Classification of logos, solution to logo similarity #1 challenge

For my final solution that is in my repository, there are only 3 Python files that matter. The rest are just previous trials, which were unsatisfactory.

The first important file is the a2.py. The purpose is to extract logos from the parquet file with the websites. In my previous trials, I noticed that when I was downloading the images, after trying to cluster them, I got clusters with the same (almost identical logo), even if the websites were different. My thought process was that, the parquet file contained websites of subsidiaries of the same company, or websites from different countries of the same company.

Upon realising this, I thought the most interesting problem to solve is how to extract the company name from a website link, and then check if that exists in any other website link. If it does, the logo is with a very high change, the same, so it is unnecessary to extract it more than once.

Therefore, the a2.py website has the role of websites of the same company and extract the logo successfully from one of them, skipping the rest. In addition, I only try to download the logo from the element which has the keyword “logo” inside its element’s class/id from the HTML. I save the logos in a company_logos folder.

The second important file is advanced-logo-similarity.py. Here, I assume it’s possible that even after i eliminated subsidieries of the same company with a2.py, it is still possible that I missed a few. So I thought to try to compare the images with each other and see if they are similar in a proportion of at least 80%. If they are, they are most likely, the same logo. I save the unique logos in the logo_clusters folder.

Finally, I do the clustering of the unique images from the logo_clusters folder. I do this with the logo_clustering.py method. Here, I use Hierarchical Clustering to group the logos in at most 10% * number_of_logos clusters. The algorithm doesn't just count logos, but tries to find clusters that have meaningful size, represent diverse but related logos and capture visual similarity across different companies. At the end, a report of the significant clusters is shown.

To run the project, first install the libraries:

```
pip install -r requirements.txt
```

Then, run the files in this order:

```
python a2.py
python advanced-logo-similarity.py --input company_logos --output logo_clusters

python logo_clustering.py -m hierarchical
```