ΥΣ19 **Artificial Intelligence II (Deep Learning for Natural Language Processing)**
Fall Semester 2020
Homework 4
25% of the course mark
Announced: December 23, 2020
Due: January 31, 2021 at 23:59

1. This exercise is about developing a document retrieval system to return titles of scientific papers containing the answer to a given user question. You will use the first version of the COVID-19 Open Research Dataset (CORD-19) in your work (articles in the folder *comm_use_subset*).

   For example, for the question "What are the coronaviruses?", your system can return the paper title "Distinct Roles for Sialoside and Protein Receptors in Coronavirus Infection" since this paper contains the answer to the asked question.

   To achieve the goal of this exercise, you will need first to read the paper Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, in order to understand how you can create *sentence embeddings.* In the related work of this paper, you will also find other approaches for developing your model. For example, you can using Glove embeddings, etc. In this link, you can find the extended versions of this dataset to test your model, if you want. You are required to:

   (a) Preprocess the provided dataset. You will decide which data of each paper is useful to your model in order to create the appropriate embeddings. You need to explain your decisions.

   (b) Implement at least 2 different sentence embedding approaches (see the related work of the Sentence-BERT paper), in order for your model to retrieve the titles of the papers related to a given question.

   (c) Compare your 2 models based on at least 2 different criteria of your choice. Explain why you selected these criteria, your implementation choices, and the results. Some questions you can pose are included here. You will need to provide the extra questions you posed to your model and the results of all the questions as well.

   (40/100 marks)

2. Expand the best system you created in Exercise 1 to retrieve the relevant passages of the papers as well. Provide your questions and the respective passages your model returns. One way to model the results is to provide the article and the passage highlighted. Explain your implementation choices.

(20/100 marks)

3. Build a BERT-based model which returns “an answer”, given a user question and a passage which includes the answer of the question. For this question answering task, you will use the SQuAD 2.0 dataset which has been discussed in the lecture “Textual Question Answering”. You should start with the BERT-base pretrained model “bert-base-uncased” and fine-tune it to have a question answering task as explained in the lecture on BERT. Note that this has been done already by the BERT team and it is available publicly, but we would like you to try to implement this by yourself. If you copy from the BERT code for this task, your mark for this exercise will be 0.

   For more information about question answering systems, you can read the post “How to Build an Open-Domain Question Answering System?” and the survey “Recent Trends in Deep Learning Based Open-Domain Textual Question Answering Systems”.

(40/100 marks)

Your solutions should be implemented in PyTorch and we expect your reports for each one of the exercises to be well-documented.