Textual Data Analysis - Exercise 2

- What are the key processing steps to create the FineWeb data from crawl sources, and what is the most important tool or method to implement each?

> First the text is extracted from the `WARC` files by using `trafilatura` library which is open source. Next step was filtering the material. In this step the data was subjected to a URL filtering to get rid of adult content, after that the data was filtered to contain only english texts. Next step was deduplication in which the dataset was cleaned from duplicate entries. After these steps the dataset also was subjected to filters used previously on `C4`. These contained *heuristic* filters. Next step was to invent and generate additional heuristic filters on top of the `C4` ones.

- Does the processing to create the FineWeb data omit any of the processing steps discussed on the lecture, and does it add any? (what?)

> All of the steps of the creation of `FineWeb` was introduced in the lecture. However the paper does not mention anything about *Personal Information Masking* or *Content quality filtering*

- What are the key differences between the FineWeb and FineWeb-edu datasets, and what are the key steps to create the latter from the former?

> The key differences in `FineWeb` and `FineWeb-edu` datasets are that the latter is comprised of a subset of texts contained in the former dataset. The aim of the `-edu` was to include only highly educational texts. To filter highly educational texts from the larger general dataset the study group used synthetic data to develop classifier to identify educational texts. Then trained an linear regression model to indentify educational texts from the larger dataset.