What's done?

Crawling:

1. Code to download warc files in each assigned segment (analyzing max 5 warc files in each segment).
2. Check each html file if it has <img> tag.
    a. Check if the image source is valid HTML.
    b. Check if the image is a **valid type (jpeg, jpg, png)**
    c. Check if an image with the same name is present already and if it has the same size, if yes then **skip adding this entry to csv**.
    d. FIlter image on size, checking if downloaded image is **at least 400 x 400**
    e. Check if the image contains nsfw.
        i. Domain filtration using adult, hacking, malware,mixed_adult, phishing list from **https://github.com/olbat/ut1-blacklists**
        ii. Couldn't use **open_nsfw_python3**, because it is built with caffe and I wasn't able to make run on neither google colab or on the local system.
        iii. Used **nudeNet** instead to filter out such images in initial dataset set creation.
3. Created image directory with **segmentName_warcFileNumber/ImageName** and a csv file with header **[my_uuid, image_url, image_filename, alt_text, context, web_url, warc_segment, warc_fn]**.

Filtering:

For each entry in the csv check alt_text and context.

1. Replace non-unicode characters in alt_text as well as context.
2. Filtered data entry with context which is shorter.
3. Check if they are still valid after this.
    a. Using **langdetect** library to check if it is a valid language.
        i. I'm considering all languages which use Unicode characters as valid languages.
    b. After this, I am using **SentenceTransformer with bert-base-nli-mean-tokens** embeddings and checking **cosine similarity** of alt_text and context text, and considering top 5 similar paragraphs into the CSV.
4. I'm printing 10 examples of language filtration, and total stats at the end.

Visualizing:

For each entry in the filtered csv:

1. I'm using the **detectron2** module to get the feature's bounding boxes and plot them on the image. Using this in-place of **fasterCnn** because it also required caffe.
2. I'm displaying 10 such examples.