Introduction to Natural Language Processing

BUS 243F: Spring 2023
Thursday 9:35 am – 12:25 pm

Instructor: Yeabin Moon (yeabinmoon@brandeis.edu)

Office Hours: After class in my office (*Sachar International Center 209B*), or by appointment online through the following link.

TA: Jiawei Fan (jiaweifan@brandeis.edu)
TA Hours: 2:00 pm – 3:30 pm Friday, through zoom.

Course Description

Natural language processing (NLP) is becoming increasingly widespread. Applications of NLP have become embedded in our everyday lives, and these applications are based somewhere between formal linguistics and statistical physics. Especially over the past decade, neural network approaches have become the de facto standard for many NLP tasks. This course aims to provide a survey of these foundations, but we will take NLP in a narrow sense to cover the text analysis only. The course assumes a background in multivariate calculus, linear algebra, and proficiency in Python. The goal of this course is to enable you to build your language applications using the Python framework.

Success in this two-credit course is based on the expectation that students would need to study for about three hours for every hour of in-class time. Hence, students will spend a *minimum of 9 hours* of study time per week in preparation for this class.

Learning Goals

With this course, you will

1. Study the concepts from NLP and thy it is challenging
2. Understand the numerical representations of natural language
3. Examine data structures and a range of algorithms used in NLP
4. Dig into natural language libraries such as NLTK and Gensim

Main Reference

We will use ***Natural Language Processing in Action*** by Hannes Hapke, Hobson Lane and Cole Howard (Manning, 2019) as a main reference. There is a partial list of useful books that will be touched during the course.

- Steven Bird, Ewan Klein, Edward Loper, *Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit*
- Sowmya Vajjala, Bodhisattwa Majumder, Anuj Gupta, Harshit Surana, *Practical Natural Language Processing*

You have an **online access** for all the references listed above through Brandeis Library. Other useful reference is *Introduction to Natural Language Processing* by Jacob Eisenstein for avid students of mathematical exposition.

Prerequisites

1. Competency in Python (**Bus215f**)
   - All class exercises will be using Python. You should be familiar with NumPy, pandas and data structures in Python. Note that you should be fine if you have ample experience in coding with a different language.
2. Calculus, Linear Algebra, Probability, and Statistics (**Econ213a**)
   - You should know college-level calculus and the basics of probabilities
3. Machine Learning (***recommended***)
   - If you have basic machine learning or deep learning experience, the course would be much easier. *You can take it without knowing them*. If you need a top-bottom textbook treatment, I highly recommend: "*Hands-on machine learning with scikit-learn and TensorFlow*" by Geron Aurelien

Class Participation

There is no such thing as a stupid question. Dialogue is not only strongly encouraged, it is critical to your understanding of the material. Vocalizing your questions often helps you solidify what you do and do not understand. It also provides me important feedback on the areas in which we need to spend more time. During lectures, I will encourage questions, and I will solicit input. If I call on you, please relax, I am NOT trying to intimidate you or embarrass you in any way. I am trying to encourage active listening and keep you engaged in the course. This will greatly assist you in learning the material. If you do not know the answer, I will move on to another student. Hence, attendance is *mandatory* for this class.

Course Requirements

First, there are three assignments that will promote both your theoretical understanding and practical skills. All assignments contain both written parts and programming parts. Second, there are weekly in-class quizzes except for the first class. The readings are mandatory for the courses, and each quiz will test the least comprehension of the reading materials. Finally, there will be one final exam.

You can submit your late work with a 10 percent penalty if you can make it within ONE day after the deadline.

The grade consists of

1. Three assignments: 60 % (15% each for the first two, 30% for the last one)
2. Five in-class quizzes: 35 %
3. Participation / Attendance: 5 %

Course Plan

The class covers the basic building blocks used in NLP. We will mainly examine the practical use cases and delve into theories where necessary. Each week will be dedicated to one concept. However, some additional concepts would be introduced due to the compact module-class structure. The following outline provides a high-level overview of the course. The mandatory readings are indicated by *, and you must be prepared before the class meeting. The additional readings will be posted on the latte one week in advance. The in-class quizzes primarily focus on the weekly readings.

