LING/C SC/PSYC 438/538

Lecture 1

Sandiway Fong

Contents

- Syllabus
- Questions about the Syllabus
- Homeworks 1 and 2 (*possibly the null homework*)
- Some real intro next time!

Syllabus

Description of Course

- An introductory level course at the advanced level for computational linguistics. Required core course for the Master's Human Language Technology (HLT) program.

Course Pre-requisites

- 438: LING 388 or familiarity with one or more of the following: formal languages, syntax,
- data structures, or compilers.
- 538: no formal pre-requisites.

Syllabus

Instructor and Contact Information

- Instructor: Sandiway Fong, Douglass 311.
- Contact email: sandiway@email.arizona.edu (all homework to be submitted here).
- Instructor: Sandiway Fong, Dept. of Linguistics Office: Douglass 311

Hours:

- make appointments by email: meet online
- Zoom during class time (*best if you have quick Qs*)

Meet:

- on Zoom (watch your email!) and Facebook Messenger (sandiway) with Panopto slides
- No class on Thursday Nov 26th (Thanksgiving)
  - *updated as necessary…*

Syllabus

Course Format and Teaching Methods

- Lecture with slides. Panopto videos (when available) for lecture review.
- All homeworks will be introduced and reviewed in Zoom class.

Course Objectives

Topics covered include:

- Introductory programming relevant to computational linguistics in two or more programming languages. We will use Perl, Python and Prolog this semester.
- Introduction to a range of topics in computational linguistics, see detailed list of topics later below.

Syllabus

Course Learning Outcomes

After completing this course, students will:

1. Have acquired the ability to read and write programs in two or more programming languages.

- Relates to Linguistics Department HLT program outcome #1.

2. Be familiar with basic concepts, techniques and applications in computational linguistics.

- Relates to Linguistics Department HLT program outcome #2 and Linguistics Department Undergraduate program outcome #1.

3. **538-only**: be able to present and explain advanced concepts in computational linguistics. (See chapter presentation requirement.)
4. Be equipped to take more advanced classes in computational linguistics, e.g. 581 (Spring) or 439/539 (Statistical NLP).

Syllabus

Absence and Class Participation Policy

- I expect you to attend lectures (though attendance will not be taken).
- The UA's policy concerning Class Attendance, Participation, and Administrative Drops is available at: http://catalog.arizona.edu/policy/class-attendance-participation-and-administrative-drop.
- Tell me ahead of time so we can make alternative arrangements in the case of missed homeworks. **No homework will be accepted late. Explained below.**
- Absences pre-approved by the UA Dean of Students (or Dean Designee) will be honored. See: https://deanofstudents.arizona.edu/absences.
- The UA policy regarding absences for any sincerely held religious belief, observance or practice will be accommodated where reasonable, http://policy.arizona.edu/human-resources/religious-accommodation-policy.

Syllabus

Required Text

- 438: None.
- 538: *Speech and Language Processing*, Jurafsky & Martin, 2nd edition, Prentice Hall 2008.

Required or Special Materials

- All required software will be available online at no cost to the student.
- However, students are expected to either have a laptop/desktop capable of handling homework and classwork, or make use of UA lab computers (?)
- Mac, PC (Windows 10) or Linux.

Syllabus

Assignments and Examinations: Schedule/Due Dates

- All homeworks will be introduced **and reviewed** in Zoom class.
- Homework submissions by email to me only.
- Late homework will be not accepted since all homeworks will be solved/reviewed in class.
- Quick homeworks are normally due at midnight before the next class, and are generally assigned in class on a **Tuesday** and due **Wednesday** midnight (before **Thursday** class).
- Homeworks not categorized as quick are normally assigned in class on a **Thursday** and due the following **Monday** midnight (before **Tuesday** class). As deemed appropriate by the instructor, some longer homeworks may have an extended due date.
- Students can expect a total of 10-14 homeworks over the course.
- An in-class quick quiz or two maybe be scheduled. If so, students will be notified ahead of time during the prior class.

Syllabus

Final Examination or Project

- No examinations, e.g. mid-term or final, are scheduled for this course.

Grading Scale and Policies

- **438**:
  - 100% of the grade comes from the homework assignments.
- **538**:
  - 75% of the grade comes from the homework assignments (possibly a superset of the 438 assignments), 25% of the grade comes from a textbook chapter presentation.
- Requests for incomplete (I) or withdrawal (W) must be made in accordance with University policies, which are available at http://catalog.arizona.edu/policy/grades-and-grading-system#incomplete and http://catalog.arizona.edu/policy/grades-and-grading-system#Withdrawal respectively.

Syllabus

Scheduled Topics/Activities

- Topics will be drawn from the following:
- – Programming Languages: Perl and Python
- – Regular Expressions (Theoretical and practical)
- – Automata (Finite State) and Transducers (Finite State)
- – Programming Language: Prolog (definite clause grammars)
- – NLTK (Natural Language Toolkit)
- – Part of Speech (POS) Tagging
- – Stemming (Morphology)
- – Edit Distance (Spelling)
- – Grammars (Regular, Context-free)
- – Parsing (Syntax trees, algorithms)
- – *and more ...*

Syllabus

Code of Academic Integrity

- You may discuss homework questions with anyone.
- You may look things up on the web and use answers found therein; however, you must write it up yourself (in your own words/own code *etc*.).
- You must cite all (web) references and your classmates (in the case of shared discussion).
- Students are encouraged to share intellectual views and discuss freely the principles and applications of course materials.
- However, graded work/exercises must be the product of independent effort unless otherwise instructed.
- Students are expected to adhere to the UA Code of Academic Integrity as described in the UA General Catalog. See: http://deanofstudents.arizona.edu/academic-integrity/students/academic-integrity.

