# process_html.py

## Requirements:

python 3.9.5
pip:

- lxml v4.6.3
- beautifulsoup4 v4.10.0
- langdetect v1.0.9
- markdownify v0.11.6
- torch v2.1.0
- transformers v4.34.1
- langchain 0.0.333

## Usage:

1. Install required packages
2. Change XPATH_JSON_PATH, TOKENIZER_PATH, MODEL_PATH as necessary
3. Run the script with “python process_html.py [-h] html url outfile”
   Positional args:
   a. html - path to html file
   b. url - source url of html file
   c. outfile - output JSON file