Assignment 2 – Natural Language Processing 2024 – 2025 (JM2050-M-6)

Instructions

- In this assignment, you will work on a topic modeling task. The goal of the task is to gain insights into the content of a corpus of news articles on Nvidia.
- Assignments are to be in groups of 3 students.
- Use Python for the assignment.
- Deliverables:
    - The file Team-XX.pdf (Where XX is your group number, e.g. Team-07.pdf) that contains your report with answers to the questions.
    - Python code that you have developed to solve the assignment. If you developed multiple files, place them in a zip file named Team-XX-code.zip
    - Submit all deliverables electronically through CANVAS. Make sure that you submit everything in a single zip file that contains the solutions to the exercise (report and source code). Name it A2-XX.zip. (where XX is your group number, e.g. A2-07.zip)
- Deadline for submission of solutions is November 7th, 2024, 22:00 hrs.
- Be concise and to the point.
- Use correct terminology.

Questions regarding the assignment can be asked through Canvas, or during the sessions on Thursday mornings.

Good luck!

The management team was interested with your model predicting Nvidia's stock price movements. They are now intrigued whether text data is valuable to improve the company's performance. However, they do not understand yet which information was in this text data that you used for your predictions, so they asked your team to explain this to them. Therefore, they approached you with the question to analyze the content of the news articles on Nvidia.

Luckily, you know about several topic modeling algorithms that can help you to do this. You decide to use the same dataset and focus on all articles with Nvidia in the content. The dataset should still be on the corporate server, but in case it already got deleted accidentally, you download it again. You will use topic modeling to gain insights onto the news coverage on Nvidia and apply the three algorithms LDA, FLSA-W and BERTopic.

Since your managers are unfamiliar with topic modeling, briefly explain in a report to the management team how these algorithms work and support your arguments with references to scientific literature. Also present the outcomes of the topic model in an easily understandable way and discuss the applicability of the models within the company. Because the managers' workload is very high during this quarter, keep your report to the point (at most five A4 pages, which excludes figures, tables, and references).

Consider at least the following points for your report:

1. How do these three algorithms extract useful information from news on Nvidia?
2. Train the three topic models. How did you select the number of topics? Create an elbow plot for one of your FLSA-w models, to see if this can help the select the number of topics. (this will be discussed in the Topic Modeling tutorial)
3. Based on the found number of topics, run multiple iterations of the algorithms to improve the topics' quality. Thus, train a model and assess the topics based on your own assessment. Do the topics contain typos, or meaningless words? Then, update the corpus and retrain the topic model. Explain how the iterations affect the output. How many iterations did you do? And why? Was this the same for each algorithm? If not, explain what explains the difference. Add the produced topics for each iteration to your report's appendix.
4. Evaluate the output qualitatively. What remarks can you make based on your own assessment about the topic quality? How do the algorithms compare?
5. Evaluate the output quantitatively. Which metric did you choose? And why? How do the findings from the quantitative and qualitative assessment compare to each other?
6. For BERTopic, include a Topic Similarity Matrix for one of you BERTopic models and discuss briefly what it shows. (this will be discussed in the Topic Modeling tutorial)
7. Present the output of your final topic model; use graphs and figures where appropriate.
8. Use ChatGPT to assign labels to topics and interpret the topics. Explain how you have optimized your prompt. Reflect on how the produced labels compare to your interpretation, are they meaningful?
9. What are the limitations of your topic modeling approach?
10. How could your topic model be applied in a financial company?
11. Any other relevant details.