1. **Introduction (March 16)**
   (a) The foundations of NLP: Learning and Search
      - Chapter 1 *
      - NLP application overview
      - Structure of NLP applications

   (b) Python and Math Reviews
      - Google Colab
      - Object Oriented Programming
      - Regular Expressions
      - Bayes in math

2. **Terms, Document, and Corpus (March 23)**
   (a) Introduction to Text Representation
      - Chapter 2 *

   (b) Token, document, and corpus
      - Introduction to Matrix representation
      - One-hot encoding revisit
      - Effect of Tokenizer

   (c) Text Normalization
   (d) Text classification exercise
      - Naïve Bayes

Assignment 1 deadline (11:59 AM, March 27)

3. **Text Representation (March 30)**
   (a) Count based representation
      - Chapter 3 *

   (b) Discussion: how to represent a large text data
      - Park, M., Leahey, E. & Funk, R.J. Papers and patents are becoming less disruptive over time. Nature 613, 138–144 (2023). * https://doi.org/10.1038/s41586-022-05543-x

   (c) Probabilistic classification
      - Introduction to Language model
        i. Zipf's Law

ii. Estimate Zipf's law coefficients

(d) Discussion on TF-IDF

- Introduction to Information retrieval
- Semantic analysis approach

Spring break

4. Semantic Analysis 1 (April 20)

(a) Meaning in words

- Chapter 4 *

(b) Understanding Dimensionality Reduction

- Linear Algebra application
- Introduction to Principal Component Analysis

(c) Discussion on Topic vectors

- Feature extraction
- Hard discussion on semantic analysis
  - Mere representation?

(d) Sentiment analysis revisit

- Theory vs. Practice

Assignment 2 deadline (11:59 AM, April 29)

5. Semantic Analysis and Introduction to Neural networks (April 27)

(a) Text representation with topic vectors

- Assess information loss
- Topic modeling in practice
  - Latent Dirichlet allocation

(b) Neural network sketch

- Chapter 5 *
- XOR problems
- Deep learning structure and Backprop

(c) Modern representation of text: Language model

- Discussion on input vector representations
- Distributional semantics

(d) Gensim in practice

- Introduction to Pretrained text data
- End-to-end guide
- Don't try this at home?

6. Modern Text Analysis (May 2, Brandeis Days)

(a) Introduction to Word2vec

- Chapter 6 and 7 *

- Soft discussion
    i. Mikolov, T., Chen, K., Corrado, G. and Dean, J., 2013. *Efficient estimation of word representations in vector space* *
    ii. Negative sampling

(b) Gensim word vectors revisit

- Glove: Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "*Glove: Global vectors for word representation*." Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014.
    i. SVD revisit

(c) Word2vec vs. LSA

- modern NLP model pipeline

Assignment 3 deadline (11:59 AM, May 15)

The course plan is subject to change due to a snow day/delayed start/early closing. If this situation happens, the class will be held on zoom. I will announce it accordingly.

Accommodations

Brandeis seeks to create a learning environment that is welcoming and inclusive of all students, and I want to support you in your learning. Live auto transcription is available for all meetings or classes hosted on Zoom and you can turn it on or off to support your learning. Please check for Zoom updates to take advantage of this new feature. To learn more, visit the Zoom Live Transcription webpage. For questions, contact help@brandeis.edu

If you think you may require disability accommodations, you will need to work with Student Accessibility Support (SAS) (781-736-3470, access@brandeis.edu). You can find helpful student FAQs and other resources on the SAS website, including guidance on how to know whether you might be eligible for support from SAS. If you already have an accommodation letter from SAS, please provide me with a copy as soon as you can so that I can ensure effective implementation of accommodations for this class.

Academic Integrity

Every member of the University community is expected to maintain the highest standards of academic integrity. A student shall not submit work that is falsified or is not the result of the student's own effort. Infringement of academic integrity by a student subjects that student to serious penalties, which may include failure on the assignment, failure in the course, suspension from the University or other sanctions. Please consult Brandeis University Rights and Responsibilities for all policies and procedures related to academic integrity. Students may be required to submit work via TurnItIn.com or similar software to verify originality. A student who is in doubt regarding standards of academic integrity as they apply to a specific course or assignment should consult the faculty member responsible for that course or assignment before submitting the work. Allegations of alleged academic dishonesty will

be forwarded to the Department of Student Rights and Community Standards. Citation and research assistance can be found at Brandeis Library Guides - Citing Sources.

Classroom Health and Safety

- Register for the Brandeis Emergency Notification System. Students who receive an emergency notification while attending class should notify their instructor immediately. In the case of a life-threatening emergency, call 911. As a precaution, review this active shooter information sheet.
- Brandeis provides this shuttle service for traveling across campus or to downtown Waltham, Cambridge and Boston.
- On the Brandeis campus, all students, faculty, staff and guests are required to observe the university's policies on physical distancing and mask-wearing to support the health and safety of all classroom participants. Review up to date COVID-related health and safety policies regularly

Student Support

Brandeis University is committed to supporting all our students so they can thrive. If you want to learn more about support resources, the Support at Brandeis webpage offers a comprehensive list that includes these staff colleagues you can consult, along with other support resources:

- The Care Team
- Academic Services (undergraduate)
- Graduate Student Affairs
- Directors of Graduate Studies in each department, School of Arts & Sciences
- Program Administrators for the Heller School and International Business School
- University Ombuds
- Office of Equal Opportunity