Regular Expressions In [1]: **import** pandas **as** pd import regex as re # 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.' repl=":" Sample_Text='Python Exercises, PHP exercises.' $x = re.sub("\s|\,|\.",repl, Sample_Text)$ print(x) # Expected Output: Python:Exercises::PHP:exercises: Python:Exercises::PHP:exercises: In [3]: # 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...']} data = {'SUMMARY': ['hello, world!', 'XXXXX test', '123four, five:; six...']} df = pd.DataFrame(data) df['SUMMARY'].str.replace('[A-Z0-9\W]',r' ', regex=True) # Expected outputhello world # 0 # 1 test # 2 four five six hello world Out[3]: four five six Name: SUMMARY, dtype: object In [14]: # 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. sample_text = "On January 15th, 2023, we celebrated a milestone with a record-breaking achievement. The team achieved a success rate of 95%, surpassing our target of 90%. This accordance is a success of the same achieved as success of the same achieved achieved as success of the same achieved achieved achieved as success of the same achieved achieved achieved achieved achieved ac def at_least_4_characters(string): pattern=r"\b\w{4,}\b" #at least 4 rec_pattern=re.compile(pattern) match=rec_pattern.findall(sample_text) return match matches_=at_least_4_characters(sample_text) print(matches_) ['January', '15th', '2023', 'celebrated', 'milestone', 'with', 'record', 'breaking', 'achievement', 'team', 'achieved', 'success', 'rate', 'surpassing', 'target', 'This', 'accompli shment', 'positive', 'tone', 'upcoming', 'year', 'Looking', 'forward', 'more', 'achievements', 'months', 'ahead'] In [19]: # Question 3-Sample_text = "On January 15th, 2023, we celebrated a milestone with a record-breaking achievement. The team achieved a success rate of 95%, surpassing our target of 90%. This acc def words(string): pattern = re.compile($r'\b\w{4,}\b'$) matches = pattern.findall(string) **return** matches result = words(Sample_text) print(result) ['January', '15th', '2023', 'celebrated', 'milestone', 'with', 'record', 'breaking', 'achievement', 'team', 'achieved', 'success', 'rate', 'surpassing', 'target', 'This', 'accompli shment', 'positive', 'tone', 'upcoming', 'year', 'Looking', 'forward', 'more', 'achievements', 'months', 'ahead'] # 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. str1="On January 15th, 2023, we celebrated a milestone with a record-breaking achievement. The team achieved a success rate of 95%, surpassing our target of 90%. This accomplishment string_pattern = $r'' b w{3,5} b''$ regex_pattern=re.compile(string_pattern) result_1=regex_pattern.findall(str1) print(result_1, "\n") # or Sample_text = "On January 15ths, 2023, we celebrated a milestone with a record-breaking achievement. The teams achieved a success rate of 95%, surpassing our target of 90%. This ac def words(string): pattern = re.compile($r'\b\w{3,5}\b'$) matches = pattern.findall(string) return matches result_2= words(Sample_text) print(result_2) ['15th', '2023', 'with', 'The', 'team', 'rate', 'our', 'This', 'has', 'set', 'tone', 'for', 'the', 'year', 'more', 'the', 'ahead'] ['15ths', '2023', 'with', 'The', 'teams', 'rate', 'our', 'This', 'has', 'set', 'tone', 'for', 'the', 'year', 'more', 'the', 'ahead'] In [523...] # 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)"] Sample_Text=["example(.com)", "hr@fliprobo(.com)", "github(.com)", "Hello(Data Science World)", "Data (Scientist)"] for i in Sample_Text: pattern= re.sub(r"\(|\)",'',i) print(pattern) # Expected Output: # example.com # hr@fliprobo.com # github.com # Hello Data Science World # Data Scientist example.com hr@fliprobo.com github.com Hello Data Science World Data Scientist In [518... | # 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)"] with open("sample_text.txt", "w") as file: file.write("[\"example (.com)\", \"hr@fliprobo (.com)\", \"github (.com)\", \"Hello (Data Science World)\", \"Data (Scientist)\"]") with open("sample_text.txt", "r") as file: file_text = file.read() pattern=r'\([^()]*\)' match = re.sub(pattern, '', file_text) print(match) # 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. ["example ", "hr@fliprobo ", "github ", "Hello ", "Data "] In [34]: # Question 7- Write a regular expression in Python to split a string into uppercase letters. sample_text = "ImportanceOfRegularExpressionsInPython" x= re.findall(r'[A-Z][a-z]+', sample_text) # x= re.split(r"(?=[A-Z])", sample_text) print(x) # Expected Output: ['Importance', 'Of', 'Regular', 'Expression', 'In', 'Python'] ['Importance', 'Of', 'Regular', 'Expressions', 'In', 'Python'] In [49]: # Question 8- Create a function in python to insert spaces between words starting with numbers. Sample_Text="RegularExpression1IsAn2ImportantTopic3InPython" $z=re.sub(r'([A-Z\d])([A-Z])', r' \1\2', Sample_Text)$ # OR # z=re.sub(r'(\d)([A-Za-z])', r' \1\2', Sample_Text) # OR $\# z=re.sub(r'(?