ECPR Methods Summer School:
Big Data Analysis in the Social Sciences

Pablo Barberá
London School of Economics
pabloobarbera.com

Course website:
pabloobarbera.com/ECPR-SC105
Discovery in Large-Scale Text Datasets
Overview of techniques

- **Descriptive analysis:**
  - What are the *characteristics* of this corpus? How do some documents compare to others?
  - Keyness, collocations, readability scores, document similarity...

- **Clustering and scaling documents:**
  - What are the main *themes* in this corpus? How do different documents relate to words differently?
  - Topic models (LDA, STM), scaling methods (wordscores, wordfish, PCA)

- **Clustering and scaling words:**
  - What are the semantic relationships between *words*?
  - Word embeddings
Topic models
Overview of text as data methods

Fig. 1 in Grimmer and Stewart (2013)
Topic Models

- Topic models are algorithms for discovering the main “themes” in an unstructured corpus.
- Can be used to organize the collection according to the discovered themes.
- Requires no prior information, training set, or human annotation – only a decision on \( K \) (number of topics).
- Most common: Latent Dirichlet Allocation (LDA) – Bayesian mixture model for discrete data where topics are assumed to be uncorrelated.
- LDA provides a generative model that describes how the documents in a dataset were created.
  - Each of the \( K \) topics is a distribution over a fixed vocabulary.
  - Each document is a collection of words, generated according to a multinomial distribution, one for each of \( K \) topics.
Latent Dirichlet Allocation

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.
Illustration of the LDA generative process

Figure 2. Illustration of the generative process and the problem of statistical inference underlying topic models

(from Steyvers and Griffiths 2007)
Topics example

<table>
<thead>
<tr>
<th>Topic 247</th>
<th>Topic 5</th>
<th>Topic 43</th>
<th>Topic 56</th>
</tr>
</thead>
<tbody>
<tr>
<td>word</td>
<td>prob.</td>
<td>word</td>
<td>prob.</td>
</tr>
<tr>
<td>DRUGS</td>
<td>.069</td>
<td>RED</td>
<td>.202</td>
</tr>
<tr>
<td>DRUG</td>
<td>.060</td>
<td>BLUE</td>
<td>.099</td>
</tr>
<tr>
<td>MEDICINE</td>
<td>.027</td>
<td>GREEN</td>
<td>.096</td>
</tr>
<tr>
<td>EFFECTS</td>
<td>.026</td>
<td>YELLOW</td>
<td>.073</td>
</tr>
<tr>
<td>BODY</td>
<td>.023</td>
<td>WHITE</td>
<td>.048</td>
</tr>
<tr>
<td>MEDICINES</td>
<td>.019</td>
<td>COLOR</td>
<td>.048</td>
</tr>
<tr>
<td>PAIN</td>
<td>.016</td>
<td>BRIGHT</td>
<td>.030</td>
</tr>
<tr>
<td>PERSON</td>
<td>.016</td>
<td>COLORS</td>
<td>.029</td>
</tr>
<tr>
<td>MARIJUANA</td>
<td>.014</td>
<td>ORANGE</td>
<td>.027</td>
</tr>
<tr>
<td>LABEL</td>
<td>.012</td>
<td>BROWN</td>
<td>.027</td>
</tr>
<tr>
<td>ALCOHOL</td>
<td>.012</td>
<td>PINK</td>
<td>.017</td>
</tr>
<tr>
<td>DANGEROUS</td>
<td>.011</td>
<td>LOOK</td>
<td>.017</td>
</tr>
<tr>
<td>ABUSE</td>
<td>.009</td>
<td>BLACK</td>
<td>.016</td>
</tr>
<tr>
<td>EFFECT</td>
<td>.009</td>
<td>PURPLE</td>
<td>.015</td>
</tr>
<tr>
<td>KNOWN</td>
<td>.008</td>
<td>CROSS</td>
<td>.011</td>
</tr>
<tr>
<td>PILLS</td>
<td>.008</td>
<td>COLORED</td>
<td>.009</td>
</tr>
</tbody>
</table>

**Figure 1.** An illustration of four (out of 300) topics extracted from the TASA corpus.

(from Steyvers and Griffiths 2007)

