Introduction to Computational Linguistics
Course Code LIN 4930/LIN 6932, Spring 2020
Instructor: Kevin Tang

January 6, 2020

General Information

Course description

Computational linguistics is the study of natural language from a computational perspective. It encompasses both applied (engineering) and theoretical (cognitive) issues, ranging from speech and language technology to formal aspects of theoretical linguistic models.

Have you ever wondered how your GPS can pronounce street names and how you can pronounce new words? Have you wondered how Amazon can process billions of reviews? Ever wished to automatically process large corpora (big data), and discover linguistic structures therein? Do you want to model our linguistic intuitions of grammaticality? Then this course is for you!

In this class, we will survey various topics and tasks in computational linguistics. While we will cover some of the basics of Natural Language Processing (which we will consider a separate subfield), this class will not focus on one specific approach (such as deep learning). Students in this class are expected to have a background in either computer science or linguistics, but not necessarily both. Expect this class to be difficult at times and easy at others. We hope to offer something new and interesting for everyone.

Objectives

On completion of this course, you should:

- Be familiar with computational linguistic topics, tools, and resources, and how they are applied in research in both computational linguistics and other subfields
- Have a rough sense of the state of the art in this subfield
- Be able to conceptualize problems from the perspective of computational linguistics

Time and place

Where: AND 0134 (Anderson Hall)

When: Monday, Wednesday and Friday: 12:50–13:40 (Period 6)

Instructor information

Instructor: Kevin Tang

Email:

- tang.kevin@ufl.edu

Offices:

- 4017 Turlington Hall, Gainesville, FL 32611-5454

Office hours:

- Fridays 13:55–15:50 or by appointment.
- For information on what are office hours and how to make use of them properly see: http://lsc.cornell.edu/wp-content/uploads/2015/10/What-Are-Office-Hours.pdf.

Requirements

Prerequisites

LIN3010 (Introduction to Linguistics).

Note on prior programming experience: Programming is not the focus of this course, but knowing how to program is an essential skill needed to do computational linguistics. **The first one/two weeks will cover a brief introduction to programming in Python, with a particular focus on learning how to process written text. In addition, I expect you to learn Python using the Python textbook and online resources by yourself.** This introduction will provide you with some of the tools needed to tackle subsequent programming assignments, which will involve implementing and analyzing language models of increasing complexity. Because this programming section will move very quickly and not comprehensive and programming assignments will be complex, prior programming experience is preferred. Moreover, the course will also include a survey of Python Natural Language Toolkit (NLTK).

If you don't have programming experience but are still interested in taking the course, please talk to me.

Course website

- There is a Canvas page for this course.
- Course url: https://ufl.instructure.com/courses/393317 or https://ufl.instructure.com/courses/395015 or find it under 'Courses'
- The name of the course: LIN 4930 or LIN 6932
- Let me know if you are not on the site.

Textbook and Reading list

Main textbook (Second edition):

- Daniel Jurafsky and James H. Martin (Sept. 2008). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (2nd edition)*. Prentice Hall. ISBN: 0131873210
- Make sure you get the second edition! Within the second edition, the hardcover "US edition" is preferred, but you can also buy the "international edition" (which might be cheaper). There's also an ebook version of the second edition (which again might be cheaper). They differ mostly only in a few exercises at the end of each chapter.
- Jurafsky and Martin are in the process of producing a new edition of this textbook, and we will also make use of draft chapters from the revised version. These are available at https://web.stanford.edu/~jurafsky/slp3/

Supplementary textbook (Freely available online):

- Steven Bird et al. (2009). *Natural language processing with Python: analyzing text*

with the natural language toolkit. O'Reilly Media, Inc
- Freely available online: https://www.nltk.org/book/
- Dickinson, Brew, and Meuers (2013), *Language and Computers*
- Available for reading online through the UF library (and download up to 51 pages)

Python textbook (Available at UF):

- Magnus Lie Hetland (2008). *Beginning Python from Novice to Professional*. Wiley
- Available for download as an ebook through the UF library
- Additionally, there are numerous online resources for python, and you may find them a convenient supplement to the Hetland book, especially if you're an experienced programmer. https://docs.python.org/3/tutorial/, and https://www.codecademy.com/learn/learn-python If you have a specific problem or question, Stack Overflow is a good place to look for answers: stackoverflow.com. I recommend searching the site for your question before asking it yourself.
- Readings from the textbooks will be supplemented by other readings and materials throughout the semester (made available on the Canvas website for the course).

