

# MMAI 5400 Assignment 1 -- Web Scraping

For this assignment you will scrape reviews from Trustpilot.com (in the second assignment you will classify the sentiment of those reviews).

## Submission

This assignment should be submitted as Python 3 code and uploaded to Canvas. The submission should be a single `.py` file, and **not** a Jupyter Notebook. The due date is on May 26 8:30am.

The code will be tested and should produce the output specified below.

## Task

Your task is to scrape reviews from Trustpilot. You chose a company for the reviews, for example, Skype. Make sure that the company has at least 500 reviews. The reviews should be written to a `csv` file with the following columns: `companyName` , `datePublished` , `ratingValue` , `reviewBody` .

Example:

<code>companyName</code>	<code>datePublished</code>	<code>ratingValue</code>	<code>reviewBody</code>
Skype	2021-01-12T17:06:39+00:00	3	It shows ...
Skype	2021-01-10T16:58:00+00:00	1	Disgusting...
...	...	...	...

## Steps

### Manual

1. Open trustpilot.com in a browser and search on a company.
2. This will show you the reviews that you will extract and the URL to use in the Python script.

Example: if the company is Skype then the URL will be <https://www.trustpilot.com/review/www.skype.com>

### Python code

1. Use the `requests` module to download the `html` for URL.
2. Extract the total number of reviews. For example the Skype page: `<h2 class="headline">Reviews &nbsp;&nbsp;<span class="headline__review-count">1,292</span></h2>`
3. Iterate over the review pages.

Example:

```
Python for i in range(num_pages): page = requests.get(URL + '?page=' + str(i + 1))
```

4. From each page extract the reviews.
5. From each review, store the following to the CSV file:
  - **companyName**, e.g. Skype.
  - **datePublished**, the date when the review was published.
  - **ratingValue**, the the numerical value of the rating.
  - **reviewBody**, the review text.
6. The final `csv` file should have at least 500 rows and four columns ("`companyName`", "`datePublished`", "`ratingValue`", and "`reviewBody`").

The six steps above should all be coded in the submitted `.py` file. The information written to the `.csv` file should be extracted from the `html` source with `BeautifulSoup`. The `.py` file should run as a script and save the `.csv` file to the present working directory. That is, it should be possible to run your script from a terminal like this: `python <your_review_scraper>.py`, and from `lpython/Jupyter` with `%run <your_review_scraper>.py`.

**Good luck!**