1. INTRODUCTION

What are Regular Expressions?

Introduction

Regular expressions are a compact way of searching for text in a document. So let's imagine we have a document that has millions of lines of text in it, including a number of different email addresses, and we want to extract those addresses from the document. Unfortunately, the addresses come in different formats, including:

- `DamianTGordon@MyMail.com`
- `Damian.Gordon@MyMail.com`
- `Damian.T.Gordon@MyMail.co.uk`

So they all have the "@" symbol in their text, and they also have zero, one, or more full stops (".") before the "@" symbol. They also have at least one, but maybe more full stops after the "@" symbol.

To add to the complexity, there are also other phrases in the document that look like emails but aren't, that we don't want to extract, including incorrect email addresses:

- `DamianTGordon@MyMail` (No domain name, e.g. `.COM`)
- `@MyMail.com` (No name before the @ symbol)

As well as some typical text that looks like an email address:

- `Can we meet@2pm?`
- `We are meeting@boardroom.`

So a regular expression is a special code we can use to describe the rules of what defines a valid email address (and we know there's a few accepted email address formats), and also how to recognise something that isn't a valid address. The good news is … the regular expression below describes a valid email address according to the rules we stated above, and I know it looks complicated, but don't worry we'll be explaining each part of this code over the following weeks:

```
^[A-Za-z+.]+@[A-Za-z.]+[A-Za-z]+$
```

This code should work in almost any programming language, but it is worth noting that different programming languages implement regular expressions in slightly different ways, so a regular expression that works in one language might not always work exactly the same way in another, but we'll specify them to work in as many languages as possible, and we'll note when different languages behave differently.

Some Terminology

A Regular Expression is often called a ***RegEx***, and a collection of Regular Expressions are called ***RegExes***. There are two types of characters used in regular expressions:

- **Metacharacters**: These are characters that have a special meaning.
- **Literal characters**: These are the actual characters we want to match.