<=\D)(\d)', r' \1', Sample_Text)$ print(z) # Expected Output: RegularExpression 1IsAn 2ImportantTopic 3InPython RegularExpression 1IsAn 2ImportantTopic 3InPython In [51]: # Question 9- Create a function in python to insert spaces between words starting with capital letters or with numbers. Sample_Text="RegularExpression1IsAn2ImportantTopic3InPython" $z=re.sub(r'([A-Z0-9])', r' \1', Sample_Text)$ # OR # z=re.sub(r'(\d)([A-Za-z])', r' \1 \2', Sample_Text) print(z) # Expected Output: RegularExpression 1 IsAn 2 ImportantTopic 3 InPython Regular Expression 1 Is An 2 Important Topic 3 In Python In [52]: # 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 df=pd.read_csv(r'https://raw.githubusercontent.com/dsrscientist/DSData/master/happiness_score_dataset.csv') pattern= $r'(\w{6})'$ mask = df["Country"].str.extract(pattern).rename(columns={0: 'first_five_letters'}) mask first_five_letters Out[52]: 0 Switze Icelan 2 Denmar 3 Norway 4 Canada 153 Rwanda 154 NaN 155 NaN 156 Burund 157 NaN 158 rows × 1 columns In [63]: # Question 11- Write a Python program to match a string that # contains only upper and lowercase letters, numbers, and underscores. target_string= "Abdul_Kalam123 was an Indian@123 aerospace scientist also known as the Missile Manof India, Valid_String123" pattern= r'[a-zA-Z0-9_]+' result = re.findall(pattern, target_string) print ("Match_object:", result) Match_object: ['Abdul_Kalam123', 'was', 'an', 'Indian', '123', 'aerospace', 'scientist', 'also', 'known', 'as', 'the', 'Missile', 'Manof', 'India', 'Valid_String123'] # Question 12- Write a Python program where a string will start with a specific number. target_string= "123Abdul Kalam123 was an Indian123 aerospace scientist also known as the Missile Manof India, Valid_String123" pattern= $r'^123[w+]+[a-zA-Z0-9]+'$ result = re.findall(pattern, target_string) print ("Match_object:", result) Match_object: ['123Abdul'] In [74]: # Question 13- Write a Python program to remove leading zeros from an IP address ip_address = "0192.012.045.007" $pattern = (r'\b0+(\d+)\b')$ cleaned_address = re.sub(pattern,r'\1', ip_address) print(cleaned_address) 192.12.45.7 In [75]: # 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. string="Virat is a cricket player , He was born on November 5th, 1988 , He is the former captain of the Indian national cricket team" pattern = r'\b(?:January|February|March|April|May|June|July|August|September|October|November|December)\s+\d{1,2}(?:st|nd|rd|th)?,\s+\d{4}\b' # matches = re.findall(pattern, string) import re # 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." # pattern = $r'' b([A-Z][a-z]+) d\{1,2\}(?:st|nd|rd|th)? d\{4\}b''$ matches = re.findall(pattern, string) print(matches) ['November 5th, 1988'] In []: # \b: Denotes a word boundary to ensure that the pattern is not part of a larger word. # (?:January|February|March|April|May|June|July|August|September|October|November|December): Non-capturing group for month names. #\s+: Matches one or more whitespace characters (space). # $\{1,2\}$: Matches one or two digits for the day number. # ,: Matches the comma between the day number and the year. # \s+: Matches one or more whitespace characters. # $\d{4}$: Matches exactly four digits for the year. In [538... | # 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'. import re with open('sample_text.txt', 'w') as file:#"w"= write file.write('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.') with open('sample_text.txt', 'r') as file:#"r"=read text_data = file.read() $\# pattern = r' b([A-Z][a-Z]+) s+ d\{1,2\}(?:st|nd|rd|th)? s+ d\{4\} b'$ pattern = r'\b(?:January|February|March|April|May|June|July|August|September|October|November|December)\s+\d{1,2}(?:st|nd|rd|th)?\s+\d{4}\b' dates = re.findall(pattern, text_data) print(dates) # Expected Output- August 15th 1947 # Note- Store given sample text in the text file and then extract the date string asked format. ['August 15th 1947'] In [250... # Question 15- Write a Python program to search some literals strings in a string. pattern= "fox|dog|horse" Sample_text='The quick brown fox jumps over the lazy dog and horse.' matches = re.findall(pattern, Sample_text) print(matches) # Searched words : 'fox', 'dog', 'horse' ['fox', 'dog', 'horse'] In [261... | # 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.' search= re.search('fox', Sample_text) print(search) # Searched words : 'fox'(16th) place <re.Match object; span=(16, 19), match='fox'> In [277... # Question 17- Write a Python program to find the substrings within a string. x='Python exercises', 'PHP exercises', 'C# exercises' pattern = 'exercises' **for** match **in** x: search= re.findall(pattern, match) search= re.search(pattern, match) for loctaion print(search) # Pattern : 'exercises'. ['exercises'] ['exercises'] ['exercises'] In [302... # Question 18- Write a Python program to