# AIT590-001 Optional Individual Lab 3

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

## Named Entity Recognition and De-Identification in SpaCy

**Tools** (as shown in the class):

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** (as shown in the class):

1) **Dr. Liao's code examples/tutorials**
2) Methods and algorithms in the lecture notes
3) Many Internet resources

**Text Data Location**: any online news article webpage as shown in the code examples

For example, **NBC4-Washington News**: https://www.nbcwashington.com/news/local/latest-updates-how-many-coronavirus-covid-diagnosed-confirmed-cases-test-deaths-fatalities-dc-maryland-virginia-dmv/2230095/?amp=&from=singlemessage&isappinstalled=0

**New York Times**: https://www.nytimes.com/2020/03/31/world/coronavirus-live-news-updates.html?action=click&module=Spotlight&pgtype=Homepage

**Tasks (10 points, Extra Credit):**

Follow the code examples and tutorials as shown in class to finish the following tasks:

1 **(4 points) Named Entity Recognition (NER):**

1.1 Copy the code examples to scrape the webpage in BeautifulSoup (0.5 point)

1.2 Write the code for **NER in SpaCy** (2.5 points)

1.2.1 Count all the named entities in the document (0.2 points)

1.2.2 Count the most frequent tokens for the entire document (0.2 points)

1.2.3 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. (1.5 points)

1.2.4 Extract part-of-speech and lemmatize **these consecutive sentences** (0.2 points)

1.2.5 Get and print the entity annotation for each token of the Kth sentence (0.5 points)

1.2.6 Visualize the entities and dependencies of Kth sentence (0.4 points)

1.2.7 Visualize all the entities in the document (0.5 points)

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 **(1 point)** 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. **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.
3. **Save to HTML**:
   Go to **File**-> **Export Notebook As…**->**Export Notebook to HTML**, and save your work into HTML file.
4. **Submission**:
   Write your code with two separate file names **"AIT590_YourFullName_Lab3.ipynb",** then export to corresponding **HTML files**. Go to the Blackboard **/Course Content/Optional Individual Labs/** to submit both files with **ONE zipped file** since blackboard does not allow you to submit HTML file separately.