Experiment 3

*Start Date: Nov. 16, 2022
*Report Due Date: Nov. 30, 2022

Notes:

- Each experiment must be done individually. You can search material through Internet but remember to mark it.
- Write an experiment report to describe and analyze the experimental observations.
- You should pack your files including **source code, report, readme file, and other related files** in one .zip/.rar/.7z file.
- Please submit on the Web Learning platform. Do NOT Email or Wechat your report to the instructor or TAs.
- No late report is accepted. No exceptions.
- You can write the report using English or Chinese.

Task:

Learn about how to identify similar sentence pairs using the BERT model.

Goal:

In this experiment, the dataset contains nearly 10,000 pairs of query sentences and requires the identification of similar queries using natural language processing techniques. You are expected to analyze the data, design your model and try some practical tricks to improve the performance.

Data:

Please check the following train and test dataset. You can download it via this link.

Query1	Query2	Label
剧烈运动后咯血, 是怎么了?	剧烈运动后咯血是什么原因?	1
剧烈运动后咯血, 是怎么了?	剧烈运动后咯血，需要就医吗?	0

Table 1: An illustration of data.

Experiment Step:

- Do some EDA (Explore Data Analyze) work to analyze the data (e.g. the distribution of sentence length, keywords).
- Build your BERT based classification model using Transformers package from Hugging Face.
- Try **two or more** practical tricks to improve the performance of your model, such as data augmentation, adversarial training, ensemble modeling, etc. You can follow the demo here but DO NOT copy it directly.
- Analyze the experimental results and state your conclusion about this task.