Course requirements (tentative, and subject to revision)

1. **Lectures and Reading:** The content of the course will be presented through lectures and reading assignments. You should attend lectures and complete the reading assignment by the date of the lecture with which the reading is associated. The material in the lectures and readings will not be identical, so you will need to both attend lectures and do the readings to succeed. Homework and quizzes will presume familiarity with both the readings and the
2. **Homework assignments:** There will be a number of homework assignments (on the order of 3 or 4) in this course, which will involve considerable programming. These assignments will require you to implement algorithms from computational linguistics, test them on data sets, and suggest and explore potential improvements. I will ask that you do your programming for these assignments in Python: not only is this the language that you used for the pre-requisite to this course, but in addition there are a number of tools for doing this work that you will find useful.
3. **Lab exercises**: Over the course, there will be a number of lab sessions. They will test your knowledge of the material we discuss in class.
4. **Final Project**: You will work in a group of 2-3 people on a final project that builds in some way on the material we cover in class and/or connects with related literature. The amount of work that I expect from you will, of course, vary by the number of people who contribute, but you may find it helpful and interesting to work with others who have backgrounds different from yours. The range of allowable topics is very flexible: it may relate to a personal passion of your (e.g., comic books) or to an area of research. I would encourage you to try to find a topic that inspires you (and your partners), since this is something that you will be spending a good deal of time on.
   Once you have identified a topic with your group, you should schedule an appointment with me to finalize the idea. Your project must include a computational implementation of some sort, and must be presented in a written report that explains what you have done, how it relates to past work, and what you have learned from your results. You should also explain the contributions of each of the participants (if you

are working in a group of 2 or 3). There will also be a group presentation component.

Grading (tentative, and subject to revision)

Please do not negotiate with me about your grade. For instance, if there is anything extra you could do to improve your grade, e.g. extra credits; if I could add points to any of graded components; if I could talk to you about the points deducted.

- Class participation: 10%
  - Active in-class participation is a requirement of the course. 'Active participation' means that you should regularly ask thoughtful questions in class during lectures and tutorials, and participate with the group exercises during tutorials. If you are habitually absent from class, leave early without letting us know ahead of time, or are otherwise disengaged (e.g. on your smartphone), that will negatively affect your participation grade.
  - If you miss a class, it is up to you to borrow notes from someone, ask other students about changes to the reading/homework schedule, etc. Please don't ask me to go over what we did in class.

- Homework (approx. 4-5 pieces): 45%
  - Assignments will be uploaded to the Canvas course website.
  - Assignments must be submitted via Canvas.
  - Late assignments will NOT be accepted, except under extreme circumstances.
  - Emailed assignments will NOT be accepted, except under extreme circumstances.
  - In general, I will distribute the assignments one week before they are due, in class and/or on e-Learning.
  - I will use automatic checks for overlap between your code and other students' code.
- Lab work: 15%
  - Completion of lab exercises. You should submit the answers of the lab exercises. Generally, the deadline is a week after each of the lab sessions.

- Final project: 30%
  - Project presentations/demos: you are also required to present your work as a group. The exact format is to be determined. It will likely be a poster session open to the public.
  - Submission of an individual write-up of your final project (max. 10 pages).
  - Graduate students will be graded more vigorously.

Grading scale

A: 92-100, A-: 88-91.9,
B+ 85-87.9, B: 81-84.9, B-: 78-80.9,
C+: 75-77.9, C: 71-74.9, C-: 68-70.9,
D+: 65-67.9, D: 61-64.9, D-: 58-60.9,
E: Below 58

Expectations

I expect everyone to come to class and be actively engaged. I am confident that you will find it easier to master the course material by hearing it presented and also by asking questions when you don't understand something. I do not wish to see you being distracted by social media, email, and the web, therefore please avoid using your laptop, smartphone, iPad, or the like during class, except if they are needed for a class activity, such as note-taking.

Any evidence of plagiarism on problem sets will result in disciplinary penalties. In this course specifically, I expect you to do your own programming for the homework assignments. You will not learn anything if you simply copy and submit a classmate's code or code you find on the internet as your own. However, if you are stuck on a programming problem or a non-programming part of an assignment, you are free, and indeed encouraged, to consult with your classmates (or with resources on the web) about the problems you are having. Sharing ideas with others is extraordinarily helpful in figuring things out, and understanding a topic more deeply (both for the question asker and answerer). Once you have finished your discussions, you must write up your own code and answers, and the product you turn in should represent your work alone and not something copied from the work of your classmate. You should also note on each your assignments who or which internet resources you have consulted with. On the final project, you may work freely with the members of your group, though again I expect you to give credit to any other resources you consult.

Academic honesty

You are required to abide by the Student Honor Code. Any violation of the academic integrity expected of you will result in a minimum academic sanction of failing grade on the assignment or assessment. Any alleged violations of the Student Honor Code will result in a referral to Student Conduct and Conflict Resolution. Please review the Student Honor Code and Student Conduct Code at sccr.dso.ufl.edu/policies/student-honor-code-student-conduct-code/.

Students should be aware of their faculty's policy on collaboration, should understand how to properly cite sources, and should not give nor receive an improper academic advantage in any manner through any medium. No student may work or collaborate with another person on any academic activity in this course. Should group work be assigned or this class policy change, I will provide that in writing on the individual assignment instructions.

Remember you are bound by the UF Honor Pledge:

> We, the members of the University of Florida community, pledge to hold ourselves and our peers to the highest standards of honesty and integrity by abiding by the Student Honor Code. On all work submitted for credit by Students at the University of Florida, the following pledge is either required or implied: "On my honor, I have neither given nor received unauthorized aid in doing this assignment."