Syllabus

UA Nondiscrimination and Anti-harassment Policy

- The University is committed to creating and maintaining an environment free of discrimination; see http://policy.arizona.edu/human-resources/nondiscrimination-and-anti-harassment-policy.

Subject to Change Statement

- Information contained in the course syllabus, other than the grade and absence policy, may be subject to change with advance notice, as deemed appropriate by the instructor.

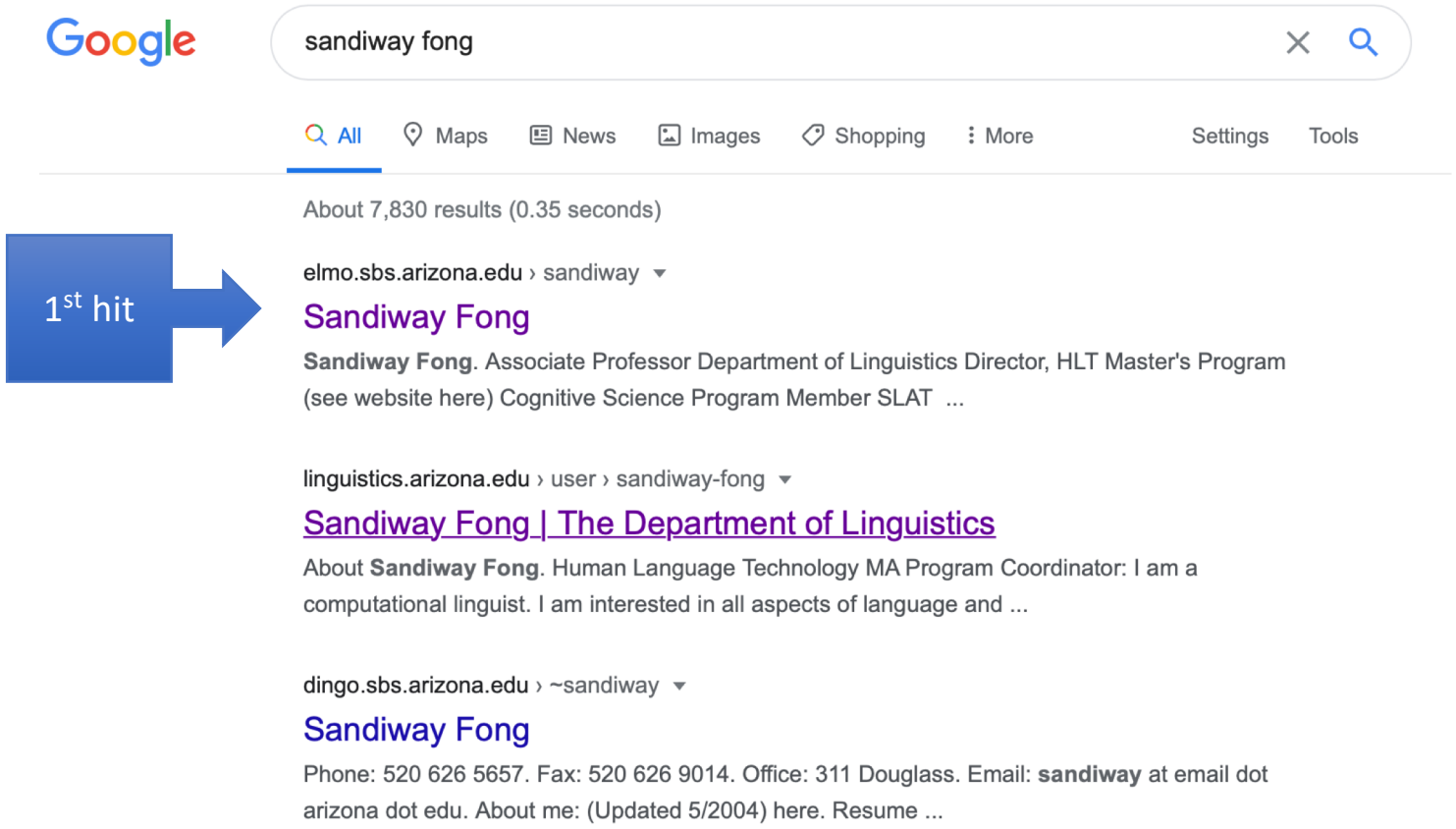
Syllabus

- Questions?

Course website

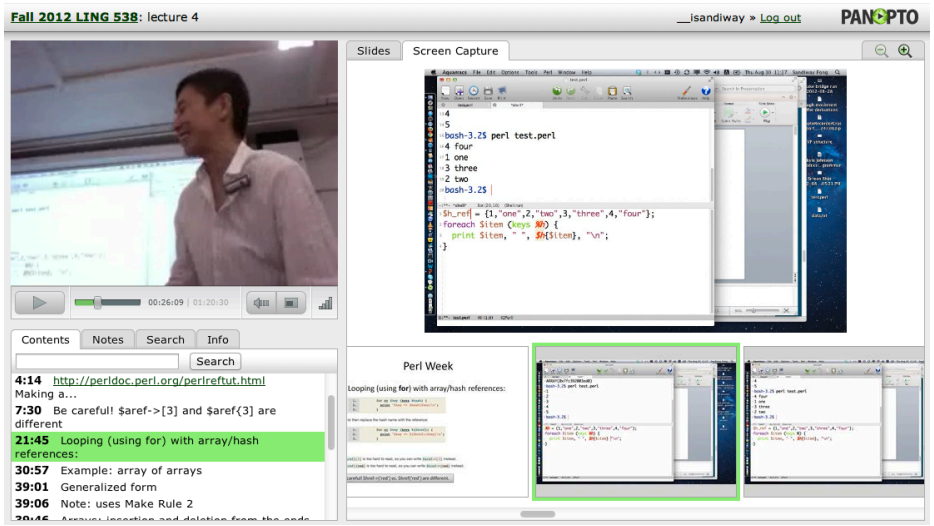
- Download lecture slides from my homepage
  - http://elmo.sbs.arizona.edu/~sandiway/#courses
  - available from just before class time
    - (afterwards, please look again for updates and corrections)
  - in .pptx (good for animations) and .pdf formats

