

LAB – II

Compiler Design

Syntax Analysis for Natural Language

(Basics of NLTK Package in Python)

MAR 2022 – JUNE 2023

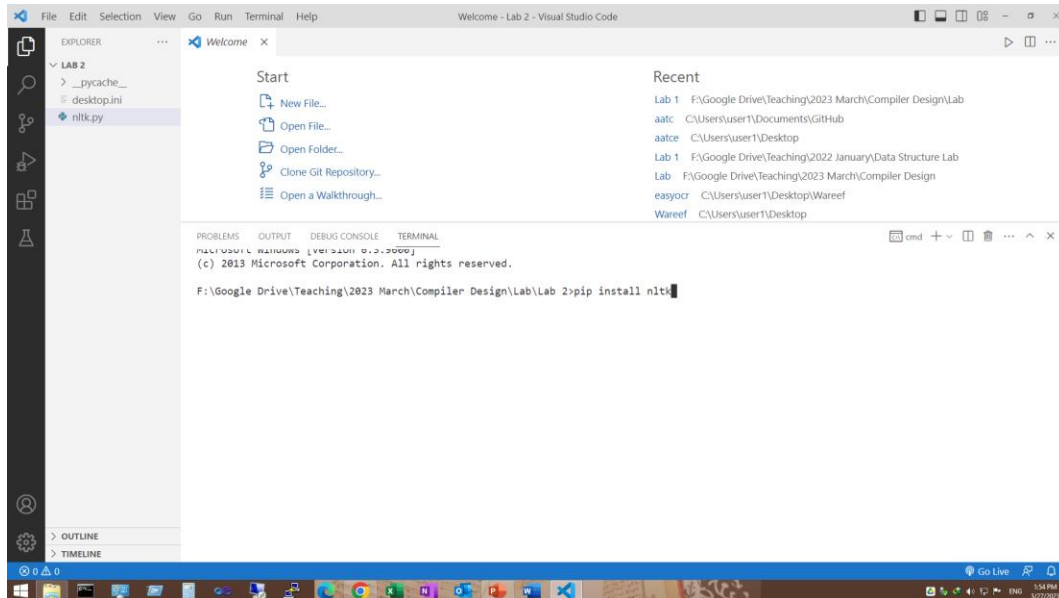
DEPARTMENT OF COMPUTER SCIENCE
COLLEGE OF COMPUTING AND INFORMATION TECHNOLOGY
SHAQRA UNIVERSITY
P.O. BOX 15572

Dr. Nayyar Ahmed Khan / Dr. Nouf Altamami

EMAIL:

