SI485i
Natural Language Processing

Set 1
Intro to NLP

Fall 2013 : Chambers
Assumptions about You

• You know…
  • how to program Java
  • basic UNIX usage
  • basic probability and statistics (we’ll also review)

• You will learn…
  • computational approaches to manipulating and understanding language
  • basic learning algorithms
  • how to build practical systems
Early NLP

• **Dave**: Open the pod bay doors, HAL.
• **HAL**: I’m sorry Dave. I’m afraid I can’t do that.
Commercial NLP

There are many approaches, out
State of the Art NLP

- **Speech recognition**: audio in, text out
  - SOTA: 0.3% error for digit strings, 5% dictation, 50% TV

- **Text-to-speech**: text in, audio out
  - SOTA: Very intelligible, but often bad prosody

- **Information extraction**: text in, DB record out
  - SOTA: 40–90% field accuracy, all depending on details

- **Parsing**: text in, sentence structure out
  - SOTA: Over 90% dependency accuracy for formal text

- **Question answering**: text in, question answer out
  - SOTA: 70%+ for factoid questions, otherwise challenging

- **Machine translation**: language A to language B
  - SOTA: Now often usable for gisting purposes; not great
So what is NLP?

- Go beneath the surface of words
  - Don’t just manipulate word strings
  - Don’t just keyword match on search engines

- **Goal**: recover *some* aspect of the *structure* in language (groups of words move together)
- **Goal**: recover *some* of the *meaning* in language (words map to real-world things)
Any creative work is a collaboration between the capital—a Author and Reader. All are dead and lifeless until they're experienced and interpreted by someone!

That someone is me, baby!

And if I'm an adult and I re-read a book I loved as a child, I'll get different things out of it, just as I'll get different things out of it when I read it again as an old man. Even if I'm the author of the text, my reactions to it will be different! I know this because I remember my high school essays were AMAZING, but when I read them now, SOMEHOW, they're pages and pages of ultimate embarrassment.

Shakespeare needs tons of notes to be readable today!

Exactly!

So we agree that even with an identical text, the meaning of a work is constructed by the Reader, and that meaning changes with time as the readers themselves change. And this applies to books, movies, games, paintings: anything creative!

Sure. We are agreed! So what's your point?

My point? Utahraptor, if we wait till we're old to experience all the creative works we want to, we're doomed. We simply won't get the same enjoyment than what we'd get now! Put simply, Utahraptor, my point is this: wanna take the day off and play video games?
NLP is hard. (news headlines)

1. Minister Accused Of Having 8 Wives In Jail
2. Juvenile Court to Try Shooting Defendant
3. Teacher Strikes Idle Kids
4. Miners refuse to work after death
5. Local High School Dropouts Cut in Half
6. Red Tape Holds Up New Bridges
7. Clinton Wins on Budget, but More Lies Ahead
8. Hospitals Are Sued by 7 Foot Doctors
9. Police: Crack Found in Man's Buttocks
NLP needs to adapt.

The Bamboozle @TheBamboozle
#ballin “@TheGunzShow: @thebamboozle being able to show tens of thousands of ppl that i'm a baller. oh and TREOS!!”
Expand

Tracy Trinita @TracyTrinita
RT @flo_ren: Tommorow, 6pm, at GPPS Hosana will be held the Revival Night with @TracyTrinita for more info u can ask me lockerz.com/s/233083091
View photo

Sketch - #JO_LADY @ItzSketch
3) #Friday House Party [ Today ] ... Housee Party At lekki .. Gate Fee: Free For the first 3 people to download #Jo_Lady tonight.
NLP needs to adapt.

http://xkcd.com/1083/
NLP is also a Knowledge Problem

Where is *A Bug’s Life* playing in *Mountain View*?

A Bug’s Life is playing at the Century 16 Theater.

When is it playing there?

It’s playing at 2pm, 5pm, and 8pm.

OK. I’d like 1 adult and 2 children for the first show. How much would that cost?

But we need domain knowledge, discourse knowledge, world knowledge, linguistic knowledge.
What will we do?

- Language Modeling
  - Build probabilities of words and phrases
- Document Classification
  - Identify some hidden property of documents
- Sentiment Analysis
  - Learn to extract the emotion and mood from language
- Parsing
  - Identify the syntax of language
- Information Extraction
  - Automatically pull out valuable nuggets of information