Advanced Data Mining and Language Technologies - HW2

[Luca Becchetti, Fabrizio Silvestri]

Academic Year 2023-2024

In the following, we introduce the second homework. It will consist of 3 parts: as the program progresses, you will be able to solve tasks that we will release in due course.

Rules The rules published in the Exam file are always valid, in addition:

1. Each homework must be completed in groups of up to two people, the same of HW1.
2. **Use Google Colab to develop your solutions**. Code should be designed for GPU execution. **Important**: homeworks that do not use *pytorch lightning* or adhere to PEP8 *formatting rules* will not be accepted.
3. **A complete notebook must operate within Colab's runtime constraints.** Homeworks that do not comply with Colab's limitations will not be accepted.
4. Each notebook must automatically download all necessary datasets from a designated repository.
5. The code needs to be well written and annotated. You have "text" cells in Colab: explain what you are doing and why.
6. You are **not** permitted to "copy" code fragments from online sources. All solutions must be original. Copying entire models, training pipelines, or any elements central to the main homework's code is strictly prohibited. If you decide not to abide by this rule and you incorporate a code fragment taken from an online source, you **must** cite the URL of the source in your code comments and justify why you had done so. Failure to do so will result in the homework being considered invalid.
7. Copying project solutions from other students is, of course, considered cheating and is strictly forbidden.
8. Do not send e-mails directly to the teacher or the teaching assistants with questions about the project; use Google Classroom instead. Your doubt can be the doubt of your colleagues. The professor or the teaching assistants will answer to you there.
9. You can use your own personal GPU. Download the network weights and load them using *gdown* into the colab you submit.

Submission As with the first assignment, only one of the group will have to turn it in. You must submit a Google Colab that meets the requirements written above. Name your file as *surname1-idNumber1—surname2-idNumber2.ipynb* .

Evaluation The homework will be evaluated based on the following criteria:

1. Code Quality: correctness, readability (length, comments, unnecessary repetitions)
2. Quality of Textual Responses: correctness, clarity, etc.
3. Quality of Produced Visualizations (tables/figures/etc.): correctness, clarity, etc.
4. Quality of the Report: correctness, clarity, etc.
5. Timely Submission: delays in submission will result in point deductions
6. Plagiarism: copying code/text from peers or online sources will result in significant point deductions.

1 Task 1

Word2Vec: This task consists of creating embeddings with word2vec, fasttext and glove and combine them to train a classifier. You must implement your own version of LSTM using pytorch lightning.
Dataset Amazon Reviews. The goal of the analysis is to classify positive (3-4-5 stars) and negative (1-2 stars) reviews based on the review content. Choose the category you like the most.

Metrics: Accuracy, Precision, Recall, F1-score, Loss. Provide a plot of Accuracy and Loss.

1.1 Theory

1. How would a Deep Learning model (of the kind we have seen) behave in the case where a word was never seen during training? Answer on both practical and theoretical aspects.
2. Seq2seq can be done using CNNs? How deal with variable lenghts

Use at most 3 sentences for each answer.

2 Task 2

Q&A: You have to download a dataset from HF, choose a pre-processing (if you want to do it, otherwise not), choose 2 models, and fine-tune this task. One fine-tuning has to be done without using LoRa, while the other has to be done using LoRa and everything has to run within the Colab limits.
Datasets: News Q&A, MED Q&A, Arxiv Q&A. Choose the dataset you like the most.
Metrics: Plots of the training and validation loss (maybe it is useful to use wandb, but it's your choice).

2.1 Theory

1. You have an LLM that generates text and you want to generate the word *Ferrari* within a sentence. How can you do this? Answer on both practical and theoretical aspects.
2. How might biases in training data affect the output of LLMs, and what strategies can be employed to mitigate these biases?

Use at most 3 sentences for each answer.

3 Task 3

For this section, only theoretical part will be asked.

3.1 Theory

1. Discuss the potential advantages of using RAG over purely generative models. In what scenarios might RAG provide significant benefits?
2. Can RAG models handle cases where the needed information is very rare or not present in the training data?
3. Which are some improvements over plain RAG that could potentially enhance its performance or applicability?

Use at most 3 sentences for each answer.