nayyar@su.edu.sa ---- naltamami@su.edu.sa

Installation of NLTK Package



```
File Edit Selection View Go Run Terminal Help
Welcome - Lab 2 - Visual Studio Code

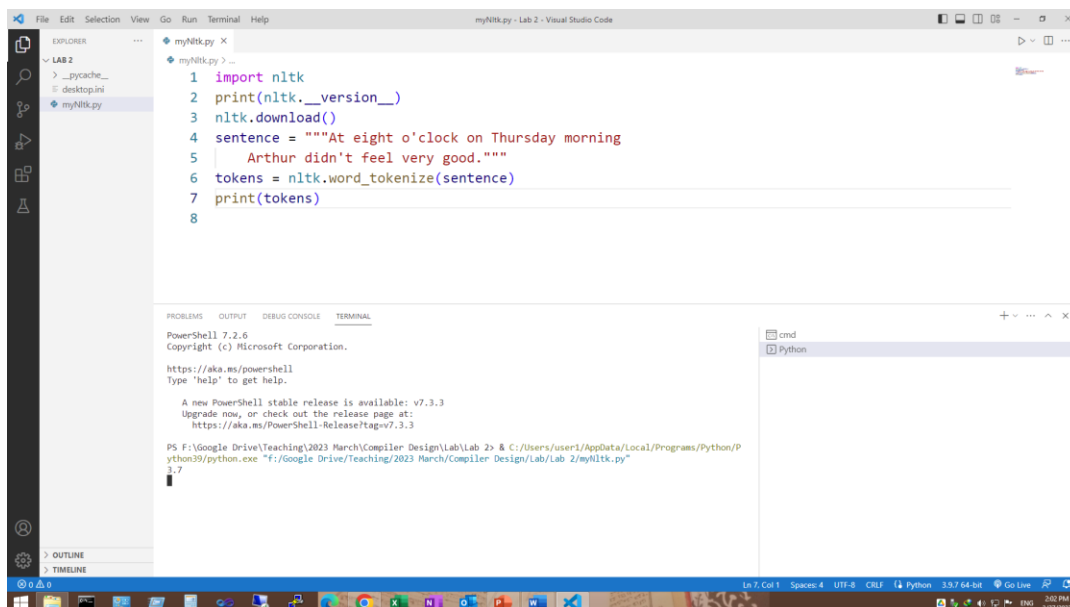
EXPLORER
LAB 2
  _pycache_
  desktop.ini
  nltk.py

Start
  New File...
  Open File...
  Open Folder...
  Clone Git Repository...
  Open a Walkthrough...

Recent
  Lab 1 F:\Google Drive\Teaching\2023 March\Compiler Design\Lab
  aatc C:\Users\user1\Documents\GitHub
  aatcc C:\Users\user1\Desktop
  Lab 1 F:\Google Drive\Teaching\2023 January\Data Structure Lab
  Lab F:\Google Drive\Teaching\2023 March\Compiler Design
  easyocr C:\Users\user1\Desktop\Wareef
  Wareef C:\Users\user1\Desktop

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL
Microsoft Windows [Version 10.0.20000]
(c) 2013 Microsoft Corporation. All rights reserved.

F:\Google Drive\Teaching\2023 March\Compiler Design\Lab\Lab 2>pip install nltk
```



```
File Edit Selection View Go Run Terminal Help
myNltk.py - Lab 2 - Visual Studio Code

EXPLORER
LAB 2
  _pycache_
  desktop.ini
  myNltk.py

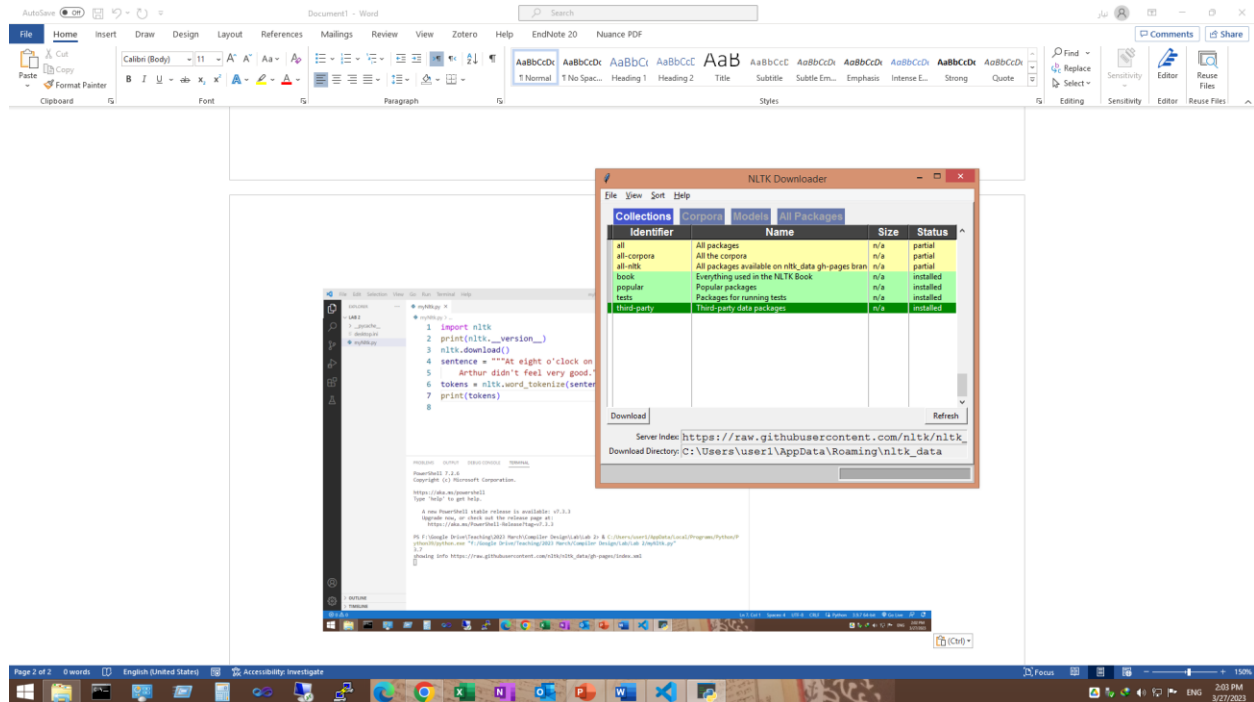
myNltk.py
1 import nltk
2 print(nltk.__version__)
3 nltk.download()
4 sentence = """At eight o'clock on Thursday morning
5 Arthur didn't feel very good."""
6 tokens = nltk.word_tokenize(sentence)
7 print(tokens)
8

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL
PowerShell 7.2.6
Copyright (c) Microsoft Corporation.

https://aka.ms/powershell
Type 'help' to get help.

A new PowerShell stable release is available: v7.3.3
Upgrade now, or check out the release page at:
https://aka.ms/PowerShell-Release?tag=v7.3.3

PS F:\Google Drive\Teaching\2023 March\Compiler Design\Lab\Lab 2> & C:\Users\user1\AppData\Local\Programs\Python\Python39\python.exe "F:\Google Drive\Teaching\2023 March\Compiler Design\Lab\Lab 2\myNltk.py"
3.7
```



