11-711: Algorithms for NLP, Fall 2017

Instructors: Taylor Berg-Kirkpatrick and Robert Frederking
Lecture: Tuesday and Thursday 1:30pm-2:50pm, DH 1212
Recitation: Friday 1:30pm-2:20pm, MM 103
Office Hours: Taylor - Friday 12pm-1pm, GHC 6403
Bob - by appointment

TAs: Hieu Pham, Nikita Srivatsan and Maria Ryskina
Office Hours: Hieu - Monday 2pm-3pm, GHC 6418
Nikita - Wednesday 2pm-3pm, GHC 6418
Maria - Thursday 3:30pm-4:30pm, GHC 6603
Forum: Piazza

Announcements

12/6/17: Reminder: there is no recitation this Friday (Dec 8).
12/2/17: Slides from the final recitation posted here: HMM Aligner Recitation Slides.
11/29/17: Taylor will hold extra office hours on Thurs (Nov 30) at 11am at GHC 6403 (Taylor's office).
11/28/17: Hieu's OH next Monday (Dec 4) is canceled. He will hold make-up OH this Friday (Dec 1) at 10am.
11/28/17: Hieu released a great note about deriving EM algorithm for word alignment from scratch: Latent Models for Word Alignment. (updated 12/1). *Disclaimer: this is work in progress. Errata can be found on Piazza.*
11/22/17: There are no OH or recitation this Wed-Fri (Nov 22-24) because of Thanksgiving break.
11/17/17: Slides from the tenth recitation posted here: EM Recitation Slides. (updated 11/28)
11/13/17: Project 4 has been released. It is due Dec 4 at 11:59pm ET.
11/9/17: There is no recitation or OH this Friday (Nov 10) because of 50th Anniversary celebration.
11/6/17: Project 3 now due Sunday 11/12 by 11:59pm.
11/3/17: Slides from the ninth recitation posted here: P3 Interface Recitation Slides.
10/30/17: Project 3 has been released. It is due Nov 10 at 11:59pm ET.
10/28/17: Notes from the eighth recitation (coarse-to-fine part) posted here: Coarse-to-fine Recitation Notes.
10/23/17: Parsing scores for a correctly implemented parser have been added to the Project 2 page.
10/22/17: Project 2 now due Monday 10/30 by 11:59pm.
10/16/17: There is no recitation or OH this Friday (Oct 20) because of mid-semester break.
10/14/17: Taylor will hold extra office hours this week on Wed at 12:30pm at GHC 6403 (Taylor's office).
10/10/17: Project 2 has been released. It is due Oct 27 at 11:59pm ET.
10/7/17: Slides from the sixth recitation posted here: PCFG Recitation Slides. (Powerpoint version included for fans of the animation!)
9/29/17: Slides from the fifth recitation posted here: CRF Recitation Slides. (typos fixed 10/2)
9/28/17: Notes from the fourth recitation posted here: HMM Recitation Notes.
9/22/17: Slides from the fourth recitation posted here: HMM Recitation Slides.
9/21/17: A sample writeup (but for a different assignment) is available here: Sample writeup.
9/21/17: Canvas is up. See project submission instructions below.
9/19/17: Project 1 now due Saturday 9/23 by 11:59pm. Submission details to follow
9/15/17: Slides from the third recitation posted here: Implementation Tricks Slides.
9/8/17: Notes from the second recitation posted here: KN Recitation Notes. (updated 9/10)
9/1/17: Project 1 has been released.
9/1/17: Slides from the first recitation posted here: Project Setup Recitation Slides.
8/31/17: Piazza link posted above.
8/25/17: First lecture will be on Tuesday 8/29 at 1:30pm in DH 1212.

Course Description

This course will explore current statistical techniques for the automatic analysis of natural (human) language data. The dominant modeling paradigm is corpus-driven statistical learning, with a split focus between supervised and unsupervised methods. This term we are making Algorithms for NLP a lab-based course. Instead of homeworks and exams, you will complete four hands-on coding projects. This course assumes a good background in basic probability and a strong ability to program in Java. Prior experience with linguistics or natural languages is helpful, but not required. There will be a lot of statistics, algorithms, and coding in this class.

Slides, materials, and projects for this new iteration of Algorithms for NLP are mainly borrowed from Dan Klein at UC Berkeley.

Project Submission

Submit projects using the class Canvas site.

1. Prepare a directory named 'project' containing no more than 3 files: (a) a jar named 'submit.jar', (b) a pdf named 'writeup.pdf', and (c) an optional jar named 'best.jar'. The jar named 'submit.jar' should contain your implementation of the core project that passes the basic requirements. For example, for project 1, the jar named 'assign1-submit.jar' is all that you would need to turn in -- renaming it 'submit.jar'. The pdf 'writeup.pdf' should contain your writeup for the project. Finally, the file 'best.jar' is an optional additional jar that implements the core project, but need not pass spot-checks. Include this last jar if you wish to demonstrate an improvement over the basic project, possibly using approximations are alternative models.

