Natural Language Processing for the Social Sciences (GR5067)
Course Syllabus

Contacts:

- Instructor
  - Patrick Houlihan, pjh2144@columbia.edu
  - Office Hours: Monday 06:00pm-8:00pm
    - Please contact me in advance to coordinate
- Class: Monday 8:10pm-10:00pm EST
- Class Location: 413 Kent Hall
- Teaching Assistants:
  - Lin Zhu lz2808@columbia.edu
  - Sibo Geng sg4010@columbia.edu
  - Zeya Ahmad za2291@columbia.edu

Course Description:

The course centers on the techniques of Natural Language Processing using the Python programming language.

Course Goals:

1. To gain a background in the basics of Python
2. Understanding of Natural Language Processing
3. Ability to perform sentiment analysis on text through Machine Learning

Prerequisites: Students expected to have a basic understanding of mathematics and basic familiarity with the Python Programming Language. Please have the following installed on your laptops prior to the start of the first day of class:

- Install Python IDE – **version 3.0**
  - Anaconda distribution of Python
- Download and install NLTK
  - Follow section 1.2 of nltk book

Individual Evaluations: Student grades based on the following criteria:

Homework 50%
Project:
- Paper 30%
- Presentation 20%

A detailed grading rubric is below on next page of this syllabus:

		Category/Expectation	Meets (100%)	Below (75%)	Unsatisfactory (50% with submission 0% no submission)	Total Points
Homework		Specification	The program works and produces the correct results and displays them correctly. It also meets most of the other specifications.	The program produces correct results but does not display them correctly	The program is producing incorrect results.	50
Project	Presentation	Knowledge of Material	Presenter was fully prepared with understanding of the parameters.	Presenter was unprepared for the talk, demonstrating little understanding of how the topic relates to the Missing backup material where conclusions were inferred	Did not show up for presentation without valid excuse	10
		Content	Excellent visuals		Did not show up for presentation without valid excuse	10
	Paper	Content	An accurate and complete explanation of key concepts and theories is made. Enough detail is presented to allow the reader to understand the content and make judgments about it. Quality of paper is submission worthy to a journal.	The explanation is sufficiently inaccurate, incomplete, or confusing that the reader gains little information from the report. It appears that little attempt has been made to help the reader understand the material.	Did not submit a report	30

Homework collaboration policy: You may brainstorm and think through solutions with a small number of your classmates. However, you must write up your solutions entirely on your own. If you have used collaborators, you must state their names clearly next to your name on your write-up. Finally, copying solutions from the Internet or other textbooks is strictly prohibited.

Homework programming assignments: Each programming assignment MUST be completed and full reproducible using **Python 3.0**. You first need to install Python on your machine, if not already available. We recommend the Anaconda distribution:

- Download the Anaconda distribution that conveniently installs Python, the Jupyter Notebook, and other commonly used packages for scientific computing and data science.
- Please familiarize yourselves with this environment as soon as possible.
- Download and install NLTK

Course materials:

Text:
Dan Jurafsky and James H. Martin, Speech and Language Processing (3rd Ed)
https://web.stanford.edu/~jurafsky/

Natural Language Processing with Python (https://www.nltk.org/book/)

Reference: scikit-learn (https://scikit-learn.org)

List of topics:

- Python
- Natural Language Processing
- Machine Learning

Research project:

You will work on a team project that will involve the use of Python and NLP techniques. The project will count as 50% of your total grade, where you are required to provide a final presentation (presentations will take place the last two classes of the semester). Some ideas, but not limited to, for projects are below.

A. Leverage New York Times API to extract key words, like 'trade war' and plotting against stock market movements
B. Compare corpuses extracted between Republican vs Democratic speeches
C. Detecting sentiment changes for geopolitical policy changes

---

Detailed description of the course:

If it were not for the mass spend in R&D during the mid to late 1990s, we would still be hearing that crackling sound of the dial-up internet connection. Facebook and Google would be mere fractions of what they are today and Amazon would most likely be selling just books. Thankfully, through that massive investment in technology during the internet revolution, technological progression enabling the analysis of terabytes worth of data possible. From harnessing graphics processing units, GPUs, to perform deep learning to deploying entire architectures in the Amazon Cloud, both beautiful and amazing to say the least. You are amidst the biggest technological revolution your generation has ever experienced, hop aboard, and learn the skill-sets and get into that practitioners mind-set to make a significant impact within the world of big data!

The goal of this class is to arm students with the necessary skill-set to use the techniques of Natural Language Processing. This course will provide students with a foundational knowledge base of Natural Language Processing, NLP, and leveraging machine learning for sentiment analysis and prediction using the Python programming language. Students will learn vital data wrangling techniques with an emphasis on *text analytics*. The intent is to not only expose students to NLP techniques but also place students into the mind-set of a practitioner to build a real working system through modules they create during both in-classroom and homework exercises.

Course Schedule:

Session	Date	HW	Chapter	Topic(s)
1	9/12		NLTK: CH0	Course Introduction, Project Description, What is Natural Language Processing
2	9/19		NLTK: CH1, CH3 SLP: CH2	Syntax, Variables, Operators, Regex, Datetime, Escape Characters, GitHub
3	9/26	HW1	NLTK: CH2, CH3.1, CH4	Round Robin Project Topics, Sets, Dictionary, Lists, For, While, Do, I/O Read/Write
4	10/3		NLTK: CH4.4	Numpy, Pandas, Classes and Objects
5	10/10		NLTK: CH3.6, CH3.7	NLTK, Text Tokenization, Lemmatization, Stemming
6	10/17	HW2		Document Similarity, Web Scraping, Text Clustering
7	10/24		SLP: CH3, CH6, CH19	Word2Vec, Count Vector, Co-Occurrence, N-Gram
8	10/31		SLP: CH6, CH8	TF-IDF, Bag of Words, Parts of Speech Tagging, Word Embeddings, Project Check-In
9	11/14	HW3	link	Topical Extraction, Latent Dirichlet Allocation
10	11/21		SLP: CH4, CH21	Sentiment Analysis, Expert, Semi-Supervised, Lexicons
11	11/28		SLP: CH4, CH5	Scikit-learn, Machine Learning Algos, Training, Validation, Test, Grid, x-Fold CV
12	12/5	HW4	SLP: CH4, CH5	Predicting Sentiment, Predictive Streaming Analytics
13	12/12			Individual Project Presentations