**COSI 114b: Fundamentals of NLP I**  
*Last updated: 8/27/2020*

**Instructor**

Prof. Constantine Lignos (he/him/his),  [lignos@brandeis.edu](mailto:lignos@brandeis.edu), office hours TBD (held via Zoom)

The Zoom link for class will be available on LATTE shortly before the first class session. You are welcome to “shop” the first class if you are not enrolled, but you will need to email the instructor for the link.

---

**High-Level Course Summary**

**What will I learn from this course?**

You’ll learn fundamental concepts in the sciences of natural language processing (NLP) and computational linguistics (CL), specifically how to represent language quantitatively and implement probabilistic models for common NLP tasks (classification, tagging, etc.) from scratch in Python.

After taking this course, you’ll be prepared to take more advanced courses in NLP/CL. COSI 114a and 114b form a two-course introductory sequence. If you’re not interested in further NLP/CL courses, you may want to wait for a potential future COSI course focusing on interdisciplinary applications of NLP, which will have fewer prerequisites (like the discontinued LING 131a).

You won’t learn about deep learning for NLP (covered in COSI 114b, 134a, and 137b), how to use NLP toolkits (e.g., spaCy, NLTK), machine learning toolkits (e.g. scikit-learn), deep learning frameworks (e.g., PyTorch, TensorFlow), or how to apply NLP at scale (“big data,” analytics, etc.).

**What background do I need to take this course?**

See the “Course Information” section. **If you do not meet the prerequisites, do not enroll in the class without emailing the instructor.**

**Can I take the course if I can’t attend during class/recitation times?**

Yes. See “Modality” in the “Course Information” section.

**How does this compare to LING 131a (discontinued) and COSI 114b?**

**COSI 114a is a new course** that covers some of the same material as the discontinued course LING 131a but at greater technical depth. The focus on probabilistic models and implementing concepts from scratch. COSI 114b is a more advanced course that you should take after 114a. You should not take this course if you have already taken LING 131b or COSI 114b.
Description and Objectives

This course explores the computational properties of natural language and the foundations of the algorithms used to process it. Students will develop an understanding of basic natural language processing (NLP) methods by implementing language analysis and classification algorithms in Python. Topics will include corpus statistics, text classification, language modeling, and the computational techniques needed to support these tasks such as finite-state methods and generative models (e.g., naive Bayes, hidden Markov models).

The coursework consists of regular readings, homework assignments, and quizzes. Attendance at lecture and the recitation sections are an essential part of learning. Excessive absences will result in a reduced engagement grade. **However, students unable to attend due to time zone differences will be excused from attending synchronously.** Please contact the instructor one week in advance if you plan to miss class or recitation due to an athletic, religious, or other planned event. If your wellness or unforeseen circumstances prevent you from coming to class, please inform the instructor as soon as is feasible to make a plan to make up any missed work.

Success in this four-credit course is based on the expectation that students will spend a minimum of nine hours of study time per week in preparation for class (readings, homework, etc.).

Learning Goals

Upon completion of the course, students will understand the fundamentals of how natural language can be represented using computational structures. Students will be able to use Python to implement the tools necessary to classify text using statistical NLP methods, specifically generative models such as naive Bayes and hidden Markov models. Students will be able to use existing libraries, such as finite state toolkits, to perform additional NLP tasks.

Course Information

**Prerequisites** may be met by satisfying one of following three conditions:

1. Enrollment in the MS in Computational Linguistics program or MS in Computer Science (MS3) program. (Note that this does not include students in the MS in Computer Science for Non-Majors (MS4) program, who must qualify under one of the requirements below.)
2. Taking COSI 10a (introductory Python programming) and LING 100a.
3. Taking COSI 12b (intermediate-level programming) and having Python programming experience.

If you do not meet the prerequisites, you should not enroll in the course and may be administratively removed from the course if you do not receive instructor permission. Instructor permission will be granted only in cases where it is clear that the student has **mastered** the prerequisite material in another way. If you are a transfer student or graduate student and have completed coursework at another university equivalent to the prerequisites, please email the instructor indicating the relevant courses and attach an unofficial transcript showing that you completed them.
**Textbook:** No textbook purchase is required. Readings will be linked or distributed through LATTE. Many readings will come from *Jurafsky and Martin, Speech and Language Processing (3rd edition)* (the 3rd edition is unpublished but available online).

**Modality:** This course will be offered remotely for fall 2020. **Class and recitation sections will be held via Zoom at the time specified in the course listing. Based on demand we may make additional recitation times available to support students in more time zones.** We will make any necessary accommodations for students who are unable to synchronously attend the class and/or recitation times due to time zone differences. Recordings of class sessions will be made available shortly after each class session. There is no requirement to ever attend class sessions synchronously, but synchronous attendance is strongly encouraged.

