

LING 1330/2330 Computational Linguistics

Fall 2018, University of Pittsburgh

Meetings: Tue & Thu 4:30pm - 5:45pm **Classroom:** 363 Cathedral of Learning

Description

This is a course designed to introduce students who have been exposed to linguistics to real-world applications of computational linguistics. The students will first learn the fundamentals of how computers are used to represent and process textual and spoken information. They will then be introduced to the challenges of real-world language engineering problems and learn how they are handled with the latest language technologies. The topics include: spell-checking, machine translation, part-of-speech tagging, parsing, document classification, and corpus building and exploration. Students will be given hands-on training on the basics of text processing using Python and will have a chance to work with NLTK, a popular natural language processing application suite. This course is designed specifically for students in the humanities; computer science majors (who are not linguists) are encouraged to take [CS 1671](#) or [CS 1571](#) instead.

Prerequisites

LING 1000 Introduction to Linguistics is the only prerequisite for this course. Prior knowledge of **Python** or other programming languages is *not required but highly recommended*. [CS 0008 "Introduction to Computer Programming with Python"](#) or [CS 0155 "Data Witchcraft"](#) will give you a good preparation.

Notes for Incoming Students

Do you have **little/no programming experience**? Please read the following.

1. I strongly recommend you to **get a head-start on Python** this summer. The class moves very fast, and you will be a lot more comfortable knowing a bit going in. The [Python 3 Notes](#) I put together has everything you need.
2. Students with no prior programming experience do complete this course with great success, but they typically put in more hours than their experienced peers (which is understandable). For this reason, I encourage you to **sign up for an additional 1-credit recitation**. It will be run as LING 1901 Independent Study with the instructor, on a S/NC basis. 3 recitation sessions are being planned, all on Friday: (1) 10am -- 11:15am, (2) 11:30am -- 12:45pm, (3) 1pm -- 2:15pm. Please leave one of these slots open in your schedule.

Class Home Page

- [Class Schedule Table](#)
- [Policies Page](#)
- [Recitation Log](#)



[CourseWeb Class Home](#)

Assignments:

- Exercise [#1](#), [#2](#), [#3](#)
- Homework [#1](#)

Learning Python:

- [Python 3 Notes](#)
- [FAQ](#)
- [Text samples](#)

Python Resources:

- Python 3 Quick Reference [\[link\]](#)
- Think Python: How to Think Like a Computer Scientist [\[link\]](#)

Regular Expressions:

- regexpal [\[link\]](#)
- RE syntax [\[link\]](#)
- Python RE HOWTO [\[link\]](#)

NLTK & NLP:

Students are required to **bring their own laptop** to class. It should be running one of the following operating systems: Windows 10 (7 & 8 are also fine), Mac OS-X, and Linux (any distribution). Mobile or cloud-based machines such as Android/Apple tablets or Chromebooks are not suited.

Instructors

Who	Pitt email	Office hours	Location
Na-Rae Han	naraehan	Mon 11am-1pm & Wed 11am-noon	G17 CL
Katherine Kairis (TA)	kak275	Mon & Wed 3-5pm	2832 CL (linguistics grad/undergrad office)
Daniel Zheng (TA)	daniel.zheng	Mon & Wed 7-9pm	2832 CL (linguistics grad/undergrad office)

**Note: We are also available to meet by appointment.

Textbooks

[1] [Language and Computers](#). Markus Dickinson et al. Wiley-Blackwell. 2012.

[2] Python tutorial: [Python 3 Notes](#)

[3] [Natural Language Processing with Python](#). (updated edition based on Python 3 and NLTK 3) Steven Bird et al. O'Reilly Media.

Course Organization

Each meeting will comprise two parts: lecture and lab. In the first half of the class, topics presented in the textbook [1] *Language and Computers* will be covered in a lecture-and-discussion format. In the second half, students will get hands-on training on the basics of text processing using Python and [Natural Language Toolkit \(NLTK\)](#). Friday recitations (optional) will focus on the programming aspect: additional Python exercises, upcoming homework reviews, and individual help will be offered.