Course website

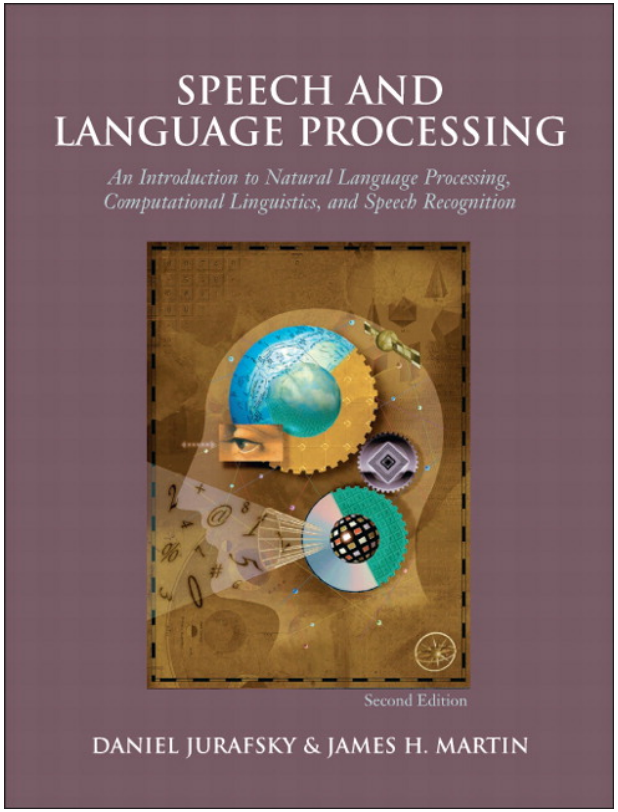


Panopto

- Lectures will be recorded using the panopto system
  - accessible via the course webpage and your browser
  - **sometimes crashes**
  - (video, laptop screen, synchronized slides, keyword search)



Textbook (J&M)



2008 (2nd edition)

Nearly 1000 pages
(*maybe more than a full year's worth…*)
25 chapters
Divided into 5 parts

I. **Words**
II. Speech – *not this course*
III. Syntax
IV. Semantics and Pragmatics
V. Applications

Book chapters

- 1. Introduction
- 1.1. Knowledge in Speech and Language Process
- 1.2. Ambiguity
- 1.3. Models and Algorithms
- 1.4. Language, Thought, and Understanding
- 1.5. The State of the Art
- 1.6. Some Brief History
- 1.6.1. Foundational Insights: 1940s and 1950s
- 1.6.2. The Two Camps: 1957–1970
- 1.6.3. Four Paradigms: 1970–1983
- 1.6.4. Empiricism and Finite-State Models Redux
- 1.6.5. The Field Comes Together: 1994–1999
- 1.6.6. The Rise of Machine Learning: 2000–2008
- 1.6.7. On Multiple Discoveries
- 1.6.8. A Final Brief Note on Psychology
- 1.7. Summary
- Bibliographical and Historical Notes
- I. Words
- 2. Regular Expressions and Automata
- 2.1. Regular Expressions
- 2.1.1. Basic Regular Expression Patterns
- 2.1.2. Disjunction, Grouping, and Precedence
- 2.1.3. A Simple Example
- 2.1.4. A More Complex Example
- 2.1.5. Advanced Operators

1. Introduction
1.1. Knowledge in Speech and Language Processing
1.2. Ambiguity
1.3. Models and Algorithms
1.4. Language, Thought, and Understanding
1.5. The State of the Art
1.6. Some Brief History
1.6.1. Foundational Insights: 1940s and 1950s
1.6.2. The Two Camps: 1957–1970
1.6.3. Four Paradigms: 1970–1983
1.6.4. Empiricism and Finite-State Models Redux: 1983–1993
1.6.5. The Field Comes Together: 1994–1999
1.6.6. The Rise of Machine Learning: 2000–2008
1.6.7. On Multiple Discoveries
1.6.8. A Final Brief Note on Psychology
1.7. Summary
Bibliographical and Historical Notes

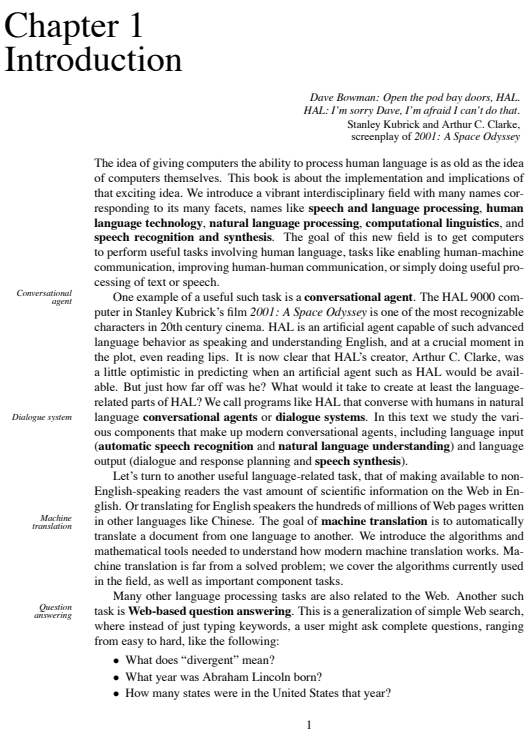
- 4.3.2. Unknown Words: Open Versus Closed Vocabulary Tasks
- 4.4. Evaluating N-Grams: Perplexity

