NPFL097 Assignment - Latent Dirichlet Allocation[1]

In this assignment, you get the document collection data from the US political blog "Daily Kos". Your task is to compute some statistics on this dataset, to implement the Latent Dirichlet Allocation (LDA) topic model (presented in the previous lectures), and to provide some characteristics of the model you learned.

The data are already preprocessed for the task. The `training_data.tsv` and the `test_data.tsv` contain word counts in the document collection. Each line consists of a triple

`<document_id>TAB<word_id>TAB<number_of_occurences>`.

The file `dictionary.txt` maps the word IDs to words. The stop words (the most frequent English words), which generally worsen the LDA performance, were already removed from the data.

1 Questions

1. Using the training data, compute the maximum likelihood estimation over the words across the documents. Plot a histogram showing 20 most probable words. (0.5 pts)
2. What is the log-probability of the test data if we use the MLE model from question 1)? Note that the test set contains words which are not contained in the training set. Explain. (0.5 pts)
3. Instead of MLE, do the Bayessian inference model on the training data using a symmetric Dirichlet prior with a concentration parameter 0.1. Provide a formula for the predictive distribution. (0.5 pts)
4. Using the Bayessian model from question 3), compute the log probability and the per-word perplexity of the test data. (0.5 pts)
5. Compute the per-word perplexity of the test data using uniform multinomial dictribution over the dictionary. Compare this value to the previously computed perplexities and explain. (0.5 pts)
6. Implement the Latent Dirichlet allocation topic model as described in the previous lectures. Set the hyperparameters $\alpha = 0.1$, $\gamma = 0.1$ and set number of topics $K = 20$. Plot the topic posteriors of the document 1 as a function of the number of Gibbs sweeps, up to 20 sweeps. Comment on these. (4 pts)
7. Compute the word entropy for each of the topics as a function of the number of Gibbs sweeps. (1 pts)
8. Show histograms of the most frequent 20 words of a three chosen topics after 20 Gibbs sweeps. (0.5 pts)
9. Compute the per-word perplexity of the test data for the state after 20 Gibbs sweeps, and compare it to the previously computed perplexities. Are 20 Gibbs sweeps adequate? (1 pts)
10. Try to change the number of topics K, the hyperparameters α and β and the number of Gibbs sweeps. How the performance changes? (1 pts)

[1] This assignment was inspired by one in the course taught by Carl Edward Rasmussen in Cambridge University. (`http://mlg.eng.cam.ac.uk/teaching/4f13/`)