**Regular Expressions**

**Question 1-** Write a Python program to replace all occurrences of a space, comma, or dot with a colon.

**Sample Text-** 'Python Exercises, PHP exercises.'

**Expected Output:** Python:Exercises::PHP:exercises:

**SOL :**

import re

print(re.sub("[ ,.]",":","Python Exercises, PHP exercises."))

**Question 2-** Create a dataframe using the dictionary below and remove everything (commas (,), !, XXXX, ;, etc.) from the columns except words.

**Dictionary-** {'SUMMARY' : ['hello, world!', 'XXXXX test', '123four, five:; six...']}

**Expected output-**

0 hello world

1 test

2 four five six

**SOL:**

import re

import pandas as pd

dictionary = {'SUMMARY':['hello, world!','XXXXX test','123four,five:; six...']}

df=pd.DataFrame(dictionary)

pat = "[\d\W'XXXXX']"

for col in df.columns:

df[col]=df[col].str.replace(pat," ",regex=True)

print(df)

**Question 3-** Create a function in python to find all words that are at least 4 characters long in a string. The use of the re.compile() method is mandatory.

**SOL:**

import re

def fourchar(text):

pat = re.compile(r"\b\w{4}\b")

return pat.findall(text)

text="The best way to learn is by doing yourself"

foundchar = fourchar(text)

print(foundchar)

**Question 4-** Create a function in python to find all three, four, and five character words in a string. The use of the re.compile() method is mandatory.

**SOL:**

import re

def search\_word(text):

nchar\_word=re.compile(r"\b\w{3,5}\b")

word=re.findall(nchar\_word,text)

return word

text = "This is a sample text for testing, to check words with three to five characters"

print(search\_word(text))

**Question 5-** Create a function in Python to remove the parenthesis in a list of strings. The use of the re.compile() method is mandatory.

**Sample Text:** ["example (.com)", "hr@fliprobo (.com)", "github (.com)", "Hello (Data Science World)", "Data (Scientist)"]

**Expected Output:**

example.com

hr@fliprobo.com

github.com

Hello Data Science World

Data Scientist

**SOL:**

import re

def remove\_par():

sample = ["example (.com)", "hr@fliprobo (.com)", "github (.com)", "Hello (Data Science World)", "Data (Scientist)"]

rspace1=re.compile(r"[\(\)]")

rspace2=re.compile(r" \.")

for item in sample:

test =re.sub(rspace1,"",item)

test1=re.sub(rspace2,".",test)

print(test1)

remove\_par()

**Question 6-** Write a python program to remove the parenthesis area from the text stored in the text file using Regular Expression.

**Sample Text:** ["example (.com)", "hr@fliprobo (.com)", "github (.com)", "Hello (Data Science World)", "Data (Scientist)"]

**Expected Output:** ["example", "hr@fliprobo", "github", "Hello", "Data"]

**Note-** Store given sample text in the text file and then to remove the parenthesis area from the text.

**SOL:**

import re

sampletext= '"example (.com)", "hr@fliprobo (.com)", "github (.com)", "Hello (Data Science World)", "Data (Scientist)"'

def writefile(strsample):

with open("data.txt","w") as wr:

wr.write(strsample)

writefile(sampletext)

with open("data.txt","r") as d:

text = d.read()

print(text)

modtext = re.sub(r"\([^)]+\)","",text)

with open("data.txt","w") as wr:

wr.write("\n")

wr.write(modtext)

print(modtext)

**Question 7-** Write a regular expression in Python to split a string into uppercase letters.

**Sample text:** “ImportanceOfRegularExpressionsInPython”

**Expected Output:** [‘Importance’, ‘Of’, ‘Regular’, ‘Expression’, ‘In’, ‘Python’]

**SOL:**

import re

samplstr= "ImportanceOfRegularExpressionsInPython"

initsplit = re.split(r"(?=[A-Z])",samplstr)

print(initsplit)

**Question 8-** Create a function in python to insert spaces between words starting with numbers.

Sample Text: “RegularExpression1IsAn2ImportantTopic3InPython"

Expected Output: RegularExpression 1IsAn 2ImportantTopic 3InPython

**SOL:**

import re

sampletext = "RegularExpression1IsAn2ImportantTopic3InPython"

def numsplit(string):

nsplit=re.sub(r"\d"," ",sampletext)

return nsplit

print(numsplit(sampletext))

**Question 9-** Create a function in python to insert spaces between words starting with capital letters or with numbers.

