Running and validating topic models

Wouter van Atteveldt

CCS Hannover, Feb 2018
Programme

Tuesday:
- AM: Intro, Text analysis with R
- PM: Topic Modeling: application and validation

Wednesday:
- AM: Technical Details
- PM: Structural Topic Modeling

Thursday:
- AM: Linguistic processing; visualization
What are topic models?

- Some words and documents are quite similar to each other
  - company, business, form, enterprise, corporation
  - Works by Asimov (and papers on framing ;-
    ))
- Can we automatically figure out what words and documents form clusters?
  - topic = cluster of words and documents
Less informal definition

- DTM is very high dimensional space
- But many rows/columns correlate very strongly
- Topics: latent 'factors' that 'explain' the data in fewer dimensions
- Topic modeling as dimensionality reduction
  - (cf. factor analysis, PCA, hierarchical clustering etc.)
Graphical interpretation

(see handout Graphical Interpretation, top part)
LDA topic modeling

- Technical details will follow
- **Generative process**: LDA ‘models’ how an authors writes a document
  - pick a topic, grab random words from topic
- Fitting a topic model
Generative process

- For each document $d$:
  - Choose the number of words
  - Choose a mixture of topics
  - For each word:
    - Choose a topic from the mixture
    - Choose a word from the topic

Mixture model: Each document has multiple topics, each word can be in multiple topics
Plate notation

\[ \alpha \rightarrow \theta_d \rightarrow Z_{d,n} \rightarrow W_{d,n} \rightarrow \beta_k \rightarrow \eta \]

- Dirichlet parameter
- Per-document topic proportions
- Per-word topic assignment
- Observed word
- Topics
- Topic hyperparameter

Intro: What are topic models?
Plate notation

1. Random word distribution $\beta_i$ for all K topics
2. For each document $d$:
   1. Draw random topic proportions $\theta_d$
   2. For each word $n$:
      1. Draw actual topic $Z_{d,n}$ from $\theta_d$
      2. Draw actual word $W_{d,n}$ from $\beta_{Z}$
Latent Dirichlet Allocation

- $\beta_i$ and $\theta_d$ are drawn from *dirichlet distributions*
  - Distributions over proportions, i.e. over multinomial distributions
- $\beta_i$ and $\theta_d$ are *latent*:
  - We can’t measure them directly
  - They are assumed to explain the selection of manifest words
- How to determine the value of the latent variables?
Fitting the model

- Generative model 'assumes' that we know topics and want documents
- But we have documents and want topics!
  - $\beta$: Which words belong to each topic?
  - $\theta$: Which documents belong to each topic?
- Given the model, we can generate new data ($W$)
- Find the model that maximizes likelihood of our actual text
Examples

Some examples of topic models "in the wild"
Nuclear technology


- Topic model of NY Times coverage of nuclear technology
- Similar results as Gamson & Modigliani (1989)?
FIGURE 3
Occurrence of topics that have a strong temporal component
FIGURE 4
Occurrence over time of detailed ($K = 25$) topics that constitute the *Accidents/ Danger* topic from the $K = 10$ model
Scientific topics


