A New CAPTCHA Approach

To complete your web registration, please prove that you’re human:

WHEN LITTLEBOY’S MOTHER DIED IN THE ORIGINAL "LAND BEFORE TIME," DID YOU FEEL SAD?

☐ YES
☐ NO

(BOTS: NO LIVING)

Office hours, schedule, assigned readings, assignments
Everything will be posted there

Read the administrative section of the webpage
~7 assignments (in a variety of languages)
4 quizzes
final project for the last 3-4 weeks
teams of 2-3 people
Readings

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

Assignment 1 posted
Won’t cover all material until next Monday
Due Monday 1/30 at 1pm

Academic Honesty Collaboration
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

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

Getting to know each other

What’s your name?
What year?
What has been your favorite CS class?
What’s been your least favorite CS class?
Why are you taking this class?

Videos before class
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

Where have you seen NLP used?

“A growing number of businesses are making Facebook an indispensable 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?

Google’s search knows about over 130 trillion pages
In the last four years, Google’s search knowledge of pages has grown by more than 100 trillion new pages.

[Image]

https://searchengineland.com/googles-search-indexes-have-130-trillion-pages-documents-060308

Web is just the start...

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

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?

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.
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.

Identifying mispellings is challenging especially in the dessert.

Moderately difficult

• morphological analysis/stemming

• speech recognition

• text classification

SPAM

sentiment analysis
NLP problems and applications

Moderately difficult continued

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

I eat sushi with tuna

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 am good at grammar.
- speech synthesis

Hard (many of these contain many smaller problems)

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 herself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport.

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>
NLP problems and applications

Summarization

A company that acts as a middle man between content companies and Internet service providers is shutting Comcast Corp., the nation's largest broadband provider, of anti-competitive behavior (article 5) Comcast Corp. and N: 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 worsen further under a Comcast-controlled NBC Universal, the Writers Guild of America East warned Wednesday in letters to key Washington officials echoing 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 4) Despite flinch - 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 5) Susan Fox talked on Wednesday with two senior staff members of FCC Commissioner Meredith Artwell Baker (article 1)

Natural language understanding

- Text to 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?"

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
- sentence splitters
- word segmenters/tokenizers

Where are we now?

Many of the “easy” and “medium” problems have reasonable solutions
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 jist of most content, but still no where near a human translation

better for some types of text

http://translate.google.com

Many commercial versions...

syrtran

language weaver

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?
: David.last_name@pomona.edu

Where are we now?
Information retrieval/query answering

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

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

What is the fifteenth president of the United States?

- United States
- President: 15

Name: James Buchanan

Many others systems
- TREC question answering competition
- language computer corp
- answerbus
- ...
Where are we now?

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

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

Ukraine crisis: Russian forces reportedly seen in rebel cities Donetsk and Luhansk
Summary from multiple countries, from articles in English

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