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

We'll cover the following

- Solution

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>
<table style="width:100%">
<tr align="center"><td>1</td> <td>England</td> <td>English</td></tr>
<tr align="center"><td>2</td> <td>Japan</td> <td>Japanese</td></tr>
<tr align="center"><td>3</td> <td>China</td> <td>Chinese</td></tr>
<tr align="center"><td>4</td> <td>Middle-east</td> <td>Arabic</td></tr>
<tr align="center"><td>5</td> <td>India</td> <td>Hindi</td></tr>
<tr align="center"><td>6</td> <td>Thailand</td> <td>Thai</td></tr>
</table>
```

```
</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.

main.py

data.html

```
import re

with open('data.html', 'r') as myfile:
    data=myfile.read().replace('\n', '')

result=re.findall(r'<td>\w+</td>\s<td>(\w+)</td>\s<td>(\w+)</td>',data)
print(result)
```

Expected output :

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