NLTK in Python Basics

NLTK is a standard python library with prebuilt functions and utilities for the ease of use and implementation. It is one of the most used libraries for natural language processing and computational linguistics.

Tokenization is the process of breaking text up into smaller chunks as per our requirements.

Sentence Tokenization

Word Tokenization

Punctuation Remover

Stop Words Remover

Stemming

Code to demonstrate the Sentence Tokenizer

```
import nltk
print(nltk.__version__)
nltk.download()
sentence = """At eight o'clock on Thursday morning
    Arthur didn't feel very good."""
tokens = nltk.word_tokenize(sentence)
print(tokens)
```

Code to demonstrate the Paragraph Tokenizer

```
from nltk.tokenize import sent_tokenize
para="""Cake is a form of sweet food made from
flour, sugar, and other ingredients, that is
usually baked.In their oldest forms, cakes were
modifications of bread, but cakes now cover a wide
range of preparations that can be simple or
elaborate, and that share features with other
desserts such as pastries, meringues, custards, and
pies.The most commonly used cake ingredients
include flour, sugar, eggs, butter or oil or
margarine, a liquid, and leavening agents, such as
baking soda or baking powder."""
tokenized_para=sent_tokenize(para)
print(tokenized_para)
print(type(tokenized_para))
```

Code to demonstrate the Punctuations Removal

```
from nltk.tokenize import RegexpTokenizer
tokenizer = RegexpTokenizer(r'\w+')
result = tokenizer.tokenize("Wow! I am excited to
learn Compiler Designing")
print(result)
```

Stop words removal from paragraph:

```
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
to_be_removed = set(stopwords.words('english'))
para="""Cake is a form of sweet food made from
flour, sugar, and other ingredients, that is
usually baked.
In their oldest forms, cakes were modifications of
bread, but cakes now cover a wide range of
preparations
that can be simple or elaborate, and that share
features with other desserts such as pastries,
meringues, custards,
and pies."""
tokenized_para=word_tokenize(para)
print(tokenized_para)
modified_token_list=[word for word in
tokenized_para if not word in to_be_removed]
print(modified_token_list)
```

Stemming the user input given in the NLTK Package

```
from nltk.stem import PorterStemmer
from nltk.tokenize import word_tokenize
stemmer = PorterStemmer()
content = """Cake is a form of sweet food made from
flour, sugar, and other ingredients, that is
usually baked.In their oldest forms, cakes were
modifications of bread, but cakes now cover a wide
range of preparations
that can be simple or elaborate, and that share
features with other desserts such as pastries,
meringues, custards, and pies."""
tk_content=word_tokenize(content)
stemmed_words = [stemmer.stem(i) for i in
tk_content]
print(stemmed_words)
```

Some reference websites to study:

<https://realpython.com/nltk-nlp-python/>

<https://www.guru99.com/download-install-nltk.html>

<https://www.guru99.com/tokenize-words-sentences-nltk.html>

<https://www.nltk.org/book/ch01.html>

Error Resolution:

<https://bobbyhadz.com/blog/python-no-module-named-nltk>

General Questions to answer:

1. What is Tokenization?

.....

.....

.....

.....

.....

.....

2. What is the difference between Lexemes and Tokens?

.....

.....

.....

.....

.....

.....

3. What is Semantic Analysis? Try to write a program to split sentence in grammatical elements (Ex. Noun/Verb/Adjective/Articles etc.)

.....

.....

.....

.....

.....

.....