CS 124/LINGUIST 180
From Languages to Information
Dan Jurafsky
Stanford University
Introduction and Course Overview
From languages to information

• For humans, going from the largely unstructured languages of the web to information is effortlessly easy
• But it’s hard for computers!
• To build the next generation of intelligent agents to make decisions on your behalf
  • Answering your routine email
  • Booking your next trip to Fiji
• they need to be able to go *from languages to information*
What this course is about

- Automatically extracting meaning and structure from:
  - Human language text
  - Speech
  - Web pages
  - Social networks (and other networks)
  - Genome sequences
- Interacting with humans on-line via language
  - Question Answering
  - Personal Assistants
  - Chatbots
Commercial World
Extracting information
Hi Dan, we’ve now scheduled the curriculum meeting. It will be in Gates 159 tomorrow from 10:00-11:30.

-Chris
Sentiment Analysis

Emotional Spell-Check
Blog Analytics

- Data-mining of blogs, discussion forums, message boards, user groups, and other forms of user generated media
  - Product marketing information
  - Political opinion tracking
  - Social network analysis
  - Buzz analysis (what’s hot, what topics are people talking about right now).
Livejournal.com: 

*count of I, me, my* on or after Sep 11, 2001


Graph from Pennebaker slides
September 11 LiveJournal.com study: count of “we”, “us”, “our”

Extracting Social Meaning from Speech

- **Uncertainty** (students in tutoring)
- **Annoyance**
  - callers to dialog systems:
- **Deception**
- **Emotion**
- **Intoxication**
- **Flirtation, Romantic interest**
  - McFarland, Jurafsky, Ranganath
What do flirters do?

- **Women when flirting:**
  - use negation (*don’t, no, not*)
  - raise pitch ceiling
  - laugh at themselves
  - say “I”
  - use “like”

- **Men when flirting:**
  - raise their pitch floor
  - laugh at their date (teasing?)
  - say “you” and “you know”
  - don’t use words related to academics
Unlikely words for male flirting

– academia
– interview
– teacher
– phd
– advisor
– lab
– research
– management
– finish
How well did I click with this person? (1-10)

• both parties talk about the woman
  – women use *I*,
  – men use *you*
• man supports woman’s face
  – men use *appreciations* and *sympathy*,
  – men *accommodate* women’s laughter
  – men interrupt with *collaborative completions*
• woman is engaged
  – women raise their pitch, vary loudness and pitch
  – women avoid hedges

Women are the empowered party in speed dating!
Speed date lessons

• Don’t talk about your advisor
• Focus on the empowered party
• Watch your pitch:
  • flirting women raise pitch ceiling
  • flirting men raise pitch floor
Analyzing Restaurant Reviews


900,000 Yelp reviews online

A very bad (one-star) review:

The bartender... absolutely horrible... we waited 10 min before we even got her attention... and then we had to wait 45 - FORTY FIVE! - minutes for our entrees... stalk the waitress to get the cheque... she didn't make eye contact or even break her stride to wait for a response ...
What is the language of bad reviews?

- Negative sentiment language
  - horrible aweful terrible bad disgusting
- Past narratives about people
  - waited, didn’t, was
  - he, she, his, her,
  - manager, customer, waitress, waiter
- Frequent mentions of we and us
  - ... we were ignored until we flagged down one waiter to go get our waitress ...
Other narratives with this language

A genre using:
- Past tense, we/us, negative, people narratives

Texts written by **people suffering trauma**
- People who write after tragedies
- Past tense: distancing from events
- Use of “we”: seeking solace in community

**1-star reviews are trauma narratives!**
- The lesson of reviews:
  - It’s all about personal interaction
What about positive reviews?

Sex, Drugs, and Dessert

- *addicted* to pepper shooters
- garlic noodles... my *drug of choice*
- the fries are *like crack*

- *orgasmic pastry*
- *sexy food*
- *seductively seared fois gras*
Computational Biology: Comparing Sequences

Sequence comparison is key to
- Finding genes
- Determining function
- Uncovering the evolutionary processes
High school dating

What is the structure of social relations?
Imagine a graph of high school
• people are nodes
• links are romantic relationships
What will the shape of this graph be?
A densely connected graph?
A line?
A cycle?

