1 Introduction

In `LDAvis`, we visualize the fit of an LDA topic model to a corpus of documents. The data and model are described as follows:

Data:

- D documents in the corpus
- n_d tokens in document d, for $d = 1...D$ (denoted `doc.length` in our package's R code)
- $N = \sum_d n_d$ total tokens in the corpus
- W terms in the vocabulary
- M_w is defined as the frequency of term w across the corpus, where $\sum_w M_w = N$ (denoted `term.frequency` in our package's R code)

Model:

- K topics in the model
- For document $d = 1...D$, the length-K topic probability vector, $\boldsymbol{\theta}_d$, is drawn from a Dirichlet($\boldsymbol{\alpha}$) prior, where $\alpha_k > 0$ for topics $k = 1...K$.
- For topic $k = 1...K$, the length-W term probability vector, $\boldsymbol{\phi}_k$, is drawn from a Dirichlet($\boldsymbol{\beta}$) prior, where $\beta_w > 0$ for terms $w = 1...W$.
- The probability model states that for the j^{th} token from document d, a latent topic, z_{dj}, is drawn, where $\mathrm{P}(z_{dj} = k) = \theta_{dk}$ for document $d = 1...D$, token $j = 1...n_d$, and topic $k = 1...K$.
- Then, the j^{th} token from the d^{th} document, Y_{dj}, is drawn from the vocabulary of terms according to $\mathrm{P}(Y_{dj} = w \mid z_{dj}) = \phi_{(z_{dj},w)}$, for document $d = 1...D$, token $j = 1...n_d$, and term $w = 1...W$.

A number of algorithms can be used to fit an LDA model to a data set. Two of the most common are the collapsed Gibbs sampler (Griffiths and Steyvers, 2004) and variational Bayes (Blei et al 2003).

Our interactive visualization tool, `LDAvis`, requires five input arguments:

1. $\boldsymbol{\phi}$, the $K \times W$ matrix containing the estimated probability mass function over the W terms in the vocabulary for each of the K topics in the model. Note that $\phi_{kw} > 0$ for all $k \in 1...K$ and all $w \in 1...W$, because of the priors. (Although our software allows values of zero due to rounding). Each of the K rows of $\boldsymbol{\phi}$ must sum to one.

2. $\boldsymbol{\theta}$, the $D \times K$ matrix containing the estimated probability mass function over the K topics in the model for each of the D documents in the corpus. Note that $\theta_{dk} > 0$ for all $d \in 1...D$ and all $k \in 1...K$, because of the priors (although, as above, our software accepts zeroes due to rounding). Each of the D rows of $\boldsymbol{\theta}$ must sum to one.

3. n_d, the number of tokens observed in document d, where n_d is required to be an integer greater than zero, for documents $d = 1...D$. Denoted `doc.length` in our code.

4. `vocab`, the length-W character vector containing the terms in the vocabulary (listed in the same order as the columns of ϕ).

5. M_w, the frequency of term w across the entire corpus, where M_w is required to be an integer greater than zero for each term $w = 1...W$. Denoted `term.frequency` in our code.

In general, the prior parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are specified by the modeler (although in some cases they are estimated from the data), n_d and M_w are computed from the data, and the algorithm used to fit the model produces point estimates of $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$. When using the collapsed Gibbs sampler, we recommend using equations 6 and 7 from Griffiths and Steyvers (2004) to estimate $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$. These are the "smoothed" estimates of the parameters that incorporate the priors, rather than, for example, the matrices containing the counts of topic assignments to each document and term, which are a common output of Gibbs Sampler implementations that don't necessarily incorporate the priors. Two popular packages for fitting an LDA model to data are the R package `lda` (Chang, 2012) and the JAVA-based standalone software package `MALLET` (McCallum, 2002). Our package contains an example of using the `lda` package to fit a topic model to a corpus of movie reviews, available in the `inst/examples/reviews` directory of `LDAvis`.

2 Definitions of visual elements in `LDAvis`

Here we define the dimensions of the visual elements in `LDAvis`. There are essentially four sets of visual elements that can be displayed, depending on the state of the visualization. They are:

1. **Default Topic Circles:** K circles, one to represent each topic, whose areas are set to be proportional to the proportions of the topics across the N total tokens in the corpus. The default topic circles are displayed when no term is highlighted.

2. **Red Bars:** $K \times W$ red horizontal bars, each of which represents the estimated number of times a given term was generated by a given topic. When a topic is selected, we show the red bars for the R most *relevant* terms for the selected topic, where $R = 30$ by default (see Sievert and Shirley (2014) for the definition of *relevance*).

