AIT590-001 Optional Individual Lab1

Due Date: Please check the class schedule on blackboard.

NLP - Text Summarization for Webpage

Tools (as shown in the class):

1) **Jupyter Lab** (Desktop or online) or Desktop **Jupyter Notebook** or any **Python IDEs**
2) **Python 3**
3) **NLTK** (https://www.nltk.org/)
4) **Gensim** (https://radimrehurek.com/gensim/)
5) **BeautifulSoup** (https://www.crummy.com/software/BeautifulSoup/) for **web scraping**

Coding Resources (as shown in the class):

1) **Dr. Liao's code examples/tutorials** (BeautifulSoup for web scraping, NLTK, Gensim, etc.)
2) Methods and algorithms in the lecture notes
3) Many internet sources

Text Data Location: https://en.wikipedia.org/wiki/Natural_language_processing

Tasks (10 points, Extra Credit):

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

1 **(0.25 point)** Use the **web scraping** technique with **BeautifulSoup** as shown in class to get the text data from the specified data location on the Wikipedia webpage.

2 **(2 points)** Process the text data and must include:
- 2.1 Tokenize the words (0.25 point)
- 2.2 Remove the stop words, punctuation, and **digit numbers**. (0.50 point)
- 2.3 write a function to **lemmatize the words (1 point)**
- 2.4 Calculate the word distribution using FreqDist (0.25 point)
- 2.5 List and plot the top 15 words (0.25 point)

3 **(5 points)** Summarize the text as shown in class:
- 3.1 **Calculate word frequency** using two different methods:
  - 3.1.1 **Weighted frequency** (please specify the math formula) **(1 point)**
    (Must use FreqDist to get original word frequencies)
  - 3.1.2 **TF-IDF with NLTK** (NLTK does not have TF-IDF) **(2 points)**
- 3.2 **Score the sentences (1 point)**
- 3.3 **Build a summary** (based on ratio, sentence or word count, etc.) **(1 point)**

4 **(0.25 point)** Summarize the same text data using **Gensim** with **TextRank**

5 **(1.50 point)** Compare 3.1.1 and 3.1.2 methods with Gensim-TextRank. What's different and why? Could your methods be improved? And How to improved? Please clearly explain.

6 **(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 webpages 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 work with two file names **"AIT590_YourFullName_Lab1.ipynb"** and **"AIT590_YourFullName_ Lab1.HTML"**. 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.