Adding the course?

http://www.cs.pomona.edu/classes/cs159/
- Office hours, schedule, assigned readings, assignments
- Everything will be posted there

Read the “administrivia” handout!
- ~7 assignments (in a variety of languages)
- 4 quizzes
- final project for the last 3-4 weeks
- teams of 2-3 people
- class participation
- Readings

Academic Honesty Collaboration
**Administrivia**

Assignment 0 posted already
- Shouldn’t take too long
- Due Friday by 5pm

Assignment 1 posted soon
- Won’t cover all material until next Monday
- Due Monday 2/4

**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 machine learning techniques
- some recent advances in neural networks for language processing
- NLP applications

**Videos before class**

**Requirements and goals**

Requirements
- Competent programmer
- Some assignments in Java, but I will allow/encourage other languages after the first few assignments
- 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 indispensable part of hanging out their shingles. Small businesses are using …”

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

Google’s search knows about over 130 trillion pages

In less than four years, Google's search knowledge of pages have grown by more than 130 trillion pages.

https://searchengineland.com/google-search-knowledge-base-130-trillion-page-documents-303720
Web is just the start...

- 1/24/19
- 4

- e-mail
- ~200-300 million e-mails a day
- corporate databases
- ~500 million tweets a day
- Blogs: ~200 million different blogs

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 Escape Room playing in the Claremont Area?
System: Escape Room is playing at the Edwards in La Verne.
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?

Why is NLP hard?

**User:** Where is Escape Room playing in the Claremont Area?
**System:** Escape Room is playing at the Edwards in La Verne.
**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?

**Why is NLP hard?**

**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?

Lots of problems of varying difficulty

- Easier
  - Word segmentation: where are the words?
    - I would’ve liked Dr. Dave to finish early.
    - But he didn’t.
NLP problems and applications

Lots of problems of varying difficulty

Easier
- Speech segmentation
- Sentence splitting (aka sentence breaking, sentence boundary disambiguation)
  I would've liked Dr. Dave to finish early. But he didn't.
- Language identification
  Soy un maestro con queso.

Easier continued
- truecasing
  I would've liked Dr. Dave to finish early. But he didn't.
- spell checking
  Identifying mispellings is challenging especially in the dessert.
- OCR

Moderately difficult
- morphological analysis/stemming
- speech recognition
  smarter smarter smart
- text classification

Moderately difficult continued
- text segmentation: break up text by topics
- part of speech tagging (and inducing word classes)
- parsing

I eat sushi with tuna

I would've liked Dr. Dave to finish early. But he didn't.
NLP problems and applications

Moderately difficult continued

- word sense disambiguation

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.

- grammar correction

We are good at grammar.

- speech synthesis

Information extraction

IBM hired Fred Smith as president.

<table>
<thead>
<tr>
<th>person</th>
<th>company</th>
<th>position</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fred Smith</td>
<td>IBM</td>
<td>president</td>
</tr>
</tbody>
</table>

Machine translation

The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport.

Information extraction

IBM hired Fred Smith as president.

Summarization

A company that acts as a middle-man between content companies and Internet service providers is accusing Comcast Corp., the nation's largest broadband provider, of anti-competitive behavior. (article 1) Comcast Corp. and NBC Universal made new promises to the Federal Communications Commission that the companies hope will help get the regulatory agency to approve the proposed deal between the media giants. (article 5) At issue is the cable operator's decision to offer the Tennis Channel on a specialty tier of sports networks as opposed to its widely distributed basic tier. (article 3) The quality of television news could deteriorate further under a Comcast-controlled NBC Universal, the Writers Guild of America East warned Wednesday in letters to key Washington officials opposing the government's review of the proposed merger. (article 5) With regulatory approval still weeks if not months away, Comcast and NBC Universal have extended the term of their merger agreement to March of next year. (article 3) Democrat Michael Coffin from the joint venture would put too much control of content into the hands of a company that also controls how consumers access the Internet and television. (article 3) Susan Fox talked on Wednesday with two senior staff members of FCC Commissioner Meredith Attwell Baker (article 1).
NLP problems and applications

Natural language understanding
- Text => semantic representation (e.g. logic, probabilistic relationships)

Information retrieval and question answering

“How many programmers in the child care department make over $50,000?”

“How was the fourteenth president?”

“How did he die?”

NLP problems and applications

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.

Where are we now?

Many of the “easy” and “medium” problems have reasonable solutions

- spell checkers
- natural language
- sentence splitters
- word segmenters/tokenizers

Where are we now?

Parsing

Stanford Parser (http://nlp.stanford.edu:8080/parser)
Where are we now?

Machine translation

How is it?

- 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

http://translate.google.com

Many commercial versions...

- Systran
- Language Weaver

Where are we now?

Information extraction

- Structured documents (very good!)
  - www.dealtime.com
  - www.google.com/shopping
- AKT technologies
  - Lots of these
  - FlipDog
  - WhizBang! Labs
  - ...
  - Work fairly well

Where are we now?

CMU’s NELL (Never Ending Language Learner)

http://rtw.ml.cmu.edu/rtw/
Where are we now?

Why do people do this?

dianchuk@cs.posnau.edu

Where are we now?

Information retrieval/query answering

How are search engines?
What are/aren’t the good at?
How do they work?

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

Question answering
• wolfram alpha
Where are we now?

Question answering: Wolfram Alpha

Many others systems
- TREC question answering competition
- Language computer corp
- Answerbus

Question answering

The Science Behind IBM’s Watson

IBM Watson

Where are we now?

Question answering

Summary

NewsBlaster
(Columbia)

http://newsblaster.cs.columbia.edu/

Ukraine crisis: Russian forces reportedly seen in rebel cities Donetsk and Luhansk

Summary from multiple countries, from articles in English

Pro-Russian rebels refused their demand for full independence Monday, saying they would accept Ukraine’s sovereignty in exchange for autonomy and a deal that would allow Moscow’s armed forces to take a deal to a new round of peace talks. 

The Crimean peninsula, which was annexed by Russia in 2014, is a key to the ongoing conflict. The Ukrainian government has said it will not negotiate with Moscow until its demands are met. 

Russian forces have been building up near the Ukrainian border, and tension has increased in recent weeks. 

The United Nations has called for an immediate withdrawal of all foreign military forces from Ukraine, and the European Union has imposed sanctions against Russia.

http://www.tumblr.com

http://www.politicalscenarios.com

http://www.thenationalinterest.org

http://www.bbc.com

http://www.cnn.com

http://www.cbsnews.com

http://www.nbcnews.com

http://www.foxnews.com

http://www.huffingtonpost.com

http://www.washingtonpost.com

http://www.reuters.com

http://www.aljazeera.com

http://www.nytimes.com

http://www.nytimes.com

http://www.npr.org

http://www.cbsnews.com

http://www.huffingtonpost.com

http://www.aljazeera.com

http://www.nytimes.com

http://www.npr.org
**Where are we now?**

Voice recognition
- pretty good, particularly with speaker training
  - Apple OS/Siri
  - Android/Google
  - Alexa, Google Assistant, etc...
  - IBM ViaVoice
  - Dragon Naturally Speaking

Speech generation
- The systems can generate the words, but getting the subtle nuances right is still tricky
  - Apple OS
  - http://translate.google.com

**Other problems**

Many problems untackled/undiscovered

“That’s What She Said: Double Entendre Identification”
- ACL 2011