## Web Scraping

Web scraping is a process of retrieving data through automated means. We will be scraping data from a slack channel. Visual studio code will be used to write code and find messages that have the word 'feedback' in them and scrape the message.

The programming language that will be used to scrape messages is c# language.

## C# web scraping tools

Before writing any code, the first step is choosing the suitable C# library or package. These C# libraries or packages will have the functionality to download HTML pages, parse them, and make it possible to extract the required data from these pages.

I will use Html Agility Pack for its ability to download web pages directly or using a browser. This package is tolerant of malformed HTML and supports XPath.

## Setting up development environment

Using visual studio code I installed .Net 5.0

## Project structure and dependencies

Code to create console application using CLI -  dotnet new console

Install Html Agility Pack – dotnet add package HtmlAgilityPack

Install CSV Helper – dotnet add package CsvHelper

## Download and parse web pages

```
// Parses the URL and returns HtmlDocument object
static HtmlDocument GetDocument(string url)
{
    HtmlWeb web = new HtmlWeb();
    HtmlDocument doc = web.Load(url);
    return doc;
}
```

## Parsing the HTML: Getting page links

```
public HtmlNodeCollection SelectNodes(string xpath);
public HtmlNode SelectSingleNode(string xpath);

HtmlDocument doc = GetDocument(url);
HtmlNodeCollection linkNodes = doc.DocumentNode.SelectNodes("//h3/a");
```

```
Uri(Uri baseUri, string? relativeUri);


static List<string> GetMessageLinks(string url)
    {
        var messageLinks = new List<string>();
        HtmlDocument doc = GetDocument(url);
        HtmlNodeCollection linkNodes =
doc.DocumentNode.SelectNodes("//h3/a");
        var baseUri = new Uri(url);
        foreach (var link in linkNodes)
        {
            string href = link.Attributes["href"].Value;
            messageLinks.Add(new Uri(baseUri, href).AbsoluteUri);
        }
        return messageLinks;
    }

static void Main(string[] args)
{
    var messageLinks = GetMessageLinks("http://slack.com/features channels");
    Console.WriteLine("Found {0} links", messageLinks.Count);
}
```

To run this code, open the terminal and navigate to the directory which contains this file, and type in the following: dotnet run