Syllabus

- I’m gonna assume you don’t know how to program at all (*yet*)
  - we’re going to use Perl and Python
  - good to learn both …
  - good to be polyvalent
- Topics: selected chapters from J&M
  - Chapters 1–6, skip Speech part (7–11), 12–25

Homework: Reading

- Chapter 1 from JM
  - **homework 1:**
    - **READ IT before next time!**
    - **in-class Quick Quiz on Thursday**
  - available online https://www.cs.colorado.edu/~martin/SLP/Updates/1.pdf
- Whole book is available as an e-book
  - www.coursesmart.com



Chapter 1
Introduction

Dave Bowman: Open the pod bay doors, HAL.
HAL: I'm sorry Dave, I'm afraid I can't do that.
Stanley Kubrick and Arthur C. Clarke,
screenplay of *2001: A Space Odyssey*

The idea of giving computers the ability to process human language is as old as the idea of computers themselves. This book is about the implementation and implications of that exciting idea. We introduce a vibrant interdisciplinary field with many names corresponding to its many facets, names like **speech and language processing**, **human language technology**, **natural language processing**, **computational linguistics**, and **speech recognition and synthesis**. The goal of this new field is to get computers to perform useful tasks involving human language, tasks like enabling human-machine communication, improving human-human communication, or simply doing useful processing of text or speech.

Conversational agent

One example of a useful such task is a **conversational agent**. The HAL 9000 computer in Stanley Kubrick's film *2001: A Space Odyssey* is one of the most recognizable characters in 20th century cinema. HAL is an artificial agent capable of such advanced language behavior as speaking and understanding English, and at a crucial moment in the plot, even reading lips. It is now clear that HAL's creator, Arthur C. Clarke, was a little optimistic in predicting when an artificial agent such as HAL would be available. But just how far off was he? What would it take to create at least the language-related parts of HAL? We call programs like HAL that converse with humans in natural language **conversational agents** or **dialogue systems**. In this text we study the various components that make up modern conversational agents, including language input (**automatic speech recognition** and **natural language understanding**) and language output (dialogue and response planning and **speech synthesis**).

Dialogue system

Machine translation

Let's turn to another useful language-related task, that of making available to non-English-speaking readers the vast amount of scientific information on the Web in English. Or translating for English speakers the hundreds of millions of Web pages written in other languages like Chinese. The goal of **machine translation** is to automatically translate a document from one language to another. We introduce the algorithms and mathematical tools needed to understand how modern machine translation works. Machine translation is far from a solved problem; we cover the algorithms currently used in the field, as well as important component tasks.

Question answering

Many other language processing tasks are also related to the Web. Another such task is **Web-based question answering**. This is a generalization of simple Web search, where instead of just typing keywords, a user might ask complete questions, ranging from easy to hard, like the following:

- What does "divergent" mean?
- What year was Abraham Lincoln born?
- How many states were in the United States that year?

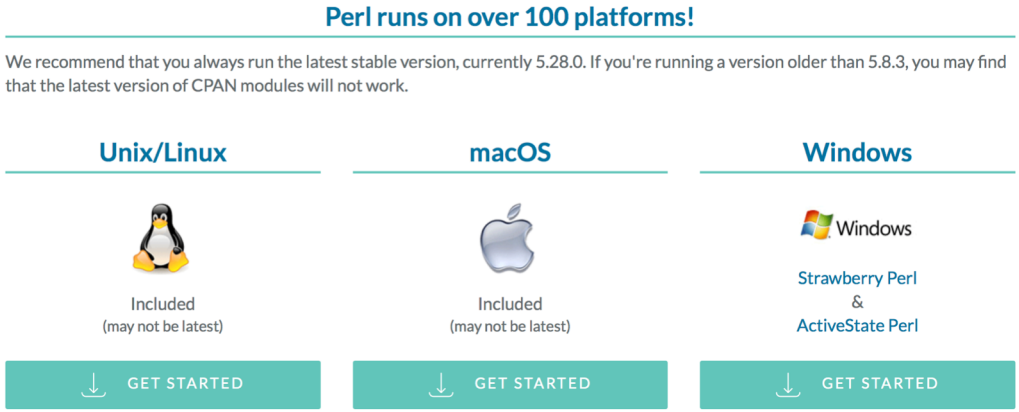
1

Homework 2

- Install Perl and Python (version 3, **not** 2.7)
  - could be the null homework …

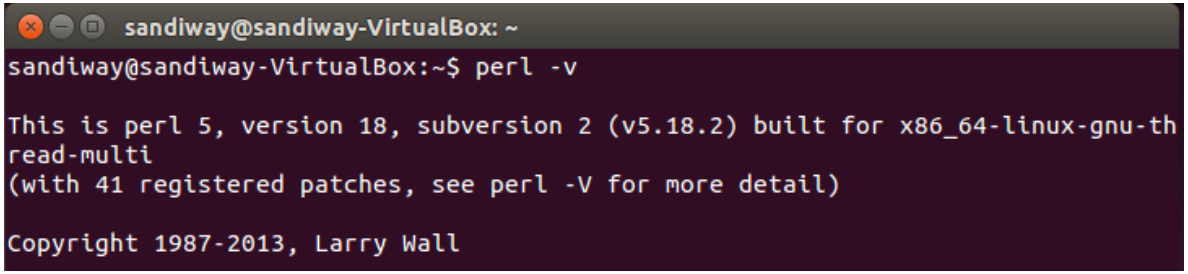
Homework: Install Perl

- Install Perl on your laptop
  - should be pre-installed on macs and Linux, check your machine from the Terminal/command line
  - on Windows PCs, if you don't already have it, it's freely available here
  - http://www.activestate.com/ (*don't pay, get the free version*) **or**
  - www.perl.org