Often \( K \) is quite large!
Latent Dirichlet Allocation

- Document = random mixture over latent topics
- Topic = distribution over n-grams

Probabilistic model with 3 steps:

1. Choose $\theta_i \sim \text{Dirichlet}(\alpha)$
2. Choose $\beta_k \sim \text{Dirichlet}(\delta)$
3. For each word in document $i$:
   - Choose a topic $z_m \sim \text{Multinomial}(\theta_i)$
   - Choose a word $w_{im} \sim \text{Multinomial}(\beta_{i,k=z_m})$

where:
- $\alpha$ = parameter of Dirichlet prior on distribution of topics over docs.
- $\theta_i$ = topic distribution for document $i$
- $\delta$ = parameter of Dirichlet prior on distribution of words over topics
- $\beta_k$ = word distribution for topic $k$
Latent Dirichlet Allocation

Key parameters:

1. $\theta = \text{matrix of dimensions } N \text{ documents by } K \text{ topics where } \theta_{ik}$ corresponds to the probability that document $i$ belongs to topic $k$; i.e. assuming $K = 5$:

   $\begin{array}{ccccc}
   \text{T1} & \text{T2} & \text{T3} & \text{T4} & \text{T5} \\
   \text{Document 1} & 0.15 & 0.15 & 0.05 & 0.10 & 0.55 \\
   \text{Document 2} & 0.80 & 0.02 & 0.02 & 0.10 & 0.06 \\
   \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
   \text{Document } N & 0.01 & 0.01 & 0.96 & 0.01 & 0.01 \\
   \end{array}$

2. $\beta = \text{matrix of dimensions } K \text{ topics by } M \text{ words where } \beta_{km}$ corresponds to the probability that word $m$ belongs to topic $k$; i.e. assuming $M = 6$:

   $\begin{array}{cccccccc}
   \text{W1} & \text{W2} & \text{W3} & \text{W4} & \text{W5} & \text{W6} \\
   \text{Topic 1} & 0.40 & 0.05 & 0.05 & 0.10 & 0.10 & 0.30 \\
   \text{Topic 2} & 0.10 & 0.10 & 0.10 & 0.50 & 0.10 & 0.10 \\
   \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
   \text{Topic } k & 0.05 & 0.60 & 0.10 & 0.05 & 0.10 & 0.10 \\
   \end{array}$
Plate notation

\[ \beta = M \times K \] matrix where \( \beta_{im} \) indicates \( \text{prob(topic}=k \) for word \( m \)

\[ \theta = N \times K \] matrix where \( \theta_{ik} \) indicates \( \text{prob(topic}=k \) for document \( i \)
Validation

From Quinn et al, AJPS, 2010:

1. **Semantic validity**
   - Do the topics identify coherent groups of tweets that are internally homogenous, and are related to each other in a meaningful way?

2. **Convergent/discriminant construct validity**
   - Do the topics match existing measures where they should match?
   - Do they depart from existing measures where they should depart?

3. **Predictive validity**
   - Does variation in topic usage correspond with expected events?

4. **Hypothesis validity**
   - Can topic variation be used effectively to test substantive hypotheses?
Example: open-ended survey responses


- Data: General Social Survey (2008) in Germany
- Responses to questions: Would you please tell me what you associate with the term “left”? and would you please tell me what you associate with the term “right”?
- Open-ended questions minimize priming and potential interviewer effects
- Sparse Additive Generative model instead of LDA (more coherent topics for short text)
- $K = 4$ topics for each question
Example: open-ended survey responses

