This report, produced as a part of the requirement for the Directed studies, has the different experiments been carried out using the softwares used in the state of the art for topic modelling using LDA. Also the replication of the results using the softwares based on different sources are presented.

Author: Arun Jayapal

Supervisor: Dr. Martin Emms

November 27, 2014
1 Introduction

Latent Dirichlet Allocation is a statistical model used to discover the topics from a collection of documents. All the technical details about this model are discussed in detail in the report submitted along with this. This document reports the experiments performed on different packages which are widely used for topic discovery. The experiments were carried out based on the results reported from some particular source. The results of the experiments carried out are summarized along with the reproduction of results based on the sources. In addition to the results of the experiments, a list of softwares in use with brief description about each of them are listed.
2 Topics result reproduction

<table>
<thead>
<tr>
<th>Software</th>
<th>Dataset</th>
<th>Paper used</th>
<th>Reproduce result</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDA-C</td>
<td>Subset of AP dataset (2249 documents)</td>
<td><a href="#">Paper here</a></td>
<td>No, the results are not reproducible. This could be because of different reasons (1) The parameter settings used to produce the results were not specified in the paper (2) Randomization of hyper-parameters $\alpha$ (3) Usage of distributions. Yet to try the models stored at the file download. This has been developed based on Gibbs sampling. All the references to the paper are mentioned at the following link.</td>
</tr>
<tr>
<td>GibbsLDA++</td>
<td>Model trained on their own dataset. With the dataset, models and their outputs are also provided. We could compare the results we get with the results they got (is provided for reference). Further to this, wikipedia dataset’s 200 topics are also provided (although the dataset that they worked on is not provided).</td>
<td><a href="#">Link</a></td>
<td>I was able to run the exact procedure as they have mentioned in their tutorial and was able to reproduce only 67% of their results. Few reasons for not being able to reproduce 100% are (1) Randomization involved in initializations and (2) Usage of distributions (3) Randomization involved in Gibbs sampling.</td>
</tr>
<tr>
<td>Software</td>
<td>Dataset</td>
<td>Paper used</td>
<td>Reproduce result</td>
</tr>
<tr>
<td>----------</td>
<td>---------</td>
<td>------------</td>
<td>------------------</td>
</tr>
<tr>
<td>Mallet</td>
<td>A small dataset has been provided with the software package. Further to this, they have also provided the output of the dataset which they got.</td>
<td>I did not come across any particular paper on this. But a short tutorial describing the outputs with the program settings are provided <a href="http://www.cs.princeton.edu/~blei/papers/BleiNgJordan2003.pdf">here</a></td>
<td>The outputs could be reproduced to a substantial level. The dataset was used to produce 10 topics. Most of the top words of the topics produced by LDA on my system are found in the topics produced in the tutorial as well. One advantage of using mallet is the implementation is multi-threaded, which makes it faster.</td>
</tr>
</tbody>
</table>

---

3 The percentages were calculated using a overlap metric $\frac{T_x \cap T_y}{\min(T_x, T_y)}$. This could be used only to compare the similarity between topics. A different metric called Hellinger distance has been used widely to measure the similarity between documents based on topic distribution.
3 Available Softwares

The available softwares according to my knowledge are listed here.

1. LDA-C
   - Implemented in C
   - Provides implementation of Variational Expectation version of LDA
   - Developed by Blei

2. GibbsLDA++
   - Implemented in C++
   - Provides implementation of Gibbs sampling version of LDA
   - Developed by Xuan-Hieu Phan and Cam-Tu Nguyen

3. Gensim
   - Implemented in Python
   - Provides implementations based on the papers “Online learning for Latent Dirichlet Allocation” and “Online variational inference for the Hierarchical Dirichlet Process”
   - Developed by Radim Řehůřek

4. Matlab topic modelling toolbox 1.4
   - Available from [http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm](http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm)
   - Implemented in Matlab
   - Provides implementation of Gibbs sampling version of LDA (based on the original paper “Finding Scientific topics”)
   - Developed by Mark Steyvers

5. Mallet topic modelling
   - Available from [http://mallet.cs.umass.edu/topics.php](http://mallet.cs.umass.edu/topics.php)
   - Implemented in JAVA
   - Provides implementation based on the paper “Probabilistic Topic Models” by Steyvers and Griffiths (2007)
   - Developed by Andrew McCallum
   - Further to the original implementation of Mallet, there is a package to be used from R, a complete tutorial on its usage from R is provided at [Topic modelling in R](http://mallet.cs.umass.edu/topics.php)
6. Stanford Topic Modelling Toolbox

- Available from [http://nlp.stanford.edu/software/tmt/tmt-0.4/](http://nlp.stanford.edu/software/tmt/tmt-0.4/)
- Implemented in Scala
- I have never tried using this, but their claim is to generate word usage across topics, time and other data groupings.