Week 11 – Session 2: File Input & Output

In this laboratory session you are going to write a C++ parser using File I/O operation. Use this Github repo: `https://github.com/eiramlan/HTMLParser`

Challenge 1: HTML Parser

We are going to create a simple C++ parser that will read a HTML file from a news portal, extract the key HTML elements, and then then produce a summarise document of the news items (heading) and the corresponding links. For this exercise, you can use the sourceNews.html file provided in the github project.

Task 1 Details, details, details - duration 15 minutes

Similar to Lab 1 Challenge, we will begin this task with the designing our parser. Take a look at the sourceNews.html file. Identify the important HTML elements that we might use in designing our parser. Remember, most of the elements in the file are essential for UI/UX of the news portal but not critical to the important that we want to extract.

- What data should be read into the application?
- How the information is structured in the file?
- Identify important HTML elements that we might need
- Check the sample output file for guidance

Task 2 Setting up - duration 10 minutes

Now that you have decided on what information you want to extract from our sample output file, you can write your function definition. Before you begin, clone the repository into your VStudio using any of the suitable methods. Inside your skeleton codes, you will see three functions that are required to complete your HTML parser.

```
vector<string> GetDataFile(string url);
vector<string> ParseData(vector<string> s, string pattern);
bool WriteExtractNews(vector<string>& news, vector<string>& urls);
```

You need to write the function definitions for all three.

Task 3 Read from the HTML source - duration 15 minutes

Write a function definition for:

```
vector<string> GetDataFile(string url);
```

This function will accept a URL of the source material, read each line into a vector object, and will return the object.

Task 4 Creating your own parser - duration 45 minutes

Write a function definition for:

```
vector<string> ParseData(vector<string> s, string pattern);
```

This function will accept a vector object containing string to be processed, and a string pattern representing the HTML elements that you want to parse. Depending on your implementation, this pattern can be a symbolic representation, or this pattern can actually be a regular expression.

For instance, if you decided not to use regex in extracting the HTML elements, you can use the string pattern as identifier to tell the function on which element that will be parsed. On the other hand, if you decided to use regex, then you should pass the regex as pattern.

The function act as a utility and can be called multiple times to process various HTML elements necessary to extract your news items.

Task 5 Write output to a file - duration 15 minutes

Write a function definition for:

```
bool WriteExtractNews(vector<string>& news, vector<string>& urls);
```

This function will accept a vector of string containing the news topics and urls. You need to write the topic titles and urls to a file. Please refer the demo output for details.