Computational Linguistics: Use and Meaning
LINGUIST 492B

Instructor: Brian Dillon (brian@linguist.umass.edu)
TA: Alan Zaffetti (azaffett@ umass.edu)

Meets: TuTh 11:30 – 12:45
Location: Hasbrouck Laboratory 228
Office: N436 Integrative Learning Center (Brian)
TBD (Alan)
Office hours: Monday, 1-3pm (Brian); TBD (Alan)

Course overview:

This course is a one-semester course on statistical natural language processing (NLP). Statistical NLP is perhaps the dominant paradigm in current computational linguistics, and refers to a broad class of statistical techniques for processing natural language. This course will familiarize you with the range of techniques that are being applied in contemporary NLP. This course has three goals:

1) Develop Python programming skills.
2) Introduce basic probability theory.
3) Introduce fundamental algorithms and techniques in natural language processing.

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 (*Introduction to Computational Linguistics*, Bhatt). If you are concerned that you do not have sufficient background, please contact either Brian or Alan. We will cover a range of topics in intermediate Python programming designed to build upon these skills in the context of statistical NLP.

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 supervised classification (Naïve Bayes), 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.

Course textbooks:

There are two textbook resources that we will draw upon for this class.

1) *Think Python: How to think like a computer scientist* by Allen B. Downey will be used for reading and for homework exercises in the first section of the course. It is freely available in HTML formal at the following web address:

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

This book provides a comprehensive introduction to the Python programming language, which we will use throughout this class. If you are unfamiliar with the language (or even if you are!), it will be a great resource.

PLEASE NOTE: Think Python in HTML format and PDF format have different exercise numbering. **All assignments from Think Python refer to the exercise numbering in the HTML Format.**

2) *The foundations of statistical natural language processing* by Christopher Manning and Hinrich Schütze. This is an introduction to statistical NLP, and we will cover a subset of the many topics this book covers. A digital copy of this book may be accessed at http://library.umass.edu using your UMass OIT login.

Web site & Email policy:
The course website is hosted on Moodle; log into http://moodle.umass.edu with your SPIRE NetID to access course materials and readings.

If you need to talk about course work or have questions of any sort, you are encouraged to see either of the instructors. You are also welcome to email if you have minor questions, and we will respond to email messages within 24 hours Monday through Friday.

Grading:

Your grade will be composed of:

Homework assignments	*(6)*	60%
Midterm Examination		20%
Final Examination		20%

The grade breakdown is as follows:

93-100	A	77-79	C+	0-59	F
90-92	A-	73-76	C		
87-89	B+	70-72	C-		
83-86	B	67-69	D+		
80-82	B-	60-66	D		

Extra Credit:
You may drop your lowest homework score in exchange for participation in a half hour experiment in the Linguistics department. To participate in a linguistics experiment, please visit **https://xlingumass.youcanbook.me/**

If you are concerned about your progress in class, you need to contact me **well in advance of the end of the semester** to notify me of your concerns. **End-of-semester requests for grade adjustments will not be considered under any circumstances.**

Homework assignments:
There will be six homework assignments given throughout the semester, and they will be due one week from their assignment date. All homework will require hands-on programming in Python. The assignments are intended to give you hands-on experience with the concepts discussed in class. **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.**

Midterm:
There will be one midterm exam given in this course. Your score on this exam will count for 20% of your grade. **The midterm is scheduled for Thursday, 3/12**. If you cannot be in class on that date, it is your responsibility to notify the instructors at least a week ahead of time and make arrangements for a makeup.

Final Exam:
There will be one final exam given in this course. The final will only cover the second half of the course; it is not cumulative. Your score on this exam will count for 20% of your grade.

Attendance policy:
Consistent attendance (and participation in class) is expected, and is necessary to keep ahead of class material. In keeping with the University's policies, any student who needs to miss a class due to a religious holiday will be allowed to make up the work they miss, provided that s/he notifies me at least a week in advance of any expected absence. It is your responsibility to contact me to make the necessary arrangements in a timely fashion.

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**. If you reuse portions of someone else's code for your assignments, you must be sure it is properly attributed to its source.

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 I can make any accommodations possible.

Schedule:

Date	Topic	Reading	Assignment
1/19	Introduction	*TP* Chap 1-3	
1/21	Python Review	*TP* Chap 8, 10, 11	
1/26	Python Review	*TP* Chap 7	**HW#1: Webscraping with twitter**
1/28	Python review	*TP* Chap 14	
2/2	Python review	*MS* Chap 4	
2/4	Probability theory	*MS* Chap 2	
2/9	Probability theory		**HW#1 DUE** **HW#2: Quantifying twitter**
2/11	Probability theory		
2/16	**NO CLASS**	-	-
2/18	*n*-gram models	*MS* Chap 6	
2/23	*n*-gram models		**HW#2 DUE** **HW#3: *n*-gram models**
2/25	*n*-gram models		
3/1	Supervised classification	*MS* Chap 7	
3/3	Supervised classification	*Alan teaches*	
3/8	Midterm Review		**HW#3 DUE**
3/11	**MIDTERM EXAM**		
3/15	**NO CLASS**		
3/17	**NO CLASS**		
3/22	Supervised classification	*MS* Chap 8	**HW #4: Word sense disambiguation**
3/24	Unsupervised classification		
3/29	Hidden Markov Models	*MS* Chap 9	
3/31	Hidden Markov Models		
4/5	Hidden Markov Models		**HW #4 DUE** **HW #5: POS Tagging**
4/7	Deterministic Parsing	Wolf & Gibson (2003)	

4/12	Deterministic Parsing	
4/14	Probabilistic CFGs	
4/19	Probabilistic CFGs	**HW #5 DUE** **HW #6 Parsing**
4/21	Probabilistic CFGs	
4/26	Final Review	