Use python package RegEx to process text

Overview

Nowadays, it becomes more and more popular to obtain some information from web pages. Especially, we are working on a SearchExperts project, which is expected to extract important information like phone, emails and so on given a web page.

Python RegEx can be used to check if a string contains the specified search pattern. I am planning to use that package for this project, so I would like to introduce it to you, which may help the classmates to know this wonderful python package.

In this paper, I would like to introduce what is the regular expressions and show you some typical examples.

What are regular expressions?

A lot of information has some patterns, which is called Regular expression.

Officially, a Regular Expression is used for identifying a search pattern in a text string. It also helps in finding out the correctness of the data and even operations such as finding, replacing and formatting the data is possible using Regular Expressions.

For example, the phone number like 615-525-3139, which has 3 digits – 3 digits – 4 digits. Email has very specific patterns like: peterzhangon@gmail.com, where @is mandatory.

The main functions and Examples

Python re module provides a lot of functions for the users to search a string given a pattern.

Function	Description
findall	Returns a list containing all matches
search	Returns a Match object if there is a match anywhere in the string
split	Returns a list where the string has been split at each match
sub	Replaces one or many matches with a string

The typical usage

1. **Return a list containing every occurrence of "an":**

```
import re

txt = "Natural language is designed to make human communication efficient."
x = re.findall("an", txt)
print(x)
```

```
['an', 'an']
```

2. **A Match object can bring back a lot of information.**

```
import re

txt = "A dog is chasing a boy on the playground"
x = re.search("dog", txt)

print("The first 'dog' is located in position:", x.start())
```

```
The first 'dog' is located in position: 2
```

3. **Split the string at every white-space character:**

```
import re

txt = "A dog is chasing a boy on the playground"
x = re.split("\s", txt)
print(x)
```

```
['A', 'dog', 'is', 'chasing', 'a', 'boy', 'on', 'the', 'playground']
```

4. **To replace the matches with given text**

```
import re

txt = "A dog is chasing a boy on the playground"
x = re.sub("dog", "cat", txt)
print(x)
```

```
A cat is chasing a boy on the playground
```

5. **Find a phone number given a string**

```
import  re

text = "789 615-525-3139  123"
ret = re.findall(r"[\d]{3}-[\d]{3}-[\d]{3}", text)
print ( ret )
```

```
['615-525-313']
```

6. **Find emails**

```
import  re

text = "If you have any quesiton, please contact peterzhangon@gmail.com or aijun2@illinois.edu"


emails = re.findall(r"[a-z0-9\.\-+_]+@[a-z0-9\.\-+_]+\.[a-z]+", text)
print(emails)
```

```
['peterzhangon@gmail.com', 'aijun2@illinois.edu']
```

Summary

In this paper, I showed you some basics of Python Regex, which is used to process texts to get what you wanted.

Hope it will save you a lot of time.

References

https://www.w3schools.com/python/python_regex.asp

https://www.edureka.co/blog/python-regex/

https://towardsdatascience.com/web-scraping-basics-82f8b5acd45c

https://ozenero.com/python-regular-expression-to-extract-phone-number-example