Natural Language Processing, 2024/25

Group Assignment

released on 19/04/2025

In this assignment you will analyse **publicly available datasets**, applying techniques you have learnt about in class to perform various **NLP tasks**.

Please read carefully the project description below.

Details of the assignment:

- The assignment must be completed in **groups of 4** (minimum) **or 5** (maximum) students.
- The project involves developing a **Python notebook** to analyse and build models on a particular dataset as discussed below. The notebook should be **self-explanatory**, with **clear descriptions** of the analysis performed and the **conclusions** drawn.
- In addition to a notebook, each group should create **5-minute** (maximum) **screen-capture video** where the members of the group present their notebook. Don't prepare slides for the presentation. We just want to see you discuss the most interesting parts of your notebook.

Due date for the assignment:

- The assignment is due **on Sunday the 25th of May** at **18:00** via **WeBeep**.
- Only **one member** from each group should hand-in the notebook (ideally in both **.ipynb** and .**html** format), but the **names of all group members should be listed** at the start of the notebook.
- **Important**: the notebook should also contain a **link to the 5-minute video**.
- Note that on Monday the 27th and Wednesday the 29th of May, groups will have to answer questions **about their notebooks** during the practical sessions. (The schedule for this will be released closer to the date.)

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 NLP techniques you have learnt in class** to analyse **one** of the **datasets** described below.

- Note: some of the datasets are quite large, so **you may need to sample** a small percentage of the data and work that.

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 data:

- What is the structure of the dataset? What type of task was the dataset collected for?
- What type of documents does it contain? How many are there? How long are they on average and what is their distribution?
- How big is the vocabulary of the collection? How big is the vocabulary of a document on average?

Play around with documents using code from the early parts of the course. For example, you could:

- Cluster the documents, visualise the clusters and to try to understand what types of groups are present.
- Index the documents so that you can perform keyword search over them.
- Train a Word2Vec embedding and investigate the properties of the resulting embedding.

2. Training models:

Each dataset has been created with a particular task in mind. You don't necessarily need to tackle that particular problem, but you do need to train some model(s) on the data:

- train ML models (e.g. a linear classifier, an LSTM and/or a Transformer) to perform a particular task on the data;
- if possible, try to fine-tune a pretrained models on the same task and compare their performance;
- try an LLM on the task, comparing one, few and zero-shot performance;
- and perhaps even try to fine-tune a small LLM on the task (if it makes sense to do so).

3. Possible extensions:

Depending on the dataset chosen there will be many **additional investigations** that you could perform, for example:

- investigate another task on the same dataset,
- investigate the same task on a related dataset,
- use **text-to-speech** and **speech-to-text** models to **create a voice interactive chatbot**,
- 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.

- We limit the **number of groups** working **on each dataset** to **maximum 10**. So please **add your group** to the **list for the chosen task** in this document:
https://docs.google.com/document/d/1-PVBySZxDIw-C3QyOUF2pWNJZ6DGYH79Ck3cVSaK1n4/edit?usp=sharing
- If you are **looking for a group** or for new members to join your group, please **add your name** and contact detail **to the table at the bottom** of the page.

Sentiment Analysis:

1. **Bittensor Subnet 13 Reddit Dataset**

- https://huggingface.co/datasets/smmrokn/reddit_dataset_44
- large dataset of (90 million) reddit posts and comments
- could be used for sentiment analysis or some other classification

Question Answering and Reading comprehension:

2. **ReAding Comprehension dataset from Examinations (RACE)**

- https://huggingface.co/datasets/ehovy/race
- Large-scale reading comprehension dataset with over 28K passages and almost 100K questions from English middle and high school examinations in China.

3. **Retrieval-Augmented Generation (RAG) Dataset 12000**
- https://huggingface.co/datasets/neural-bridge/rag-dataset-12000
- Dataset contains 12K questions and answers along with the context (text that would have been found by a RAG model) that contains the information for answering the question. The answer has been generated automatically by GPT-4.

4. **RAG-Instruct**
- https://huggingface.co/datasets/FreedomIntelligence/RAG-Instruct
- Dataset contains 40K questions and answers along with the documents (returned by a search based on the question) containing information to answer the question. Answer were generated automatically by GPT-4o.

Program Code Generation:

5. **BigO(Bench)**
- https://huggingface.co/datasets/facebook/BigOBench
- Dataset contains 3K coding problems and 1M solutions, where the aim is to train a chatbot to convert a text description of a programming problem into computer code.

Chatbot Knowledge Distillation:

6. **EagleSFT**
- https://huggingface.co/datasets/nyuuzyou/EagleSFT
- This dataset contains over 500K pairs of human questions and machine-generated responses (mostly Russian (99%) but also English (1%)) for training a chatbot to mimic another model.

7. **WildChat-1M**
- https://huggingface.co/datasets/allenai/WildChat-1M
- Collection of 1M conversations between users and ChatGPT, with demographic information (state, country, hashed IP addresses, request headers).

Medical Question Answering:

8. **PubMedQA**
- https://huggingface.co/datasets/qiaojin/PubMedQA
- Dataset of 273K medical questions with answers extracted from PubMed publications.

Handling Tables in Documents:

9. **ArXiv-tables**
- https://huggingface.co/datasets/staghado/ArXiv-tables
- Small dataset of 1.3K tables extracted from ML papers published on arXiv, including LaTeX source and images. This dataset could be used to experiment with tabular data retrieval.

Cooking!

10. **RecipeNLG**
- https://huggingface.co/datasets/mbien/recipe_nlg
- Millions of recipes! Lots of different task that might be possible from classification to RAG to fine-tuning a cooking bot.

Multimodal Medical:

11. NIH-CXR14-BiomedCLIP-Features Dataset
- https://huggingface.co/datasets/Yasintuncer/NIH-CXR14-BiomedCLIP-Features
- Dataset contains over 100k chest X-ray images and their corresponding textual reports, and CLIP embeddings for both. Note that the images need to be downloaded from a different website

12. PubMedVision
- https://huggingface.co/datasets/FreedomIntelligence/PubMedVision
- Dataset contains over 600K pairs of medical image and chatbot responses describing the image using GPT-4V.

Multimodal Question Answering:

13. ScienceQA
- https://huggingface.co/datasets/derek-thomas/ScienceQA
- Dataset contains over 20K examples consisting of images with textual questions, sourced from an online science learning platform.

Chatbot Reasoning:

14. Reasoning Required Dataset
- https://huggingface.co/datasets/davanstrien/reasoning-required
- This dataset contains 5K chatbot responses when prompted to reason about an input text taken from the web. The idea is to use this dataset to improve the reasoning performance of another model.

15. OpenCodeReasoning
- https://huggingface.co/datasets/nvidia/OpenCodeReasoning
- Dataset contains a huge number (over 700K) examples of problems, reasoning and code produced as an answer the problem. The answers have been generated automatically by the Deepseek-R1 model.

16. DeepMath-103K
- https://huggingface.co/datasets/zwhe99/DeepMath-103K
- Dataset contains over 100K maths questions along with the corresponding answer and three different "solutions" (i.e. reasonings/explanations) generated by the Deepseek-R1 model.

Games:

17. Strategic Game Chess
- https://huggingface.co/datasets/laion/strategic_game_chess
- Massive number (3.2B) chess games played by a chess engine (Stockfish) against itself, along with the final result of the game. Dataset could potentially be used to teach chatbot to play chess well.