Table 1: Top scoring words associated with each topic, and English translations

<table>
<thead>
<tr>
<th>Left topic 1: <strong>Parties</strong> (proportion = .26, average lr-scale value = 5.38)</th>
</tr>
</thead>
<tbody>
<tr>
<td>linke, spd, partei, linken, pds, politik, kommunisten, parteien, grünen, punks</td>
</tr>
<tr>
<td>the left, spd, party, the left, pds, politics, communists, parties, greens, punks</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Left topic 2: <strong>Ideologies</strong> (proportion = .26, average lr-scale value = 5.36)</th>
</tr>
</thead>
<tbody>
<tr>
<td>kommunismus, links, sozialismus, lafontaine, rechts, aber, gysi, linkspartei, richtung, gleichmacherei</td>
</tr>
<tr>
<td>communism, left, socialism, lafontaine, right, but, gysi, left party, direction, levelling</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Left topic 3: <strong>Values</strong> (proportion = .24, average lr-scale value = 4.06)</th>
</tr>
</thead>
<tbody>
<tr>
<td>soziale, gerechtigkeit, demokratie, soziales, bürger, gleichheit, gleiche, freiheit, rechte, gleichberechtigung</td>
</tr>
<tr>
<td>social, justice, democracy, social, citizen, equality, equal, freedom, rights, equal rights</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Left topic 4: <strong>Policies</strong> (proportion = .24, average lr-scale value = 4.89)</th>
</tr>
</thead>
<tbody>
<tr>
<td>sozial, menschen, leute, ddr, verbinde, kleinen, einstellung, umverteilung, sozialen, vertreten</td>
</tr>
<tr>
<td>social, humans, people, ddr, associate, the little, attitude, redistribution, social, represent</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Right topic 1: <strong>Ideologies</strong> (proportion = .27, average lr-scale value = 5.00)</th>
</tr>
</thead>
<tbody>
<tr>
<td>konservativ, nationalsozialismus, rechtsradikal, radikal, ordnung, politik, nazi, recht, menschen, konservative</td>
</tr>
<tr>
<td>conservative, national socialism, right-wing radicalism, radical, order, politics, nazi, right, people, conservatives</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Right topic 2: <strong>Parties</strong> (proportion = .25, average lr-scale value = 5.26)</th>
</tr>
</thead>
<tbody>
<tr>
<td>npd, rechts, cdu, csu, rechten, parteien, leute, aber, verbinde, rechtsradikal</td>
</tr>
<tr>
<td>npd, right, cdu, csu, the right, parties, people, but, associate, right-wing radicals</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Right topic 3: <strong>Xenophobia</strong> (proportion = .25, average lr-scale value = 4.55)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ausländerfeindlichkeit, gewalt, ausländer, demokratie, nationalismus, rechtsradikalismus, diktatur, national, intoleranz, faschismus</td>
</tr>
<tr>
<td>xenophobia, violence, foreigners, democracy, nationalism, right-wing radicalism, dictatorship, national, intolerance, fascism</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Right topic 4: <strong>Right-wing extremists</strong> (proportion = .23, average lr-scale value = 4.90)</th>
</tr>
</thead>
<tbody>
<tr>
<td>nazis, neonazis, rechtsradikale, rechte, radikale, radikalismus, partei, ausländerfeindlich, reich, nationale</td>
</tr>
<tr>
<td>nazis, neonazis, right-wing radicalists, rightists, radicals, radicalism, party, xenophobia, rich, national</td>
</tr>
</tbody>
</table>

Note: “proportion” indicates the average estimated probability that any given response is assigned to a topic. “average lr-scale value” is the mean position on the left-right scale (from 0 to 10) of individuals whose highest probability belongs to that particular topic.

Example: open-ended survey responses

Fig. 6: Left-right scale means for different subsamples of associations with left (dashed = sample mean, bars = 95% CIs)

Fig. 7: Left-right scale means for different subsamples of associations with right (dashed = sample mean, bars = 95% CIs)

Example: open-ended survey responses

Fig. 9: Systematic relationship between associations with “left” and “right” and characteristics of respondents

Note: Each line indicates a 95% confidence interval (and 66% confidence interval in darker color) for the coefficient of eight different regressions of topic usage (in a scale from 0 to 100) at the respondent level on seven individual-level characteristics. The line on the bottom right corner (second row, second plot), for example, shows that individual a one-category change in age is associated with around one percentage point increase in the probability that the individual associated “right” with political parties.

Example: topics in US legislators’ tweets

- Data: 651,116 tweets sent by US legislators from January 2013 to December 2014.
- 2,920 documents = 730 days \times 2 \text{ chambers} \times 2 \text{ parties}
- Why aggregating? Applications that aggregate by author or day outperform tweet-level analyses (Hong and Davidson, 2010)
- $K = 100$ topics (more on this later)
- Validation: \text{http://j.mp/lda-congress-demo}
Choosing the number of topics

- Choosing $K$ is “one of the most difficult questions in unsupervised learning” (Grimmer and Stewart, 2013, p.19)
- We chose $K = 100$ based on cross-validated model fit.

