Submission for
Exercise task 2
of UTU course TKO_8964-3006
Textual Data Analysis
by Botond Ortutay

Instructions:

Web crawl-based corpora

Read the paper The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale (Penedo *et al.* 2024), which introduces the FineWeb and FineWeb-Edu datasets (version 1). Answer the following questions:

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

Answers:

Disclaimer on material usage and generative AI: All of the answers use the article linked in the instructions. The second answer (on comparing the creation of FineWeb to the creation of a typical crawl based dataset as discussed on the lecture) also uses the January 13th lecture and the related slides as a source. As all of the used material is explicitly mentioned in the instructions I don't feel a need to provide a reference listing.

No generative AI tool has been used in any of the answers below in any way.

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?

The creation of FineWeb started by reformatting crawl data to a more suitable format. The Common Crawl data they used was available in two different dataformats: WARC and WET. Out of these WARC is a more "raw" dataformat consisting of full HTML pages, including HTML code and even some request metadata. WET on the other hand is pre-parsed and plaintext. Out of these two formats FineWeb is based on the WARC files. The justification is that there are too many artifacts in the WET files left from the parsing process such as menu elements etc. The FineWeb team decided that it'd be better to just process the WARC files themselves. For this they used the trafilatura library.

The parsed text then went through two different processes: filtering and deduplication. Filtering was done in two parts: base filtering before deduplication and everything else after deduplication. Base filtering consisted of three different processes: removing adult content, limiting the content to the English language and applying the quality and repetition filters from another similar dataset generation project (MassiveText). Adult content was removed through a simple URL-blocklist and the English language categorization was done using the fastText language classifier, requiring an English score of at least 0,65 for all the included materials.

Deduplication was done using a fuzzy method known as MinHash. MinHash works by running the documents through multiple different hashing algorithms divided into buckets. If any two hash buckets of different documents match, then one of them is marked as a duplicate and thus removed. Deduplication was also completed on a per-crawl basis. The FineWeb team tried to do deduplication for the whole dataset at once, but then the last crawls in the process would suffer, since most of the useful data in them was marked as duplicate and as such advertisemenst and other less useful content got highlighted, affecting the performance of the whole dataset.

After base filtering and deduplication the FineWeb team noticed that one of the earliest datasets developed for training large scale LLMs (C4) still beat FineWeb on certain benchmarks. To fix this they decided to do additional filtering. Firstly they added all of the filters used by C4, except the terminal punctuation filter, which they deemed to exclude too much data.

Then they also developed their own heuristic filters. Firstly they collected statistics on the data in both high and low quality datasets. Then they analyzed how these different statistics could predict quality; whether a difference between the two datasets could be noticed or not based on these statistics. Wherever a difference was noticable a new heuristic filter was considered. In particular the FineWeb team highlighted three important filters they found: In low quality materials they found that

1. Less than 12% of lines end in punctuation
2. More than 10% of chars in dupl. lines
3. More than 67% of lines have less than 30 chars

So in short: FineWeb was created by:

1. Extracting documents from the common crawl data (WARC)
2. Applying base filtering to the data
3. Deduplicating the data
4. Applying additional features from the data, which were
    - copied from the C4 project
    - self developed based on statistical analysis

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

To answer this question I want to refer to page 25 of the 1st lecture's slides, which provides a list of steps to follow when generating a dataset from crawl data. These steps include: 1. HTML text extraction, excluding boilerplate text 2. Exact duplicate removal 3. Near-duplicate removal 4. Heuristic quality filtering 5. Language model-based quality filtering 6. URL-based filtering 7. Content-based toxicity filtering 8. Content quality filtering 9. Personal information masking

FineWeb's creation indeed started by taking the crawl data (WARC format), and performing HTML text extraction, so that the data would be in more usable format.

On the other hand exact duplicate removal wasn't performed as such. FineWeb's deduplication stage instead used MinHash, which they describe as a "fuzzy hashing based deduplication method". Basically this is the same thing as near-duplicate removal, as it is not concerned about exactly matching two documents, instead focusing on comparing the output of different hashing algorithms on the documents. so basically: FineWeb uses near-duplicate removal, and it doesn't use exact duplicate removal.

