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

Automatically extracting meaning and structure from:
- Human language text and speech (news, social media, etc.)
- Social networks
- Genome sequences

Interacting with humans via language
- Dialog systems/Chatbots
- Question Answering
- Recommendation Systems
Commercial World

Logos of Apple, Amazon, Google, Microsoft, Facebook, Pearson Knowledge Technologies, Twitter, and SoundHound.
1. Extracting information from language
Information Retrieval

6,586,013,574 web searches every day (by one estimate)

Text-based information retrieval is thus likely the most frequently used piece of software in the world

How does it work? Can you build an IR engine?

*Programming Assignment 4: Search!*
Extracting Sentiment and Social Meaning

Lots of meaning is in **connotation**

"connotation: an idea or feeling that a word invokes in addition to its literal or primary meaning."

Extracting connotation is generally called **sentiment analysis**
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 you do for fun? Dance?
Uh, dance, uh, I like to go, like camping. Uh, snowboarding, but I'm not good, but I like to go anyway.

You like boarding.
Yeah. I like to do anything. Like I, I'm up for anything.
Really?
Yeah.

Are you open-minded about most everything?
Not everything, but a lot of stuff-
What is not everything [laugh]
I don't know. Think of something, and I'll say if I do it or not. [laugh]
Okay. [unintelligible].

Skydiving. I wouldn't do skydiving I don't think.
Yeah I'm afraid of heights.

F: Yeah, yeah, me too.
M: [laugh] Are you afraid of heights?
F: [laugh] Yeah [laugh]
What do flirters do?

**Women when flirting:**
- raise pitch ceiling
- laugh at themselves
- say “I”

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

Rajesh Ranganath, Dan Jurafsky, and Daniel A. McFarland. 2013. Detecting friendly, flirtatious, awkward, and assertive speech in speed-dates. Computer Speech and Language. 27:1, 89-115
Unlikely words for male flirting

academia
interview
teacher
phd
advisor
lab
research
management
finish
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 awful 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 a waiter to get our waitress ...
Other narratives with this language

A genre using:

Past tense, we/us, negative, people narratives

Texts written by people suffering trauma
  - James Pennebaker lab
  - Past tense as distancing
  - 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 evolutionary processes

This is also how spell checkers work!

We'll learn: edit distance algorithms (Quiz 1)
Social Networks

The network formed by your friends or other relations offline or online

◦ Can we compute properties of these networks?
◦ Extract information from them?
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
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).
2. 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
- ★★★★★ 11
- $15.06 Prime

Consider the Fork: A History of How We Cook and Eat
- Bee Wilson
- Paperback
- ★★★★★ 217
- $11.28 Prime

Fifty Foods That Changed the Course of History (Fifty Things That Changed the...)
- Bill Price
- Hardcover
- ★★★★★ 2
- $23.10 Prime

More tracks like this.com

Get more Spotify music recommendations, based on your favourite tracks.

Results are drawn from the listening habits of 40 million active last.fm subscribers.

Side A. Insert track or Spotify link

<table>
<thead>
<tr>
<th>Track Name</th>
<th>Spotify link</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shake it Off</td>
<td>Taylor Swift</td>
</tr>
</tbody>
</table>

Side B. Track list

Out of the Woods by Taylor Swift

Sorry, Spotify doesn't have that track

Drag me to
Personal Assistants
Why is language interpretation hard?
Ambiguity

Resolving ambiguity is hard
Ambiguity

Find at least 6 meanings of this sentence:

I 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 possessive 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

Grammer: 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
More difficulties: Non-standard language

Great job @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either❤

And neologisms:

- unfriend
- retweet
- bromance
Making progress on this problem...