Image drawn by Mark Newman
The Structure of Romantic and Sexual Relations at "Jefferson High School"

Each circle represents a student and lines connecting students represent romantic relations occurring within the 6 months preceding the interview. Numbers under the figure count the number of times that pattern was observed (i.e., we found 63 pairs unconnected to anyone else).
Interacting with humans via language
Question Answering: IBM’s Watson
Recommendation Engines

Customers Who Bought This Item Also Bought

- A Curious History of Food and Drink
  - Ian Crofton
  - Hardcover
  - $15.06 Prime

- Consider the Fork: A History of How We Cook and Eat
  - Bee Wilson
  - Paperback
  - $11.28 Prime

- Fifty Foods That Changed the Course of History (Fifty Things That Changed the...)
  - Bill Price
  - Hardcover
  - $23.10 Prime

More tracks like this.com

Side A. Insert track or Spotify link
- Track Name: Shake it Off
- Artist Name: Taylor Swift
- Spotify link: [Link]

Side B. Track list
- Out of the Woods
  - Taylor Swift
  - [Link]
Personal Assistants
Zuckerberg’s personal assistant challenge last year

Every year, I take on a personal challenge to learn new things and grow outside my work at Facebook. My challenges in recent years have been to read two books every month, learn Mandarin and meet a new person every day.

My personal challenge for 2016 is to build a simple AI to run my home and help me with my work. You can think of it kind of like Jarvis in Iron Man.

I'm going to start by exploring what technology is already out there. Then I'll start teaching it to understand my voice to control everything in our home -- music, lights, temperature and so on. I'll teach it to let friends in by looking at
Characteristics of the task
Ambiguity

- Resolving ambiguity is a crucial goal throughout string and language processing or social network interpretation
Ambiguity

Find at least 6 meanings of this sentence:

\[ I \text{ made her duck} \]
Ambiguity

Find at least 6 meanings of this sentence:

_**I made her duck**_

- I cooked waterfowl for her benefit (to eat)
- I cooked waterfowl belonging to her
- I created the (plaster?) waterfowl she owns
- I caused her to quickly lower her head or body
- I recognized the true identity of her spy waterfowl
- I waved my magic wand and turned her into undifferentiated waterfowl
Ambiguity is Pervasive

I caused her to quickly lower her head or body
   **Part of speech:** “duck” can be a Noun or Verb
I cooked waterfowl belonging to her.
   **Part of speech:**
   “her” is possive pronoun (“of her”)
   “her” is dative pronoun (“for her”)
I made the (plaster) duck statue she owns
   **Word Meaning**: “make” can mean “create” or “cook”
Ambiguity is Pervasive

**Grammar:** make can be:

**Transitive:** (verb has a noun direct object)
- I cooked [waterfowl belonging to her]

**Ditransitive:** (verb has 2 noun objects)
- I made [her] (into) [undifferentiated waterfowl]

**Action-transitive** (verb has a direct object + verb)
- I caused [her] [to move her body]
Ambiguity is Pervasive: Phonetics!!!!!

- I mate or duck
- I’m eight or duck
- Eye maid; her duck
- Aye mate, her duck
- I maid her duck
- I’m aid her duck
- I mate her duck
- I’m ate her duck
- I’m ate or duck
- I mate or duck
Why else is natural language understanding difficult?

<table>
<thead>
<tr>
<th>non-standard English</th>
<th>segmentation issues</th>
<th>idioms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Great job @justinbieber! Were SOO PROUD of what you've accomplished! U taught us 2 #neversaynever &amp; you yourself should never give up either♥</td>
<td>the New York-New Haven Railroad the New York-New Haven Railroad</td>
<td>dark horse get cold feet lose face throw in the towel</td>
</tr>
<tr>
<td>neologisms</td>
<td>world knowledge</td>
<td>tricky entity names</td>
</tr>
<tr>
<td>unfriend retweet bromance</td>
<td>Mary and Sue are sisters. Mary and Sue are mothers.</td>
<td>Where is A Bug’s Life playing ... Let It Be was recorded ... ... a mutation on the <em>for</em> gene ...</td>
</tr>
</tbody>
</table>

But that's what makes it fun!
Making progress on this problem...