Homework: Install Perl

- Ubuntu (Terminal):

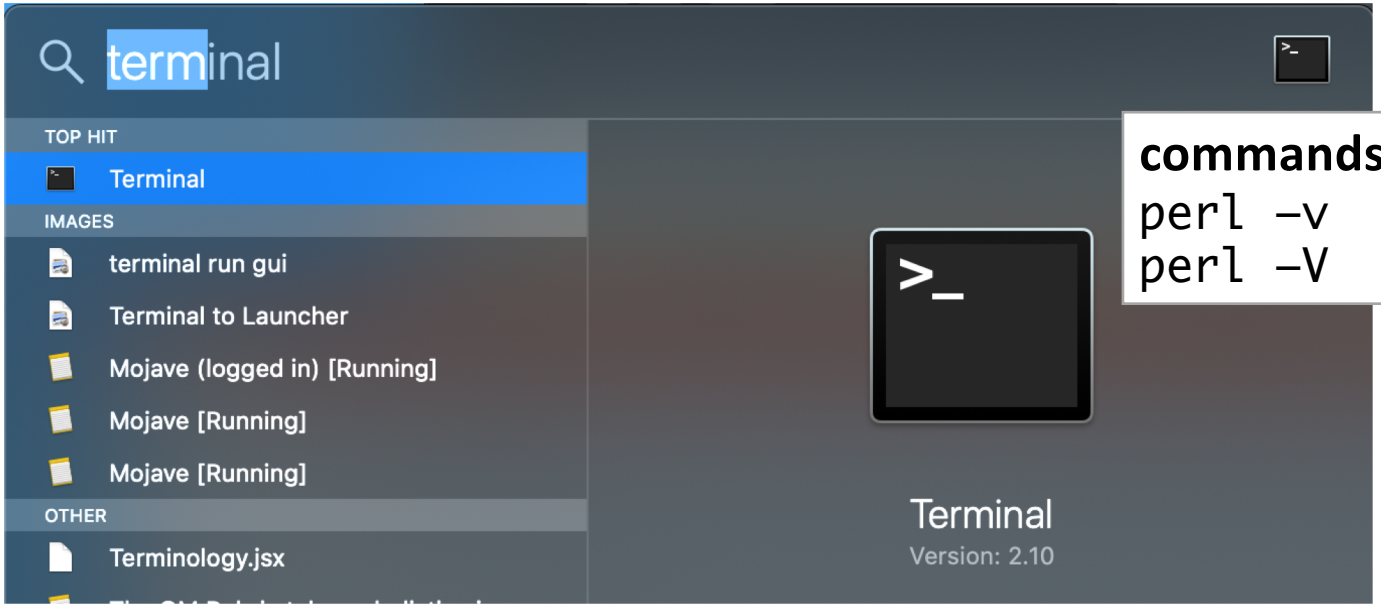


commands:

```
perl –v
which perl
```

Homework: Install Perl

- Mac (Terminal): (*complete path specified here*)
  - `/usr/bin/perl`

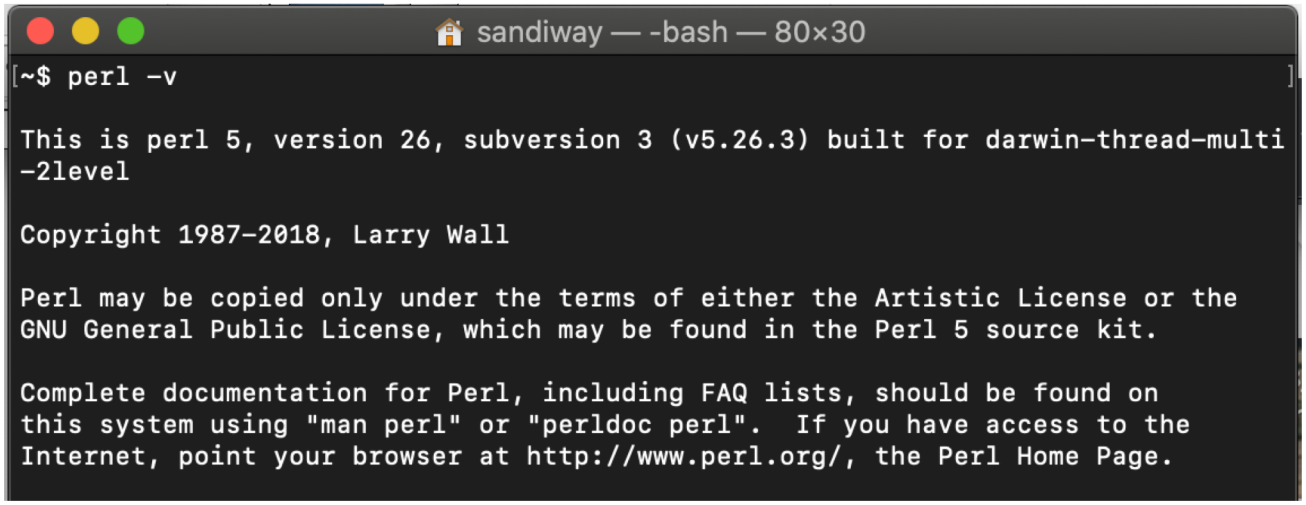


commands:

```
perl –v
perl –V        (much more verbose)
```

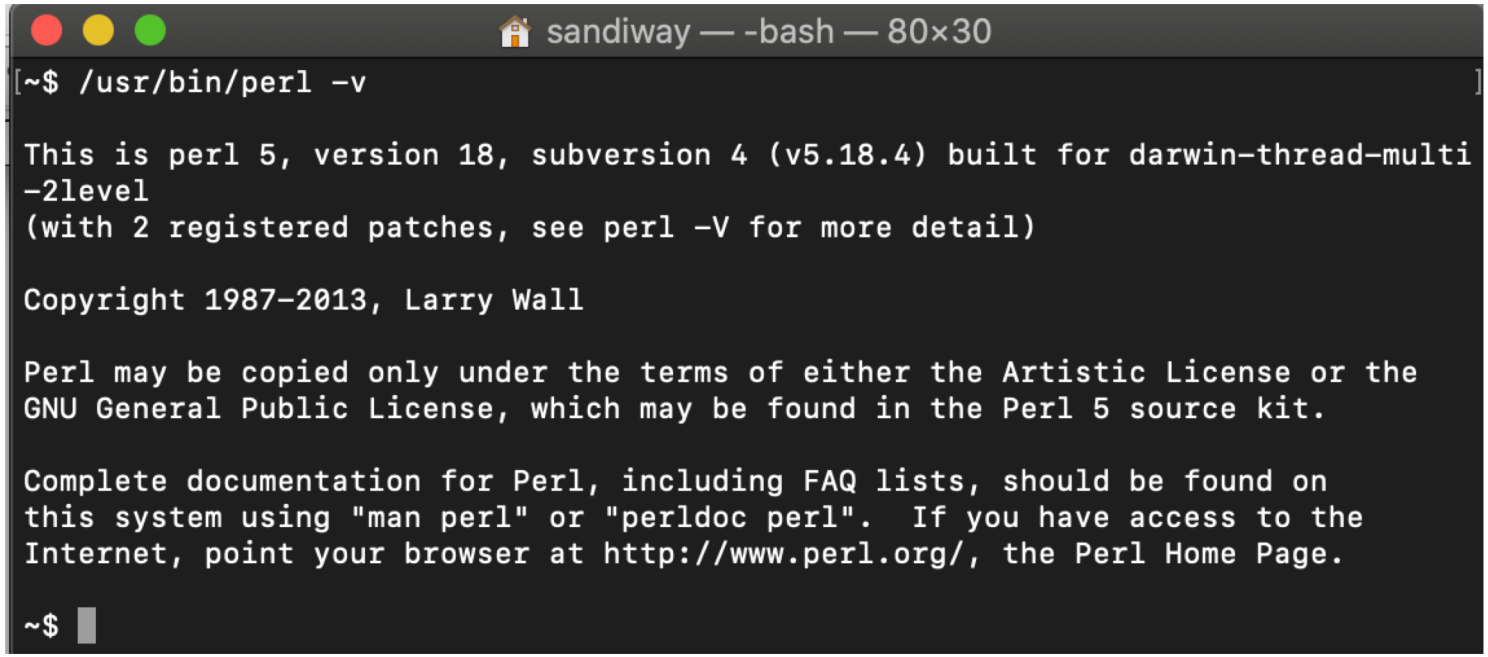
Homework: Install Perl

- Actually, on my Mac laptop, typing perl at the Terminal gives …



