

# 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](mailto: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 26<sup>th</sup> (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/~sandaway/#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



A screenshot of a Google search results page. The search query "sandiway fong" is entered in the search bar. The "All" tab is selected, showing approximately 7,830 results. The first result is highlighted with a blue box and an arrow pointing to it, labeled "1st hit". The result is from elmo.sbs.arizona.edu, titled "Sandiway Fong". Below it is a snippet: "Sandiway Fong. Associate Professor Department of Linguistics Director, HLT Master's Program (see website here) Cognitive Science Program Member SLAT ...". The second result is from linguistics.arizona.edu, titled "Sandiway Fong | The Department of Linguistics". Below it is a snippet: "About Sandiway Fong. Human Language Technology MA Program Coordinator: I am a computational linguist. I am interested in all aspects of language and ...". The third result is from dingo.sbs.arizona.edu, titled "Sandiway Fong". Below it is a snippet: "Phone: 520 626 5657. Fax: 520 626 9014. Office: 311 Douglass. Email: sandiway at email dot arizona dot edu. About me: (Updated 5/2004) here. Resume ...".

1<sup>st</sup> hit →

sandiway fong

All Maps News Images Shopping More Settings Tools

About 7,830 results (0.35 seconds)

elmo.sbs.arizona.edu › sandiway ▾

**Sandiway Fong**

Sandiway Fong. Associate Professor Department of Linguistics Director, HLT Master's Program (see website here) Cognitive Science Program Member SLAT ...

linguistics.arizona.edu › user › sandiway-fong ▾

**Sandiway Fong | The Department of Linguistics**

About Sandiway Fong. Human Language Technology MA Program Coordinator: I am a computational linguist. I am interested in all aspects of language and ...