**Sample Text:** “RegularExpression1IsAn2ImportantTopic3InPython"

**Expected Output:** RegularExpression 1 IsAn 2 ImportantTopic 3 InPython

**SOL:**

import re

sampletext = "RegularExpression1IsAn2ImportantTopic3InPython"

def numsplit(string):

nsplit=re.sub(r"(?=[A-Z|0-9])"," ",sampletext)

return nsplit

print(numsplit(sampletext))

**Question 10-** Use the github link below to read the data and create a dataframe. After creating the dataframe extract the first 6 letters of each country and store in the dataframe under a new column called first\_five\_letters.

**Github Link-**  <https://raw.githubusercontent.com/dsrscientist/DSData/master/happiness_score_dataset.csv>

**SOL:**

import pandas as pd

df=pd.read\_csv('happiness\_score\_dataset.csv')

df1=pd.DataFrame(data=df,columns=['Country'])

df2 = df1['Country'].str[:6]

df2.to\_csv('data\_file.csv')

**Question 11-** Write a Python program to match a string that contains only upper and lowercase letters, numbers, and underscores.

**SOL:**

import re

def validstr(string):

pats= r"[A-Za-z0-9\_]+$"

validstr = re.match(pats,enterstr)

return validstr

enterstr = input("Please enter a sentence :")

if validstr(enterstr):

print("Valid:Matched the pattern")

else:

print(f"invalid:Does not matched the pattern.Pattern is upper and lowercase letters, numbers, and underscores ")

**Question 12-** Write a Python program where a string will start with a specific number.

**SOL:**

import re

pats=r"123"

strs = input("Please enter a sentence :")

checks = re.match(pats,strs)

if checks:

print(f"Congrats, pattern matches")

else:

print(f"Sorry, pattern does not match")

**Question 13-** Write a Python program to remove leading zeros from an IP address

Still working in this.

**Question 14-** Write a regular expression in python to match a date string in the form of Month name followed by day number and year stored in a text file.

**Sample text :**  ' On August 15th 1947 that India was declared independent from British colonialism, and the reins of control were handed over to the leaders of the Country’.

**Expected Output-** August 15th 1947

**Note-** Store given sample text in the text file and then extract the date string asked format.

**SOL:**

import re

strdate=[]

patdate = r"(January|Feburary|March|April|May|June|July|August|September|October|November|December)\s+(\d+)(?:st|nd|rd|th)\s+(\d{4})"

with open("stringdate.txt","r")as r:

datestr = r.read()

searchdate = re.findall(patdate,datestr)

for item in searchdate:

joinstr="".join(str(item))

covstr=re.sub(r"[\(\'\),]","",joinstr)

with open("stringdate.txt","a") as r:

r.write("\n")

r.write(covstr)

print(covstr)

**Question 15-** Write a Python program to search some literals strings in a string.

**Sample text :** 'The quick brown fox jumps over the lazy dog.'

**Searched words :** 'fox', 'dog', 'horse'

**SOL:**

def searchlit(text, search\_words):

fowords = [word for word in swords if word in text]

return fowords

sampletext = 'The quick brown fox jumps over the lazy dog.'

swords = ['fox', 'dog', 'horse']

result = searchlit(sampletext, swords)

print(f"Sample text: '{sampletext}'")

print(f"Searched words: {swords}")

print(f"Found words: {result}")

**Question 16-** Write a Python program to search a literals string in a string and also find the location within the original string where the pattern occurs

**Sample text :** 'The quick brown fox jumps over the lazy dog.'

**Searched words :** 'fox'

**SOL:**

def searchloc(stext, sword):

loc = stext.find(sword)

if loc != -1:

return f"'{sword}' found at index {loc} in the text."

