**College of Computer Science and Engineering**

**Department of Computer Science and Artificial Intelligence**

**CCAI-413: Natural Language Processing**

**Lab#2 Morphological analysis**

# Objectives

* Apply different text search techniques
* Apply text tokenization
* Apply text stemming
* Apply Arabic text stemming

# Lab Tool(s)

<https://www.kaggle.com/>

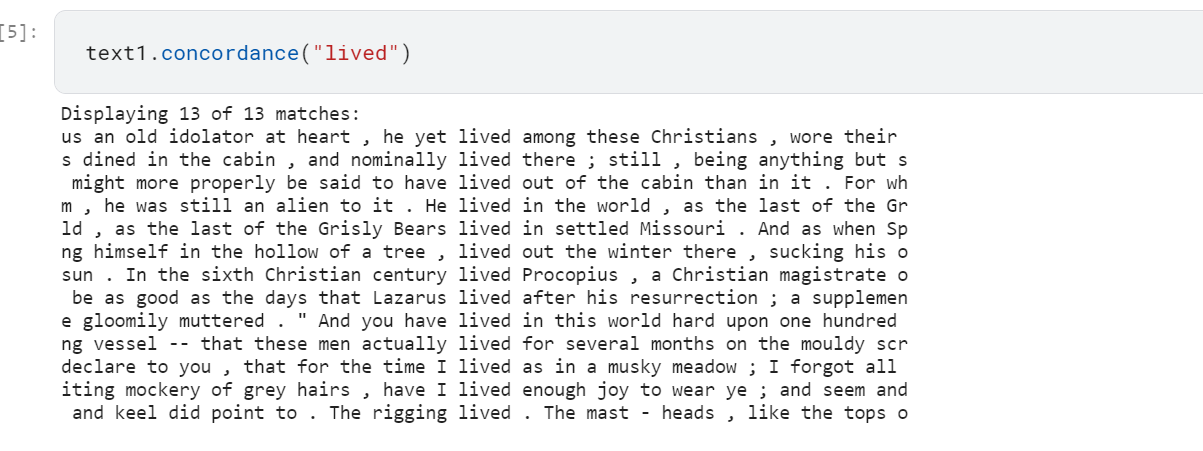
# Searching Text

Type the following commands to import NLTK and the book data at the Python prompt

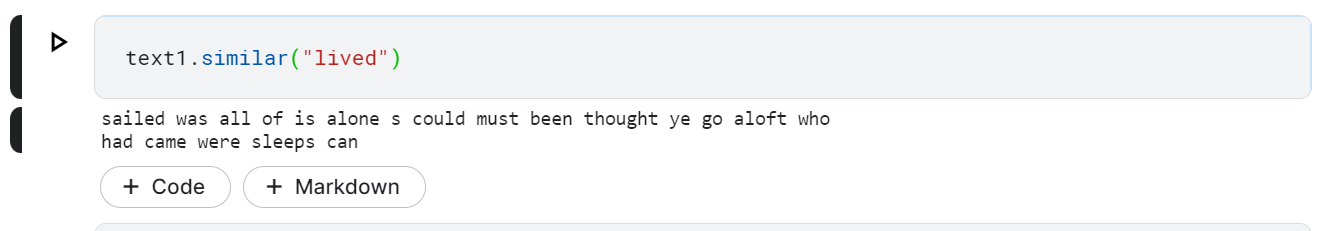
Graphical user interface, text, application

Description automatically generated

The function *concordance ()* shows the occurrence of a given word with some context. The following figure presents how the command is used to search about the word “lived”:

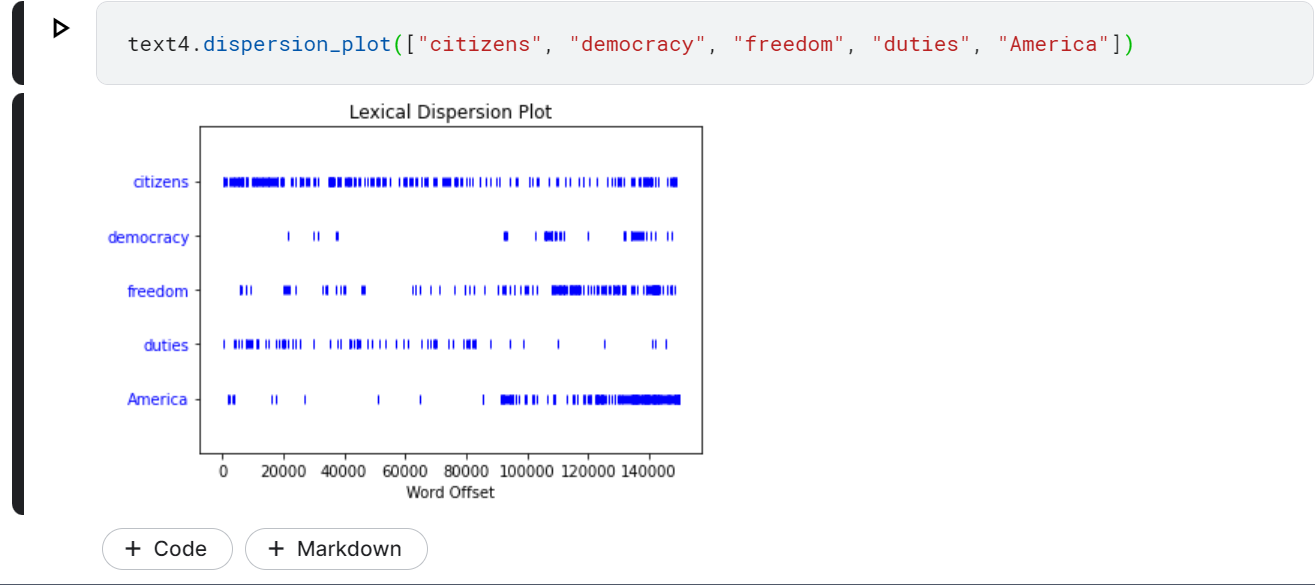


To find out what other words appear in the same context, we can use the *similar ()* function.