- **BUT**: “there is often a negative relationship between the best-fitting model and the substantive information provided”.
- GS propose to choose $K$ based on “substantive fit.”
Extensions of LDA

1. Structural topic model (Roberts et al, 2014, AJPS)

Why?
- Substantive reasons: incorporate specific elements of DGP into estimation
- Statistical reasons: structure can lead to better topics.
**Structural topic model**

- **Prevalence**: Prior on the mixture over topics is now document-specific, and can be a function of covariates (documents with similar covariates will tend to be about the same topics)

- **Content**: distribution over words is now document-specific and can be a function of covariates (documents with similar covariates will tend to use similar words to refer to the same topic)
Dynamic topic model

Source: Blei, “Modeling Science”
Dynamic topic model

Source: Blei, “Modeling Science”
Figure 5. Two topics from a dynamic topic model. This model was fit to *Science* from 1880 to 2002. We have illustrated the top words at each decade.
Word embeddings
Beyond bag-of-words

Most applications of text analysis rely on a bag-of-words representation of documents

- Only relevant feature: frequency of features
- Ignores context, grammar, word order...
- Wrong but often irrelevant

One alternative: word embeddings

- Represent words as real-valued vector in a multidimensional space (often 100–500 dimensions), common to all words
- Distance in space captures syntactic and semantic regularities, i.e. words that are close in space have similar meaning
  - How? Vectors are learned based on context similarity
  - Distributional hypothesis: words that appear in the same context share semantic meaning
- Operations with vectors are also meaningful
### Word embeddings example

<table>
<thead>
<tr>
<th>word</th>
<th>$D_1$</th>
<th>$D_2$</th>
<th>$D_3$</th>
<th>...</th>
<th>$D_N$</th>
</tr>
</thead>
<tbody>
<tr>
<td>man</td>
<td>0.46</td>
<td>0.67</td>
<td>0.05</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>woman</td>
<td>0.46</td>
<td>-0.89</td>
<td>-0.08</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>king</td>
<td>0.79</td>
<td>0.96</td>
<td>0.02</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>queen</td>
<td>0.80</td>
<td>-0.58</td>
<td>-0.14</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

### Diagrams

- **Male-Female**
  - king
  - man
  - woman
  - queen

- **Verb tense**
  - walking
  - walked
  - swam
  - swimming

- **Country-Capital**
  - Spain: Madrid, Rome, Berlin
  - Italy: Rome, Berlin
  - Germany: Berlin
  - Turkey: Ankara
  - Russia: Moscow, Ottawa
  - Canada: Ottawa
  - Japan: Tokyo
  - Vietnam: Hanoi, Beijing
  - China: Hanoi, Beijing
word2vec (Mikolov 2013)

- Statistical method to efficiently learn word embeddings from a corpus, developed by Google engineer
- Most popular, in part because pre-trained vectors are available
- Two models to learn word embeddings:
Figure 4: *Distances by core countries.* Plot of Euclidian distances between US and Russia (gray), and US and China (maroon).
Course logistics

ECTS credits:

- **Attendance:** 2 credits (pass/fail grade)
- Submission of **at least 3 coding challenges:** +1 credit
- Submission of **class project:** +1 credit
  - Due by August 27th via email to P.Barbera@lse.ac.uk
  - Goal: analysis of Big Data using techniques covered in class
- Examples:
  - Topic model of newspaper articles
  - Network analysis of social media data
  - Application of supervised learning methods
  - ...anything that is useful for your research!
- 5 pages max (including code) in Rmarkdown format
- Graded on a 100-point scale

If you wish to obtain more than 2 credits, please indicate so in the attendance sheet
Some final reminders...

1. You can download all your code, challenges, and data from RStudio Server:
   → Export > download as .zip file
   ▶ Server will be deactivated tonight at 10pm

2. Materials (but not solutions) will remain on course website

3. Please complete the teaching evaluations!

4. How you can contact me after the course:
   ▶ P.Barbera@lse.ac.uk
   ▶ www.pablobarbera.com
   ▶ @p_barbera