Homework: Install Perl

- On my Mac laptop, typing /usr/bin/perl –v at the Terminal gives …



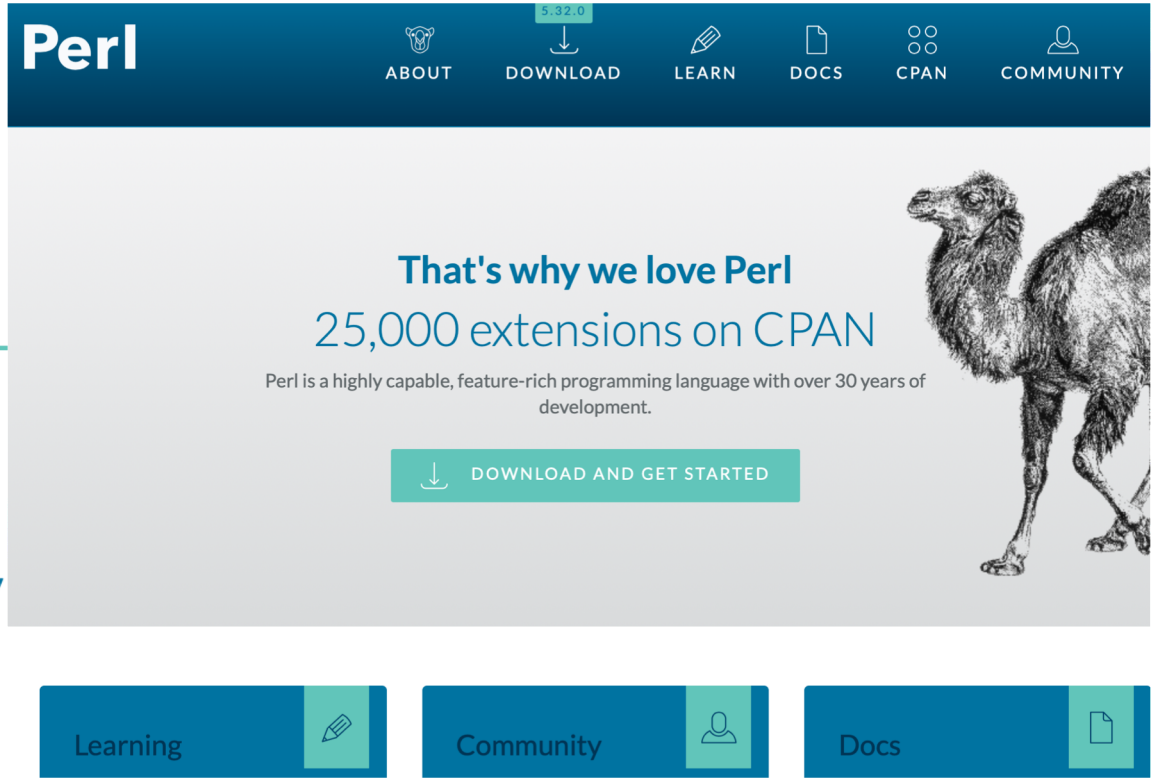
Homework: Install Perl

https://www.perl.org/learn.html

start here next time

Get Started

- learn.perl.org
- A brief introduction
- Free online Perl books
- Join your local community
- Books and More



Learning Perl

- Learn Perl
  - https://perldoc.perl.org/perlintro.html

Home > Overview > perlintro