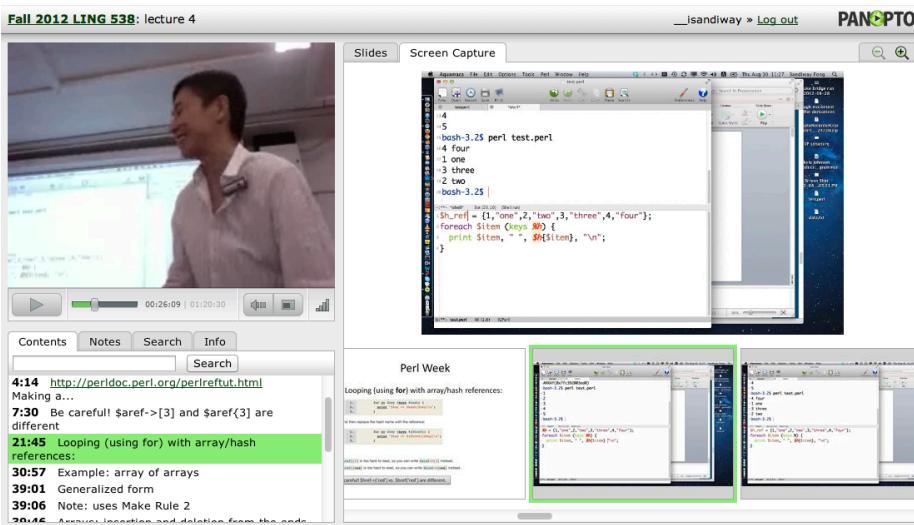
dingo.sbs.arizona.edu › ~sandiway ▾

**Sandiway Fong**

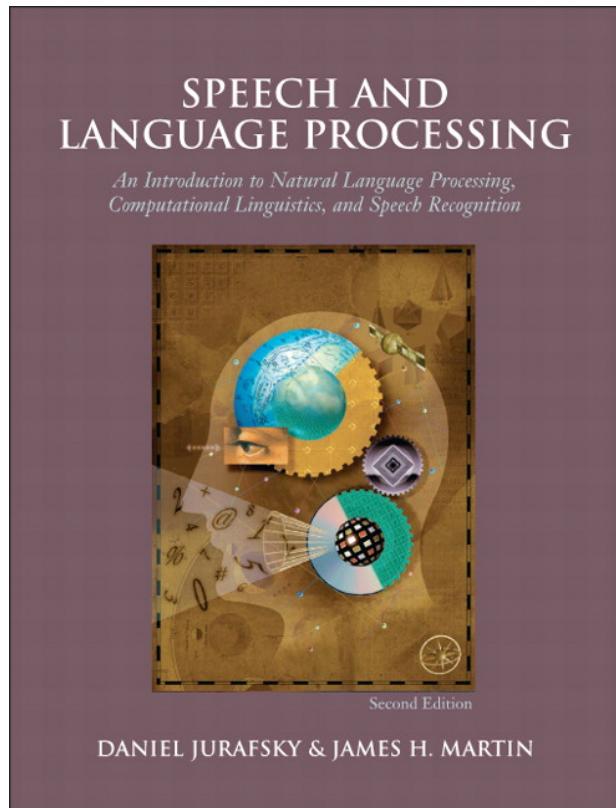
Phone: 520 626 5657. Fax: 520 626 9014. Office: 311 Douglass. Email: sandiway at email dot arizona dot edu. About me: (Updated 5/2004) here. Resume ...

# 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 (2<sup>nd</sup> 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 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
  - 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
  - 4.3.2. Unknown Words: Open Versus Closed Vocabulary Tasks
  - 4.4. Evaluating N-Grams: Perplexity

**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**

# 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](http://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*

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.

*Machine translation*

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?

# 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](http://www.perl.org)

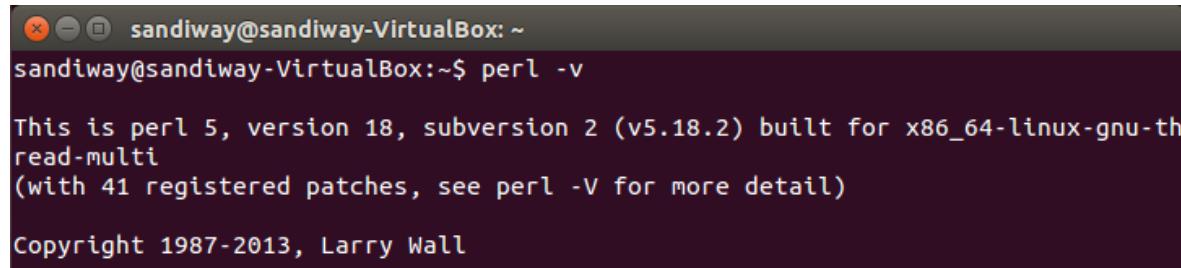
## Perl runs on over 100 platforms!

We recommend that you always run the latest stable version, currently 5.28.0. If you're running a version older than 5.8.3, you may find that the latest version of CPAN modules will not work.

Unix/Linux	macOS	Windows
 Included (may not be latest)	 Included (may not be latest)	 Windows  Strawberry Perl & ActiveState Perl
<a href="#">GET STARTED</a>	<a href="#">GET STARTED</a>	<a href="#">GET STARTED</a>