Assignment Schedule

1. As a rule, there will always be a form of assignment between classes. There are two types: homework assignments and programming exercises. They are administered via CourweWeb and due before the beginning of the next class.
2. **Homework Assignments** (40-60 points): These are assigned on most Thursdays. There will be around 11. These will comprise questions on lecture topics as well as programming problems.
3. **Programming Exercises** (20 points): They are designed to help you learn and practice the programming aspect of the course. As long as you are keeping up with the course contents, you should be able to complete them in 1-2 hours, possibly more if you are new to programming. These will be given when there is no homework assignment, mostly on Tuesdays.

- Natural Language Toolkit (NLTK) Project Home [\[link\]](#)
- NLTK Book, Python3 Edition [\[index\]](#) [\[navigation panel\]](#)
- NLTK HOWTOs [\[link\]](#)
- FSNLP Ch.4 Corpus-Based Work Links [\[link\]](#)
- Speech and Language Processing 3rd Ed. [\[link\]](#)

Data Resources:

- NLTK Corpora Index [\[page\]](#)
- Natural Language Corpus Data by Peter Norvig [\[link\]](#)
- Google Books Ngram Viewer Data [\[link\]](#) (*Slow? Try Firefox.*)
- COCA *n*-gram lists at BYU [\[link\]](#)
- SUBTLEXus Word Frequency for American English [\[link\]](#)
- BNC database and word frequency lists by Adam Kilgariff [\[link\]](#)
- Tagged and Cleaned Wikipedia and its Ngram [\[link\]](#)

4. **Readings and Previews:** In addition to the homework/exercise assignment, you will have book chapters and Python tutorials for the upcoming class to study beforehand.
5. I will make every effort to post new assignments one week in advance. However, I might need to make some adjustments depending on the progress we make in classroom. Therefore, non-immediate assignments should be considered a DRAFT until it is finalized, which will happen 30 minutes post-class.
6. Detailed assignment schedule is [found on the Class Schedule page](#).

Exams, Requirements, Grading and Policies

Please read the [Course Policies page](#).

Go to: [LING 1330/2330 home page](#)

Course Policies

Course Requirements and Grading

Grading scale:

97-100%	A+	87-89.99	B+	77-79.99	C+	67-69.99	D+	0-59.99	F
93-96.99	A	83-86.99	B	73-76.99	C	63-66.99	D		
90-92.99	A-	80-82.99	B-	70-72.99	C-	60-62.99	D-		

Requirements and grade calculation:

Requirement	Weight	Total points	Note
Midterm Exam	20%		
Final Exam	30%		
Assignments	45%	HW: 40-60 points each Exercise: 20 points each	
Attendance & Participation	5%	80 points (attendance) + 20 points (participation)	1 missed class allowed; -7 points for each absence thereafter

Late work penalties:

Assignments are due 30 minutes before each class via CourseWeb. Late assignments will be penalized as follows: 20% if submitted 1 class late, 40% if submitted 2 classes late; no submission will be accepted beyond that point.

Attendance:

Attendance is mandatory. Roll calls will be taken on a regular basis, which will be reflected on your attendance score. For information on excusable absences, see [this section below](#).

Participation:

The participation grade will be based on students' engagement level in classroom and also their activity in the CourseWeb forum.