**Technology needs:** Attending class and recitation sessions will require connecting to Zoom and using breakout rooms (see [Zoom’s supported platforms for breakout rooms](https://zoom.us/system_requirements)). Completing homework assignments will require access to a computer capable of running a Python editor and a terminal running Bash (installed on many systems but also available via Git Bash) and connecting to Brandeis COSI department machines via SSH. The teaching staff will provide instructions on LATTE for using these tools and support students in using them. A computer running MacOS or Windows is recommended, but we will do our best to accommodate students who only have access to computers running ChromeOS.

**Communication:** [LATTE](https://latte.brandeis.edu) is the Brandeis learning management system. Login using your UNET ID and password. Further information about the course can be found there once the course begins, including a detailed schedule, readings, homework assignments, and additional policies. Students must regularly check LATTE and read and follow all policies posted there. All course announcements will be made using LATTE. The “Contacting the Course Staff” page on LATTE provides information regarding the best way to contact the instructor or TAs and the expected response times.

**Office hours:** Office hours will be conducted using Zoom and additional synchronous communication tools such as Slack. See LATTE for times, relevant links, and other information.

**Class recording:** Class sessions will be recorded for educational purposes. You may decline to be recorded; if so, please contact the instructor to identify suitable alternatives for class participation. These recordings will be deleted within two months after the end of the semester. If you can be personally identified in a recording, no other use is permitted without your formal permission. You may not record classes on your own without the instructor’s express permission, and may not share the URL and/or password to anyone not enrolled in the course. Your behavior in these recordings, and in this class as a whole (including online forums such as LATTE and Slack), must fulfill Brandeis standards:

> Brandeis University is committed to providing its students, faculty and staff with an environment conducive to learning and working, where all people are treated with respect and dignity ([Brandeis Business Conduct Policy](https://brandeis.edu/business/conduct/policy.html) p. 2, 2020).

> You must refrain from any behavior toward members of our Brandeis community, including students, faculty, staff, and guests, that intimidates, threatens, harasses, or bullies ([Student Rights & Responsibilities](https://brandeis.edu/studentrights.html), p. 11, 2020 ed.).
Evaluation

Homework Assignments: 65%
Quizzes: 30%
Engagement: 5% (attendance and contribution to discussion, extra credit can be given)

Homework assignments and quizzes will be spread across the semester, with the final homework assignment due the final week of classes. Homework assignments must be submitted following the instructions on LATTE. Homework may be turned in late following the “Late Homework” policy on LATTE which specifies how “grace days” can be used to turn in homework late without a grade penalty.

There will be no midterm or final exams. Midterm grades will be provided by October 8th, 2020 reflecting homework assignments and quizzes turned in up to that point.

Course Plan

The following is a rough outline of the topics covered in this course:

1. Basics of working with text in Python
2. Characterizing language data using corpus statistics
3. Computing and working with probabilities
4. Finite-state automata
5. Word-based language models
6. Text classification basics
7. Using naive Bayes for classification
8. Part of speech (POS) tagging and sequence models
9. Implementing a POS tagger using hidden Markov models
10. Finite state transducers
11. Implementing spelling correction using finite state transducers
12. Finite state morphology
13. Distributed representations and word embeddings

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 honesty 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 (see section 20 of R&R). Please consult Brandeis University Rights and Responsibilities for all policies and procedures related to academic integrity. Homework assignments will be analyzed using software to verify originality. A student who is in doubt regarding standards of academic honesty as they apply to a specific course or assignment should consult the instructor 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.
Accommodations

Brandeis seeks to welcome and include all students. If you are a student who needs accommodations as outlined in an accommodations letter, please talk to the instructor and present your letter of accommodation as soon as you can. In order to provide quiz accommodations, accommodation letters must be received more than 48 hours in advance.

Accommodations cannot be provided retroactively. If you have questions about documenting a disability or requesting accommodations, please contact Student Accessibility Support at 781-736-3470 or access@brandeis.edu.

Privacy

This class requires the use of tools (e.g., GitHub, source code analysis tools) that may disclose your coursework and identity to parties outside the class. To protect your privacy, you may choose to use a pseudonym/alias rather than your name in submitting such work. You must share the pseudonym/alias with the instructor and teaching assistants as needed. Alternatively, with prior consultation, you may submit such work directly to the instructor.

Student Support

Brandeis University is committed to supporting all our students so they can thrive. The following resources are available to help with the many academic and non-academic factors that contribute to student success (finances, health, food supply, housing, mental health counseling, academic advising, physical and social activities, etc.). Please explore the many links on the Support at Brandeis page to find out more about the resources that Brandeis provides to help students achieve success.