# Homework: Install Perl

- Ubuntu (Terminal):



```
sandiway@sandiway-VirtualBox: ~
sandiway@sandiway-VirtualBox:~$ perl -v

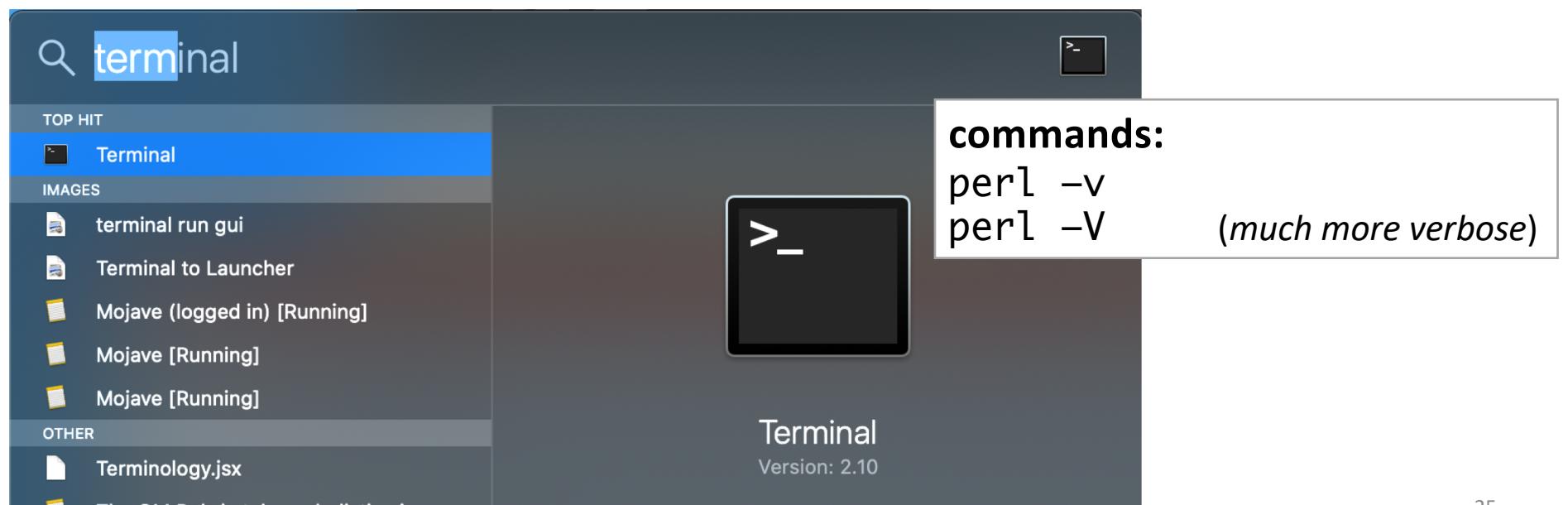
This is perl 5, version 18, subversion 2 (v5.18.2) built for x86_64-linux-gnu-thread-multi
(with 41 registered patches, see perl -V for more detail)

Copyright 1987-2013, Larry Wall
```