Collaboration:

If done properly, working together on assignments lead to a better learning outcome for all parties

Class Home Page

- [Class Schedule Table](#)
- [Policies Page](#)
- [Recitation Log](#)



[CourseWeb Class Home](#)

Assignments:

- Exercise [#1](#), [#2](#), [#3](#)
- Homework [#1](#)

Learning Python:

- [Python 3 Notes](#)
- [FAQ](#)
- [Text samples](#)

Python Resources:

- Python 3 Quick Reference [\[link\]](#)
- Think Python: How to Think Like a Computer Scientist [\[link\]](#)

Regular Expressions:

- regexpal [\[link\]](#)
- RE syntax [\[link\]](#)
- Python RE HOWTO [\[link\]](#)

NLTK & NLP:

- Natural Language Toolkit (NLTK) Project Home [\[link\]](#)
- NLTK Book, Python3 Edition [\[index\]](#)

involved. If done improperly, however, it negatively affects learning AND results in cheating. For your learning benefit, I am allowing group work provided that the following conditions are met:

1. Equal contribution: one student's contribution must not exceed 150% of other's.
2. Individual work before a study group: do not show up to a study session without having worked on the assignment on your own beforehand.
3. Individual work after a study group: do not write up your homework assignments while working in group, which leads to copying other's answers. Always finish up your answers by yourself afterwards, using your own words.

Recitation Grade

If you are new to programming and enrolled in the S/NC 1-credit recitation through LING 1901 Independent Study, your grade will be based on attendance. Out of the 12 recitation sessions, attend at least 9 in order to receive the S ("Satisfactory") grade.

Important notes

- As you can see from the course schedule, this class moves in a rapid pace. **It is critical that you attend class regularly and keep up with the course content.** If you do not have good time and project management skills, this is not the class for you.
- There is a form of assignment due at every class, and the lab section in particular will build on what was learned previously. If you miss a class, make sure to review and catch up before the next class. If you show up without catching up, you *will* be lost.
- When doing assignments, make sure to closely follow the provided instructions. The ability to follow instructions is a critical programming skill! Also, when programming, get in the habit of taking your time to get the details right. It will save you much frustration and confusion downstream: in writing a structured program, a seemingly minor detail that was glossed over in a haste has a way of resurfacing and tripping you.
- Get help early: approach the instructor or the TA when you first have trouble understanding a concept or homework problem. Please note that certain problems are not easily communicated via email, especially when the lack of understanding of the very nature of the problem is the first stumbling block. It will be easier to come and see instructors in person.
- Turn off your cell phone in class. Non-class-related activities, including Facebook, twitter, email, web surfing, are strictly forbidden. Violations will incur a penalty on your participation score.

Remediation and Justification

The exams will be given on the specified dates. Absolutely no early, late, makeup exams will be given. Exceptions may be made, at the instructor's discretion, in the case of true emergencies, for which explicit documentation must be provided. Penalties for late work or absences may be waived for documented emergencies. For planned, justified, absences or late work, advance notice must be provided. Valid reasons for missing class include the following:

[\[navigation panel\]](#)

- NLTK HOWTOs [\[link\]](#)
- FSNLP Ch.4 Corpus-Based Work Links [\[link\]](#)
- Speech and Language Processing 3rd Ed. [\[link\]](#)

Data Resources:

- NLTK Corpora Index [\[page\]](#)
- Natural Language Corpus Data by Peter Norvig [\[link\]](#)
- Google Books Ngram Viewer Data [\[link\]](#) (*Slow? Try FireFox.*)
- COCA *n*-gram lists at BYU [\[link\]](#)
- SUBTLEXus Word Frequency for American English [\[link\]](#)
- BNC database and word frequency lists by Adam Kilgarriff [\[link\]](#)
- Tagged and Cleaned Wikipedia and its Ngram [\[link\]](#)

- A religious observance: advance written notice is required
- A medical emergency (including mental health emergencies): submit a signed letter from a doctor
- A family emergency (death, birth, natural disaster, etc): submit a signed letter from an authority

Academic Integrity

Cheating/plagiarism will not be tolerated. Students suspected of violating the University of Pittsburgh Policy on Academic Integrity, noted below, will be required to participate in the outlined procedural process as initiated by the instructor. A minimum sanction of a zero score for the quiz, exam or paper will be imposed. (For the full Academic Integrity policy, go to <https://www.provost.pitt.edu/info/ai1.html>.) In this class, working together on assignments is allowed strictly under a [set of conditions](#) noted above; failing to meet them will amount to a violation of academic integrity and will be handled accordingly.