else:

return f"'{sword}' not found in the text."

stext = 'The quick brown fox jumps over the lazy dog.'

sword = 'fox'

result = searchloc(stext, sword)

print(f"Sample text: '{stext}'")

print(f"Searched word: '{sword}'")

print(result)

**Question 17-** Write a Python program to find the substrings within a string.

**Sample text :** 'Python exercises, PHP exercises, C# exercises'

**Pattern :** 'exercises'.

**SOL:**

sampletext="Python exercises, PHP exercises, C# exercises"

pattern = "exercises"

patlen=len(pattern)

start=0

textlen=len(sampletext)

while start < textlen:

textindex = sampletext.find(pattern,start)

start = textindex

if start == -1:

print("End of search")

break

start=start + 1

patlen=patlen + textindex

print(f"Pattern found index {textindex} and word is {sampletext[textindex:patlen]}")

patlen=len(pattern)

**Question 18-** Write a Python program to find the occurrence and position of the substrings within a string.

**SOL:**

sampletext="Python exercises, PHP exercises, C# exercises"

pattern = "exercises"

patlen=len(pattern)

start=0

patcount=0

textlen=len(sampletext)

while start < textlen:

textindex = sampletext.find(pattern,start)

start = textindex

if start == -1:

print(f"No of occurrence = {patcount}")

print("End of search")

break

start=start + 1

patlen=patlen + textindex

print(f"Pattern found index {textindex} and word is {sampletext[textindex:patlen]}")

patcount+=1

patlen=len(pattern)

**Question 19-** Write a Python program to convert a date of yyyy-mm-dd format to dd-mm-yyyy format.

**SOL:**

from datetime import datetime

sampledate= input(f"Please enter date in format YYYY-MM-DD : ")

oldate=datetime.strptime(sampledate,'%Y-%m-%d')

cfdate = oldate.strftime('%d-%m-%Y')

print(f"Updated date format {cfdate}")

**Question 20-** Create a function in python to find all decimal numbers with a precision of 1 or 2 in a string. The use of the re.compile() method is mandatory.

**Sample Text:** "01.12 0132.123 2.31875 145.8 3.01 27.25 0.25"

**Expected Output:** ['01.12', '145.8', '3.01', '27.25', '0.25']

**SOL:**

import re

sampletext= "01.12 0132.123 2.31875 145.8 3.01 27.25 0.25"

pattern=re.compile(r"\b\d+\.\d{1,2}\b")

modstr=re.findall(pattern,sampletext)

print(modstr)

**Question 21-** Write a Python program to separate and print the numbers and their position of a given string.

**SOL:**

inputstr=input(f"Please enter string with char and number : ")

number=[]

numdex=[]

for index,char in enumerate(inputstr):

if char.isdigit():

number.append(int(char))

numdex.append(index+1)

print(f"NUMBER {number}")

print(f"POSITION {numdex}")

**Question 22-** Write a regular expression in python program to extract maximum/largest numeric value from a string.

**Sample Text:** 'My marks in each semester are: 947, 896, 926, 524, 734, 950, 642'

**Expected Output:** 950

**SOL:**

import re

sampletext = "My marks in each semester are:947, 896, 926, 524, 734, 950, 642"

pattern = re.compile(r"\b\d+\b")

def maxmarks(sampletext,pattern):

numfind=re.findall(pattern,sampletext)

return max(map(int,numfind))

print(f"My highest marks : {maxmarks(sampletext,pattern)}")

**Question 23-** Create a function in python to insert spaces between words starting with capital letters.

**Sample Text:** “RegularExpressionIsAnImportantTopicInPython"

**Expected Output:** Regular Expression Is An Important Topic In Python

**SOL:**

import re

sampletext = "RegularExpressionIsAnImportantTopicInPython"

def numsplit(string):

nsplit=re.sub(r"(?=[A-Z])"," ",sampletext)

return nsplit

print(numsplit(sampletext))

