Advance Topics in Machine Learning, 2024
Natural Language Processing -- Group Assignment

(released on 5/12/2024)

In this assignment you will analyse **publicly available text datasets**, applying the techniques you have learnt in class to train **machine learning models** to perform **NLP tasks**.

Please read **carefully** the project description below:

Details of the assignment:

- The assignment must be completed in the assigned **groups**.
- It involves making a **Python notebook** that demonstrates the usage of various ML models on the data as discussed below.
- Your notebook should be **self-explanatory**, with **clear descriptions** of the analysis performed and the **conclusions** drawn.

Due date and presentations:

- The assignment is due **on Monday the 13th of January 2024** at 23:30 (via email to mark.carman@polimi.it, alessandro.manenti@usi.ch, and tommaso.marzi@usi.ch). Only **one member** from each group needs to hand-in the notebook, but the **names of all group members should be listed** at the start of the notebook.
- On **Tuesday the 14th of January 2024**, each group will have **10 minutes** to **present their notebook** and tell us what they have done during the exam session. (Please don't prepare any slides, we just want to see your notebook.)

The assignment will be marked based on the:

(i) appropriateness of methods applied and depth of the **analysis**,
(ii) clarity of the description in the **notebook**, and
(iii) quality of the **presentation**.

THE TASKS

The aim of the assignment is to **apply the techniques you have learnt in class** to **investigate one** of the **datasets** described below, **build** either a **chatbot or question-answering system** with it, and then make the system **voice interactive**, by connecting it to a speech-to-text and text-to-speech models.

The tasks listed below are **only suggestions**. You don't need to perform exactly these task (for some datasets they may not even be appropriate), nor should you limit yourself to the set of tasks listed:

1. Investigate dataset:

Briefly describe the chosen dataset:

- What type of documents does it contain and how many documents are there?
- Calculate and visualise statistics for the collection, e.g. distributions over document length and vocabulary size. [Note: if you are unsure how to do calculate these values, look at some of the earlier notebooks in the tutorial folder.]

Play around with documents using some of code from the course. You could, for example:

- train a Word2Vec embedding on the documents and investigate its properties.
- index the documents so that you can perform keyword search over them.

2. Train and evaluate models:

Each of the datasets comes with a particular task that you can perform, so:

- train a model to perform that task by fine-tuning a Transformer (and possibly also an LSTM or linear model) on the training data;
- test pre-trained models on the task if they already exist [Hint: check the Hugging-Face website (https://huggingface.co/) for state-of-the-art implementations];
- investigate the effectiveness of Large Language Models (LLMs) together with zero-shot and/or few-shot learning on the task;
- evaluate the different methods and compare their performance across a representative test set.

3. Add voice interactivity:

Make use of text-to-speech and speech-to-text models to add voice interaction to the best performing chatbot / question answering system found above.

- Investigate how effective and reliable the voice interactive components are.
- If they are not particularly reliable, how might you change them to make them more robust?

4. Potential extensions:

Depending on the dataset chosen there will be many additional investigations you can perform:

- For instance, you might improve the performance of a model on the task by finding additional training data from a similar/related dataset.
- Try to turn use the model developed to perform a different task (e.g. use a question answering system as a chatbot or vice-versa) and evaluate it on that task.
- Try to personalize the system for your users (i.e. yourselves), by adding your own data to the training set, performing reinforcement learning from human feedback, trying to modify the speech-to-text model to work better on you voices, etc., etc.

THE DATASETS

Each **group** should choose **ONE** of the following datasets to work on:

1. Medical Meadow Medical Flashcards:

- **Website**: https://huggingface.co/datasets/medalpaca/medical_meadow_medical_flashcards
- **Paper:** https://arxiv.org/pdf/2304.08247.pdf
- **Description**: Information on medical curriculum flashcards has been given to GPT-3.5 and used to create medical knowledge question answer pairs.
- **Task**: Medical Question Answering (i.e. train a model to answer medical questions.)

2. ELI5 (long form question answering)

- **Website:** https://facebookresearch.github.io/ELI5/explore.html
- **Subset of ELI5 dataset** (with direct download): https://huggingface.co/datasets/eli5_category
- **Paper:** https://arxiv.org/abs/1907.09190
- **Description**: Long form question answering dataset consisting of a question and long responses generated by either a human, a generative model or an abstractive model.
- **Task**: Long form question answering, generated content detection.

3. OpenAssistant-Guanaco

- **Website**: https://huggingface.co/datasets/timdettmers/openassistant-guanaco
- **Paper:** https://arxiv.org/abs/2304.07327
- **Description**: The dataset consists of multilingual human-written simulated conversations between a person and a chatbot assistant, (where the responses from the assistant were actually written by real people via crowdsourcing).
- **Task**: Fine-tune a chatbot

4. BeerQA

- **Website**: https://beerqa.github.io/
- **Paper**: https://arxiv.org/abs/2010.12527
- **Description**: Open-domain (varying-hop) question answering dataset, where in order to successfully answer a question, information from multiple Wikipedia pages must be aggregated together.
- **Task**: Question answering from Wikipedia.

5. Gridspace Stanford Harper Valley (Spoken Dialog)

- **Website**: https://github.com/cricketclub/gridspace-stanford-harper-valley
- **Paper**: https://arxiv.org/abs/2010.13929
- **Description**: Spoken dialog dataset containing audio of conversations between humans, simulating calls to the Harper Valley Bank call centre.
- **Task**: Transcribe audio, fine-tune chatbot.