• The task is difficult! What tools do we need?
  • Knowledge about language
  • Knowledge about the world
  • A way to combine knowledge sources

• How we generally do this:
  • probabilistic models built from language data
    • $P(\text{"maison"} \rightarrow \text{"house"})$ high
    • $P(\text{"L’avocat général"} \rightarrow \text{"the general avocado"})$ low
  • Luckily, rough text features can often do half the job.
Models

- Finite state machines
- Markov models
- Alignment
- Vector space models of word and document meaning
  - "word embeddings"
- Network models
- Take cs224N (this quarter!) for
  - Neural networks
Machine Learning

Machine learning based classifiers that are trained to make decisions based on features extracted from the context

Simple Classifiers:
   Naïve Bayes
   Logistic Regression (MaxEnt)
   Decision Trees
   Neural Networks

Sequence Models:
   Hidden Markov Models
   Maximum Entropy Markov Models
   Conditional Random Fields
   Recursive Neural Networks (RNNs, LSTMs)
Course logistics in brief

- **Instructor:** Dan Jurafsky
- **TAs:** Ziang Xie (head TA)
  - Monica Agrawal
  - William Chen
  - Berk Coker
  - Catherine Dong
  - Gaspar Garcia
  - Raghav Gupta
  - Brad Huang
  - Matt Lamm
  - Rafael Musa
  - Kevin Wu
- **Time:** TuTh 3:00-4:20, Bishop Auditorium
- **cs124.stanford.edu**
The flipped classroom

- http://www.knewton.com/flipped-classroom/
Why the flipped classroom

- Attention span: everyone spaces out during long lectures

- “the class started 1:00. The student sitting in front of me took copious notes until 1:20. Then he just nodded off... motionless, with eyes shut for about a minute and a half, pen still poised. Then he awoke and continued his rapid note-taking as if he hadn’t missed a beat.”

- Student remembered only the first 15-20 minutes
Why the flipped classroom (2)

- **Mastery learning**: Learn until you master
  - read the textbook and watch the video until you understand
  - metaphor: learning to ride a bike
- **Active learning**: Be in charge of your learning
  - Obviously most important: programming assignments
  - Active learning (“constructivism”), learning by doing
- **Collaborative learning**: Learn from each other
  - Use class time for group activities, worked problems
  - “Small group active learning”
cs124: Semi-flipped classroom

• **Lectures on video:**
  • You’re expected to do a first pass through the core material on your own through online videos and textbook chapters
  • On average about 90 minutes of video content each week
  • Some people watch it speeded up

• **Some lectures live:**
  • 7+1 (review) lectures are required (on final exam, no videos)
  • I will also re-lecture (double-cover) a few of the videos
    • some people like the engagement of in-class lectures

• **In-class group sessions (“active learning”)**
Logistics More Specifically

- Online Video Lectures with embedded quizzes (before class)
- Weekly online Review Quizzes (Tue of following week)
- Roughly weekly Python homeworks (Fri of following week)
- Final Exam (Tuesday March 21 3:30-6:30)
- Class sessions: All encouraged; **8 live lectures required**
  - Full lectures
  - Mini-lectures
  - Group worked problems (required for full participation credit)
Why embedded quizzes: “summative” vs “formative” assessment

Summative assessment
• Final exams: goal is grading

Formative assessment
• Along the way: goal is for you to find out what you don’t know
The Open Platform: EdX!

- https://lagunita.stanford.edu/about
- https://open.edx.org/about-open-edx
Coming up next class (Thursday)

Unix for poets

No? Would you believe ... Bell Labs patent department typists?

Watch the first half of this week’s videos (“Basic Text Processing”) before class!
Syllabus

- http://web.stanford.edu/class/cs124