

## Some popular LDA implementations

- Mallet (Java): <http://mallet.cs.umass.edu/topics.php>
- Gensim (Python): <https://radimrehurek.com/gensim/>
- Blei's group (C):  
<https://www.cs.princeton.edu/~blei/lda-c/index.html>

2. Data download
3. Corpus preprocessing

① Open and work with the notebook 'create\_arxiv\_corpus'

## 4. Topic modeling of corpuses

### Topic extraction

- 1 Download source code from  
`https://www.cs.princeton.edu/~blei/lda-c/index.html`
- 2 From the terminal, cd into folder 'lda-c-dist', and compile the code:  
`> make`
- 3 Now, you can extract topics using a command like  
`> lda est [initial alpha] [k] [settings] [data] [random/seeded/*] [directory]`  
Read the readme.txt file or ask the teachers to get more information about the meaning of each of these input parameters

### Topic visualization

- 1 Use the provided python script 'visualize\_topics.py':  
`> python visualize_topics.py`
- 2 If you have any time left, ask your teacher for assistance about graphic visualizations with D3.