Introduction to NLP

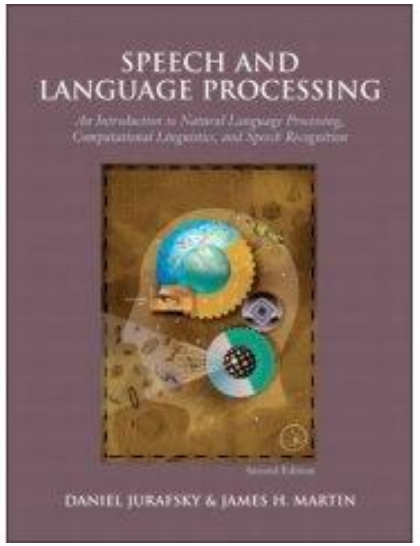
Class Logistics

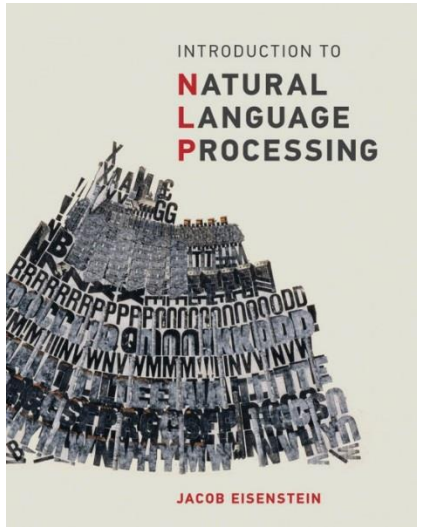
CPSC 477/577

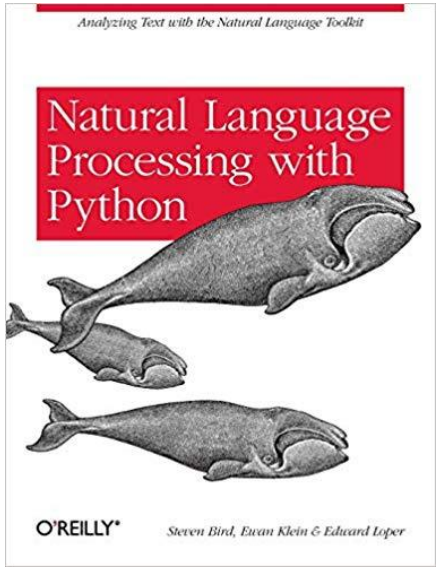
- Instructor:
  - Dragomir Radev
  - dragomir.radev@yale.edu
- Class times:
  - TTh 1-2:15
  - Location: Zoom
- Teaching staff:
  - Aditya Chander, Evan Cudone, Aarohi Srivastava, Michael Linden
- Office hours
  - Drago: F 1-2:15 (same Zoom link) or by appointment via email
  - Others: TBA

Course Readings

- Speech and Language Processing
  - Daniel Jurafsky and James Martin
  - Third edition, 2019
  - http://web.stanford.edu/~jurafsky/slp3/
- Introduction to Natural Language Processing
  - Jacob Eisenstein
  - First edition, 2019
  - https://github.com/jacobeisenstein/gt-nlp-class/blob/master/notes/eisenstein-nlp-notes.pdf
- Additional readings:
  - Natural Language Processing using NLTK (Bird et al.) http://www.nltk.org
  - AAN http://aan.how







Course Dates

Feb	2 4	9 11	16 18	23 25	
Mar	2 4	÷ 11	16 18	23 25	30
Apr	1	6 .	13 15	20 22	27 29
May	4 6				

- **No class** on March 9 (Tue) and April 8 (Thu)
- Official "middle of the term" date
  - Friday March 30
- Midterm
  - TBA (not necessarily on March 30)
- Final exam
  - TBA (during the week of May 14 – May 19)

https://yalecollege.yale.edu/academics/academic-calendar/2020-2021-academic-calendar

Structure of the Course

- Background
  - linguistic, mathematical, and computational
- Computational models
  - morphology, syntax, semantics, discourse, pragmatics
- Core NLP technology
  - parsing, part of speech tagging, text generation, semantic analysis
- Applications
  - text classification, sentiment analysis, text summarization, question answering, machine translation, etc.
- Neural Networks and Deep Learning
  - distributed semantics, sequence to sequence methods, attention, transformers