- Topic model of PNAS abstracts
- Are topics meaningful? Overlap with manual class designations?
- Can we identify 'hot' topics
A generalized\textsuperscript{1} fundamental\textsuperscript{146} theorem\textsuperscript{267} of natural\textsuperscript{250} selection\textsuperscript{250} is derived\textsuperscript{233} for populations\textsuperscript{250} incorporating\textsuperscript{149} both genetic\textsuperscript{250} and cultural\textsuperscript{250} transmission\textsuperscript{25}. The phenotype\textsuperscript{3} is determined\textsuperscript{17} by an arbitrary\textsuperscript{3} number\textsuperscript{287} of multiallelic\textsuperscript{3} loci\textsuperscript{3} with two\textsuperscript{271}-factor\textsuperscript{60} epistasis\textsuperscript{250} and an arbitrary\textsuperscript{149} linkage\textsuperscript{3} map\textsuperscript{3}, as well as by cultural\textsuperscript{250} transmission\textsuperscript{25} from the parents\textsuperscript{250}. Generations\textsuperscript{250} are discrete\textsuperscript{60} but partially\textsuperscript{273} overlapping\textsuperscript{146}, and mating\textsuperscript{250} may be nonrandom\textsuperscript{250} at either the genotypic\textsuperscript{250} or the phenotypic\textsuperscript{250} level\textsuperscript{199} (or both). I show\textsuperscript{25} that cultural\textsuperscript{250} transmission\textsuperscript{25} has several\textsuperscript{173} important\textsuperscript{173} implications\textsuperscript{17} for the evolution\textsuperscript{250} of population\textsuperscript{250} fitness\textsuperscript{250}, most notably\textsuperscript{230} that there is a time\textsuperscript{72} lag\textsuperscript{72} in the response\textsuperscript{213} to selection\textsuperscript{250} such that the future\textsuperscript{257} evolution\textsuperscript{250} depends\textsuperscript{105} on the past selection\textsuperscript{250} history\textsuperscript{250} of the population\textsuperscript{250}. 
Intro: What are topic models?
### Intro: What are topic models?

#### Running Topic Models

**Cold topics**

<table>
<thead>
<tr>
<th>Year</th>
<th>Mean $\theta_j$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1991</td>
<td>0.015</td>
</tr>
<tr>
<td>1993</td>
<td>0.01</td>
</tr>
<tr>
<td>1995</td>
<td>0.009</td>
</tr>
<tr>
<td>1997</td>
<td>0.007</td>
</tr>
<tr>
<td>1999</td>
<td>0.006</td>
</tr>
<tr>
<td>2001</td>
<td>0.005</td>
</tr>
</tbody>
</table>

Topics: CDNA, Amino acid, Sequence, Acids, Protein, Mass, Chromatography, Polyepitope

#### Hot topics

<table>
<thead>
<tr>
<th>Year</th>
<th>Mean $\theta_j$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1991</td>
<td>0.015</td>
</tr>
<tr>
<td>1993</td>
<td>0.01</td>
</tr>
<tr>
<td>1995</td>
<td>0.009</td>
</tr>
<tr>
<td>1997</td>
<td>0.007</td>
</tr>
<tr>
<td>1999</td>
<td>0.006</td>
</tr>
<tr>
<td>2001</td>
<td>0.005</td>
</tr>
</tbody>
</table>

Topics: Species, Global, Climate, CO2, Water, Environmental, Years, Marine, Carbon, Diversity, Ocean, Extinction

---

**Valuing Topic Models**

Wouter van Atteveldt
Running a topic model

library(topicmodels)
dtm <- convert(dfm, to = "topicmodels")
m <- LDA(dtm, method = "Gibbs", k = 10, control=list(alpha=.5))

terms(m, 5) # top terms per topic
posterior(m)$topics # topics per document
posterior(m)terms # topics per term
m@wordassignments # topics per (actual) word

(shortest workshop ever! :))
Hyperparameters

- \( k = \) number of topics
- \( \alpha = '\text{dispersion}' \) parameter
  - high \( \alpha = \) many topics per documents
  - low \( \alpha = \) fewer topics per document
  - default = \( 50/k \)
  - I prefer lower \( \alpha \), e.g. \( 5/k \) or \( 1/k \)

(more detail tomorrow!)
Hands-on

- Get your data
- Run some topic models
- Inspect terms, documents
- Try out different settings of alpha and k
- Try out different preprocessing (stopwords, lemma, filter, etc)
- What do you think?
How to validate topic models?