**commands:**  
perl -v  
which perl

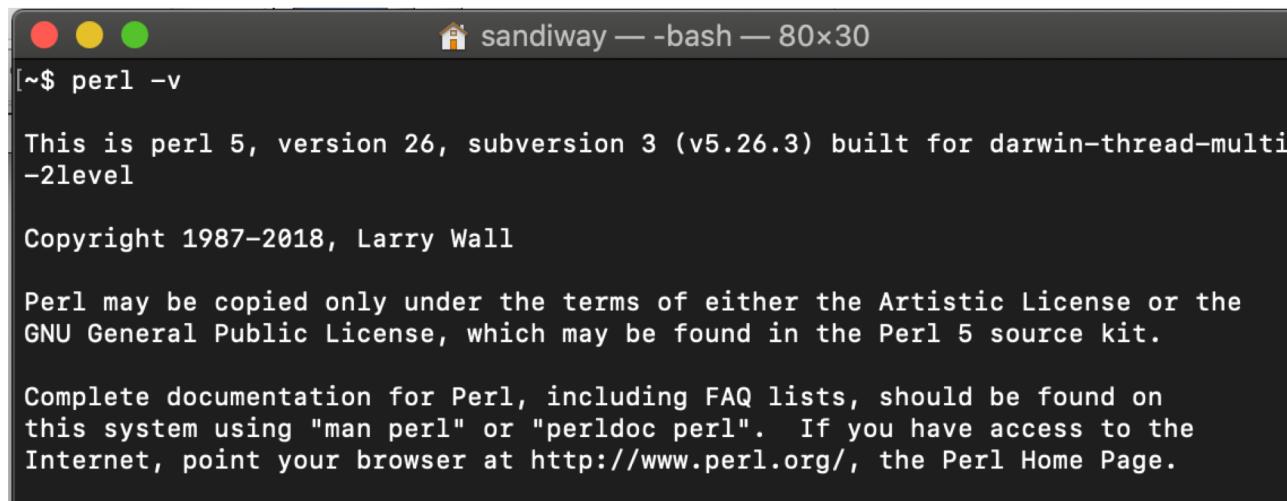
# Homework: Install Perl

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



# Homework: Install Perl

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



```
[~$ perl -v
This is perl 5, version 26, subversion 3 (v5.26.3) built for darwin-thread-multi-2level
Copyright 1987–2018, Larry Wall

Perl may be copied only under the terms of either the Artistic License or the
GNU General Public License, which may be found in the Perl 5 source kit.

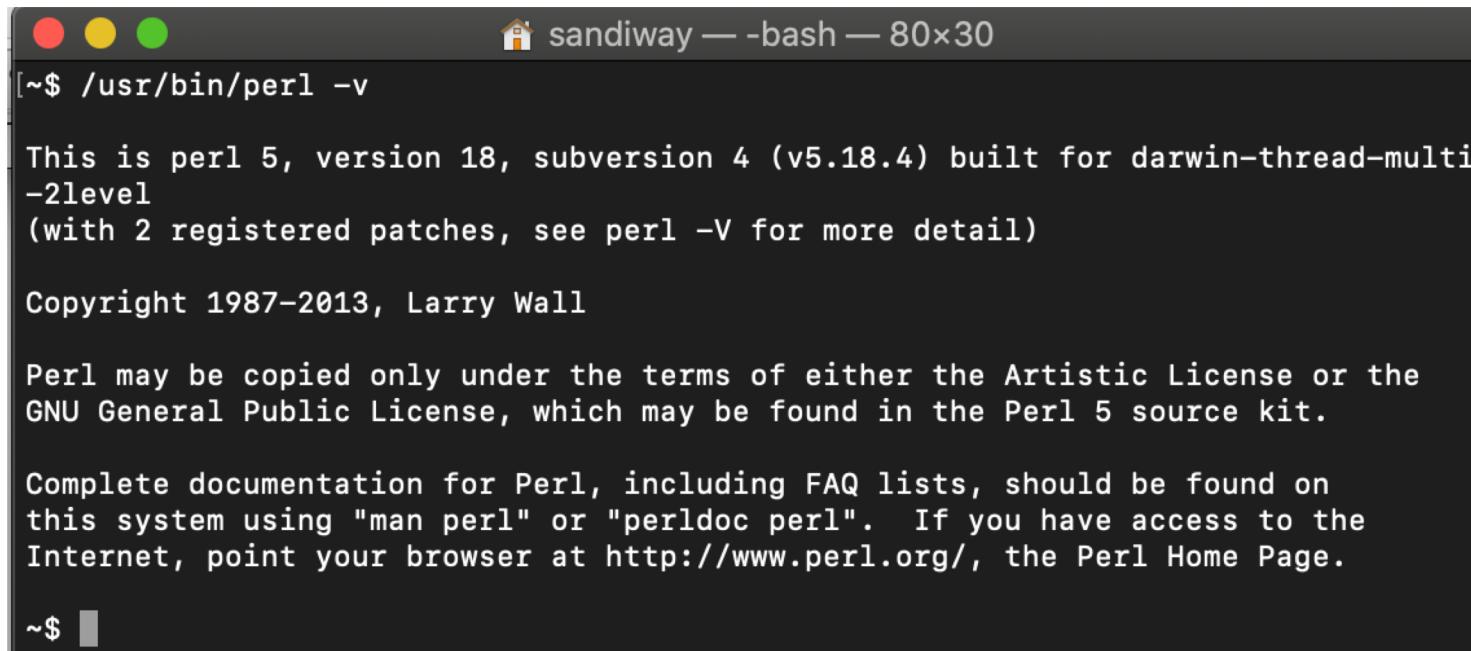
Complete documentation for Perl, including FAQ lists, should be found on
this system using "man perl" or "perldoc perl". If you have access to the
Internet, point your browser at http://www.perl.org/, the Perl Home Page.
```

```
[~$ which perl
/opt/local/bin/perl
~$ ]
```



