UTD Summer Workshop on NLP with Python

Assignment 3

Upload your completed Python program file to the MS Teams "Assignments" → Assignment 3" folder. Be sure to include your name in the comments at the top of the file in the following format:

```
################################################################################
#
#          FILE:
#            filename.py
#        AUTHOR:
#            Your Name
#  DESCRIPTION:
#            Assignment 3
#            Description of your program, what it does
# DEPENDENCIES:
#            Created with Python 3.10.11 (Python version)
#            Any dependencies, i.e. extra libraries required to run (like
#            datetime, NLTK, or re)
#
################################################################################
```

Extra help references:

- Python re reference – https://docs.python.org/3/library/re.html
- Requests
  - https://pypi.org/project/requests/
- Beautiful Soup
  - https://pypi.org/project/beautifulsoup4
- NLTK references
  - Tokenize – https://www.nltk.org/api/nltk.tokenize.html
  - WordNet –
    - Sample usage

Submit a single file Python program that does the following:

1. Using Python, download the source HTML for one or more webpages of your choosing. You may assign the HTML source directly to a variable without first saving it to a file. Otherwise, you may save the source code to a local file, then open it using Python's built-in open() function.

2. Use the Beautiful Soup library to extract just the text without any HTML markup. Note that your resulting text may still have non-relevant phrases and words, such as text menus from the website. You may use techniques discussed in lecture to remove these if you desire.

3. Once you have the plain text, tokenize it into sentences.

4. Then ***for each sentence***:

    4.1. Display a visual separator, such as: `print("=" * 80)`

    4.2. Display the plain text of the sentence followed by a list of 3-tuples made from word tokenizing the sentence. The 3-tuples should be a (lowercase word, POS, stem). Stemming should be performed using the Porter Stemmer in WordNet.

```
John came from the store.
[('john', 'NNP', 'john'), ('came', 'VBD'), ('from', 'IN'), ('the', 'DT'),
('store', 'NN')]
```