Disabilities

If you have a disability for which you are or may be requesting an accommodation, you are encouraged to contact both your instructor and the Office of Disability Resources and Services, 140 William Pitt Union, 412-648-7890, as early as possible in the term. Disability Resources and Services will verify your disability and determine reasonable accommodations for this course.

E-mail Communication Policy

Each student is issued a University e-mail address (username@pitt.edu) upon admittance. This e-mail address may be used by the University for official communication with students. Students are expected to read e-mail sent to this account on a regular basis. Failure to read and react to University communications in a timely manner does not absolve the student from knowing and complying with the content of the communications. The University provides an e-mail forwarding service that allows students to read their e-mail via other service providers (e.g., Hotmail, AOL, Yahoo). Students that choose to forward their e-mail from their pitt.edu address to another address do so at their own risk. If e-mail is lost as a result of forwarding, it does not absolve the student from responding to official communications sent to their University e-mail address. To forward e-mail sent to your University account, go to <https://accounts.pitt.edu>, log into your account, click on Edit Forwarding Addresses, and follow the instructions on the page. Be sure to log out of your account when you have finished. (For the full E-mail Communication Policy, go to <https://www.bc.pitt.edu/policies/policy/09/09-10-01.html>.)

Go to: [LING 1330/2330 home page](#)

Class Schedule

*Class schedule is subject to revision throughout the semester.

W	Date	Lab & Recitation	Lecture	Assignment (due next class)
1	8/28 (T)	Intro to IDLE, <code>print()</code> , special characters, + [Lab1.pdf] (Tutorials 1 , 2 , 3 , 4 , 6)	Course introduction	Exercise 1
	8/30 (Th)	<code>input()</code> , conditionals, variable assignment, string operations, <code>.split()</code> [Lab2.pdf] (Tutorials 5 , 7 , 16 , 17 , 18 , A1 , A2 , A3)		Exercise 2
	8/31 (Fri)	[Log] Setup check, practice		
2	9/4 (T)	<code>for</code> , <code>while</code> loop, List & string indexing [Lab3.pdf] (Tutorials 14 , 15 , A6 , 8)		Exercise 3 : palindrome
	9/6 (Th)	User-defined function, void vs. returning functions (Tutorials 19 , A7)	Ch.1 Encoding language	HW 1 : Encoding, pig Latin
	9/7 (Fri)	Control flow, indexing, function practice		
3	9/11 (T)	Dictionary and list methods (Tutorials 9 , 11)	Ch.2 Writers' aids: spell checkers	Exercise 4
	9/13 (Th)	Data types & conversion, more list methods, mutability, sorting, importing (Tutorials 9 , 10 , A4 , A5 , A8 , A9)		HW 2: Spell checkers, text stats

Class Home Page

- [Class Schedule Table](#)
- [Policies Page](#)
- [Recitation Log](#)



[CourseWeb Class Home](#)

Assignments:

- Exercise [#1](#), [#2](#), [#3](#)
- Homework [#1](#)

Learning Python:

- [Python 3 Notes](#)
- [FAQ](#)
- [Text samples](#)

Python Resources:

- Python 3 Quick Reference [[link](#)]
- Think Python: How to Think Like a Computer Scientist [[link](#)]

Regular Expressions:

- `regex`pal [[link](#)]
- RE syntax [[link](#)]
- Python RE HOWTO [[link](#)]

NLTK & NLP:

- Natural Language Toolkit (NLTK) Project Home [[link](#)]

