



**THE GEORGE
WASHINGTON
UNIVERSITY**
WASHINGTON, DC

DATS 6203 – Natural Language Processing
CRN 74856
Mondays 3:30 PM – 6:00 PM

INSTRUCTOR:

Name: Amir Jafari, PhD

Term: Fall 2022

Campus address: COR 207

Zoom: <https://gwu-edu.zoom.us/j/9991766531>

Github: <https://github.com/amir-jafari>

E-mail: ajafari@gwu.edu

Phone: (202)-994-1239

Office hours: Mondays 1:00 pm -3:00 pm

COURSE DESCRIPTION:

Natural Language Processing (NLP) is a rapidly evolving in data science field . The ability to analyze linguistic and textual information through out research and industry is a necessary skill. The main focus of this course is to understand theory and algorithms that are most widely used in the field. Implement the theory and use the e current techniques, strategies and toolkits for natural language processing. In this course, we learn to process big large corpora and big textual data sources. In addition to evaluate the classical and traditional approaches, we learn how to implement the state of art NLP deep learning models. Throughout this course, the classical (LSA, BOW, Naive bayes, ...), Neural Network based models such as LSTM and GRU and state of art modern NLP Transformers (BERT, XLNET, ROBERTA, ALBERT, T5,...) will be introduced. Finally, we will use all these models and pre-trained models in order to solve real world NLP tasks such as (sentiment analysis, text classification, next word prediction, next sentence prediction, question answering, chat bots, ...).

LEARNING OUTCOMES:

Students will be able to:

1. gain a foundational understanding in NLP methods and algorithms.
2. understand the text mining and text processing.
3. apply NLP classical models to practical problems.
4. apply deep learning and transformer models to practical problems.
5. evaluate the strengths and weaknesses of various NLP models.
6. gain practical experience in the NLP toolkits that are available.
7. gain extensive experience in using Python to process textual and linguistic analyses.

RESOURCES:

A- Speech and Language Processing, Third Edition, by Daniel Jurafsky and James H. Martin, 2020, [Web Link](#).

B- NLTK Essentials, by Nitin Hardeniya, 2015 [Web Link](#)

C- Neural Network Design, Second Edition, by Martin T. Hagan, Howard B. Demuth, Mark Hudson Beale and Orlando De Jesús [Web Link](#).

SOFTWARE:

Amazon Web Services (AWS) and Google Cloud Platform(GCP) virtual machines will be used heavily the course. Basic Linux knowledge such as working with terminal is needed. Python are used for homework assignments, Labs, class exercises, final project and demos.

TENTATIVE COURSE OUTLINE (SUBJECT TO CHANGE):

Week	Topic	Comments
August 29, 2022	Environment Setup	
September 5, 2022	Labor Day (no classes)	
September 12, 2022	Introduction to NLP and Python IO	
September 19, 2022	Natural Language Toolkit (NLTK)	Quiz1
September 26, 2022	Natural Language Toolkit (Spacy)	Quiz2
October 3, 2022	Regular Expression, Bag of Words (BOW)	Quiz3
October 10, 2022	Naive Bayes & Logistic Regression	Quiz4
October 17, 2022	Exam 1	
October 24, 2022	Fall Break (no classes)	
October 31, 2022	Intro to Neural Network (MLP)	
November 7, 2022	Pytorch	Quiz5
November 14, 2022	Recurrent Neural Network (LSTM, GRU)	Quiz6
November 21, 2022	Transformers & Attention	Quiz7
November 28, 2022	Masked Language Models	
December 5, 2022	Exam 2	
December 12, 2022	Final Project Presentation and Submission	

PREREQUISITES:

DATS 6202 - Machine Learning I

DATS 6103 -Introduction to Data Mining

ASSIGNMENT DESCRIPTION:

The labs and homework will be associated with each module; there will be lab exercises for each lecture. The exam will cover all the homework assignments, class exercises and quizzes. George Washington University has a Amazon Web Services (AWS) cloud account with NVIDIA compatible GPUs and I will give 50 dollar credit for Google Cloud Platform (GCP). The mini and final projects should be done on cloud platform.

ACADEMIC INTEGRITY:

The code of academic integrity applies to all courses in the George Washington School ("Academic dishonesty is defined as cheating of any kind, including misrepresenting one's own work, taking credit for the work of others without crediting them and without appropriate authorization, and the fabrication of information."). In the spirit of the code, a student's word is a declaration of good faith acceptable as truth in all academic matters. Cheating and attempted cheating, plagiarism, lying, and stealing of academic work and related materials constitute Honor Code violations. These will not be tolerated. Please become familiar with the code. All students are expected to maintain the highest level of academic integrity throughout the course of the semester. Please note that acts of academic dishonesty during the course will be prosecuted and harsh penalties may be sought for such acts. Students are responsible for knowing what acts constitute academic dishonesty. The code may be found at [HERE](#). The University's "Guide of Academic Integrity in Online Learning Environments" is available at [HERE](#) for your review.

GRADING AND EXAMINATION POLICY:

- 2 Exam - 25 pts each
- Quizzes - 25 pts
- 1 Final project - 25 pts

The three scores of quizzes, and 2 exams will be added to the final project score to obtain the total grade for the course (out of a total of 100 pts). All exams and quizzes may be in class or take home. I may collect homeworks or give a quiz (most probably there is a quiz after every 2 weeks). No make-up exams unless previous arrangements have been made. Students will be expected to attend class and prepare assignments. Habitual failure to do so will result in a reduced grade. An incomplete grade will only be given when a student misses a portion of the semester because of illness or accident. Cheating on examinations, plagiarism and other forms of academic dishonesty are serious offenses and may subject the student to penalties ranging from failing grades to dismissal.

