Introduction to Computational Linguistics
Course Code LIN 4930/LIN 6932, Spring 2021
Instructor: Kevin Tang

December 18, 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.

This course is for YOU if you have ever wondered: How your phone can do predictive texting? How does a computer understand the syntax of a sentence? How Amazon can process billions of reviews? How to automatically process large corpora (big data)? How to discover linguistic structures from language data automatically? How to model our linguistic intuitions of grammaticality?

This course is for YOU if you want to build a foundation for areas such as data science, the tech industry https://www.linguisticsociety.org/resource/webinar-linguists-and-linguistics-tech or for applying for a Master's in Computational Linguistics.

This course is for YOU if you want to create a computational linguistics project like these ones by last year's students: https://slam.lin.ufl.edu/2020/04/20/intro2compling-symposium-spring2020/.

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.

Please also check out this page to learn more about computational linguistics at UF: https://slam.lin.ufl.edu/computational-linguistics/

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
- Be able to code in Python and use JyputerNotebook.

Time and place

Where: Virtual and MAT 0118

When: Monday, Wednesday and Friday: 13:55–14:45 (Period 7)

Instructor information

Instructor: Dr. Kevin Tang

The Speech, Lexicon and Modeling Lab – https://slam.lin.ufl.edu/

Email:

- tang.kevin@ufl.edu

Offices:

- 4017 Turlington Hall, Gainesville, FL 32611-5454

Office hours:

- TBD 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).

Programming training

Programming is not the focus of this course, but knowing how to program is an essential skill needed to do computational linguistics. I have engineered a number of ways to get you all up to speed with programming.

- **The first three weeks** will cover a brief introduction to programming in Python, with a particular focus on learning how to process written text, and JyputerNotebook (an interactive computational environment).
- **Two free online courses** will be assigned as assignments in a flipped class room format.
- After these first few weeks, you should have **a basic foundation of Python** to be able to tackle non-coding related course components.
- The course will include a complementary survey of the **Python Natural Language Toolkit (NLTK)** which lowers the need of coding for certain topics of the course.

Note that the programming training will not be entirely comprehensive as one can never finish learning to code. You need to be prepared to be **highly motivated** when it comes to learning how to code using Python. Learning to code is *not* like learning a language or learning to do math (see https://news.mit.edu/2020/brain-reading-computer-code-1215). This means you should complete all the assigned online courses and exercises, and make use of the class time, Canvas discussion forums, office hours, and any supplementary resources provided (e.g., the Python textbook).

To stay motivated, it is important to consider why should you learn to code and how coding can improve your career prospect and open doors. Here are very cool animated videos on the basic concepts of coding by Karlie Kloss and TedTalk-Ed.

- "Coding is a superpower": https://www.youtube.com/watch?v=Bwiln7v0fdc

- "Karlie Kloss Explains How Computers Work": https://www.youtube.com/watch?v=_zeV-xWSJoU
- "Variables, Functions, and Arrays with Karlie Kloss": https://www.youtube.com/watch?v=MXkGONNYKHo
- "Think like a coder" (10 episodes): https://youtu.be/KFVdHDMcepw

To understand your suitability, I encourage you to check out the two free online courses that we will assign as assignments during the first three weeks. As always, you can always **reach out to me at tang.kevin@ufl.edu to discuss your suitability**.

- Note: As a UF student, the courses below are free (see instructions here: https://elearning.ufl.edu/supported-services/linkedin-learning/)
- Course 1: Python Essential Training (estimated duration: 4h 45m) by Bill Weinman https://www.linkedin.com/learning/python-essential-training-2
- Course 2: Introducing Jupyter by Josh McQuiston (estimated duration: 53m) https://www.linkedin.com/learning/introducing-jupyter/

Course website

- There is a Canvas page for this course.
- Course url: TBD 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

Programming Web Resources (Available at UF):

- Note: As a UF student, the LinkedIn courses below are free (see instructions here: https://elearning.ufl.edu/supported-services/linkedin-learning/)
- Course 1: Python Essential Training (estimated duration: 4h 45m) by Bill Weinman https://www.linkedin.com/learning/python-essential-training-2
- Course 2: Introducing Jupyter by Josh McQuiston (estimated duration: 53m) https://www.linkedin.com/learning/introducing-jupyter/
- Course 3 (Optional, but recommended): https://www.linkedin.com/learning/python-functions-for-data-science/
- 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.
- Learning how to google specific questions to debug is highly recommended (http://letmegooglethat.com/)
- 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 lectures.
2. **Quizzes:** There will be a number of short quizzes in this course which are designed to test your conceptual knowledge of computational linguistics.
3. **Lab exercises**: There will be a number of lab sessions, which will involve programming. They will test your knowledge of the material we discuss in class – a combination of programming and conceptual knowledge. **You are encouraged to work in a group with these lab exercises.** In my experience, students benefit a lot from working with each other.
4. **Homework assignments:** There will be a number of homework assignments (on the order of 3 or 4) in this course, which will involve programming. **You are encouraged to work in a group with these lab exercises.** These assignments will require you to *implement algorithms* from computational linguistics, t*est them on data sets*, and suggest and *explore potential improvements*.
5. **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)

- 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.