2. Compress the 'project' directory you created in the last step using the command 'tar cvfz project.tgz project'.

3. Click on the assignments tab of the main Canvas course site and select the assignment corresponding to the project (e.g. Assignment 1 corresponds to Project 1). Click 'Submit assignment' button to open submission portal, then click 'Choose file' and select your compressed project directory 'project.tgz' created in the previous step. Finally, click the 'Submit assignment' button below.

Project Grading

Projects out of 10 points total:
6 Points: Successfully implemented what we asked
2 Points: Submitted a reasonable write-up
1 Point: Write-up is written clearly
1 Point: Substantially exceeded minimum metrics
Extra Credit: Did non-trivial extension to project

Late Day Policy

Each student will be granted 5 late days to use over the duration of the semester. There are no restrictions on how the late days can be used (e.g. all 5 could be used on one project.) Using late days will not affect your grade. However, projects submitted late after all late days have been used will receive no credit. Be careful!

Readings

The primary recommended texts for this course are:

- Jurafsky and Martin, Speech and Language Processing, 2nd edition ONLY [amazon]
- Manning and Schuetze, Foundations of Statistical Natural Language Processing [amazon] [online]

Note that M&S is free online (may need to setup proxy). Also, make sure you get the purple 2nd edition of J+M, not the white 1st edition.

Note to Students

Take care of yourself! As a student, you may experience a range of challenges that can interfere with learning, such as strained relationships, increased anxiety, substance use, feeling down, difficulty concentrating and/or lack of motivation. All of us benefit from support during times of struggle. There are many helpful resources available on campus and an important part of having a healthy life is learning how to ask for help. Asking for support sooner rather than later is almost always helpful. CMU services are available, and treatment does work. You can learn more about confidential mental health services available on campus at: http://www.cmu.edu/counseling/. Support is always available (24/7) from Counseling and Psychological Services: 412-268-2922.

Accommodations for Students with Disabilities:

If you have a disability and have an accommodations letter from the Disability Resources office, I encourage you to discuss your accommodations and needs with me as early in the semester as possible. I will work with you to ensure that accommodations are provided as appropriate. If you suspect that you may have a disability and would benefit from accommodations but are not yet registered with the Office of Disability Resources, I encourage you to contact them at access@andrew.cmu.edu.

Syllabus [subject to substantial change!]

Week	Date	Topics	Readings	Assignments (Out)
1	Aug 29	Course Introduction	J+M 1, M+S 1-3	
	Aug 31	Language Modeling I	J+M 4, M+S 6, Chen & Goodman, Interpreting KN	P1: Language Modeling (Due
2	Sept 5	Language Modeling II	Massive Data, Bloom, Perfect, Efficient LMs	
	Sept 7	Language Modeling III		
3	Sept 12	Speech Recognition I	J+M 7	
	Sept 14	Speech Recognition II	J+M 9, Decoding	
4	Sept 19	Speech Recognition III, HMMs		
	Sept 21	POS Tagging, NER, CRFs	J+M 5, Brants, Toutanova & Manning	
5	Sept 26	Parsing I	M+S 3.2, 12.1, J+M 13	
	Sept 28	Parsing II	M+S 11, J+M 14, Best-First, A*, Unlexicalized	
6	Oct 3	Parsing III	Split, Lexicalized, K-Best A*, Coarse-to-fine	
	Oct 5	Formal Grammar		
7	Oct 10	Parsing IV		P2: PCFG Parser (Due Oct 3
	Oct 12	Parsing IV -- continued		
8	Oct 17	Parsing V		
	Oct 19	Parsing VI		
9	Oct 24	Structured Classification I	Pegasos, Cutting Plane	
	Oct 26	Structured Classification II	J+M 16, 18, 19, Adagrad, Subgradient SVM	
10	Oct 31	Guest lecture (Bhiksha Raj)		P3: Discriminative Reranker (
	Nov 2	Unsupervised transcription of language and music		
11	Nov 7	Machine Translation: Alignment I	J+M 25, IBM Models, HMM, Agreement, Discriminative	
	Nov 9	Machine Translation: Alignment II		
12	Nov 14	Machine Translation: Phrase-Based	Decoding	P4: Machine Translation (Due
	Nov 16	Machine Translation: Syntactic	Hiero, String-Tree, Tree-String, Tree-Tree	
13	Nov 21	Morphology; Features and Unification	J+M 3, J+M 15 (Note: errors in textbook)	
	Nov 23	Thanksgiving Day		
14	Nov 28	Semantics and Discourse I	J+M 17, J+M 18 (Note: errors in textbook)	
	Nov 30	Semantics and Discourse II	J+M 12.7.2	

15	Dec 5	Semantics and Discourse III	J+M 19, 20.6-20.9	
	Dec 7	Semantics and Discourse IV	J+M 21	