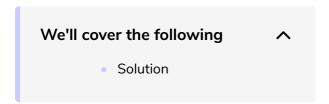
Project 2: Parsing data from a HTML file with Python and REGEX

In this project we use REGEX to find some values from a HTML file



In this project, we want to extract tabular information from a HTML file (see below). Our goal is to extract information available between and except the first numerical index (1..6).

Consider the data.html file below:

```
<html>
<head>
<style>
table, th, td {
 border: 1px solid black;
 border-collapse: collapse;
th, td {
 padding: 5px;
th {
 text-align: left;
</style>
</head>
<body>
2 Japan Japanese
3 China Chinese
4 Middle-east Arabic
5 India Hindi
6 Thailand Thai
```

```
</body>
</html>
```

If we load the HTML file onto a browser it should look like below:

1	England	English
2	Japan	Japanese
3	China	Chinese
4	Middle-east	Arabic
5	India	Hindi
6	Thailand	Thai

A HTML file with table

Solution

In this code, we first extract HTML data (data.html) and then find and extract the values from the HTML code.



Expected output:

```
[('England', 'English'),
('Japan', 'Japanese'),
('China', 'Chinese'),
('India', 'Hindi'),
('Thailand', 'Thai')]
```