- Quizzes: 10%
  - Quizzes will be available via Canvas. They are designed to test your on-going knowledge of the course content, specificially the conceptual content.

- Homework (approx. 4 pieces): 35%
  - 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.
  - Submit three files, unless mentioned otherwise (.ipynb (JyputerNotebook), .html (the output of JyputerNotebook) and .pdf (a pdf rendered version of the html). I will need the .pdf file for grading purposes (giving you comments directly in the document). If you do not submit all three files, I will not be able to grade properly and this will negatively affect your grade.
  - Length restrictions: be careful about your .html and .pdf files in terms of what is actually printed. If you should accidentally generate an excessive number of pages (e.g., you print out the whole dataset to inspect it, leading to a 300 page pdf file), I will not be able to grade properly and this will negatively affect your grade. Files submitted using the 'comment' feature will not be accepted.
  - Clarification: to do well in the HWs, you should examine the HW immediately, pay attention to the number of points assigned to each section, allocate a sufficient

amount of time per section, and ask for clarification of anything immediately via a Canvas Discussion thread. Typically, most of the points are allocated in later sections of the HWs, therefore you should persevere and do not give up.

- Lab work (approx. 5 pieces) : 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.
  - Submit three files, unless mentioned otherwise (.ipynb (JyputerNotebook), .html (the output of JyputerNotebook) and .pdf (a pdf rendered version of the html). I will need the .pdf file for grading purposes (giving you comments directly in the document). If you do not submit all three files, I will not be able to grade properly and this will negatively affect your grade. Files submitted using the 'comment' feature will not be accepted.
  - Length restrictions: be careful about your .html and .pdf files in terms of what is actually printed. If you should accidentally generate an excessive number of pages (e.g., you print out the whole dataset to inspect it, leading to a 300 page pdf file), I will not be able to grade properly and this will negatively affect your grade.
  - Clarification: to do well in the lab exercises, you should work on the lab exercises as much as you can in our lab sessions, and ask for clarification of anything during class and via a Canvas Discussion thread.

- 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 – submitted individually.
  - Submission of the group project files (the data, codes and presentation files) – submitted by your team representative.
  - 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		Introduction and Syllabus	J&M: Ch 1 (L&C Ch 1)	
		Python: Intro	LinkedIn Learning	
		Python: Intro	LinkedIn Learning	
W2		Holiday		
		Python: Intro	LinkedIn Learning	
		Python: Intro	LinkedIn Learning	
W3		Python: Data types & Files	H: Ch 2-4, (11)	
		Python: Control & Reg Exp	H: Ch 5,10 (pp.242-258), J&M(3ed): Ch 2	
		Text Normalization	J&M(3ed): Ch 2	
W4		Lab: Text Processing	BKL: Ch 1,2,3	
		Edit Distance	J&M(3ed): Ch 2	
		The Noisy Channel	J&M(3ed) Appendix B, L&C: Ch2	
W5		The Noisy Channel	J&M(3ed) Appendix B, L&C: Ch 2	
		N-Grams	J&M: Ch 4	HW 1 set
		N-Grams	J&M: Ch 4	
W6		N-Grams	J&M: Ch 4	
		N-Grams	J&M: Ch 4	
		Lab: N-Grams		HW 1 due
W7		Machine Learning: Overview	T. Mitchel. (2017)	
		Machine Learning: Overview	T. Mitchel. (2017)	
		Evaluation and Error analysis	Resnik & Lin (2010), Kummerfeld et al. (2012)	
W8		Evaluation and Error analysis	Resnik & Lin (2010), Kummerfeld et al. (2012)	
		Regression and Maximum Entropy	J&M: Ch 6.6-6.7, J&M(3rd): Ch 5	
		Regression and Maximum Entropy	J&M: Ch 6.6-6.7, J&M(3rd): Ch 5	

W9	Regression and Maximum Entropy	J&M: Ch 6.6-6.7, J&M(3rd): Ch 5	
	Regression and Maximum Entropy	J&M: Ch 6.6-6.7, J&M(3rd): Ch 5	
	Lab: Regression		
W10	Part of Speech tagging	J&M: Ch 5, J&M(3rd): Ch 8	
	Part of Speech tagging	J&M: Ch 5, J&M(3rd): Ch 8	
	Part of Speech tagging	J&M: Ch 5, J&M(3rd): Ch 8	
W11	Lab: Tagging	BKL: Ch 5	
	Vector Semantics	J&M(3rd): Ch 6	HW 3 Set
	Vector Semantics	J&M(3rd): Ch 6	Milestone 2
W12	Vector Semantics	J&M(3rd): Ch 6	
	Lab: Semantics		
	Naive Bayes Classification	J&M(3rd): Ch 4	HW 3 Due, Milestone 3
W13	Network Analysis (Guest Lecture)		HW 4 Set
	Sentiment Analysis	J&M(3rd): Ch 4	
	Sentiment, Affect, and Connotation	J&M(3rd): Ch 21	
W14	Lab: Sentiment Analysis		
	Modelling Grammar		HW 4 Due
	Modelling Intuitions		
W15	Ethics (Guest lecture)	Hovy & Spruit (2016), Bolukbasi et al (2016), Nathan et al (2007)	
	Poster session		Milestone 4

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.