# Homework: Install Perl

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



A screenshot of a Mac OS X terminal window titled "sandeway — -bash — 80x30". The window shows the command "/usr/bin/perl -v" being run and its output. The output includes details about Perl version 5.18.4, copyright information from 1987-2013, and documentation instructions.

```
[~$ /usr/bin/perl -v
This is perl 5, version 18, subversion 4 (v5.18.4) built for darwin-thread-multi-2level
(with 2 registered patches, see perl -V for more detail)

Copyright 1987-2013, Larry Wall

Perl may be copied only under the terms of either the Artistic License or the
GNU General Public License, which may be found in the Perl 5 source kit.

Complete documentation for Perl, including FAQ lists, should be found on
this system using "man perl" or "perldoc perl". If you have access to the
Internet, point your browser at http://www.perl.org/, the Perl Home Page.

~$ ]
```

# Homework: Install Perl

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

The screenshot shows the official Perl website at <https://www.perl.org/>. The top navigation bar includes links for ABOUT, DOWNLOAD (version 5.32.0), LEARN, DOCS, CPAN, and COMMUNITY. Below the navigation is a large banner featuring a camel and the text "That's why we love Perl" and "25,000 extensions on CPAN". A subtext below the banner states: "Perl is a highly capable, feature-rich programming language with over 30 years of development." A prominent green button labeled "DOWNLOAD AND GET STARTED" is visible. On the left side, a blue sidebar contains the text "start here next time" and a blue arrow pointing right, above a list titled "Get Started" which includes links to learn.perl.org, a brief introduction, free online Perl books, joining local communities, and books and more.

start here  
next time

## Get Started

- [learn.perl.org](#)
- [A brief introduction](#)
- [Free online Perl books](#)
- [Join your local community](#)
- [Books and More](#)

Perl 5.32.0

ABOUT DOWNLOAD LEARN DOCS CPAN COMMUNITY

That's why we love Perl

25,000 extensions on CPAN

Perl is a highly capable, feature-rich programming language with over 30 years of development.

