Introduction to Applied Natural Language Processing

STAT 496/696

Instructor:	Zois Boukouvalas	**Time:**	T 5:30 PM - 8:00 PM
Email:	boukouva@american.edu	**Room:**	Online (Live Sessions)
Office:	Don Myers Building - 222		

Course description: This course covers fundamental methods for analyzing textual datasets, focusing on applying classical natural language processing (NLP) methods and libraries in Python to interesting corpora. Students will gain familiarity with introductory and intermediate Python concepts to facilitate processing of text for textual analysis. Topics will include regular expressions, dictionary methods, an introduction to linguistic structure (e.g., parts of speech), bag-of-words methods and word/document embedding methods. Applications will include sentiment analysis, predictive analytics, information retrieval via clustering and outlier detection methods, and language change detection.

Overview of course topics: We will cover the following topics.

- Introduction to Python libraries and tools for NLP: NumPy, NLTK, pandas, Scikit-Learn, seaborn, matplotlib.
- String manipulation. Pre-processing text for analysis. Counting words.
- Dictionary methods. Regular expressions.
- Parts-of-speech Tagging. Named entity recognition. Stemming and Lemmatization.
- Tokenization, Vectorization, and Bag-of-word (BoW) methods.
- Latent variable methods for text data and basic topic modeling.
- Outlier detection. Clustering via k-means.
- Predictive analytics: Linear Regression and Nonlinear Regression
- Word embeddings. Advanced document representations. e.g. Word2Vec, Glove, BERT.

Learning outcomes (LO) and assessments (AS): At the end of this course, students will be able to:

STAT 696 (Graduate Level)

- LO: Develop tools and select appropriate methods for their own research problems and judge the reasonableness of the obtained research findings. AS: Main project.
- LO: Communicate both orally and verbally about NLP in a scientifically sound manner and present their research findings. AS: Project proposal and poster presentation.
- LO: Evaluate theory and critique research within the NLP field and identify the ethical questions associated with data collection and analysis. AS: Project proposal and midterm exam.
- LO: Use graphical tools to visualize and understand textual data. AS: Python Competency Test, homework assignments.
- LO: Import and analyze textual datasets from a variety of sources. AS: Homework assignments.
- LO: Scrape websites for information, and process/analyze the information. AS: Main project and homework assignments.

STAT 496 (Undergraduate Level)

- LO: Perform a rigorous literature review for their research problem and prepare poster presentations for NLP conferences. AS: Project proposal and poster presentation.
- LO: Reproduce research results from a high level peer-reviewed NLP related article, work in groups, and be part of an effective team. AS: Main project.
- LO: Evaluate theory and critique research within the NLP field and identify the ethical questions associated with data collection and analysis. AS: Project proposal and midterm exam.

- LO: Use graphical tools to visualize and understand textual data. AS: Python Competency Test, homework assignments.
- LO: Import and analyze textual datasets from a variety of sources. AS: Homework assignments.
- LO: Scrape websites for information, and process/analyze the information. AS: Main project and homework assignments.

Materials:

- **Required**: Laptop computer
- **(Recommended)**: Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit, Steven Bird, Ewan Klein, and Edward Loper.
Practical Natural Language Processing, A Comprehensive Guide to Building Real-World NLP Systems, Sowmya Vajjala, Bodhisattwa P. Majumder, Anuj Gupta, Harshit Surana, O'Reilly Media, 2020.

Software: The main software for this course is Python. If you are not familiar with Python yet, (https://learnpython.org) and (https://docs.python.org/3/tutorial) are good places to start.

Pre-requisites: STAT 415/615, basic knowledge of computer science concepts, and familiarity with the Python programming language.

Class structure and office hours: This class will be a blend of lecture, class discussion, and programming demonstrations. I want you all to be involved during class and please do not hesitate to ask questions whenever something is unclear to you. You are expected to attend all class meetings, as I believe that attending class regularly contributes greatly to your performance in the course. It is understandable that you may have to miss class on a rare occasion. You are responsible for any assignments or papers given out during any missed class. The office hours will be Monday from 3:00pm-5:00pm, or after the class, or by appointment. In addition, you are encouraged to ask me questions via email. Please schedule a meeting with me if you would like to see or discuss your grade at any point during the semester. If you are having **ANY** trouble with the class, please see me about it as soon as possible. **Do not wait until it is too late!**.

Course pages: I will use Blackboard (https://blackboard.american.edu) to post any supplementary materials, suggested readings/practice exercises, assignments, and announcements. Sometimes I may also use my personal website (https://zoisboukouvalas.github.io/).

Assignments & Grading:

Assignments (40%): During the semester I will assign, collect, and grade assignments. There will be approximately 8 formal assignments throughout the semester. You may receive assistance from other students in the class and me, but your submissions must be composed of your own thoughts, coding and words. I expect you to get ideas from online resources such as stackoverflow or github when you get stuck. Please cite your source when you do so and be specific about what you have added to it. **I will not accept late assignments.**

Python Competency Test (10%): During the second week, I will assign a hand-written exam testing your understanding of basic Python concepts.

Midterm Exam (20%): This exam is meant to gauge your progress, and to test your understanding of how to apply the methods (learned in this class) to real-world situations.

(20% Poster Presentation + 10% Project Proposal): You will have to prepare a project using the tools and methods learned in the class. You are expected to submit a mid-semester research proposal in order to get your topic approved. For the project presentations I expect you to record your talk and submit it on Blackboard. The class projects will be presented as a poster presentation. You should prepare a poster, and be prepared to give a very short explanation (10 minutes), in front of the poster, about your work. At the poster session (online), you'll also have an opportunity to see what everyone else did for their projects. You will also need to submit your poster as a PDF the day before the presentation.