Major Goals of the Class

- **Learn** the basic principles and theoretical issues underlying natural language processing
- **Understand** how to view textual data from a linguistic and computational perspective
- **Appreciate** the complexity of language and the corresponding difficulty in building NLP systems
- **Learn** techniques and tools used to develop practical, robust systems that can understand text and communicate with users in one or more languages
- **Understand the limitations** of these techniques and tools
- Gain insight into some **open research problems** in natural language processing


Start of semester
End of semester

Draft Syllabus

Introduction	Question Answering
Language Modeling	Summarization
Part-of-Speech Tagging	Dialogue and Conversational Agents
Hidden Markov Models	Machine Translation
Formal Grammars of English	Sentiment Analysis
Syntactic Parsing	Vector Semantics
Statistical Parsing	Dimensionality Reduction
Features and Unification	Word Embeddings
Dependency Parsing	Neural Networks
The Representation of Meaning	Attention
Computational Semantics	Transformers
Lexical Semantics	Recent Developments

Linguistic Knowledge

- Constituents
  - Children eat pizza.
  - They eat pizza.
  - My cousin's neighbor's children eat pizza.
  - Eat pizza!
- Collocations
  - Strong beer but *powerful beer
  - Big sister but *large sister
  - Stocks rise but ?stocks ascend
- How to get this knowledge in the system
  - Manual rules (?)
  - Automatically acquired from large text collections (corpora)

Areas of Linguistics

- Phonetics and phonology
  - the study of sounds
- Morphology
  - the study of word components
- Syntax
  - the study of sentence and phrase structure
- Lexical semantics
  - the study of the meanings of words
- Compositional semantics
  - how to combine words
- Pragmatics
  - how to accomplish goals
- Discourse conventions
  - how to deal with units larger than utterances

Mathematical Background

- Linear algebra
  - vectors and matrices
- Probabilities
  - Bayes theorem
- Calculus
  - derivatives
- Optimization
- Numerical methods

Math Background Links

- Matrix multiplication
  - https://www.intmath.com/matrices-determinants/matrix-multiplication-examples.php
- Bayes theorem
  - https://betterexplained.com/articles/an-intuitive-and-short-explanation-of-bayes-theorem/
- Derivative of the sigmoid function
  - https://beckernick.github.io/sigmoid-derivative-neural-network/

Theoretical Computer Science

- Automata
  - Deterministic and non-deterministic finite-state automata
  - Push-down automata
- Grammars
  - Regular grammars
  - Context-free grammars
  - Context-sensitive grammars
- Complexity
- Algorithms
  - Dynamic programming

Artificial Intelligence

- Logic
  - First-order logic
- Agents
  - Speech acts
- Search
  - Planning
  - Constraint satisfaction
- Machine learning
  - Neural Networks
  - Reinforcement Learning

Grading

- Assignments (50%)
  - HW0+HW1 = 2+8=10%
  - HW2 = 10%
  - HW3 = 10%
  - HW4 = 10%
  - HW5 = 10%
- Exams (45%)
  - midterm = 20%
  - final exam = 25%
- Class participation (5%)
  - In-class participation, asking questions on Piazza, answering questions, office hours

How to get the most out of the class?

- Attend the lectures and study the slides
  - Course syllabus + slides = road map
  - Some material may not be found in any of the readings
- Hands on experience
  - Implement what you've learned
- Ask questions in and after class

Questions?

- Use the right channel for communication
  - Piazza/Canvas
- In special cases (e.g., sickness, regrading), use email
  - Include [CPSC477] or [CPSC577] or [NLP Class] in the subject line
- Office Hours:
  - TBA

NLP Courses at Other Places

- Brick-and-Mortar
  - Stanford (Chris Manning, Dan Jurafsky, Richard Socher, Chris Potts)
  - Texas (Greg Durrett)
  - CMU (Graham Neubig)
  - Johns Hopkins (Jason Eisner)
  - UNC (Mohit Bansal)
  - Utah (Vivek Srikumar)
- Online
  - Manning/Jurafsky (2012, survey)
  - Michael Collins (2013, more advanced)
  - Radev (2015-2016, survey)
  - New Coursera