However FineWeb does use heuristic quality filtering. It's the last thing they described in the article. They even stated that they developed some of the heuristics they used by comparing a high quality and low quality version of the dataset.

The lecture slides mention language model-based quality filtering as one of the steps in the crawl based dataset generation process. However I'm not 100% sure on what exactly this refers to. There definitely wasn't a dedicated step for this in the FineWeb creation process. Several things in the FineWeb creation process could however be considered as "language model-based quality filtering" if we choose to define this term in that way. For example a tool called fastText was used during base filtering to exclude non-English documents. Also: MinHash (the duplicate removal method used) compares text segments from different documents through an n-gram score (which definitely is language model-based analysis). Also when creating the FineWebEdu dataset they used Llama (a large language model) to automatically detect document suitability for educational use, which by definition is language-model based quality filtering. So to summarize: language model-based quality filtering was used in the creation of FineWebEdu. When it comes to just the base FineWeb: certain tools used in the creation process did use LM-based quality analysis under the hood, but (at least if I understand things correctly) LM-based quality analysis wasn't itself a step in the process.

URL-based filtering definitely was applied to get rid of adult websites from the dataset. This by itself could count as toxicity filtering, however I wouldn't call it content based, since the content wasn't analyzed for toxicity, only certain previously known "toxic" websites were blocked.

Similarly to "language model-based quality filtering", "content quality filtering" is also up for interpretation. However as the lecture slides mention FineWebEdu's

classifier as an example on content quality filtering, I'm gonna assume that this means looking at the contents of each document and automatically ranking suitability for the purposes of the dataset using an LLM. With this interpretation it's probably not a surprise that: yes this step is used in the creation of FineWebEdu (but not in the creation of FineWeb; in the creation of FineWeb the quality filtering is mostly done using heuristics instead)

Personal information masking was also done during the FineWeb creation process to hide email- and IP-addresses although this is only mentioned very briefly in the article.

Below is a summary of all these steps, their inclusion in the FineWeb creation process and which part of the process these were included in:

Processing step (lecture examples)	**Used during FineWeb creation?**	**Included in (FineWeb processing steps)**
HTML text extraction, excl. boilerplate	yes	Text extraction (3.2)
Exact duplicate removal	no	
Near-duplicate removal	yes	Deduplication (3.4)
Heuristic quality filtering	yes	Developing additional quality filters (3.6)
Language model-based quality filtering	As a part of other steps & in FineWebEdu	
URL-based filtering	yes	Base filtering (3.3)
Content-based toxicity filtering	no	
Content quality filtering	In FineWebEdu	
Personal information masking	yes	briefly mentioned in 3.7

As for the question whether FineWeb's creation process included any processing steps not discussed on the lecture: I didn't find anything major. There are small things, such as using a tool to limit document language to English (this wasn't a discrete step in the slides (unless this is what "Language model-based quality filtering" refers to) but I do remember it being mentioned on the lecture) and applying some of C4's filters (although this probably falls under heuristic quality filtering), but again: nothing major.

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 purpose behind creating FineWeb was to create a high quality LLM-training dataset to research the dataset creation process itself, as many of the datasets used by large companies for such purpose are not publicly available and not a lot is (publicly) known about them. As a result FineWeb is a high quality general purpose dataset. On the other hand FineWeb-Edu was created with the purpose to develop a document quality classifier which could filter out documents deemed as "educational" from a larger document pool and then use this tool to create an "educational" dataset.

So first an LLM was used on a sample of more than 400 000 documents to give them all an "educational score", and then this sample was used to train their document quality classifier. Then this classifier was used to extract all the documents from FineWeb which achieved a high enough "educational score". The selected documents then formed the FineWeb-Edu dataset.

As FineWeb-Edu was formed by selecting certain high quality documents from FineWeb, it is naturally much smaller then FineWeb, containing 1,3 trillion tokens, whereas FineWeb contains 15 trillion.

As such certain topics are more represented in FineWeb-Edu, while others were cut more. The most relevant topics in FineWeb-Edu are: Education, learning & teaching (unsurprisingly); history, culture & politics **and** health, medicine & biology. The least relevant topics are: Business, finance & law; entertainment, film & theather **and** places, travel & real estate.

FineWeb-Edu also has a better domain fit for Wikipedia, academic content, arXiv and programming content, whereas FineWeb is more general.