7. REGEX USES

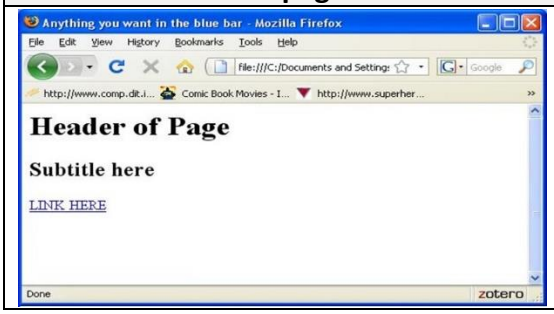
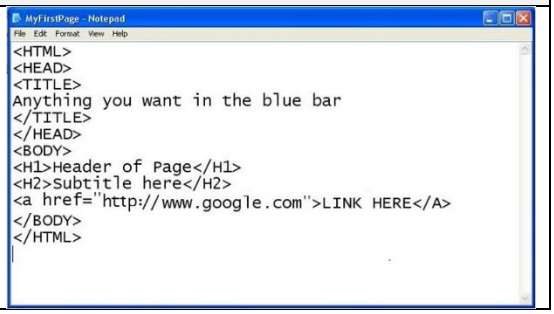
What is Web Scraping?

Introduction

Web scraping is when we write a computer program to extract information off a web page. This isn't as easy as it sounds, because when we view a web page in a web browser, one of the things that the browser is doing is, first, reading the webpage in, and, then, using the HTML instructions on the page to understand how to display the information on the screen correctly, including: which headers to put in, which images to put in, and which hyperlinks to put in. So often when we are scaping a webpage, we are trying to locate just the text on the webpage (and none of the formatting information), so we want to ignore all the HTML instructions.

Simple Example

Below is a sample webpage on the left, with the associated HTML code on the right.

Webpage	HTML
	

A web scraping process will be reading the HTML on the right, and may, for example, be attempting to extract the following text for that HTML:

- "Header of Page"
- "Subtitle here"
- "LINK HERE"

So, if we were looking for simple HTML tags, e.g. "<P>" or "<HEAD>", we can do:

```
RegEx_Pattern = "<[a-zA-Z]+>"
```

This matches a String with at least one character, and enclosed in "<" and ">".

If we are looking for closing HTML tags, e.g. "</P>" or "</HEAD>", we can do:

```
RegEx_Pattern = "<[\/a-zA-Z]+>"
```

This matches a String with a slash and at least one character, enclosed in "<" and ">".

If we are looking for the HTML for a link tag, it is typically structured as follows:

```
<A HREF="URL">TEXT</A>
```

So this can be described as follows:

```
RegEx_Pattern = "<[A-Z] [A-Z]+=\"[A-Z]+\">[A-Z]+<\/[A-Z]>"
```

This matches a String that starts with a "<", followed by a single letter, a space, a word, an equals sign, a double quote, another string, another double quote, a ">", another string, another "<", a forward slash, a single letter, and another ">".