	9/14 (Fri)	Practice list, dict, sorting		
4	9/18 (T)	Reading files, importing user-built modules (Tutorials 12 , 20 , A11)	Spell checkers and context, n-gram frequency (pp. 26 -- 28)	Exercise 5
	9/20(Th)	Writing files, file & directory path (Tutorials 13 , A10 , A11)		HW 3: n-grams
	9/21 (Fri)	Practice module importing, file IO		
5	9/25 (T)	Getting started with NLTK	Big <i>n</i> -gram data, probability	Exercise 6
	9/27 (Th)	Pickling (Tutorial A12)		HW 4: Bigram Speak
	9/28 (Fri)	NLTK check; building complex objects; working with larger texts		
6	10/2 (T)	List comprehension (Tutorials A13 , A15)	n-gram language models	Exercise 7
	10/4 (Th)	Working with corpora; processing online resources	Corpus analysis	HW 5: Corpus processing
	10/5 (Fri)	Corpus processing practice		
7	10/9 (T)	NLTK functions	HW 5 review	--
	10/11 (Th)	Midterm exam		
8	Fall break; no class			
	10/18 (Th)	NLTK Book Ch.1 Language Processing and Python , Ch.2 Accessing Text Corpora and Lexical Resources , Ch.3 Processing raw text		HW 6: NLTK
	10/19 (Fri)	NLTK resources & practice		
9	10/23 (T)	Ch.5 Categorizing and tagging words	Ch.3 Language tutoring systems 3.4 Tokenization, POS tagging	Exercise 8

- NLTK Book, Python3 Edition [[index](#)] [[navigation panel](#)]
- NLTK HOWTOs [[link](#)]
- FSNLP Ch.4 Corpus-Based Work Links [[link](#)]
- Speech and Language Processing 3rd Ed. [[link](#)]

Data Resources:

- NLTK Corpora Index [[page](#)]
- Natural Language Corpus Data by Peter Norvig [[link](#)]
- Google Books Ngram Viewer Data [[link](#)] (*Slow? Try FireFox.*)
- COCA *n*-gram lists at BYU [[link](#)]
- SUBTLEXus Word Frequency for American English [[link](#)]
- BNC database and word frequency lists by Adam Kilgariff [[link](#)]
- Tagged and Cleaned Wikipedia and its Ngram [[link](#)]

	10/25 (Th)	5.5 N-gram tagger	Building POS taggers: n-gram tagger, HMM	HW 7: n-gram tagger
	10/26 (Fri)	<i>POS tagging practice</i>		
10	10/30 (T)	HMM tagger (Jurafsky & Martin Ch.9 , Ch.10) 7.4.2 Trees	POS tagger review Ch.3 Beyond words	Exercise 9
	11/1 (Th)	Ch.8 Analyzing sentence structure 8.6 Grammar Development: Treebanks and Grammars		HW 8: CFG
	11/2 (Fri)	<i>Trees and CF grammar practice</i>		
11	11/6 (T)	Weighted Grammar Ch.6: 6.1.3 Learning to classify text	Ch.5 Classifying documents, Naive Bayes classifier	Exercise 10
	11/8 (Th)	6.5 Naive Bayes classifiers		HW 9: Who said it? (due 11/13, 11/15)
	11/9 (Fri)	<i>Classification practice</i>		
12	11/13 (T)	6.3 Evaluation	Evaluation methods	--
	11/15 (Th)	2.1 Accessing Text Corpora , 2.2 Conditional Frequency Distributions , 2.4 Lexical Resources , 2.5 Wordnet		HW 10: Corpora & NLTK
	11/16 (Fri)	<i>NLTK practice</i>		
13	11/20 (T)		Ch.4 Searching, regular expressions	Exercise 11
	Thanksgiving break			
14	11/27 (T)	Regular expressions in Python (Tutorials A14 , A15)		Exercise 12
	11/29 (Th)		Ch.7 Machine translation Jurafsky & Martin Ed.2 Ch.21 MT, 5.4 The noisy channel model	HW 11: MT
	11/30 (Fri)	<i>RE practice, string formatting</i>		

with , wrap-up

15	12/4 (T)		Ch.4 FSA: Formal language theory, Chomsky hierarchy	Exercise 13
	12/6 (Th)		Formal language theory	--
Date/Time TBD		Final exam		

*Class schedule is subject to revision throughout the semester.