA topic model in R

Friday: N15 103

Today's class

- Learn about and understand topic models (LDA in particular)
- Discover themes or topics in a corpus without fixing them in advance
- Practice LDA topic model in R

Overview of Text as Data Methods



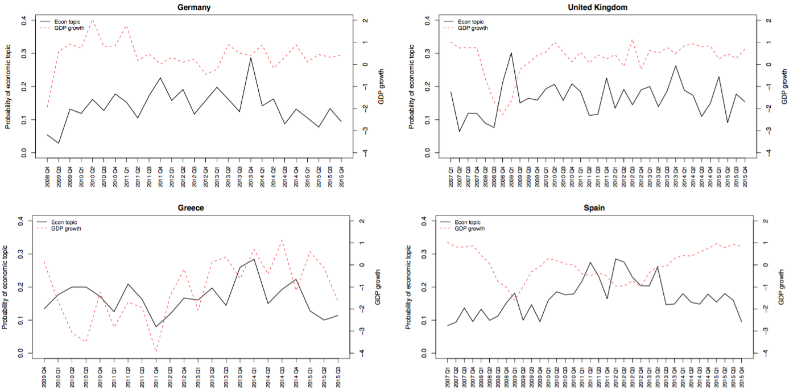
(Grimmer and Stewart, 2013, 268)

Topic models

- Why topic modeling:
  - Discover themes or topics in a corpus without fixing them in advance
  - Can be applied to large quantities of text – explore a corpus
- There are many different topics models out there, which share the following characteristics (Grimmer & Stewart, 2013):
  - Topics are represented as a probability function over words
  - Assume a generative process for observed text

Example: economy topic in EUSpeech

Figure 2: Economic topic probabilities 2007-2015



So how do we get from a bag of words to these topics?

Topics in topic models

k-th topic uses the *m-th* word. Substantively, topics are distinct concepts. In congressional speech, one topic may convey attention to America's involvement in Afghanistan, with a high probability attached to words like `troop`, `war`, `taliban`, and `Afghanistan`. A second topic may discuss the health-care debate, regularly using words like `health`, `care`, `reform`, and `insurance`. To estimate a topic, the models use the co-occurrence of words across documents.

Grimmer & Stewart (2013)

LDA topic model

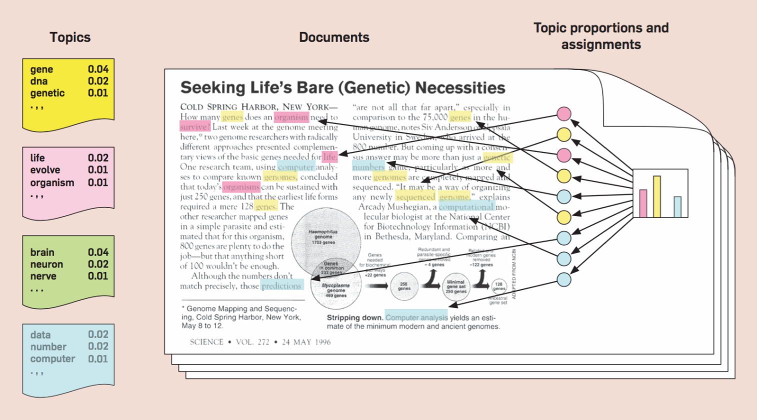
- Why focus on the LDA topic model (Blei, Ng and Jordan 2003)?
- Output of a LDA topic model is intuitive: topic proportions for each document, and word probabilities for each topic
    - Allows for statements like: 80% of words in document A is assigned to topic 1, and 10% of words in document B is assigned to topic 1
- Forms basis for extensions like structural topic model (Roberts et al 2014)

Assumptions behind the LDA topic model

- Documents exhibit multiple topics
    - “mixed membership model” (as opposed to single membership topics model)
- Number of topics is fixed in advance
- LDA is a probabilistic model:
    - Each time you run a LDA topic model on the same data with the same parameter settings you will get slightly different results
    - Problem of multimodality (Roberts et al. 2016)

Generative process of text

Figure 1. The intuitions behind latent Dirichlet allocation. We assume that some number of "topics," which are distributions over words, exist for the whole collection (far left). Each document is assumed to be generated as follows. First choose a distribution over the topics (the histogram at right); then, for each word, choose a topic assignment (the colored coins) and choose the word from the corresponding topic. The topics and topic assignments in this figure are illustrative—they are not fit from real data. See Figure 2 for topics fit from data.