1. Face validity: informal inspection, present top words
2. Goodness-of-fit/predictive likelihood measures
3. External task improvement
4. Compare to human coding of desired or derived topics
5. Formal validation of topic coherence

(cf. Grimmer & Stewart, Political Analysis)
Face validity

- Probably most common method
- Inspect/present top-n words per topic
- Better: also inspect/present top-n documents per topic
  - Context matters!
- Problem: we’re good at seeing patterns
  - And we want to see the pattern!
Aside: corpustools / tokenbrowser

- Corpustools preserves word order
- Allows for e.g. proximity search, KWIC
- Can also recreate document:

```r
#+BEGIN_SRC
devtools::install_github("kasperwelbers/tokenbrowser")
library(corpustools) library(tokenbrowser)
tc = create_corpus(sotutexts, doc_column = 'id')
tc$preprocess('token', 'feature', min_docfreq = 5, remove_stopwords = T)
m = tc$lda.fit('feature', create_feature = 'topic', K = 10, alpha = 0.01)
d = tc$get()
url = categorical_reader(d, category = d$topic)
browseURL(url)
#+END_SRC
```
Goodness-of-fit

- How likely are the actual texts according to the model
- Use separate 'test' data to prevent overfitting

```
topicmodels::perplexity(m, dtm_test)
```

FIGURE 2
Perplexity of LDA models with different numbers of topics and alpha

Quantitative Analysis of Large Amounts of Journalistic Texts
External/concurrent validity

- **External validity**: Does having the topic model improve a 'downstream' task?
  - e.g. better retrieval of relevant documents, predict dependent variable
- **Concurrent validity**: Quantitative manual coding of sample given desired (or derived) codebook, compare
  - (cf. intercoder reliability)
Validation of topic coherence

- Topics are supposed to be meaningful and coherent
- Test directly with human/crowd coders:
  - Pick 'odd word out' from topic
  - Pick 'odd topic out' for document

Odd word out

- Pick a topic
- Get top-N words for that topic
- Get random 'intruder' word
  - That has low probability in chosen topic
  - But high probability in other topic
- Can human distinguish between actual and intruder words?
Odd topic out

- Pick a document
- Get top-N topics for that document
- Pick a random topic that is unlikely in that document
- Present the topics (as top-word lists)
- Can human pick out which topic does not belong?
Figure 2: Screenshots of our two human tasks. In the word intrusion task (left), subjects are presented with a set of words and asked to select the word which does not belong with the others. In the topic intrusion task (right), users are given a document’s title and the first few sentences of the document. The users must select which of the four groups of words does not belong.
Validating topic models

Running and validating topic models

Wouter van Atteveldt
### Running and Validating Topic Models

<table>
<thead>
<tr>
<th>CORPUS</th>
<th>TOPICS</th>
<th>LDA</th>
<th>CTM</th>
<th>PLSI</th>
</tr>
</thead>
<tbody>
<tr>
<td>NEW YORK TIMES</td>
<td>50</td>
<td>-7.3214 / 784.38</td>
<td>-7.3335 / 788.58</td>
<td>-7.3384 / 796.43</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>-7.2761 / 778.24</td>
<td>-7.2647 / 762.16</td>
<td>-7.2834 / 785.05</td>
</tr>
<tr>
<td>WIKIPEDIA</td>
<td>50</td>
<td>-7.5257 / 961.86</td>
<td>-7.5332 / 936.58</td>
<td>-7.5378 / 975.88</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>-7.4629 / 935.53</td>
<td>-7.4385 / 880.30</td>
<td>-7.4748 / 951.78</td>
</tr>
<tr>
<td></td>
<td>150</td>
<td>-7.4266 / 929.76</td>
<td>-7.3872 / 852.46</td>
<td>-7.4355 / 945.29</td>
</tr>
</tbody>
</table>

Figure 3: The model precision (Equation 1) for the three models on two corpora. Higher is better. Surprisingly, although CTM generally achieves a better predictive likelihood than the other models (Table 1), the topics it infers fare worst when evaluated against human judgments.
Validating topic models

Running and validating topic models

Wouter van Atteveldt
Hands-on

- Run a topic model! :)
- Try out different settings
  - hyperparameters: K, alpha, etc.
  - preprocessing: stemming, stop words, filters
  - subsets of documents (e.g. only positive/negative reviews; or take only contrasting words)
  - Optional: Make a scree plot (if not too much data)
- Validate/interpret based on:
  - Term list
  - Documents per topic