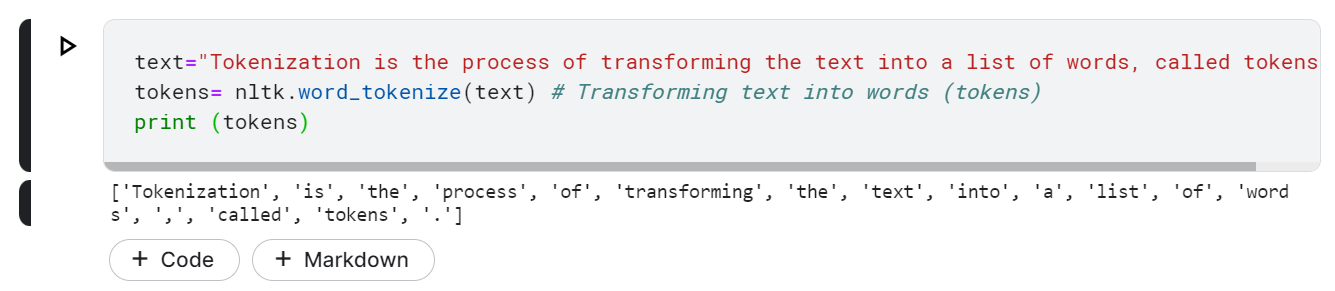
* *Your Turn!* Try searching for other words

We can also determine the location of a word in the text: how many words from the beginning it appears. This positional information can be displayed using a ***dispersion plot***. Each stripe represents an instance of a word, and each row represents the entire text.

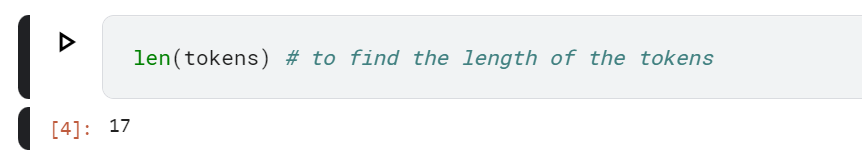


# Text Tokenization

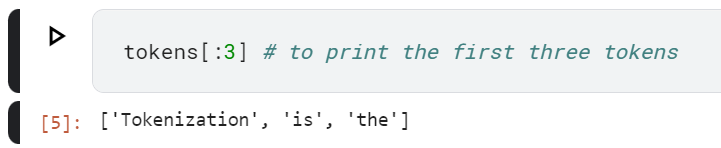
Tokenization is the process of transforming the text into a list of words and punctions, called tokens. NLTK provides a function for tokenization as follows *nltk.word\_tokenize()*.



Type the following command to find the length of the tokens:



Type the following command to print the first N tokens:



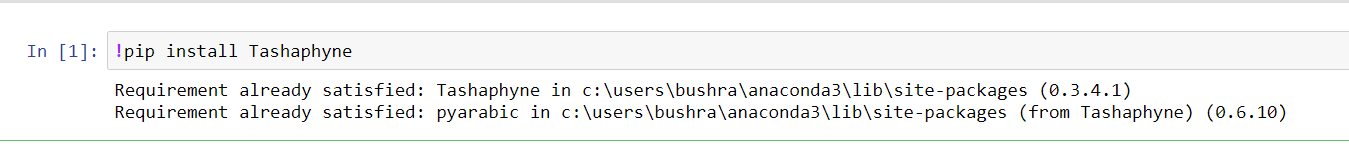
# Text Stemming

Words can be written in different forms (e.g., studying, studies, and study). Stemming is the process of removing prefixes and suffixes of a word. NLTK provides a stemmer called *PorterStemmer()*.

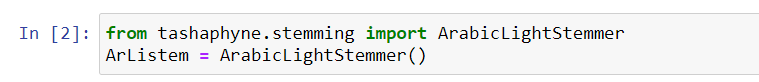


# Arabic Text Stemming

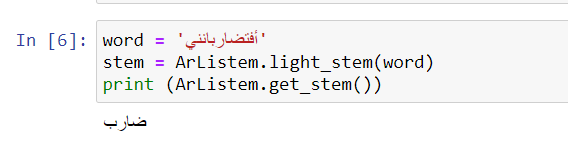
Tashaphyne is a python library that provides Arabic stemmer. To use Tashaphyne you have to install the library using the following command:



Then we should import the stemmer and create its object:

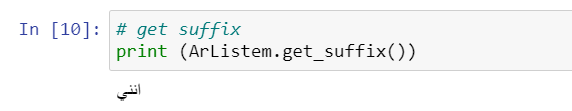


Below is an example of using the Arabic stemmer on the word “أفتضاربانني”:



# 

# 



# References:

* Bird and Klein, O’Reilly Media, Natural Language Processing with Python, 2nd Edition, 2017.
* <https://pypi.org/project/Tashaphyne/>