find the occurrence and position of the substrings within a string. x='Python exercises', 'PHP exercises', 'C# exercises' pattern = 'exercises' **for** match **in** x: search= re.search(pattern,match) print(search) <re.Match object; span=(7, 16), match='exercises'> <re.Match object; span=(4, 13), match='exercises'> <re.Match object; span=(3, 12), match='exercises'> In [293... | # Question 19- Write a Python program to convert a date of yyyy-mm-dd format to dd-mm-yyyy format. input_date="2024-01-01" pattern= $r'(\d{4})-(\d{2})-(\d{2})'$ modified_date=re.sub(pattern, r'\3-\2-\1',input_date) print(input_date) print(modified_date) 2024-01-01 01-01-2024 In [299... # 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" # pattern=r'\d+\.\d{1,2}' pattern= $r'\b\d+\.\d\{1,2\}\b'$ rec_pattern=re.compile(pattern) dec_no_pre=rec_pattern.findall(Sample_Text) print(dec_no_pre) # Expected Output: ['01.12', '145.8', '3.01', '27.25', '0.25'] ['01.12', '0132.12', '2.31', '145.8', '3.01', '27.25', '0.25'] In [323... # Question 21- Write a Python program to separate and print the numbers and their position of a given string. x="The quick 455 brown fox jumps over the lazy dog." pattern = r' d+'search= re.search(pattern,x) print(search, "\n") # or x='i ate 45 banans','I have done 65 pushups ', 'It has 34inbuilt camera' pattern = r' d+'**for** match **in** x: search= re.search(pattern, match) print(search) <re.Match object; span=(10, 13), match='455'> <re.Match object; span=(6, 8), match='45'> <re.Match object; span=(12, 14), match='65'> <re.Match object; span=(7, 9), match='34'> In [338... # 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' pattern= r'\d+' match=re.findall(pattern, Sample_Text) print(max(match)) # Expected Output: 950 950 In [373... # Question 23- Create a function in python to insert spaces between words starting with capital letters. Sample_Text="RegularExpressionIsAnImportantTopicInPython" pattern=r'([a-z])([A-Z])' match=re.sub(pattern, r'\1 \2', Sample_Text) print(match) # Expected Output: Regular Expression Is An Important Topic In Python Regular Expression Is An Important Topic In Python # Question 24- Python regex to find sequences of one upper case letter followed by lower case letters Sample_Text='Create a function in Python to insert spaces Between words starting With capital letters.' pattern=r'([A-Z][a-z]+)'match=re.findall(pattern, Sample_Text) print(match) ['Create', 'Python', 'Between', 'With'] In [385... # Question 25- Write a Python program to remove continuous duplicate words from Sentence using Regular Expression. Sample_Text="Hello hello world world" pattern = $r'\b(\w+)(?:\W+\1\b)+'$ x=re.sub(pattern, r'\1', Sample_Text) print(x) # Expected Output: Hello hello world Hello hello world In [91]: # Question 26- Write a python python using RegEx to accept string ending with alphanumeric character. sample_text1= " python326, python738, javed9876, 123javed" pattern= $r'[\w$]+'$ # pattern=r'[a-zA-z0-9]+' match=re.findall(pattern, sample_text1) print(match) ['python326', 'python738', 'javed9876', '123javed'] In [417... # 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+00B9> "acquired funds" No pattern=r"#\w+" match=re.findall(pattern, Sample_Text) print(match) # Expected Output: ['#Doltiwal', '#xyzabc', '#Demonetization'] ['#Doltiwal', '#xyzabc', '#Demonetization'] # 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 Sample_Text="@Jags123456 Bharat band on 28??<ed><U+00A0><U+00BD><ed><U+00BB><U+00B8><U+00B2>Those who are protesting #demonetization are all different party leaders." # pattern=r'<[a-zA-Z0-9-\+]+>' # or # pattern=r'<U\+\w+>' # or pattern = $r''<U'+\w{4}>"$ match=re.sub(pattern,r' ',Sample_Text) print(match) # Expected Output: @Jags123456 Bharat band on 28??<ed><ed>Those who are protesting #demonetization are all different party leaders @Jags123456 Bharat band on 28??<ed> <ed> Those who are protesting #demonetization are all different party leaders. In [439... # 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.' import re with open('Sample_Text.txt', 'w') as file:#"w"= write file.write('Ron was born on 12-09-1992 and he was admitted to school 15-12-1999.') with open('Sample_Text.txt', 'r') as file:#"r"=read text_data = file.read() # print(text_data) pattern= $r'\b\d{2}-\d{2}-\d{4}\b'$ match=re.findall(pattern,text_data) print(match) # Note- Store this sample text in the file and then extract dates. ['12-09-1992', '15-12-1999'] In [449... # 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." pattern= $r'\b\w\{2,4\}\b'$ rec_pattern=re.compile(pattern) match=rec_pattern.sub('', Sample_Text) print(match) # Expected Output: following example creates ArrayList a capacity elements. 4 elements added ArrayList ArrayList trimmed accordingly. following example creates ArrayList a capacity elements. 4 elements added ArrayList ArrayList trimmed accordingly. In []: