TENTATIVE SYLLABUS
LINGUISTICS 492B – COMPUTATIONAL LINGUISTICS: USE AND MEANING
SPRING 2020

COURSE & CONTACT INFORMATION

Meeting Time/Place: TTh 4:00 – 5:15, Hasbrouck 109
Instructor: Gaja Jarosz, jarosz@linguist.umass.edu
Office Hours: N410 ILC, Mondays 1:30-2:30 or by appointment
Teaching Assistant: Andrew Lamont, alamont@umass.edu
TA Office Hours: N431E ILC, Thursdays 2-3pm

OVERVIEW

This course is an introduction to computational linguistics, the study of natural language from a computational perspective. Computational linguistics encompasses both applied (engineering) and theoretical (cognitive) issues, and in this course you will get a taste of both. You will learn how to write programs to automatically process and analyze linguistic structure in language corpora. You will learn how formal language models (grammars) can be implemented computationally and used to represent linguistic structure at various levels. You will use these formal language models to automatically analyze (assign structure to) linguistic data and to model grammaticality, and you will see how these models can be trained using language corpora. A major focus of the course will be on statistical techniques, especially Bayesian inference and maximum likelihood learning, because statistical approaches such as these are the foundation of much current work in computational linguistics, both theoretical and applied.

GOALS AND PREREQUISITES

This course has three goals:

1) Develop python programming skills.
2) Introduce basic probability theory.
3) Introduce fundamental algorithms and techniques in computational linguistics

For (1) we assume a basic level of familiarity with the Python programming language. In particular, we assume a background in Python equivalent to the coursework in LINGUIST409/509 (*Introduction to Computational Linguistics*, Bhatt). If you are concerned that you do not have sufficient background, please contact me. We will cover a range of topics in intermediate Python programming designed to build upon these skills.

For (2) we assume basic university-level math background (i.e. you have completed R1 and R2 general education requirements). No particular background in probability theory is necessary.

Using the mathematical and programming tools we will develop in the context of this class, we will cover a range of statistical approaches to natural language processing. At the end of this class you will master basic probability theory, and know how to implement *n*-gram language models, probabilistic classification, Hidden Markov Models for POS tagging, and basic context free parsing techniques in Python. Together, these techniques form the most basic and widely used techniques in NLP, statistical and otherwise, and are increasingly of interest for linguistic and psycholinguistic research.

REQUIREMENTS

Readings

There will be a reading assignment corresponding to each topic. **The reading should be completed by the date listed**. Lectures will complement the reading – there will be a good deal of overlap, but lectures will not cover all the material from the readings. You will be responsible for both lecture and reading material.

Midterm and Final Exams* (30% total*)

There will be a midterm and final exam. The final will cover material from the second half of the course (it will not be cumulative). While the homework assignments primarily provide you with hands-on experience using computational techniques, the exams are an opportunity to demonstrate you have understood the concepts underlying these techniques.

Homework Assignments* (60% total*)

There will be seven homework assignments, and the lowest grade will be dropped. All the homework assignments involve programming, and some will require you to do some on-paper exercises and/or some math. The assignments cover a range of computational linguistics tasks including corpus processing, (statistical) language modeling and generation, and implementation of complex algorithms.

Participation* (10% total*)

We will occasionally have short in-class or take-home exercises which will not be graded, but will count towards your participation in the course. In class and forum discussion will also count toward participation credit.

COURSE TEXTBOOKS

We will be using two texts in the course.

(JM) The first is *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, which is a foundational textbook in computational linguistics. We will be using the online draft of the 3rd edition, available online here:

https://web.stanford.edu/~jurafsky/slp3/

(TP) This second text, *Think Python: How to think like a computer scientist* by Allen B. Downey, provides a comprehensive introduction to the Python programming language. It is freely available in HTML format at the following web address:

http://www.greenteapress.com/thinkpython/html/index.html

(MS) We will also read some excerpts from *The foundations of statistical natural language processing* by Christopher Manning and Hinrich Schütze. A digital copy of this book may be accessed here using your UMass OIT login:

https://ebookcentral.proquest.com/lib/uma/detail.action?docID=3339544

POLICIES

Grading

- Two Exams – 15% each (30% total).
- **Seven** assignments – 10% each, lowest grade dropped (60% total).
- Participation – 10%, in class and homework exercises, participation in class and on the forum

Lateness

Homework assignments that are late will **lose 10 points (out of 100) for each day they are late**. If your homework assignment is more than a week late, it will not be accepted, and you will receive a 0.