First there is a distribution of topics, and then the rest follows

Generative process of text

For each **document** in the corpus, **words** are generated in a two-stage process:

1. Randomly choose a distribution over topics
2. For each **word** in a **document**
    - Randomly choose a topic from the distribution over topics in step # 1
    - Randomly choose a word from the corresponding distribution over the vocabulary

Inference: from words to topics

- The only thing we observe is words
  - The term-document matrix
- Given our generative process, we infer the topic structure that is most likely to have generated the observed words
  - Move in the opposite direction of the generative process
- To start off this inferential process we need to make some assumptions:
  - Prior distribution for topics across documents
  - Prior distribution for words across topics

Prior Distribution: Dirichlet

- LDA assumes that initial (i) topic distributions across documents, and (ii) word distributions across topics follow a Dirichlet distribution
- Direchlet distribution: "distribution of multinomial distributions"
  - For D documents and K topics: D multinomial distributions of size K
  - For K topics and N words: K multinomial distributions of size N
- Hence: Latent Dirichlet Allocation

Document-topic distribution

	Topic 1	Topic 2	Topic 3	Topic K
Document 1	0.05	0.20	0.35	0.40
Document 2				
Document 3				
Document D				

$D \times K$ document-topic distribution

Word-topic distribution

	Topic 1	Topic 2	Topic 3	Topic K
Word 1	0.25	0.20	0.15	0.40
Word 2				
Word 3				
Word N				

$N \times K$ word-topic distribution

From prior to posterior

- Now we need to go from prior distribution to posterior distribution
  - Prior distribution: randomly selected Dirichlet distribution for topics-words and document-topics
  - Posterior distribution: estimate Direchlet distribution that is most likely to have generated the observed words
- This can be done in different ways: Gibbs sampling or variational inference

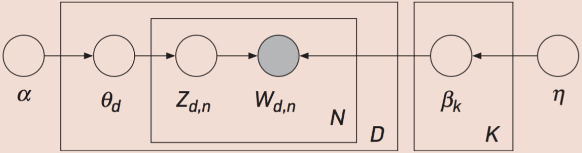
Gibbs sampling algorithm

Initialize topic assignment randomly according. For each iteration:

- For each document:
  - For each word:
    - Resample topic for word, given all other words and their topic assignments
    - Depends on (i) how many words in that document are assigned to a topic, and (ii) how often the word is assigned to each topic
- Number of iterations is determined by the researcher

Graphical model for LDA (Blei, 2012)

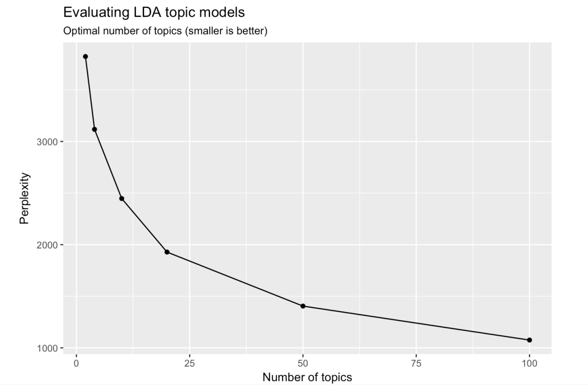
Figure 4. The graphical model for latent Dirichlet allocation. Each node is a random variable and is labeled according to its role in the generative process (see Figure 1). The hidden nodes—the topic proportions, assignments, and topics—are unshaded. The observed nodes—the words of the documents—are shaded. The rectangles are "plate" notation, which denotes replication. The *N* plate denotes the collection words within documents; the *D* plate denotes the collection of documents within the collection.



Determining k

- Number of topics k is determined by the researcher
- One approach to the right number of topics: perplexity criterion
- Perplexity is a measure of the likelihood of a hold-out test set given the model
- Procedure:
    - Estimate topic models with various values of k
    - Calculate perplexity score.
    - Choose topic model with lower perplexity

Determining k



Perplexity of topic model on Associated Press dataset. From: `http://cfss.uchicago.edu/fall2016/text02.html`

Validating topic models (Grimmer & Stewart 2013)

- Semantic validity: extent to which topics are coherent
  - Absence of random, unrelated words
  - Topics that are specific enough and not overly general
  - Can be evaluated using coders
- Predictive validity: how well does variation in topic usage correspond with predicted events
  - E.g, a terrorism topic in media reports should peak after a terrorist incident
- Convergent validity: extent to which model output can be validated with other approaches

Issues with topic models

Multimodality (Roberts et al. 2016) – the outcomes we get may depend on the random starting values

Difficult to evaluate – how do we decide on the correct model?

Topic models in R

`lda`

`topicmodels`

- Use **convert()** function in **quanteda** to export dfm:

```
#create corpus
speeches <- corpus(speeches)

#create dfm
speeches.dfm <- dfm(speeches)

#convert quanteda dfm to topicmodels dfm
speeches.lda.dfm <- convert(speeches.dfm, to = "topicmodels")
```

Extensions of LDA topic model

- Structural topic model (Roberts et al, 2014)
- `stm`
- Like LDA but with document-level covariate information
- The covariates can affect affect topical prevalence, topical content or both
- For a nice intro using **tidytext**: `https://juliasilge.com/blog/sherlock-holmes-stm/`