DOWNLOAD AND GET STARTED

Learning Community Docs

# 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

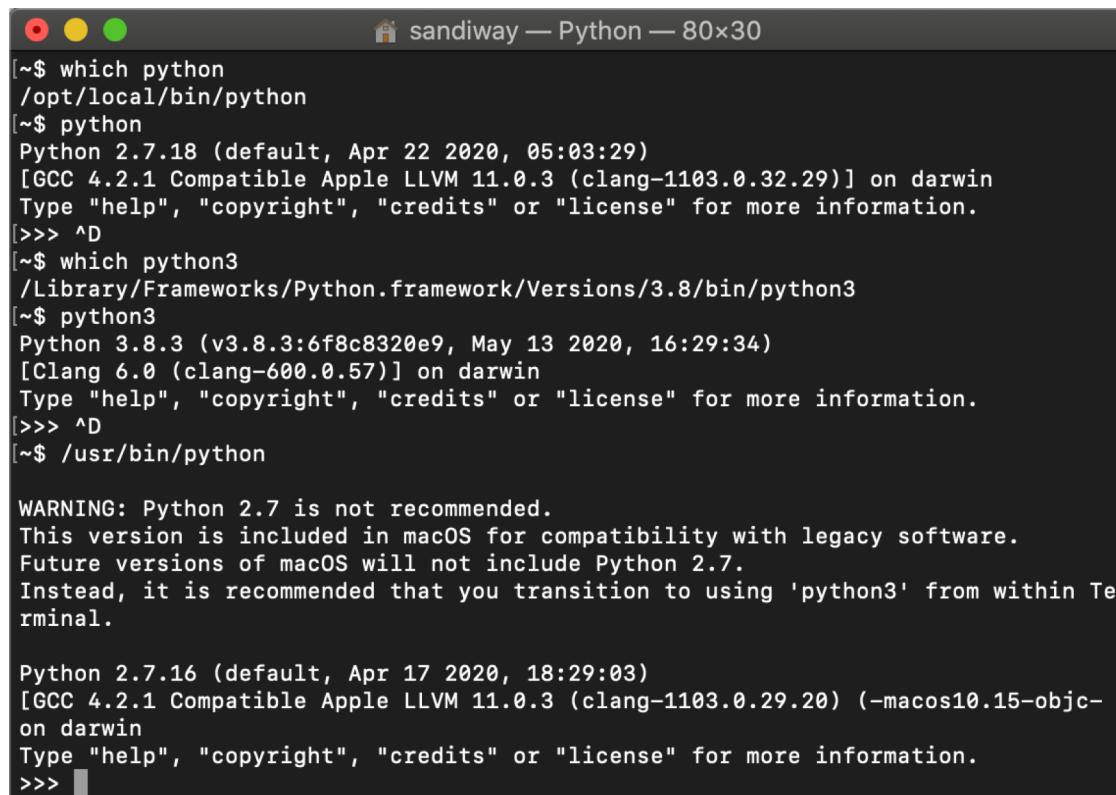
# Homework: Install Python

- [www.python.org](http://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 python
/opt/local/bin/python
[~$ python
Python 2.7.18 (default, Apr 22 2020, 05:03:29)
[GCC 4.2.1 Compatible Apple LLVM 11.0.3 (clang-1103.0.32.29)] on darwin
Type "help", "copyright", "credits" or "license" for more information.
[>>> ^D
[~$ which python3
/Library/Frameworks/Python.framework/Versions/3.8/bin/python3
[~$ python3
Python 3.8.3 (v3.8.3:6f8c8320e9, May 13 2020, 16:29:34)
[Clang 6.0 (clang-600.0.57)] on darwin
Type "help", "copyright", "credits" or "license" for more information.
[>>> ^D
[~$ /usr/bin/python

WARNING: Python 2.7 is not recommended.
This version is included in macOS for compatibility with legacy software.
Future versions of macOS will not include Python 2.7.
Instead, it is recommended that you transition to using 'python3' from within Te
rminal.

Python 2.7.16 (default, Apr 17 2020, 18:29:03)
[GCC 4.2.1 Compatible Apple LLVM 11.0.3 (clang-1103.0.29.20) (-macos10.15-objc-
on darwin
Type "help", "copyright", "credits" or "license" for more information.
>>> ]
```

# Which one is easier?

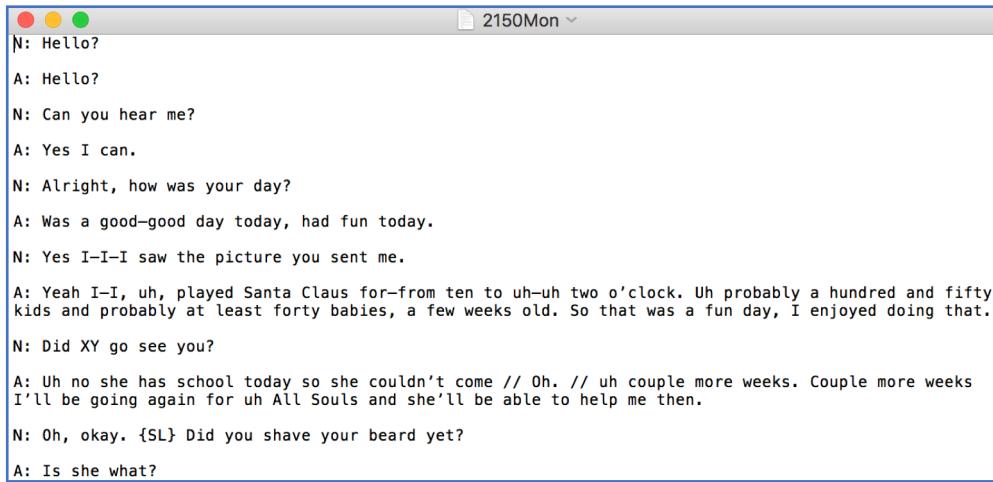
A *subjective question* ...

- 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`