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

How we generally do this:
- Probabilistic models built from language data
  \[ P(\text{“maison”} \rightarrow \text{“house”}) \quad \text{high} \]
  \[ P(\text{“L’avocat général”} \rightarrow \text{“the general avocado”}) \quad \text{low} \]
Models and Tools

Regular Expressions

Edit distance and alignment

Word embeddings
  ◦ vector/neural models of meaning

Language models (word prediction)

Machine Learning classifiers
  ◦ Naïve Bayes
  ◦ Logistic Regression

Network algorithms
  ◦ PageRank

Recommendation algorithms
  ◦ Collaborative filtering

Linguistic tools
  ◦ Sentiment lexicons
Course logistics in brief

Instructor: Dan Jurafsky
TAs: Will Hamilton (head TA)

Janette Cheng       Robin Jia       Jeff Pyke
Tim Dozat           Rafael Musa     Kelly Shen
Ashkon Farhangi     Kate Park       Stephanie Tang
Gaspar Garcia       Charissa Plattner Lucy Wang
                           Rob Voigt

Time: TuTh 3:00-4:20, 420-040
cs124.stanford.edu
Evidence Based Pedagogy!
WHAT IS THE FLIPPED CLASSROOM?

The flipped classroom inverts traditional teaching methods, delivering instruction online outside of class and moving “homework” into the classroom.

THE INVERSION

The Traditional Classroom
Teacher’s Role: Sage on the Stage

LECTURE TODAY

Reading and questions due tomorrow

Homework

The Flipped Classroom
Teacher’s Role: Guide on the Side

ACTIVITY TODAY

Watch lecture online tonight!

http://www.knewton.com/flipped-classroom/
Why the flipped classroom (1)

**Mastery learning:** Learn until you master

Benjamin Bloom, 1968
Bloom's mastery learning

Personalized, **goal-driven practice**, driven by **feedback**

1. Watch (and re-watch) lectures at your own pace and learn when it's best for you

2. Videos have embedded miniquizzes. If you get it wrong, it gives you feedback about why you misunderstood.

3. You have 2 chances at each weekly Tuesday Quizzes, so you can go back to the lecture and retake them.

4. With programming assignments you can see your performance on the training set to see what you're doing wrong!
Why the videos have 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 so you can learn
Why the flipped classroom (2)

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 (3)

**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: I expect you to:
- Watch video lectures (and/or and read textbook chapters)
- On average about 90 minutes of video content each week
- Some people watch it speeded up

Some lectures live:
- 8 lectures and 1 group session 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”)
- Optional but recommended
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 20 3:30-6:30)

Class sessions: All encouraged; 8 live lectures required

- Full lectures
- Mini-lectures
- Group worked problems
The Open Platform: EdX!

https://lagunita.stanford.edu/about
https://open.edx.org/about-open-edx
Learning Goals

At the end of this course, you will be able to:
Learning goals

Write efficient regular expressions to solve any kind of text-based extraction task
Learning goals

Apply the edit distance algorithm to all sorts of text sequence problems
Learning goals

Build a supervised classifier to solve problems like sentiment classification
Learning goals

Build a search engine
Learning goals

Build a recommendation engine
Learning goals

Build a computational model of word meaning (using lexicons and embeddings)
Learning goals

Build a chatbot
Learning goals

Understand and implement PageRank
This class is the undergrad intro to:

Win 2018: cs224N Natural Language Processing w/Deep Learning
Win 2018: cs246 Mining Massive Data Sets
Spr 2018: cs222U Natural Language Understanding
Aut 2018: cs224W Analysis of Networks

Spr 2019: cs276 Information Retrieval and Web Search
TBD: cs224S Spoken Language Processing
Syllabus

http://web.stanford.edu/class/cs124
Coming up next class (Thursday)

Unix for poets

grep

sort
PA1: Spam Lord!

Write regular expressions to spread evil throughout the galaxy!

By extracting email addresses and phone numbers from the web!

jur a fs ky at st anford dot e d u

Goes live Friday!
Action Items Before Thursday!

1) Read the syllabus webpage at cs124.stanford.edu

2) Sign up for piazza and edX
   ◦ For edX, you'll need to first sign in with your SUnet ID at suclass.stanford.edu, and then click on the EdX button at the top of the cs124.stanford.edu webpage

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

4) Download this file to your laptop
   http://cs124.stanford.edu/nyt_200811.txt.gz