The Association for Computational Linguistics (ACL)

Association for
Computational Linguistics

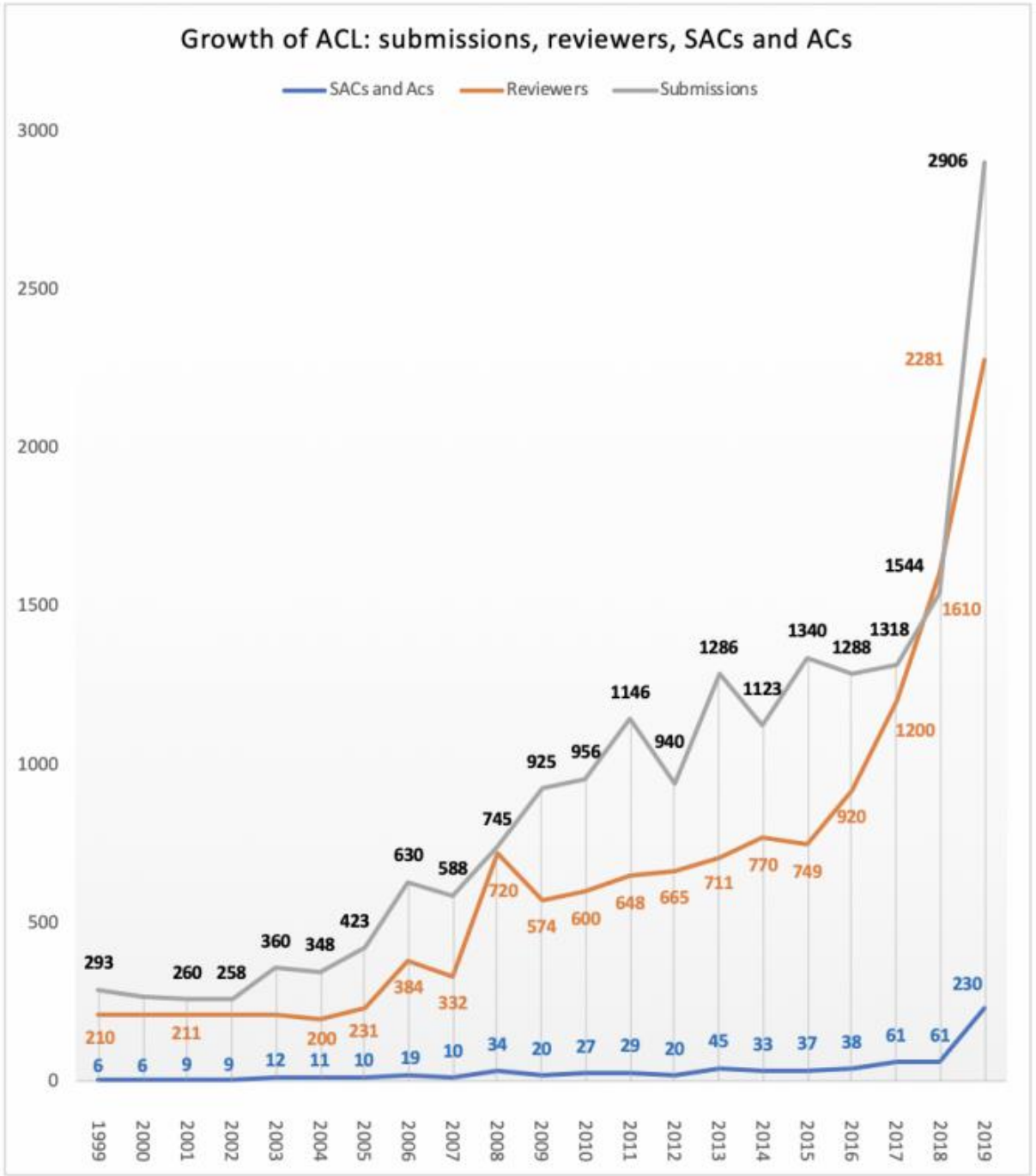
ACL 2020 Election Results

November 05, 2020 | BY webmaster

I am happy to announce the results of the elections for members of the ACL Executive Committee:

- Iryna Gurevych (Technical University (TU) of Darmstadt) has been elected as the VP-Elect.