**Question 24-** Python regex to find sequences of one upper case letter followed by lower case letters

**SOL:**

import re

sampletext="Da bbb BAD ADa Cm Ca Dog Man Ate"

pattern = re.compile(r"\b[A-Z][a-z]\*\b")

findsq=re.findall(pattern,sampletext)

print(findsq)

**Question 25-** Write a Python program to remove continuous duplicate words from Sentence using Regular Expression.

**Sample Text:** "Hello hello world world"

**Expected Output:** Hello hello world

**SOL:**

import re

sampletext= "Hello hello world world"

pattern = re.compile(r"\b(\w+)( \1)+\b")

output = pattern.sub(r'\1',sampletext)

print(output)

**Question 26-** Write a python program using RegEx to accept string ending with alphanumeric character.

**SOL:**

import re

sampletext = "how are you doing 123sample123"

pattern = re.compile(r"^.\*[a-zA-Z0-9]$")

output = re.match(pattern,sampletext)

if output:

print(f"Valid String : The string is ending with alphanumberic character")

else:

print(f"Invalid string : It is not ending with alphanumeric character")

**Question 27-**Write a python program using RegEx to extract the hashtags.

**Sample Text:**  """RT @kapil\_kausik: #Doltiwal I mean #xyzabc is "hurt" by #Demonetization as the same has rendered USELESS <ed><U+00A0><U+00BD><ed><U+00B1><U+0089> "acquired funds" No wo"""

**Expected Output:** ['#Doltiwal', '#xyzabc', '#Demonetization']

**SOL:**

import re

sampletext = """RT @kapil\_kausik: #Doltiwal I mean #xyzabc is "hurt" by #Demonetization as the same has rendered USELESS <ed><U+00A0><U+00BD><ed><U+00B1><U+0089> "acquired funds" No wo"""

pattern = re.compile(r"\s#[\w]+")

output = re.findall(pattern,sampletext)

print(output)

**Question 28-** Write a python program using RegEx to remove <U+..> like symbols

Check the below sample text, there are strange symbols something of the sort <U+..> all over the place. You need to come up with a general Regex expression that will cover all such symbols.

**Sample Text:** "@Jags123456 Bharat band on 28??<ed><U+00A0><U+00BD><ed><U+00B8><U+0082>Those who are protesting #demonetization are all different party leaders"

**Expected Output:** @Jags123456 Bharat band on 28??<ed><ed>Those who are protesting #demonetization are all different party leaders

**SOL:**

import re

sampletext = "@Jags123456 Bharat band on 28??<ed><U+00A0><U+00BD><ed><U+00B8><U+0082>Those who are protesting #demonetization are all different party leaders"

pattern = re.compile(r"<U\+\w\d+\w+>")

output = pattern.sub("",sampletext)

print(output)

**Question 29-** Write a python program to extract dates from the text stored in the text file.

**Sample Text:** Ron was born on 12-09-1992 and he was admitted to school 15-12-1999.

**Note-** Store this sample text in the file and then extract dates.

SOL:

import re

pattern = re.compile(r"\d+\-\d+-\d+")

strdate=""

with open("Q29.txt","r") as rd:

sampletext = rd.read()

exdate=re.findall(pattern,sampletext)

print(sampletext)

with open("Q29.txt","a") as ap:

ap.write("\n")

for item in exdate:

strdate="".join(item)

ap.write(f"{strdate} ")

print(strdate)

**Question 30-** Create a function in python to remove all words from a string of length between 2 and 4.

The use of the re.compile() method is mandatory.

**Sample Text:** "The following example creates an ArrayList with a capacity of 50 elements. 4 elements are then added to the ArrayList and the ArrayList is trimmed accordingly."

**Expected Output:** following example creates ArrayList a capacity elements. 4 elements added ArrayList ArrayList trimmed accordingly.

**SOL:**

import re

pattern = re.compile(r"\b\w{2,4}\b\s+")

sampletext = "The following example creates an ArrayList with a capacity of 50 elements. 4 elements are then added to the ArrayList and the ArrayList is trimmed accordingly."

modstr=re.sub(pattern,"",sampletext)

print(modstr)