

DATA 622 | HOMEWORK 1

Questions

Proceed as indicated below:

1. Retrieve the following Web Page

Use the `requests` library to fetch the content of <https://www.visitmaryland.org/>

What is the HTTP status code?

2. Extract the Main Page Text

Using `BeautifulSoup`, extract and print all the visible text from the page (ignore script and style tags).

3. Extract Headings from a Wikipedia Page

Fetch the Wikipedia page for Natural Language Processing.

Extract and print all the headings (`<h1>`, `<h2>`, `<h3>`) from the page.

4. Extract Links from a Page

From the Wikipedia page above, extract and print all the URLs (`href` values) of links found on the page.

5. Extract and Save Text

Extract the first paragraph (`<p>`) from the Wikipedia page and save it to a local text file called `nlp_intro.txt`.