- Yusuke Miyao (the University of Tokyo) has been elected as the Member at large.



https://aclweb.org/aclwiki/Conference_acceptance_rates

Research in NLP

- Conferences:
  - ACL, NAACL, EMNLP, SIGIR, AAAI/IJCAI, COLING, EACL, Interspeech, NeurIPS, ICLR, SIGDIAL
- Journals:
  - Computational Linguistics, TACL, Natural Language Engineering, Information Retrieval, Information Processing and Management, ACM Transactions on Information Systems, ACM TALIP, ACM TSLP
- University centers:
  - Stanford, Berkeley, Columbia, CMU, JHU, Brown, UMass, MIT, UPenn, Illinois, Michigan, Yale, Washington, Maryland, NYU, UNC, OSU, GA Tech, Princeton, etc.
  - Toronto, Edinburgh, Cambridge, Sheffield, Saarland, Trento, Prague, QCRI, NUS, and many others
- Industrial research sites:
  - Google, Facebook, MSR, IBM, SRI, BBN, MITRE, Baidu, Salesforce
- The ACL Anthology
  - http://www.aclweb.org/anthology
- The ACL Anthology Network (AAN)
  - http://aan.how

Students with Disabilities

- If you think you need an accommodation for a disability, please let me know at your earliest convenience.
- Some aspects of this course, the assignments, the in-class activities, and the way we teach may be modified to facilitate your participation and progress.
- I will treat any information that you provide in as confidential a manner as possible.

Student Mental Health and Wellbeing

- Yale University is committed to advancing the mental health and wellbeing of its students.
- If you or someone you know is feeling overwhelmed, depressed, and/or in need of support, services are available. Yale Counseling: **203-432-0290, 203-432-0123** (after hours)

Course Design Choices

- A wide-coverage survey course
- A mixture of traditional and neural techniques
- A non-trivial focus on linguistic issues
- A mixture of programming and written assignments
- Significant readings
- External links (tutorials)
- Fairly independent assignments

Programming environment

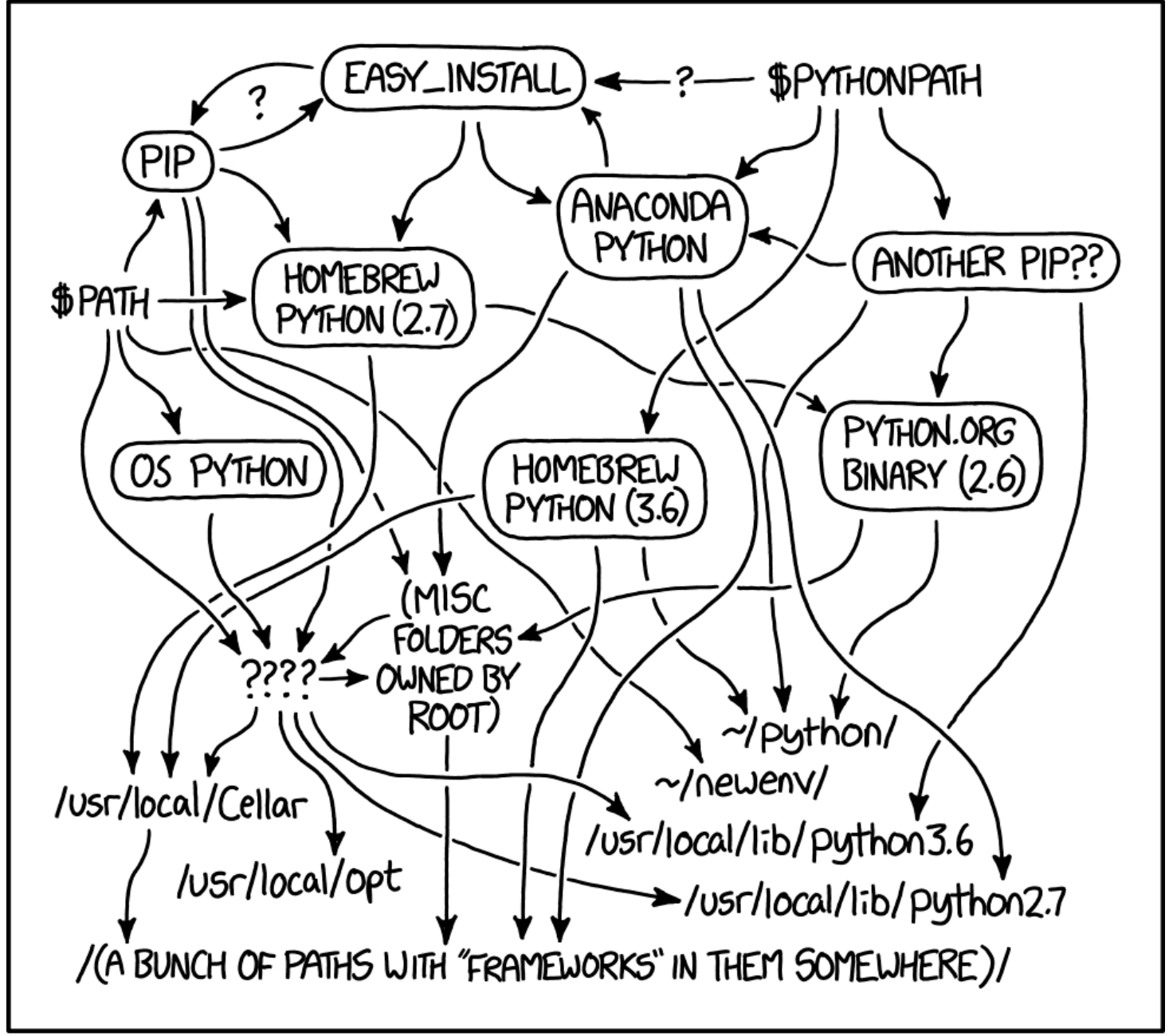
- python in a UNIX environment
- pytorch
- text manipulation
- scipy
- sklearn