Grade Concerns

If you are concerned about your progress in class, it is your responsibility to contact us **well in advance of the end of the semester** to notify us of your concerns. After the fact grade change requests will only be considered in the most extreme of circumstances.

Collaboration

Programming is a solitary activity. You won't learn how to program unless you do it yourself. Therefore, you are required to write your own code and any other responses to homework assignments. However, I do encourage you to discuss the homework with one another, ask each other for help if you get stuck, and even compare the output of your programs to make sure you have no bugs. For non-coding portions of the homework, I encourage you to discuss the questions, but again, you must write up your own solutions. **If you discussed your assignment with another student or students, list their name(s) on your assignment (otherwise you risk violating the academic honesty policy, see below).**

Academic (dis-)honesty

The University's official policies regarding academic honesty may be found here:
http://www.umass.edu/dean_students/codeofconduct/acadhonesty/

You are expected to be familiar with the University's policies on academic policy. There will be **zero tolerance** for any cases of plagiarism (representing another's words or ideas as your own work), fabrication, cheating and facilitation of other forms of academic dishonesty. You should be familiar with the definitions of each of these terms, as defined in the official policy given above.

Note: **using other people's code without proper attribution is plagiarism**.

Students with disabilities

If you have a physical or learning disability, it is your responsibility to bring it to our attention at the **beginning of the semester** so that we can make any accommodations possible.

Extra Credit

As part of this course, you are encouraged to earn up to 4 credits of experimental participation in the Experimental Linguistics labs, which are located in the Linguistics department in the Integrative Learning Center. **1 credit must be earned by mid-February, another may be earned only until mid-March, and the remaining two can be earned until the last day of classes**. To browse the range of experimental opportunities available to you, please sign onto the SONA systems website for the Linguistics department:

http://umassxling.sona-systems.com

After the Add/Drop period ends, you will receive an email inviting you to set up a SONA account. If you do not receive this email, you may go to the web address above and request an account on your own. If you do this, be sure to select any and all linguistics courses you are in this semester when you log in.

On SONA, you will be able to sign up for experiments that look interesting to you. Very short experiments will give one credit; more typically, a single experimental session will count as two experimental credits. If you are unable to sign up for an experiment, you may contact the experimenters directly through the SONA system.

If you would like more information on how to use SONA systems as a participant, here is a brief YouTube tutorial on how to use it:

https://www.youtube.com/watch?v=_1OnT2ZU6QQ

Participation in experimental research is highly encouraged, as it gives you insight into the day-to-day work of linguists studying human language. However, if you do not wish to participate in experimental research, you may receive extra credit by doing an alternative assignment. These will be made available on SONA as well.

Each point of SONA credit will translate to one percentage point on your final grade.

TENTATIVE SCHEDULE

Date	Topic	Assignment	Reading
1/21	Introduction, syllabus		TP: 1-2, 5,7, MS: 1
1/23	Python Review		TP: 8, 10-12
1/28	Python Review	HW 1 assigned (Analyzing Twitter)	TP: 3, 14
1/30	Python, Frequency, & RegExp		TP: 13 JM: Ch 2
2/4	String Similarity (Gaja away)		
2/6	N-grams, Analogy (Gaja away)	**HW1 due** HW 2 assigned (Lexical Neighbors)	MS: Ch 2
2/11	Probability		
2/13	Probability		JM: Chap 3
2/18	**NO CLASS**	**HW 2 due**	
2/20	N-grams	HW 3 assigned (Ngram Phonotactics)	
2/25	N-grams		
2/27	HMMs	**HW 3 due**	JM: 8, App. A
3/3	Review (Gaja away)		
3/5	**MIDTERM** (Gaja away)		
3/10	HMMs	HW 4 assigned (POS Tagging)	
3/12	HMMs		
3/16-3/20	*Spring Break*		
3/24	CFGs	HW 5 assigned (Writing CFGs)	JM: Ch 10
3/26	CFGs	**HW 4 due**	JM: Ch 11
3/31	CFG Parsing: CKY & Earley		
4/2	PCFGs	**HW 5 due** HW 6 assigned (Parsing PCFGs)	JM: Ch 12
4/7	Statistical Parsing		
4/9	Supervised Classification		JM: 4 (5)
4/14	Vector Semantics		JM: 6
4/16	Clustering	**HW 6 due** HW 7 assigned (Clustering)	MS: 14
4/21	More Unsupervised Learning		
4/23	TBA		
4/28	Review and Wrap-up	**HW 7 due**	
5/1	**3:30-5:30**	**FINAL EXAM**	