Probabilistic Graphical Models **Department of Computer Science**
Prof. Dr. Sophie Fellenz (previously Burkhardt) TU Kaiserslautern
Sourav Dutta

Probabilistic Graphical Models, SS 2022

Exam

Date of Issue: 08.08.2022
Deadline: 18.09.2022 - 23:59

Problem: Topic Models

Here we have provided to you the *20NewsGroup* dataset in two formats: the original (`orig`) dataset, and a slightly modified (`mod`) version of the same.

1. Download both versions of the dataset here. Perform the necessary preprocessing steps on the data.
2. Implement the code for Latent Dirichlet Allocation (LDA). Let this be `tm1`.
3. Implement the code for a non-parametric topic model. Let this be `tm2`.
4. Train `tm1` and `tm2` on both versions of the dataset.
5. Report your findings supported with metrics like perplexity and coherence for all the models. What are the differences in topics for the two dataset versions? Mention the effect of number of topics and dataset size on the results.
6. Through your analysis and comparison of all the results, try to figure out what has been modified in the `mod` dataset. Please note that you should justify your answer based on your analysis only.

Problem: Analysis and Improvements

1. You are most likely to find a difference in the topics for the two dataset variants `orig` and `mod`. Propose and explain at least one method using which you think you could improve the topic diversity and quality of the worse model. (Hint: using additional information from the data, changing priors in the model, etc.) Please note that this question requires a proper explanation as your answer and does not ask you to code an implementation. However, you would get extra marks if you provide an implementation with results to support your views.

Notes and Guidelines

- You are allowed to use *gensim* for training any topic model in this project.
- Along with your solution file, you are required to write a detailed report in the format of one of the ML conferences (e.g. *NeurIPS*, *ICML*, *ICLR*, *AISTATS*, etc.). Please use the official paper format for any such conference from the internet. It is recommended to use *Overleaf* to edit the LaTeX document.
- Your **documentation** should contain separate sections as per the guidelines. You may also follow the sections provided in the paper format as reference. Some important sections could be Introduction, related work, method, results and discussion, conclusion, and references. You also need to provide assets like proper architecture diagrams, algorithm blocks, graphical plots, and tabular results which not only support your methods and experiments but also helps the reader to easily interpret your idea.
- You may refer to references like publications/blogs on the internet for ideas and motivation. Any such sources that you use as your reference should be cited properly in your documentation. However, please note that **Plagiarism of any kind is STRICTLY FORBIDDEN**. If found guilty, you will be penalised, which could even lead to your disqualification from the exam.
- For the submission of your solution, please create a private Gitlab repository in the **TUK RHRK Gitlab account** and push your code files as well as your documentation report PDF into the repository. You need to share the Gitlab repository with us so that we are able to grade your submission. Please note that any file pushed after the deadline will not be considered eligible for your grading.
- Please refer the **Portfolio Exam Guidelines** PDF file in OLAT for more detailed description and guidelines related to the portfolio exam.
- If you have any question related to the exercise, please contact Sourav Dutta only via email or Mattermost (*dutta@cs.uni-kl.de*).