Evaluation statement

"Students are expected to provide professional and respectful feedback on the quality of instruction in this course by completing course evaluations online via GatorEvals. Guidance on how to give feedback in a professional and respectful manner is available at https://gatorevals.aa.ufl.edu/students/. Students will be notified when the evaluation period opens, and can complete evaluations through the email they receive from GatorEvals, in their

Canvas course menu under GatorEvals, or via https://ufl.bluera.com/ufl/. Summaries of course evaluation results are available to students at https://gatorevals.aa.ufl.edu/public-results/."

Accommodation

Students requesting classroom accommodation must first register with the Dean of Students Office. That office will provide documentation for me, so that requests for accommodation can be honored. Please do this as early in the term as possible.

Health and Wellness

If you or a friend is in distress, please contact umatter@ufl.edu or 352-392-1575 so that a U Matter We Care team member can reach out to the student in distress.

Course outline (tentative, and subject to revision)

Readings should be completed **before** the class date listed.

Acronyms:

- J&M: Jurafsky and Martin (2008)
- J&M(3ed): Jurafsky and Martin (3ed)
- L&C: Language and Computer (2013)
- H: Hetland (2008)
- BKL: Bird et al (2009)

Week	Date	Topic	Reading	Assignment
W1	Jan 6	Introduction and Syllabus	J&M: Ch 1 (L&C Ch 1)	
	Jan 8	Python 1: Intro	H: Ch 1	
	Jan 10	Python 2: Data types & Files	H: Ch 2-4, (11)	
W2	Jan 13	Python 3: Control & Reg Exp	H: Ch 5,10 (pp.242-258), J&M(3ed): Ch 2	
	Jan 15	Text Normalization	J&M(3ed): Ch 2	
	Jan 17	Lab: Text Processing	BKL: Ch 1,2,3	
W3	Jan 20	Holiday		
	Jan 22	Edit Distance	J&M(3ed): Ch 2	
	Jan 24	The Noisy Channel	J&M(3ed) Appendix B, L&C: Ch2	HW 1 set
W4	Jan 27	N-Grams	L&M: Ch 4	
	Jan 29	N-Grams	L&M: Ch 4	
	Jan 31	N-Grams	L&M: Ch 4	
W5	Feb 3	Lab: N-Grams		HW 1 due
	Feb 5	Machine Learning: Overview	T. Mitchel. (2017)	HW 2 set
	Feb 7	Evaluation and Error analysis	Resnik & Lin (2010), Kummerfeld et al. (2012)	
W6	Feb 10	Regression and Maximum Entropy	J&M: Ch 6.6-6.7, J&M(3rd): Ch 5	
	Feb 12	Regression and Maximum Entropy	J&M: Ch 6.6-6.7, J&M(3rd): Ch 5	
	Feb 14	Regression and Maximum Entropy	J&M: Ch 6.6-6.7, J&M(3rd): Ch 5	HW 2 Due

	Feb 17	Lab: Regression		HW 3 Set
W7	Feb 19	Part of Speech tagging	J&M: Ch 5, J&M(3rd): Ch 8	
	Feb 21	Part of Speech tagging	J&M: Ch 5, J&M(3rd): Ch 8	
	Feb 24	Part of Speech tagging	J&M: Ch 5, J&M(3rd): Ch 8	
W8	Feb 26	Lab: Tagging	BKL: Ch 5	HW 3 Due
	Feb 28	Vector Semantics	J&M(3rd): Ch 6	HW 4 Set
	March 2	Spring Break		
W9	March 4	Spring Break		
	March 6	Spring Break		
	March 9	Vector Semantics	J&M(3rd): Ch 6	
W10	March 11	Vector Semantics	J&M(3rd): Ch 6	HW 4 Due
	March 13	Lab: Semantics		HW 5 Set
	March 16	Naive Bayes Classification	J&M(3rd): Ch 4	
W11	March 18	Naive Bayes Classification	J&M(3rd): Ch 4	
	March 20	Sentiment Analysis	J&M(3rd): Ch 4	
	March 23	Sentiment, Affect, and Connotation	J&M(3rd): Ch 21	HW 5 Due
W12	March 25	Lab: Sentiment Analysis		
	March 27	Forensic linguistics: Authorship		
	March 30	Forensic linguistics: Profiling		
W13	Apr 1	Lab: Forensics		
	Apr 3	Modelling Grammar		
	Apr 6	Modelling Intuitions		
W14	Apr 8	Modelling Processing		
	Apr 10	Lab: Modelling		
	Apr 13	TBD		
W15	Apr 15	TBD		
	Apr 17	TBD		
	Apr 20	Guest Lecture (TBD)		
W16	Apr 22	Poster session		

References

Bird, Steven, Ewan Klein, and Edward Loper (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Hetland, Magnus Lie (2008). *Beginning Python from Novice to Professional*. Wiley.

Jurafsky, Daniel and James H. Martin (Sept. 2008). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (2nd edition)*. Prentice Hall. ISBN: 0131873210.