Exercise 6 - Topic Modeling

Discover Topics and Trends in Computer Science

Deadlines

The deadline for Exercise 6 is **19.12.2022, 23:59 (Zurich Time).**
The deadline for the peer review is **09.01.2022, 23:59 (Zurich Time).** You will find instructions for the peer review process at the end of this document.
The deadline for feedback to your peer reviewers is **14.01.2022, 23:59 (Zurich Time).**

Learning goals

This exercise is about topic modelling, more specifically about Latent Dirichlet Allocation (LDA) and topic modelling based on pretrained language models (PLM). By completing this exercise you should …

- … understand how topic modeling is used as a text-mining tool.
- … be able to apply LDA and PLM-based topic models.

Please keep in mind that you can always consult and use the exercise forum if you get stuck (note that we have a separate forum for the exercises).

Deliverables

We encourage you to use Colab to develop your notebooks, since there you have access to GPU time. **After you have finished the assignment, download your notebook as a .ipynb file.** That way your reviewers can view and execute your code. Or can view your already executed code.

Please hand in your code and your lab report. Hand in the following files and name them exactly in the following fashion:

- ex06_tm.ipynb
- ex06_labreport.pdf

zip it and name the zip-folder *ex06_ml4nlp1.zip*.

Please submit the lab report in PDF format. The lab report should contain a detailed description of the approaches you have used to solve this exercise. Please also include results. In this exercise description, we highlight places in green where we expect a statement about an issue in your lab report.

Please note:

- Your peers need to be able to run your code. If it does not work, you will not be able to obtain the maximum number of points.
- DO NOT submit the data files!

Data

For this exercise, you will work with the dblp: a large database containing metadata on computer science (and related) publications. You will perform topic modelling on the titles (**not on the text itself!**) of computer science publications to detect important topics and trends in the field.

Given

For the exercise, you are given this notebook. It already contains code for downloading the dataset, preprocessing, and for constructing a first topic model.

Part 1 - Topic Modelling using LDA

The given notebook limits the number of titles to a reasonable amount and divides publications into three time-periods: before 1990, from 1990 to 2009, and 2010 onwards. An LDA-based topic model for the time period “before 1990” is already implemented.

Extend the notebook to perform topic modelling on the other two time periods.

We encourage you to experiment with different numbers of topics and with different ways of preprocessing.

You can increase the number of topics generated by the topic model, but you should not go below 5.

- For each time-period assign a name to each generated topic based on the topic's top words. List all topic names in your report. If a topic is incoherent to the degree that no common theme is detectable, you can just mark it as incoherent (in other words: no need to name a topic that does not exist).
- Do the topics make sense to you? Are they coherent? Do you observe trends? Discuss in 4-6 sentences.

Part 2 - Topic Modelling using Combined Topic Models (CTMs)

Bianchi et al. 2021 propose a topic modelling method that makes use of pre-trained language models such as BERT. The authors provide a simple colab tutorial showcasing how to use the CTM library that implements their method.

Again, perform topic modelling for the 3 time-periods. This time using the CTMs. Use the same number of topics as before. You can copy and adjust code from the author's tutorial.

- Again: Assign a name to each topic based on the topic's top words (for each time-period). List all topic names in your report.
- Bianchi et al. 2021 claim that their approach produces more coherent topics than previous methods. Let's test this claim by comparing the coherence of the topics produced by CTM with the topics produced by LDA. Describe your observations in 2-4 sentences.
- Do the two models generate similar topics? Can you discover the same temporal trends (if there are any)? Discuss in 4-6 sentences.

Peer Review Instructions

If you are not already registered on Eduflow follow this link https://app.eduflow.com/join/GXHN93 and register with the E-mail address you use for OLAT. Then you should be added to the course page automatically.
As soon as the deadline for handing in the exercise expires you will have time to review the submissions of your peers. You need to do **2 reviews** to get the maximum number of points for this exercise.
Here are some additional rules:

- If you do not submit 2 reviews, the maximum number of points you can achieve is 0.75 (from a total of 1).
- Please use full sentences when giving feedback.
- Be critical, helpful, and fair!
- **All reviews are anonymous: Do not put your name into the python scripts, the lab report or the file names.**
- You must also give your reviewers some feedback. The same criteria as above apply.
- If you consistently provide very helpful feedback, you can be awarded a bonus of 0.5 in total in case you didn't achieve the full 6 points from all exercises. A maximum of 6 points from the exercises can go into the final grade.

Groups:

- You can create groups of two to solve the exercise together.
- Both students should submit the solutions separately.
- If you did not already work together for the previous exercise, write a small post in the "Groups"-thread in the exercise forum on OLAT to notify the instructors about the group.
- As a group member, you still have to review two submissions with your own eduflow account. However, you may work together in the group to write all 4 reviews.