SECURITY:

In the case of an emergency, if at all possible, the class should shelter in place. If the building that the class is in is affected, follow the evacuation procedures for the building. After evacuation, seek shelter at a predetermined rendezvous location.

DISABILITY SUPPORT SERVICES (DSS):

Any student who may need an accommodation based on the potential impact of a disability should contact the Disability Support Services office at 202-994-8250 in the Marvin Center, Suite 242, to establish eligibility and to coordinate reasonable accommodations. See [HERE](#)

The University Counseling Center (UCC Phone: 202-994-5300) offers 24/7 assistance and referral to address students' personal, social, career, and study skills problems [Web Link](#). Services for students include:

- crisis and emergency mental health consultations
- confidential assessment, counseling services (individual and small group), and referrals

UNIVERSITY POLICIES:

Students should notify faculty during the first week of the semester of their intention to be absent from class on their day(s) of religious observance. Faculty should extend to these students the courtesy of absence

without penalty on such occasions, including permission to make up examinations. Faculty who intend to observe a religious holiday should arrange at the beginning of the semester to reschedule missed classes or to make other provisions for their course-related activities. For details and policy, see “Religious Holidays” in [HERE](#).

EMAIL ETIQUETTE:

In the age of technology, when most forms of communication are electronic, it is important to adopt a proper etiquette to communicate with one another. It is asked that students use salutation when sending emails to their instructors and also make sure to SIGN their name and include their class/section at the end of the email. The instructor reserves the right NOT to reply to emails that are not properly addressed or do not have a signature. Students should also use their GWU email for any correspondence with the instructors. Students are required to check their emails daily and especially the morning before class.

COURSE CONTENT (SUBJECT TO CHANGE):

Week	Module	Topics
August 29, 2022	Software Setup	Google Cloud Platform (GCP) Amazon Web Services (AWS) Pycharm Editor
September 5, 2022	Labor Day (no classes)	
September 12, 2022	Introduction to NLP and Python IO	File Read Strings, Numpy, Dictionary, Set Counter Linux Terminal Git Python OS
September 19, 2022	Natural Language Toolkit (NLTK)	Installation Tokenization Stemming Correcting Words Custom Copra Part of Speech Tagging
September 26, 2022	Natural Language Toolkit (Spacy)	Installation Core Operations Linguistic Features Rule Based Matching Entities Matcher
October 3, 2022	Regular Expression Bag of Words (BOW)	Regx TFIDF Count Vectorizer LSA SVD BOW

Week	Module	Topics
October 10, 2022	Naive Bayes & Logistic Regression	Bayes Rule Marginal Density Maximum Likelihood Logistic Function Evaluation Metrics
October 17, 2022	Exam 1	
October 24, 2022	Fall Break (no classes)	
October 31, 2022	Intro to Neural Network (MLP)	Multi Layer Perceptron BackPropagation Optimization Stochastic Gradient Word2Vec
November 7, 2022	Pytorch	Computational Graph NN Module Tensor Data Loader
November 14, 2022	Recurrent Neural Network (LSTM, GRU)	Training Recurrent Networks Seq2seq PadPack Sequence Next Word Prediction Attention
November 21, 2022	Transformer	GLUE Next Sentence Prediction Question Answering Sequence Classification
November 28, 2022	Masked Language Models	BERT XLNET Electra Pretraining from Scratch Explainability
December 5, 2022	Exam 2	
December 12, 2022	Final Project Presentation and Submission	

AVERAGE AMOUNT TIME LEARNING PER WEEK:

Students are expected to spend a minimum of 100 minutes of out-of-class work for every 50 minutes of direct instruction, for a minimum total of 2.5 hours a week. A 3-credit course should include 2.5 hours of direct instruction and a minimum of 5 hours of independent learning or 7.5 hours per week.

ONLINE RESOURCES:

For technical requirements and support, student services, obtaining a GWorld card, and state contact information please check [HERE](#)

CLASSROOM RECORDING:

The particular class recordings will be available to students who are registered on an individual basis, upon request. Please let me know in advance if you have any medical issues or any emergencies that you will not be able to join the class.

Virtual ACADEMIC SUPPORT:

A full range of academic support is offered virtually in fall 2020. See [HERE](#) for updates. Tutoring and course review sessions are offered through Academic Commons in an online format. See [HERE](#). Writing and research consultations are available online. See [HERE](#). Coaching, offered through the Office of Student Success, is available in a virtual format. See [HERE](#). Academic Commons offers several short videos addressing different virtual learning strategies for the unique circumstances of the fall 2020 semester. See [HERE](#). They also offer a variety of live virtual workshops to equip students with the tools they need to succeed in a virtual environment. See [HERE](#)

SAFETY and SECURITY:

In an emergency: call GWPD 202-994-6111 or 911. For situation-specific actions: review the Emergency Response Handbook in [HERE](#). In an active violence situation: Get Out, Hide Out, or Take Out. See [HERE](#). Stay informed: safety.gwu.edu/stay-informed