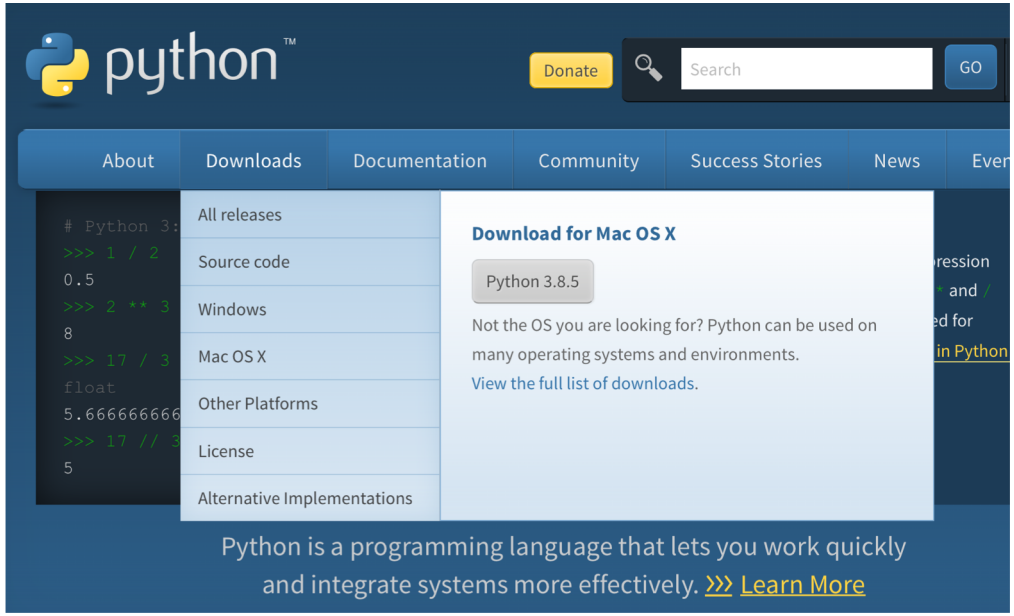
Perl 5 version 32.0 documentation

perlintro

- NAME
- DESCRIPTION
  - What is Perl?
  - Running Perl programs
  - Safety net
  - Basic syntax overview
  - Perl variable types
  - Variable scoping
  - Conditional and looping constructs
  - Builtin operators and functions

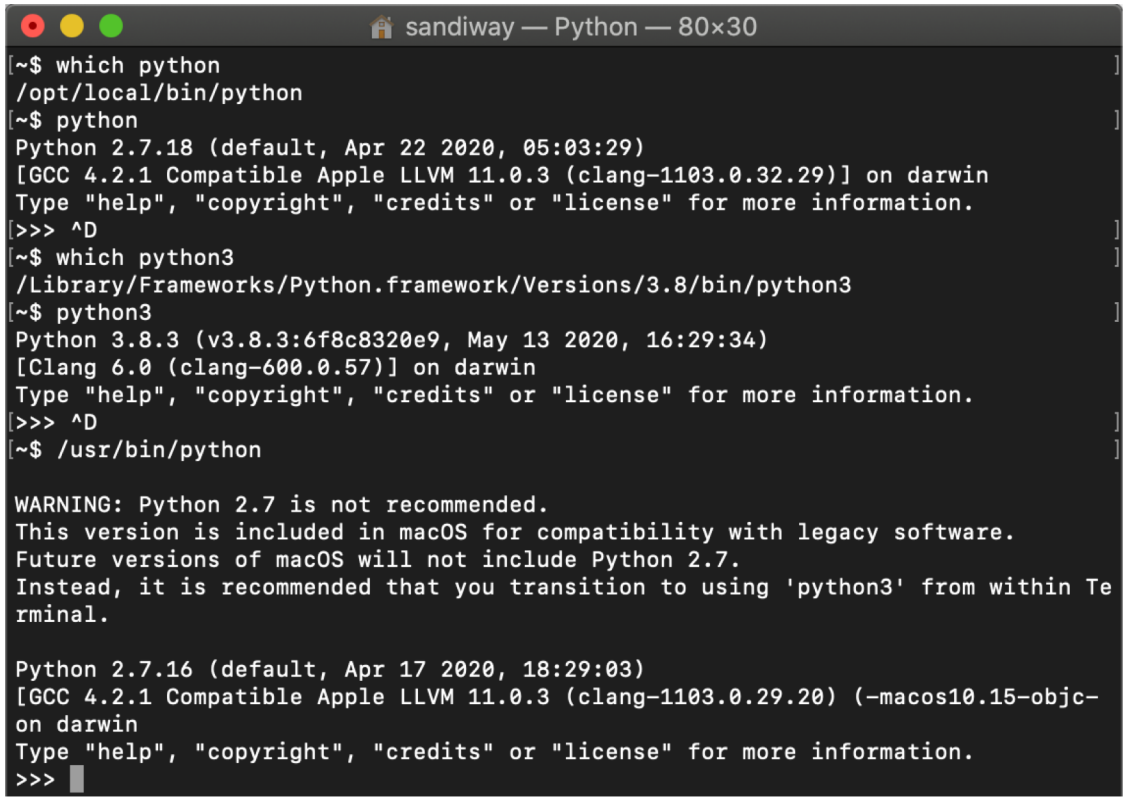
Homework: Install Python

- www.python.org
- **Note**: 3.x is not quite backwards compatible with Python 2.7!