Sample Programming Assignments

- Language Modeling and Part of Speech Tagging
- Dependency Parsing
- Vector Semantics and Word Sense Disambiguation
- Question Answering
- Deep Learning
- Machine Translation
- Sentiment Analysis
- Natural Language Interface to a Database
- Semantic Parsing

Programming Language

- The programming assignments will be in Python.
- You are expected to either know Python already or to learn it on your own.
- We will be using pytorch for most assignments.
- The code base will be installed on the **Zoo** machines.



MY PYTHON ENVIRONMENT HAS BECOME SO DEGRADED
THAT MY LAPTOP HAS BEEN DECLARED A SUPERFUND SITE.

Submitting Assignments

- In the absence of a prior emailed authorization from the instructor, you should turn in your assignments electronically by 11:59:59 PM on the due date. For each day (or fraction of a day) that your submission is late, it will be penalized 10%, for a maximum of 30%. After three days, the assignment will be given a score of zero.
- You will need to hand in the source code for the project, relevant documentation, and a script of a test run of your program to show that it actually works on the Zoo machines.

Academic Honesty

- Unless otherwise specified in an assignment all submitted work must be your own, original work. Any excerpts, statements, or phrases from the work of others must be clearly identified as a quotation, and a proper citation provided.
- Any violation of the University's policy on Academic and Professional Integrity will result in serious penalties, which might range from failing an assignment, to failing a course, to being expelled from the program.
- Violations of academic and professional integrity will be reported to Student Affairs. Consequences impacting assignment or course grades are determined by the faculty instructor; additional sanctions may be imposed.

Integrity Policies

- Collaboration policy:
  - You may discuss the course material and the textbook with other students. You may also discuss the *requirements* of the assignments. However, you cannot get help with the assignments and exams themselves in oral or written form from anyone. If you are unsure about this policy, ask the instructors.
- Honesty policy:
  - We will be using high grade plagiarism detection code
  - Do not copy other people's code or misrepresent it as yours, period.

Specifics

- Coding and write up should be done independently
- Do not show your work to anyone
- Do not look at anyone’s work
- Do not use existing web code (e.g., github) that is specifically designed to solve the problem in the assignment. If you use other github code, attribute it properly in your submission

Grading Appeals

- If you have a question about your grade on a particular assignment (or exam), write a short email to the TA in charge of that assignment.
- Please submit any such requests within a week of receiving your grade.