Who are you and why are you here?

- Name/nickname
- Dept., college and year
- Why are you taking this course?
- What topics would you like to see covered?

Administrivia

- [http://www.cs.pomona.edu/classes/cs159/](http://www.cs.pomona.edu/classes/cs159/)
- Office hours, schedule, assigned readings, assignments
- Everything will be posted there
- Read the “administrivia” handout!
- ~5 assignments (in a variety of languages)
- 4 quizzes (dates are tentative)
- In-class presentation
- Final project for the last month
- Teams of 2-3 people
- Research-like with write-up and presentation
- Class participation
- Readings
- Academic honesty and collaboration

Administrivia

- First assignment posted already
- Shouldn’t take too long
- Due Monday at the beginning of class
- CS colloquium tomorrow
  - Text simplification
  - 4:15pm Rose Hill Theatre
- CS accounts
What to expect...

- This course will be challenging for many of you
- Assignments will be non-trivial
- Content can be challenging
- But it is a fun field!
- We'll cover
  - Basic linguistics
  - Probability
  - The common problems
  - Many techniques and algorithms
  - Common learning techniques
  - Applications

Requirements and goals

- Requirements
  - Competent programmer
  - Mostly in Java, but I may allow/encourage other languages
  - Comfortable with mathematical thinking
  - We'll use a fair amount of probability, which I will review
  - Other basic concepts, like logs, summation, etc.
  - Data structures
    - Trees, hashtables, etc.
- Goals
  - Learn the problems and techniques of NLP
  - Build real NLP tools
  - Understand what the current research problems are in the field

What is NLP?

Natural language processing (NLP) is a field of computer science and linguistics concerned with the interactions between computers and human (natural) languages.

- Wikipedia

What is NLP?

The goal of this new field is to get computers to perform useful tasks involving human language...

- The book
**Key: Natural text**

“A growing number of businesses are making Facebook an indispensible part of hanging out their shingles. Small businesses are using …”

- Natural text is written by people, generally for people

**Why do we even care about natural text in computer science?**

**Why do we need computers for dealing with natural text?**

**Web is just the start…**

- e-mail
  - 247 million e-mails a day

- corporate databases
  - Blogs: 126 million different blogs

- Twitter
  - 27 million tweets a day

**Why is NLP hard?**

- Iraqi Head Seeks Arms
- Juvenile Court to Try Shooting Defendant
- Stolen Painting Found by Tree
- Kids Make Nutritious Snacks
- Local HS Dropouts Cut in Half
- Obesity Study Looks for Larger Test Group
- British Left Waffles on Falkland Islands
- Red Tape Holds Up New Bridges
- Hospitals Are Sued by 7 Foot Doctors
Why is NLP hard?

- **User**: Where is The Green Hornet playing in the Claremont Area?
- **System**: The Green Hornet is playing at the Ontario Mills theatre.
- **User**: When is it playing there?
- **System**: It’s playing at 2pm, 5pm and 8pm.
- **User**: I’d like 1 adult and 2 children for the first show. How much would that cost?

Natural language:
- is highly ambiguous at many different levels
- is complex and contains subtle use of context to convey meaning
- is probabilistic?
- involves reasoning about the world
- is highly social
- is a key part in how people interact

However, some NLP problems can be surprisingly easy.

Different levels of NLP

- **pragmatics/discourse**: how does the context affect the interpretation?
- **semantics**: what does it mean?
- **syntax**: phrases, how do words interact
- **words**: morphology, classes of words

NLP problems and applications

What are some places where you have seen NLP used?

What are NLP problems?
NLP problems and applications

- Lots of problems of varying difficulty
- Easier
  - Word segmentation: where are the words?
    - I would’ve like Dr. Kauchak to finish early. But he didn’t.
  - Speech segmentation
  - Sentence splitting (aka sentence breaking, sentence boundary disambiguation)
  - Language identification
    - Soy un profesor con queso.

Moderately difficult

- Morphological analysis/stemming
  - smarter
  - smart
  - smartest
  - smartly
  - speech recognition
  - text classification

I eat sushi with tuna

Easier continued

- Truecasing
  - i would’ve like dr. kauchak to finish early, but he didn’t.
- Spell checking
  - Identifying mispellings is challenging especially in the dessert.
  - OCR

Moderately difficult continued

- Text segmentation: break up the text by topics
- Part of speech tagging (and inducing word classes)
- Parsing

I eat sushi with tuna
As he walked along the side of the stream, he spotted some money by the bank. The money had gotten muddy from being so close to the water.

We are good at grammar.

IBM hired Fred Smith as president.

A company that acts as a middle man between content companies and Internet service providers is acquiring Comcast Corp., the nation’s largest broadband provider, of anti-competitive behavior.

IBM hired Fred Smith as president.
NLP problems and applications

- Natural language understanding
  - Text $\Rightarrow$ semantic representation (e.g. logic, probabilistic relationships)
- Information retrieval and question answering
  - "How many programmers in the child care department make over $50,000?"
  - "Who was the fourteenth president?"
  - "How did he die?"

Where are we now?

- Many of the "easy" and "medium" problems have reasonable solutions
  - spell checkers
  - sentence splitters
  - word segmenters/tokenizers

Text simplification

Alfonso Perez Munoz, usually referred to as Alfonso, is a former Spanish footballer, in the striker position.

Alfonso Perez is a former Spanish football player.

Parsing

Stanford Parser (http://nlp.stanford.edu:8080/parser/)

* Example input:*
  
  "My name is John. I am a student at Stanford University."

* Example output:*
  
  "My name is John. I am a student at Stanford University."
Where are we now?

- Machine translation
  - Getting better every year
  - Enough to get the gist of most content, but still no where near a human translation
  - Better for some types of text

- translate.google.com
- Many commercial versions...
  - Systran
  - Language Weaver

Where are we now?

- Information retrieval/query answering
  - Search engines:
    - Pretty good for some things
    - Does mostly pattern matching and ranking
    - No deep understanding
    - Still requires user to “find” the answer

Where are we now?

- Question answering
  - Wolfram Alpha
Where are we now?

- Question answering
  - Many others...
  - TREC question answering competition
  - Language computer corp
  - answerbus

Where are we now?

- Information extraction
  - Structured documents (very good!)
    - www.dealtime.com
    - www.froogle.com
  - AIX technologies
    - Lots of these
    - FlipDog
    - WhizBang! Labs
    - ...
    - work fairly well

Where are we now?

- Summarization
  - NewsBlaster (Columbia)
  - http://newblaster.cs.columbia.edu/

Where are we now?

- Voice recognition
  - pretty good, particularly with speaker training
    - Apple OS has one built in:
      - “What time is it?”
      - “Switch to finder”
      - “Hide this application”
    - IBM ViaVoice
    - Dragon Naturally Speaking

- Speech generation
  - The systems can generate the words, but getting the subtle nuances right is still tricky
  - Apple OS – select text: CMD+SHIFT+T
  - translate.google.com

Where are we now?

- Speech recognition
  - A company that uses a mobile phone between sensor computer and Internet
    - voice services: a new way to interact with computers
    - Google has a version called Google Voice
    - Amazon has a version called Alexa
    - IBM has a version called ViaVoice
  - IBM ViaVoice
  - Apple OS has one built in:
    - “What time is it?”
    - “Switch to finder”
    - “Hide this application”
  - IBM ViaVoice
  - Dragon Naturally Speaking

- Speech generation
  - The systems can generate the words, but getting the subtle nuances right is still tricky
  - Apple OS – select text: CMD+SHIFT+T
  - translate.google.com