# AIT526 Individual Lab 3

**Due Date: Please check the class schedule on blackboard.**

## Named Entity Recognition and De-Identification with SpaCy

### Tools:

1) **Jupyter Lab** (Desktop or online) or Desktop **Jupyter Notebook** or any **Python IDEs**
2) **Python 3**
3) **SpaCy** (https://spacy.io/)
4) **BeautifulSoup** (https://www.crummy.com/software/BeautifulSoup/) for **web scraping**

** Optional tools*

### Coding Resources:

1) **Dr. Liao's Code examples/tutorials/Hints**
2) Methods and algorithms in the lecture notes
3) Source of Internet

**Note that you must include **reference(s)** in the code comments when you refer others' work.*

### Text Data Location: any online news article webpage, e.g., Washington News, New York Times, etc.

### Tasks (10 points):

Please follow the code examples and tutorials to implement the following tasks:

1 **(5 points) Named Entity Recognition (NER):**
- 1.1 **(0.3 points)** Copy the code examples to scrape the webpage in BeautifulSoup
- 1.2 **(4.7 points)** Write the code for **NER in SpaCy**
  - 1.2.1 **(0.2 points)** Count all the named entities in the document
  - 1.2.2 **(0.5 points)** Count the most frequent tokens for the entire document
  - 1.2.3 **(2.0 points)** Pick a random integer **K** using Python random module, then pick **three consecutive sentences** starting with **K**th, and print these sentences. Note that you must make sure all picked sentences are in the document.
  - 1.2.4 **(0.5 points)** Extract part-of-speech and lemmatize **these consecutive sentences**
  - 1.2.5 **(0.5 points)** Get and print the entity annotation for each token of the Kth sentence
  - 1.2.6 **(0.5 points)** Visualize the entities and dependencies of Kth sentence
  - 1.2.7 **(0.5 points)** Visualize all the entities in the document

2 **(5 points) De-Identification**:

2.1 De-identify all person names (PERSON) in the webpage document with **[REDACTED]** and visualize them as shown in class.

3 You are strongly suggested to follow Python coding convention to write the code. The program should be robust and will be tested with several different text files for grading.

## SUBMISSION

1. Write all your code and answers with explanation in the Notebook.
2. In the code file, please do not forget to write your name, course #, and date in the comments.
3. **Run ALL Cells**:
   Open your IPython file in Jupyter, go to **Run**->**Run All Cells**. Please make sure all of you code has been run and print out the results.
4. **Save to HTML**:
   Go to **File**-> **Export Notebook As…**->**Export Notebook to HTML**, and save your work into HTML file.
5. **Submission**:
   a. Write your work with two file names **"AIT526_YourFullName_Lab3.ipynb"** and **"AIT526_YourFullName_ Lab3.HTML"**.
   b. **Zip** both files to **ONE zipped file** since blackboard does not allow you to submit HTML file separately.
   c. Go to the Blackboard **/Course Content/Optional Individual Labs/** to submit **ONE zipped file.**