Introduction to Regex and Web Scrapping

01 November 2023 08:32 PM

Regex or regex expression or in layman terms often called as a pattern is used when you want to extract a common pattern of data from a very large heap of organized/unorganized data.

Eg Data → { 1, doc 1, 2, doc 2, doc 3, doc 4, 2, 3, 5, 7 - - - }

Now the task is to find all correct document names with the correct number associated with them

→ If we observe carefully the common pattern for a correct document name: → "doc"
→ <number>

So the common pattern for the document name will be → doc <number>

※ In order to apply regex, the input must be a string.

⇒ Because of the above property, regex is widely used in NLP (A prime component of ML and DL)

At the very core regex is basically helping you verify the structure of a string by matching it with a pattern.

$ Real time use case of Regex

(1) Checking if input follows a pattern, like is the email correct
↳ abc@gmail.com ✓
↳ abc#.in ✗

(2) In real time data it often happens that 2 different words which mean the same thing need to be captured.

Eg iphone, Apple phone → Both refer to the phone only

(3) Extracting specific portion of a text. Eg a text column containing the address is given, and we want to extract postal code.

Mr John Smith. 132, My Street, Kingston, New York 12401
Name | Street no. | Area | City | Pincode

The above break down can be done using Regex.

(4) Retrive/replace/structure
↳ find particular word/structure or replace a word detected by pattern.

(5) Splitting a text based on the occurrence of a specific character.

$ Creating Regex Patterns in Python

Package → import re

→ Understanding patterns
- Literal
  - ↳ When you give exactly what you want in your pattern
    - ↳ Text → mabnabc ab
      Pattern → ab
  - ↳ It matches literally, i.e. case sensitive
- Metachar
  - ↳ special characters with a general representative meaning
  - ↳ These characters can be combined together to create complex patterns to extract specific text.

Depending upon the requirement literal & metachar can also be combined together to create a pattern

Eg vai\wt