

**DEPARTMENT OF COMPUTER ENGINEERING**

## CSL804 Computational Lab II

**Eighth Semester, 2021-2022 (Even Semester)**

**Name of Student :** Saurav Kumar

## Roll No. 23

**Division :** BE – CMPN

**Day/ Session :** Monday/Afternoon

**Venue :** SLRTCE Lab 305

## Experiment No. 1

**Title of Experiment :** To study and implement preprocessing of text (Tokenization, Filtration, Script Validation)

## Date of Conduction :

**Date of Submission :**

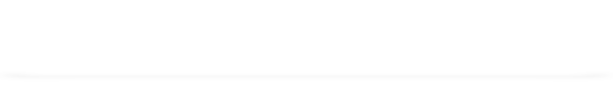
|  |  |  |
| --- | --- | --- |
| **Particulars Max. Marks Marks Obtained** | | |
| Preparedness and Efforts(PE) | **3** |  |
| Knowledge of tools(KT) | **3** |  |
| Debugging and results(DR) | **3** |  |
| Documentation(DN) | **3** |  |
| Punctuality & Lab Ethics(PL) | **3** |  |
| **Total** | **15** |  |

**Grades – Meet Expectations (3 Marks), Moderate Expectations (2 Marks), Below Expectations (1 Mark)**

**Checked and Verified by Name of Faculty :** Prof. Neelam Kulkarni

## Signature :

**Date :**

EXPERIMENT NO: 1

PREPROCESSING OF TEXT

**AIM:** To study and implement preprocessing of text (Tokenization, Filtration, Script Validation)

**SOFTWARE:** Python, NLTK, CLTK

# THEORY:

PREPROCESSING IN NLP

Natural Language Processing (NLP) is a branch of Data Science which deals with Text data. Apart from numerical data, Text data is available to a great extent which is used to analyze and solve business problems. But before using the data for analysis or prediction, processing the data is important.

To prepare the text data for the model building we perform text preprocessing. It is the very first step of NLP projects. Some of the preprocessing steps are:

* Tokenization
* To lowercase
* Remove numbers
* Replace numbers by corresponding number words
* Remove punctuation
* Remove whitespaces

# IMPLEMENTATION:

TOKENIZATION

Tokenization is the process of tokenizing or splitting a string, text into a list of tokens. One can think of token as parts like a word is a token in a sentence, and a sentence is a token in a paragraph.

CODE

import os import nltk import ssl

from nltk.tokenize import sent\_tokenize, word\_tokenize

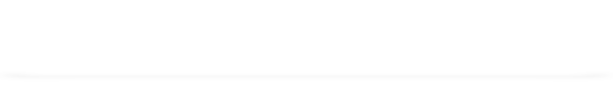
text="Assessment consists of two class tests of 20 marks each. The first class test is to be conducted when approx. 40 syllabus is completed and second class test when additional 40 syllabus is completed. Duration of each test shall be one hour."

print(sent\_tokenize(text)) print(word\_tokenize(text))

OUTPUT

Text

Description automatically generated



TO LOWERCASE

Convert the text to lowercase.

CODE

import nltk import string import re

def text\_lowercase(text): return text.lower()

input\_str = "Hey, did you know that the summer break is coming? Amazing right !! It's only 5 more days !!" print(text\_lowercase(input\_str))

OUTPUT

Text

Description automatically generated

REMOVE NUMBERS

Remove numbers from the text.

CODE

import nltk import string import re

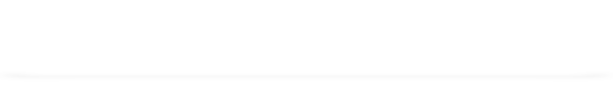
def remove\_numbers(text):

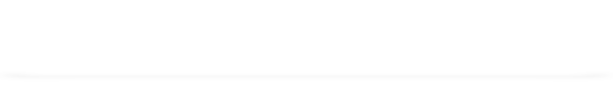
result = re.sub(r'\d+', '', text) return result

input\_str = "There are 3 balls in this bag, and 12 in the other one." print(remove\_numbers(input\_str))

OUTPUT





REPLACE NUMBERS BY CORRESPONDING NUMBER WORDS

Replace all the numbers in the text with their corresponding number words.

CODE

import nltk import inflect

p = inflect.engine()

# convert number into words def convert\_number(text):

# split string into list of words temp\_str = text.split()

# initialise empty list new\_string = []

for word in temp\_str:

# if word is a digit, convert the digit

# to numbers and append into the new\_string list if word.isdigit():

temp = p.number\_to\_words(word) new\_string.append(temp)

# append the word as it is else:

new\_string.append(word)

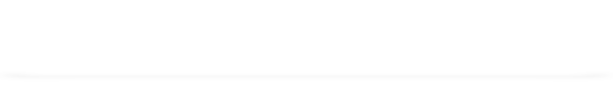
# join the words of new\_string to form a string temp\_str = ' '.join(new\_string)

return temp\_str

input\_str = 'There are 3 balls in this bag, and 12 in the other one.' print(convert\_number(input\_str))

OUTPUT





REMOVE PUNCTUATIONS

Replace all the punctuations in the text.

CODE

import nltk import string

def remove\_punctuation(text):

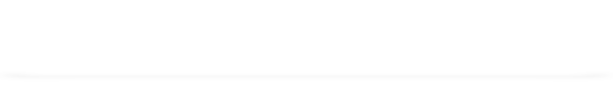
translator = str.maketrans('', '', string.punctuation) return text.translate(translator)

input\_str = "Hey, did you know that the summer break is coming? Amazing right !! It's only 5 more days !!" print(remove\_punctuation(input\_str))

OUTPUT

Text

Description automatically generated

REMOVE WHITESPACES

Replace all the whitespaces in the text.

CODE

import nltk

def remove\_whitespace(text): return " ".join(text.split())

input\_str = " we don't need the given questions" print(remove\_whitespace(input\_str))

OUTPUT



# CONCLUSION:

Thus we have studied and implemented preprocessing of text.