Natural Language Processing, 2023/24

Group Assignment

released on 18/04/2024

In this assignment you will analyse **publicly available text and/or audio datasets**, applying the techniques you have learnt in the NLP exercise classes to train various machine learning models to perform **NLP tasks**. Please read carefully the project description below:

Details of the assignment:

- The assignment must be completed in **groups.** Groups must consist of a **minimum of 3, maximum of 5** students.
- The project involves creating a **Python notebook** to analyse and build models using a particular dataset as discussed below.
- Your notebook should be **self-explanatory**, with **clear descriptions** of the analysis performed and the **conclusions** drawn.

Due date for the assignment and subsequent presentations:

- The assignment is due **on Sunday the 26th of May** at 23:00 via **WeBeep**.
- Only **one member** from each group should hand-in the notebook.
- The **names of all group members should be listed** at the start of the notebook.
- On Monday the 27th and Wednesday the 29th of May, each group will have **5 minutes** to **present their notebook** during the practical sessions. (Schedule for presentations will be released closer to the date.)
- Don't prepare slides for the presentation, 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 analyse **one** of the **text datasets** described below. The exact tasks performed may depend on the dataset chosen, but we would expect to see some of the following:

1. Preliminary analysis:

Briefly describe the dataset:

- what type of documents does it contain?
- how many documents are there?
- calculate and visualise some simple statistics for the collection, e.g. the average document length, the average vocabulary size, etc.

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

- cluster the documents and visualise the clusters to see what types of groups are present (or whether the known classes can be found);
- index the documents so that you can perform keyword search over them;
- train a Word2Vec embedding on the data and investigate the properties of the resulting embedding.

2. Training models:

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

- train a model to perform that task (by fine-tuning models on the training data);
- test pre-trained models on the task (if they already exist); and
- evaluate different models and compare their performance.

HINT: as a minimum here we would expect to see a linear classifier trained on the data (if an appropriate for the task) and compare it with deep learning model, such as BERT.

3. Possible extensions:

Depending on the dataset chosen there will be many additional investigations that you can perform, e.g.:

- investigate another task on the same dataset,
- investigate the same task on another related dataset,
- use text-to-speech and speech-to-text models to create a voice interactive system,
- create your own dialog dataset by transcribing audio conversations (e.g. using MS Teams).

THE DATASETS

Each group must choose **ONE** of the following datasets to work on.

- To prevent that every group works on the same task, we limit the total number of groups working on each dataset to 10. So once you have chosen a dataset, please **add your group for the list for the chosen task** in this document:
  https://docs.google.com/document/d/1JxGge5PvywmNuoSH9jQas3vr1xMuZRZqTeD5Y2GLf4Q/edit?usp=sharing
- If you are looking for partners to form a group, add your name and contact detail to the last table at the link above.
- Note that some of the datasets are quite large, so **you may need to sample a subset from them** to be able to work with them.

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. Cornell Movie Dialog:

- **Website**: https://huggingface.co/datasets/cornell_movie_dialog
- **Paper**: https://arxiv.org/abs/1106.3077
- **Description**: Movie dialogue scripts along with metadata information (film title, characters, etc.).
- **Task**: Film dialog generation, prediction of metadata for specific dialog.

3. AVeriTeC (Automated Verification of Textual Claims)

- **Website**: https://fever.ai/dataset/averitec.html
- **Paper**: https://arxiv.org/abs/2305.13117
- **Description:** Claims that have been fact-checked along with textual justification and evidence supporting the verdict. Other fact-checking datasets (FEVER, FEVER2, FEVEROUS) available on https://fever.ai/
- **Task**: Claim detection, claim veracity prediction, justification generation.

4. ELI5

- **Website:** https://facebookresearch.github.io/ELI5/explore.html
- **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.

5. 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

6. OpenOrca-SlimOrca

- **Website**: https://huggingface.co/datasets/Open-Orca/SlimOrca
- **Paper**: https://arxiv.org/abs/2306.02707
- **Description**: Recorded interactions between a user and the chatbot ChatGPT, that can be used to train a model to act like ChatGPT.
- **Task**: Fine-tune a chatbot to mimic ChatGPT.

7. 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.

8. FEW-NERD

- **Website**: https://huggingface.co/datasets/DFKI-SLT/few-nerd
- **Paper**: https://aclanthology.org/2021.acl-long.248/
- **Description**: Manually annotated Named Entity Extraction dataset, with fine-grained labels.
- **Task**: Named entity recognition.

9. YELP R:

- **Website**: https://huggingface.co/datasets/yelp_review_full
- **Paper**: https://arxiv.org/abs/1509.01626
- **Description**: Very large collection of product reviews with star ratings.
- **Task**: Star rating prediction, sentiment analysis

10. 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.