```
N: Hello?
A: Hello?
N: Can you hear me?
A: Yes I can.
N: Alright, how was your day?
A: Was a good-good day today, had fun today.
N: Yes I-I-I saw the picture you sent me.
A: Yeah I-I, uh, played Santa Claus for-from ten to uh-uh two o'clock. Uh probably a hundred and fifty kids and probably at least forty babies, a few weeks old. So that was a fun day, I enjoyed doing that.
N: Did XY go see you?
A: Uh no she has school today so she couldn't come // Oh. // uh couple more weeks. Couple more weeks I'll be going again for uh All Souls and she'll be able to help me then.
N: Oh, okay. {SL} Did you shave your beard yet?
A: Is she what?
```

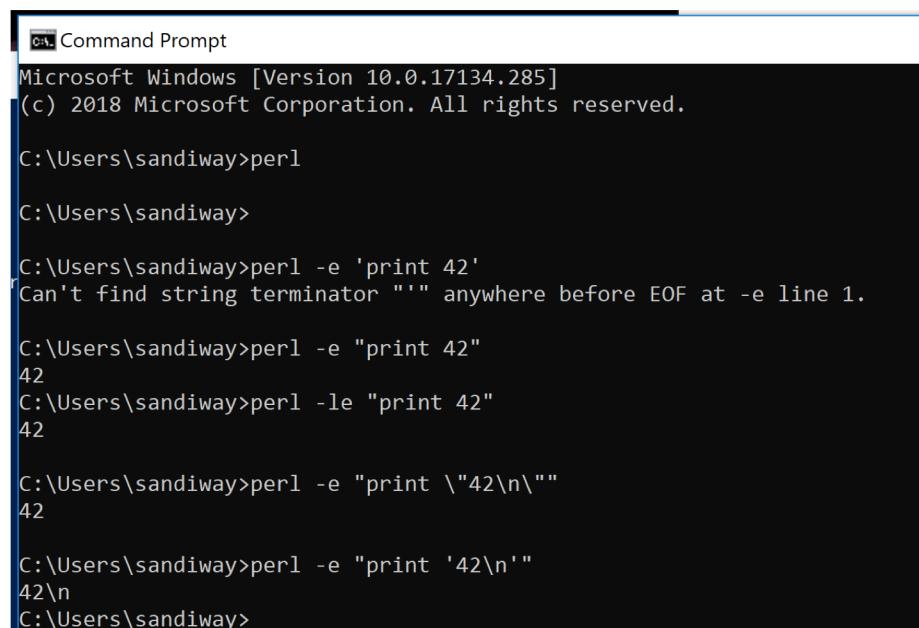
**Between the quotes we have a program!**

**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.



```
Command Prompt
Microsoft Windows [Version 10.0.17134.285]
(c) 2018 Microsoft Corporation. All rights reserved.

C:\Users\ sandiway> perl

C:\Users\ sandiway>

C:\Users\ sandiway> perl -e 'print 42'
Can't find string terminator "'" anywhere before EOF at -e line 1.

C:\Users\ sandiway> perl -e "print 42"
42
C:\Users\ sandiway> perl -le "print 42"
42

C:\Users\ sandiway> perl -e "print \"42\n\""
42
C:\Users\ sandiway> perl -e "print '42\n'"
42\n
C:\Users\ sandiway>
```

# 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.

## Windows PowerShell

```
PS C:\Users\sandiway> perl -e "print 42"
42
PS C:\Users\sandiway> perl -e 'print 42'
42
PS C:\Users\sandiway> perl -e 'print \'Hello, world\''
Hello, world
PS C:\Users\sandiway> perl -e 'print "Hello, world"'
No comma allowed after filehandle at -e line 1.
PS C:\Users\sandiway>
```