3. **Blue Bars:** W blue horizontal bars, one to represent the overall frequency of each term in the corpus. When no topic is selected, we display the blue bars for the R most salient terms in the corpus, and when a topic is selected, we display the blue bars for the R most relevant terms. See Chuang et al. (2012) for the definition of the *saliency* of a term in a topic model.

4. **Topic-Term Circles:** $K \times W$ circles whose areas are set to be proportional to the frequencies with which a given term is estimated to have been generated by the topics. When a given term, w, is highlighted, the K default circles transition (i.e. their areas change) to the K topic-term circles for term w.

Let's define the dimensions of these visual elements:

1. The area of the **Default Circle** for topic k, A_k^{default}, is set to be proportional to $N_k / \sum_k N_k$, where N_k is the estimated number of tokens that were generated by topic k across the entire corpus. The formula for N_k is:

$$N_k = \sum_{d=1}^{D} \theta_{dk} n_d.$$

   It is straightforward to verify that $\sum_k N_k = N$.

2. The width of the **Red Bar** for topic k and term w, denoted P_{kw}, is set to $\phi_{kw} \times N_k$ for all topics $k = 1...K$ and terms $w = 1...W$.

3. The width of the **Blue Bar** for term w is set to $\sum_k P_{kw}$, the total number of occurrences of term w in the corpus (note that prior to version 0.3.2 of LDAvis, this width was set to M_w, the user-supplied frequency of term w across the entire corpus).

4. The area of the **Topic-Term Circle** for term w and topic k, denoted $A_{kw}^{\text{topic-term}}$, is set to be proportional to $P_{kw} / \sum_k P_{kw}$.

3 Discussion

Here we point out a few things about `LDAvis`:

1. Note that the all the visual elements represent frequencies (of various things in the training data), rather than conditional probabilities. For example, the area of topic-term circle $A_{kw}^{\text{topic-term}}$ could have been set to be proportional to $\phi_{kw} / \sum_k \phi_{kw}$, but instead we set it to be proportional to $P_{kw} / \sum_k P_{kw}$. So, suppose the term "foo" had a 0.003 probability under, say, topic 20 and topic 45, and negligible probability under all other topics. One might expect that upon highlighting "foo", the topic-term circles would all disappear except for two equal-area topic-term circles representing topics 20 and 45. Instead, if, for example, topic 20 occurred twice as frequently as topic 45, then the topic-term circle for topic 20 would be twice as large as that for topic 45 upon "foo" being highlighted. This reflects the fact that 2/3 of the occurrences of "foo" in the training data were estimated to have come from topic 20. In other words, we reflect the underlying (and potentially variable) frequencies of the topics themselves when we compute the areas of the topic-term circles.

   The same principle holds for the red bars and blue bars – they visualize frequencies, rather than proportions, so that wider bars signify more frequent terms in the training data. We felt this was an important feature of the data to visualize, rather than building a visualization that simply displayed aspects of $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$, which are normalized, and don't reflect the frequencies of the terms and topics in the data.

2. By default, we set the dimensions of the left panel to be 530 x 530 pixels, and we set the sum of the areas of the default topic circles and the topic-term circles to be $530^2/4$, so that these circles cover at most 1/4 of the panel (in practice, because of overlapping circles, they cover less than 1/4 of the area of the panel). Likewise, the sum of the areas of the topic-term circles is set to be 1/4 of the area of the left panel of the display. This way the visualization looks OK for a range of numbers of topics, from roughly $10 \leq K \leq 100$.

3. The centers of the default topic circles are laid out in two dimensions according to a multidimensional scaling (MDS) algorithm that is run on the inter-topic distance matrix. We use Jensen-Shannon divergence to compute distances between topics, and then we use the `cmdscale()` function in `R` to implement classical multidimensional scaling. The range of the first coordinate (along the x-axis) is not necessarily equal to that of the second coordinate (along the y-axis); thus we force the aspect ratio to be 1 to preserve the MDS distances. In practice (across the examples we've seen), the ranges of the x and y coordinates are within about 10% of each other.

4 References

1. Blei, David M., Ng, Andrew Y., and Jordan, Michael I. (2003). Latent Dirichlet Allocation, *Journal of Machine Learning Research*, Volume 3, pages 993-1022.

2. Griffiths, Thomas L., and Steyvers, Mark (2004). Finding Scientific Topics, *Proceedings of the National Academy of Science*, Volume 101, pages 5228-5235.

3. Jonathan Chang (2012). lda: Collapsed Gibbs sampling methods for topic models. R package version 1.3.2. http://CRAN.R-project.org/package=lda.

4. McCallum, Andrew Kachites (2002). MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu.

5. Chuang, Jason, Manning, Christopher D., and Heer, Jeffrey (2012). Termite: Visualization Techniques for Assessing Textual Topic Models, *Advanced Visual Interfaces*.