Data_Preprocessing_NLP_Removing_Numbers

By Venkatesh_Mungi venkateshmungi1247@gmail.com

1.Remove Numbers from String using regex

Python provides a regex module that has a built-in function sub() to remove numbers from the string. This method replaces all the occurrences of the given pattern in the string with a replacement string. If the pattern is not found in the string, then it returns the same string.

In the below example, we take a pattern as r'[0-9]' and an empty string as a replacement string. This pattern matches with all the numbers in the given string and the sub() function replaces all the matched digits with an empty string. It then deletes all the matched numbers.

In [1]:
```
import re
string = "Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract or extrapolate knowledge and insights from noisy, s
pattern = r'[0-9]'
new_string = re.sub(pattern, '', string)
new_string
```

Out[1]:
```
'Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract or extrapolate knowledge and insights from noisy, structur
ed and unstructured data,[][] and apply knowledge from data across a broad range of application domains. Data science is related to data mining, machine learning, big data, comp
utational statistics and analytics.[]Data science is a concept to unify statistics, data analysis, informatics, and their related methods in order to understand and analyse actu
al phenomena with data.[] It uses techniques and theories drawn from many fields within the context of mathematics, statistics, computer science, information science, and domain
knowledge.[] However, data science is different from computer science and information science. Turing Award winner Jim Gray imagined data science as a fourth paradigm of science
(empirical, theoretical, computational, and now data-driven) and asserted that everything about science is changing because of the impact of information technology and the data
deluge.[][]A data scientist is someone who creates programming code and combines it with statistical knowledge to create insights from data.[]'
```

In [2]:
```
def remove_string(string):
    pattern = r'[0-9]'
    new_string = re.sub(pattern, '', string)
    return new_string
string = "Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract or extrapolate knowledge and insights from noisy, s
remove_string(string)
```

Out[2]:
```
'Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract or extrapolate knowledge and insights from noisy, structur
ed and unstructured data,[][] and apply knowledge from data across a broad range of application domains. Data science is related to data mining, machine learning, big data, comp
utational statistics and analytics.[]Data science is a concept to unify statistics, data analysis, informatics, and their related methods in order to understand and analyse actu
al phenomena with data.[] It uses techniques and theories drawn from many fields within the context of mathematics, statistics, computer science, information science, and domain
knowledge.[] However, data science is different from computer science and information science. Turing Award winner Jim Gray imagined data science as a fourth paradigm of science
(empirical, theoretical, computational, and now data-driven) and asserted that everything about science is changing because of the impact of information technology and the data
deluge.[][]A data scientist is someone who creates programming code and combines it with statistical knowledge to create insights from data.[]'
```

2.Remove Numbers from String using join() & isdigit()

This method uses isdigit() to check whether the element is a digit or not. It returns True if the element is a digit. This method uses for loop to iterate over each character in the string.

The below example skips all numbers from the string while iterating and joins all remaining characters to print a new string.

In [3]:
```
text = "Shivaji, also spelled Śivaji, (born February 19, 1630, or April 1627, Shivner, Poona [now Pune], India—died April 3, 1680, Rajgarh), founder of the Maratha kingdom of Indi
new_text = ''.join((x for x in text if not x.isdigit()))
new_text
```

Out[3]:
```
'Shivaji, also spelled Śivaji, (born February , , or April , Shivner, Poona [now Pune], India—died April , , Rajgarh), founder of the Maratha kingdom of India. The kingdoms secu
rity was based on religious toleration and on the functional integration of the Brahmans, Marathas, and Prabhus.'
```

In [4]:
```
def rem_str(text):
    new_text = ''.join((x for x in text if not x.isdigit()))
    return new_text

text = "Shivaji, also spelled Śivaji, (born February 19, 1630, or April 1627, Shivner, Poona [now Pune], India—died April 3, 1680, Rajgarh), founder of the Maratha kingdom of Indi
rem_str(text)
```

Out[4]:
```
'Shivaji, also spelled Śivaji, (born February , , or April , Shivner, Poona [now Pune], India—died April , , Rajgarh), founder of the Maratha kingdom of India. The kingdoms secu
rity was based on religious toleration and on the functional integration of the Brahmans, Marathas, and Prabhus.'
```

3.Remove Numbers from String using translate()

This method uses string Python library. With the help of a string object, maketrans() separates numbers from the given string. Afterward, a translation table is created where each digit character i.e. ‘0’ to ‘9’ will be mapped to None and this translation table is passed to translate() function.

The below example creates a translation table and replaces characters in string based on this table, so it will delete all numbers from the string

In [5]:
```
import string

sentence = "The history of India starts with the existence of India itself as It located in the continent of Asia, India covers 2,973,193 square kilometers of land and 314,070 squ

#digits are mapped to None
translation_table = str.maketrans('', '', string.digits)

#deletes all number
new_string = sentence.translate(translation_table)

new_string
```

Out[5]:
```
'The history of India starts with the existence of India itself as It located in the continent of Asia, India covers ,, square kilometers of land and , square kilometers of wate
r.\xa0Making it the th largest nation in the world with a total area of ,, square kilometers. Surrounded by Bhutan, Nepal, and Bangladesh to the North East, China to the North,
Pakistan to the North West, and Sri Lanka on the South East coast.'
```

In [6]:
```
def rm_st(sentence):
    translation_table = str.maketrans('', '', string.digits)
    new_string = sentence.translate(translation_table)
    return new_string

sentence = "The history of India starts with the existence of India itself as It located in the continent of Asia, India covers 2,973,193 square kilometers of land and 314,070 squ
rm_st(sentence)
```

Out[6]:
```
'The history of India starts with the existence of India itself as It located in the continent of Asia, India covers ,, square kilometers of land and , square kilometers of wate
r.\xa0Making it the th largest nation in the world with a total area of ,, square kilometers. Surrounded by Bhutan, Nepal, and Bangladesh to the North East, China to the North,
Pakistan to the North West, and Sri Lanka on the South East coast.'
```

4. Remove Numbers from String

This example uses the filter() and lambda in the generating expression. It filters or deletes all the numbers from the given string and joins the remaining characters of the string to create a new string.

In [7]:
```
strn = "The history of India starts with the existence of India itself as It located in the continent of Asia, India covers 2,973,193 square kilometers of land and 314,070 square
#Filters all digits
new_string = ''.join(filter(lambda x: not x.isdigit(), strn))
new_string
```

Out[7]:
```
'The history of India starts with the existence of India itself as It located in the continent of Asia, India covers ,, square kilometers of land and , square kilometers of wate
r.\xa0Making it the th largest nation in the world with a total area of ,, square kilometers. Surrounded by Bhutan, Nepal, and Bangladesh to the North East, China to the North,
Pakistan to the North West, and Sri Lanka on the South East coast.'
```

In [8]:
```
def rmv_str(strn):
    new_string = "".join(filter(lambda x: not x.isdigit(), strn))
    return new_string
strn = "The history of India starts with the existence of India itself as It located in the continent of Asia, India covers 2,973,193 square kilometers of land and 314,070 square
rmv_str(strn)
```

Out[8]:
```
'The history of India starts with the existence of India itself as It located in the continent of Asia, India covers ,, square kilometers of land and , square kilometers of wate
r.\xa0Making it the th largest nation in the world with a total area of ,, square kilometers. Surrounded by Bhutan, Nepal, and Bangladesh to the North East, China to the North,
Pakistan to the North West, and Sri Lanka on the South East coast.'
```

In [ ]: