Computational Linguistics: Use and Meaning
LINGUIST 492B

Instructor: Brian Dillon (brian@linguist.umass.edu)
TA: Helene Gene (hgene@ umass.edu)

Meets: TuTh 1:00 – 2:15
Location: N458 Integrative Learning Center
Office: N436 Integrative Learning Center
Office hours: Tuesdays, 11-12 (BD) ; Tu 4-5, W 2:30-3:30 (HG)

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 using the Natural Language Toolkit (NLTK) package.
2) Introduce basic probability theory
3) Introduce fundamental techniques in statistical 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 Helene. We will cover a range of topics in intermediate Python programming (including the use of NLTK) designed to build upon these skills in the context of statistical language processing.

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 and Maximum Entropy), Hidden Markov Models for POS tagging, and PCFGs in Python, using the NLTK toolkit. Together, these techniques form the most basic and widely used techniques in statistical NLP, and are increasingly of interest for linguistic and psycholinguistic research.

Course textbooks:

There are three main 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.

2) *The foundations of statistical natural language processing* by Christopher Manning and Hinrich Schütze. This is an introduction to statistical natural language processing (NLP), and we will cover a very small 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.

3) *Natural language processing with Python* by Steven Bird, Ewan Klein and Edward Loper is a text that discusses how to use the Python Natural Language Toolkit (NLTK) to accomplish a variety of NLP tasks. It may be freely accessed here:

http://nltk.org/book/

We will be using the NLTK throughout the class to get hands-on experience with the algorithms and methods discussed in Manning & Schütze, and we will also assign readings from the NLTK book.

Web site & Email policy:

The course website is hosted at on the University Moodle; check back often for the most up to date schedule and assignment information.

If you have questions about the homework or have questions about the content of the class, **you are encouraged to make use of the appropriate discussion forum on the Moodle.** The instructors and your fellow classmates will be engaged in discussion on these forums, and you are likely to receive a **much faster response** than if you email the instructors directly.

You are also welcome to email us or make an appointment to see us if you have questions that are not appropriate for discussion in a public forum, or if you have major concerns about your understanding of the material. You may expect us to respond to email messages up until 7pm on Monday through Friday. Responses to emails sent after 7pm, or on the weekend, may be significantly delayed. This means that last minute questions emailed before an assignment is due may go unanswered, so **plan ahead** and talk to us early if you think there is going to be an issue.

Homework assignments must be emailed to umasscomputerling@gmail.com by the start of class on the day they are due. Please name your files appropriately: {firstname}_{lastname}_assignment{#}.py or .txt.

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.**

Student responsibilities:

In order to get the most out of this class, you are expected to take responsibility for your own learning experience! We are here to facilitate and guide your learning, but without commitment and effort on your part, only so much can happen**.** In order to get the most out of this class, you will be expected to…

Attend and participate in class. The material covered in class will overlap with assigned readings, but not always, and you will not be able to do well in this class without attending all classes. There will be ample opportunity for discussion and hands-on work in class.

Ask questions! If you do not understand something we are covering in class, ask! If you do not understand the homework, ask! Come to office hours (with me or with Helene) if you need additional help. The classroom activities are designed to help you come to a better understanding of the material, so if you find yourself not sure of something, be sure to speak up!

Be courteous. Contribute to a polite and respectful classroom atmosphere. Turn off your cell phone. Do not talk to your neighbors, and don't engage in disruptive behavior. If you cannot adhere to common sense rules of classroom etiquette, you will be asked to leave.

Plan ahead. **Do not put things off until the last minute, or we may not be able to help you**. If you're having a problem that prevents you from attending class, be sure to tell us as soon as possible. There will be **no grade adjustments after the fact** (not for homeworks, not for exams, not for the course). You **must notify us** of any special situations that we need to be aware of well in advance of any grading.

Likewise, if you email us with urgent questions the night before an assignment is due, it may be too late for us to get help to you on time (see email policy). If the homework requires use of a particular piece of software, install it early on so you can be sure it works. Installation of software can sometimes come with its own headaches, and we may not be able to help troubleshoot last minute problems. **Plan ahead and be organized.**

Grading:

Your grade will be composed of:

Homework assignments	*(6)*	20%
In-class quizzes	*(8)*	20%
Midterm Examination		20%
Final project		30%
Participation		10%

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 either your lowest quiz or lowest homework score in exchange for participation in a half hour experiment in the Linguistics department. If you participate in two such experiments, you may drop both your lowest quiz and your lowest homework score. You may participate in a maximum of two extra credit experiments per class.

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, often using NLTK. The assignments are intended to give you hands-on experience with the concepts discussed in class. **Homework assignments that are late will lose 1 point (out of 10) 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.**

Quizzes:

A short quiz will be given almost biweekly (8 times throughout the semester). While not intended to be difficult, the quizzes will require knowledge of the core concepts discussed in the prior week's classes. Make-ups will not be given for missed quizzes unless i) the

student notifies the instructors more than a week ahead of time or ii) the student presents proof that s/he could not attend class (e.g. in the form of a doctor's note).

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 project:

For your final project, you will construct a probabilistic model of some natural language phenomenon. This project is intended to be a hands-on, real-world application of the techniques you use in this class to a problem of your choosing. To get started thinking about this, there are two types of project you might consider:

i) *Classification task*: Given some variable linguistic phenomenon, what features predict the outcome of that phenomenon? Use both generative (e.g. Naïve Bayes) and discriminative (e.g. MaxEnt / Logistic regression) techniques to explore the factors that control the variable realization of some linguistic phenomenon.

ii) *Syntactic ambiguity*: Identify a novel syntactic ambiguity (in English or another language that has a Treebank in NLTK). Induce a PCFG for that language, and discuss the role of probabilistic grammar in disambiguating the ambiguity of your choosing.

Your task in your final project is to identify a linguistic phenomenon that would be interesting to explore given these approaches, and think about a strategy for approaching the question given the specific techniques learned in class. Your final project should involve **significant *linguistic* research on the phenomenon of choice,** going beyond what we have done in class**.** You should first seek to identify some problem or question you're interested in, and **run your idea by the instructors by 3/12. We are here to help you figure out your approach and the specific tools you could be using!** You may work in groups of up to three people, and I expect the final project should involve involve approximately 2-3 assignments' worth of work **per person**. You will also be responsible for a short in-class presentation of your project and your findings on the last two days of class. **Be prepared to present by 4/23**; order will be determined randomly.

Attendance policy:

Consistent attendance (and participation in class) will be reflected in your participation grade. 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:* ignorance of the rules is not an acceptable excuse for academic dishonesty. It is worth noting that it is easier than ever to catch plagiarizing due to Google. You may have noticed lots of high profile writers getting caught plagiarizing in recent years due to this fact. In every case, it has ruined their career. There's a lesson in there: **Plagiarizing is never acceptable, not under any circumstances.**

Note: **using other people's code without proper attribution is plagiarism**. If you reuse portions of someone's else 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 my attention at the beginning of the semester so that I can make any accommodations possible.

Schedule (tentative):

Date	Topic	Reading	Assignment
1/20	Introduction	*TP* Chap 1-3	
1/22	Python review: Iteration, List Comprehension	*TP* Chap 7	
1/27	SNOW DAY!		
1/29	Python review: Tuples, Dictionaries	*TP* Chap 10-12	Quiz #1
2/3	Python review: NLTK Corpora, reading data	*NLTK* Chap 2.1-2.2	**HW #1: Python Basics**
2/5	Probability theory: joint, conditional probability	*MS* Chap 2.1	Quiz #2
2/10	Probability theory: independence, Bayes' rule	*MS* Chap 2.1	

2/12	Estimating probabilities: relative frequency, conditional frequency	*NLTK* Chap 1	Quiz #3 **HW #2: Probability Theory**
2/17	**NO CLASS; UMASS MONDAY**	-	-
2/19	*n*-gram Language Models	*MS* Chap 6	
2/24	*n*-gram Language Models	*MS* Chap 6	Quiz #4
2/26	**Guest Lecture**	*MS* Chap 2.2	**HW #3: *n*-gram models**
3/3	Classification: Feature Extraction	*NLTK* Chap 6.1-6.3	
3/5	Classification: Naïve Bayes	*NLTK* Chap 6.5	Quiz #5
3/10	Classification: Evaluation		
3/12	***MIDTERM***		
3/18	**NO CLASS; SPRING BREAK**		
3/20	**NO CLASS; SPRING BREAK**		
3/24	Classification: Maximum Entropy	*NLTK* Chap 6.6	
3/26	Classification: Maximum Entropy		**HW #4: Supervised classification**
3/31	Project RunBayes		
4/2	Hidden Markov Models	*MS* Chap 9.1-9.2	Quiz #5
4/7	Hidden Markov Models		
4/9	Hidden Markov Models		**HW #5: HMMs**
4/14	Syntax: Context Free Grammars, PCFGs	*NLTK* Chap 8.1-8.4	Quiz #6
4/16	Syntax: Chart parsing	*NLTK* Chap 8.6, 'Grammar Induction' (extras)	**HW #6: PCFGs**
4/21	Syntax: Parsing with PCFGs	*NLTK* Chap 8, extras	Quiz #7
4/23	Project Presentations		
4/28	Project Presentations		