HUL381/ELL457: Assignment 5

Maximum Marks: 6

Submission deadline: **2 May, 23:55**

This assignment involves some simple experimentation with Latent Dirichlet Allocation (LDA), the probabilisitic topic modelling approach we discussed in class. Here is how you should proceed:

1. Download an LDA implementation. The recommended one is the original one at `http://www.cs.princeton.edu/~blei/lda-c/`, but that page itself links to several other implementations as well, in case you prefer one of those.

2. Download the corpus of 2246 Associated Press newspaper articles, also available at the above link.

3. Understand both the code and the data. You don't need to worry about the internal details of the code, but just need to figure out how to use it on your given corpus, and how you can set parameters like the number of topics. It may be helpful to experiment initially with just a very small data set (say, 10 documents), to get yourself comfortable with running the code, parameter settings, and making sense of the output.

4. Now run the code on the full corpus, experimenting with the number of topics. Try at least 3 different values for the number of topics. In each case, show the top 10 most likely words for all your topics. Can you interpret any of these topics? What kinds of articles are they capturing? Also comment on how the nature of the topics changes as you change the number of topics. Based on this, can you make any suggestion as to what an 'appropriate' number of topics might be for this corpus; or do you feel there is no single correct number? Comment.

5. In your results, identify all cases of polysemy and homonymy that you can find. Discuss at least a few of these cases in your report, mentioning how the learnt topics are accounting for these cases.

6. Write up a report with all your results and discussion (including your answers to all of the questions posed above). Submit this via Moodle, at the latest by **May 2nd, 23:55**. Any late submissions will be penalised.

Collaboration policy: You are free to work with each other in terms of figuring out how to get the code to work and any other operational issues. However, your experiments, results, and report must be entirely your own. Any copying detected will be treated as plagiarism and dealt with accordingly.