Data scientists must learn to discover solutions for themselves. You should expect to have to research (use Google, stackoverflow, etc) to do your assignments. All you need to do the assignment will NOT have been provided to you in the lectures and course book. This is an essential part of becoming a data scientist!

Grading scale: Students with final percent between 93%-100% will receive an A, 90%-92% an A−, 88%-89% a B+, 83%-87% a B, 80%-82% a B−, 78%-79% a C+, 73%-77% a C, 70%-72% a C−, 60%-69% a D, and under 60% F.

Academic dishonesty: Academic dishonesty is a serious offense. As a start, you should read and understand our University's policies https://www.american.edu/academics/integrity/code.cfm. You may collaborate with other students in this class on your projects, but the work you turn in must be your own. You may use online forums such as GitHub, StackOverflow, etc. At a minimum, honesty consists of presenting your ideas clearly and in your own words, possibly orally. Here are a few typical cases that are relevant for your assignments:

1. If you use code for your work on your own, you need not notate it as such.

2. If a colleague (or online source) shows you some application or code that you like, please credit that person by name in your write-up. You should expect that I may challenge you to explain your writing in your own words, possibly orally. If you cannot successfully defend your answer, no credit will be awarded to you.

3. If you write your project in collaboration with others, as part of a group, please credit all members of your group by name in your write-up. You should expect that I may challenge you to explain your writing in your own words, possibly orally. If you cannot successfully defend your answer, no credit will be awarded to you, even if other group members are able to defend the same answer.

If you have any questions on this matter, you are expected to consult me directly for advice.

Emergency preparedness: In the event of an emergency, students should refer to the AU Web site http://www.american.edu/emergency and the AU information line at (202) 885-1100 for general university-wide information. In case of a prolonged closure of the University, I send updates to you by email and will post all announcements on Blackboard.

Support services: A wide range of services is available to support you in your efforts to meet the course requirements.

1. Mathematics & Statistics Tutoring Lab (Don Myers Building) provides tutoring in Intermediate Mathematics and Statistics. http://www.american.edu/cas/mathstat/tutoring.cfm

2. Academic Support and Access Center offers study skills workshops, individual instruction, tutor referrals, Supplemental Instruction, writing support, and technical and practical support and assistance with accommodations for students with physical, medical, or psychological disabilities. Writing support is also available in the Writing Center, Battelle-Tompkins 228.

3. Center for Diversity & Inclusion (X3651, MGC 201) is dedicated to enhancing LGBTQ, Multicultural, First Generation, and Women's experiences on campus and to advance AU's commitment to respecting & valuing diversity by serving as a resource and liaison to students, staff, and faculty on issues of equity through education, outreach, and advocacy.

4. The Office of Advocacy Services for Interpersonal and Sexual Violence (X7070) provides free and confidential advocacy services for anyone in the campus community who is impacted by sexual violence (sexual assault, dating or domestic violence, and stalking).

Additional notes:

1. I expect you to be courteous to me and your fellow classmates during the online lectures/meetings.

2. Please let me know during the first week of classes if you have any special needs that require accommodations.

Project:
The main goals of the project is to prepare students to apply NLP to real-world tasks, or to leave them well-qualified to start NLP related research. The project will have two parts. The **poster presentation** as well as the **project proposal**. The due date for the project proposal is **TBD**. Please submit this proposal on Blackboard by that time. Your proposal should be **well-written**, **well-organized**, and **reproducible** following the current standards of machine learning reproducibility in research https://arxiv.org/abs/2003.12206. This will ensure evidence of the correctness of your results and will enable other researchers to make use of your methods and results. The class projects will be presented as a poster presentation. You should prepare a poster, and be prepared to give a very short explanation (10 minutes), about your work. You will also have an opportunity to see what everyone else did for their projects. You have to submit your project presentation by **TBD**. You will also need to submit your poster as a PDF the day before the presentation.

1. Project topics: If you are looking for project ideas, please talk to me early enough, and I will be happy to brainstorm and suggest some project ideas. There are three type of projects that you can pick and these include:

- Application project. Pick an application that interests you, and explore how best to apply learning algorithms to solve it.

- Algorithmic project. Pick a problem or family of problems, and try to develop a novel variant of an existing algorithm, to solve it.

2. Evaluation: Projects will be evaluated based on:

- The technical quality of the work. Does the technical material make sense? Are the things tried reasonable? Are the proposed algorithms or applications clever and interesting?

- Significance. Did the authors choose an interesting or a "real" problem to work on, or only a small ?toy" problem? Is this work likely to be useful and/or have impact?

- The novelty of the work. Is this project applying a common technique to a well-studied problem, or is the problem or method relatively unexplored?

3. Project proposals: The project proposal is mainly intended to make sure you decide on a project topic and get feedback from me early. Your proposal should be a PDF document. Please make sure that you give the title of the project, the project category, the full names of all of your team members, and a 300-500 word description of what you planning to work on. Your project proposal should include the following information:

- Motivation: What problem are you tackling? Is this an application or a theoretical result? What makes this problem interesting and important?

- Method: What NLP techniques are you planning to apply or improve upon? Note that you have to only provide high level technical details of the techniques.

- Intended experiments: What experiments are you planning to run? How do you plan to evaluate your NLP algorithm?

If you are planning to use a real world dataset, make sure that you are including all the references and links. In addition, present at least one example of prior research on the topic and include all the information of the papers that you are planning to use for your project. Use a standard format for your references (such as APA or MLA).

4. Poster presentations: Posters will be graded on the poster quality and clarity, the technical content of the poster, as well as the knowledge demonstrated by the team when discussing their work. For a poster example please check (https://sigport.org/sites/default/files/docs/MLSP_2019.pdf). For poster templates check (https://www.overleaf.com/learn/latex/Posters).