Homework: Install Python

- On my Mac laptop, I have different versions:



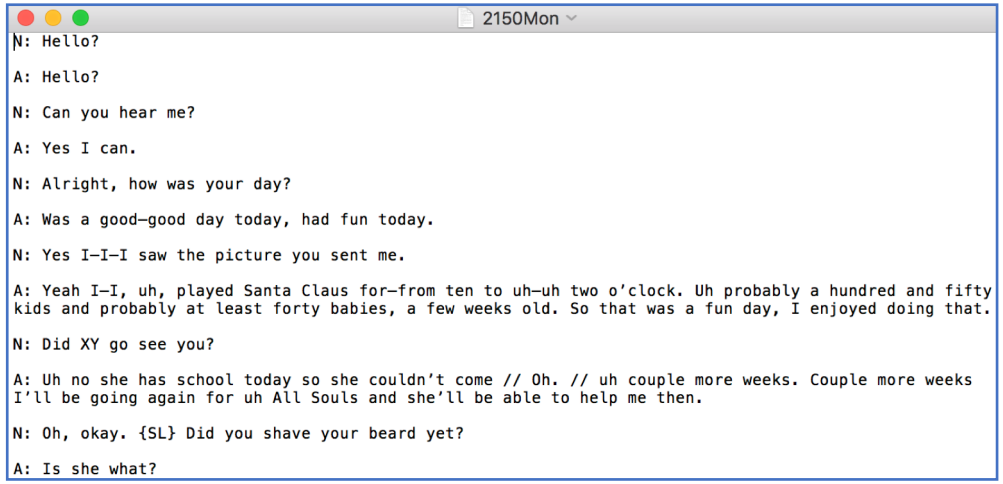
Which one is easier?

A subjective question …

- All good programmers know more than one programming language
- In NLP, Python is *de facto* overwhelmingly popular we will do both Perl and Python, see pros and cons here: https://www.tecmint.com/python-vs-perl-debate-what-should-i-learn-python-or-perl/

Simplest level of programming

- how do I automate repetitive tasks (**scripting**) on my laptop?
  - learn bash shell scripting…
- Search: how do I look for patterns (**regex**) in a text corpus?
  - `grep -Eo ':\s+Yea[^h]' 2150Mon.txt`
  - `perl -nle 'print $1 if /(.):\s+Yep/' 2149Mon.txt`



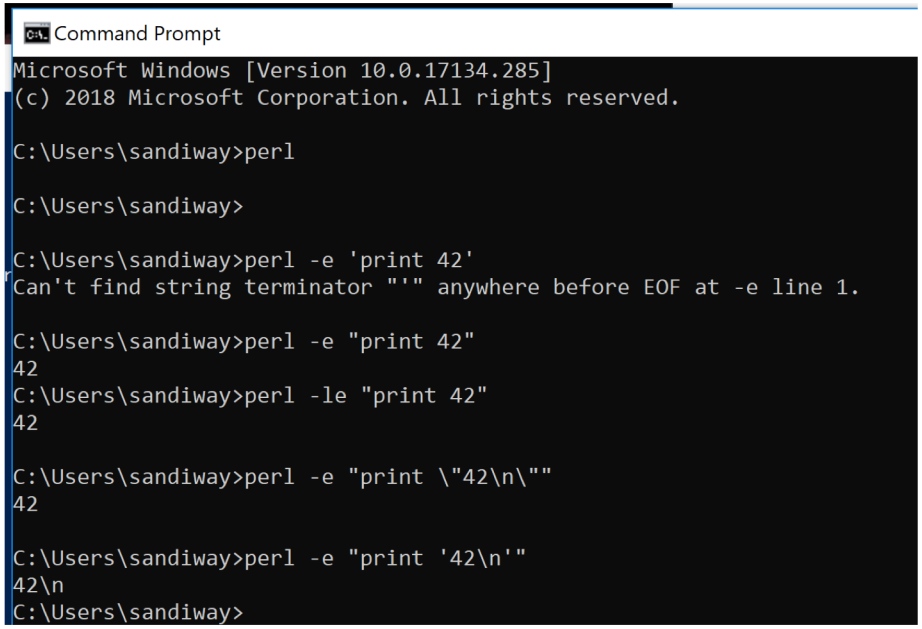
Between the quotes we have a program!

command line options: `–nle`

-n	loop over lines of the input file
-l	print newlines
-e	execute code from command line

Windows 10: command prompt

- Windows 10 command prompt
  - single quotes not recognized as quotes
  - double quotes can be used
- inside the (initial) double quotes, everything will be passed to Perl:
  - you can use single quotes
  - you can escape double quotes, i.e. \" to get a double quote.



Windows 10: PowerShell

- You can also use PowerShell in Windows 10:
  - both single and double quotes can be used as quoting characters in the shell.
  - Inside the initial quotes, you can escape any quotes that need to be passed to Perl.