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

- Apple
- Amazon
- Google
- Microsoft
- Facebook
- Twitter
Social World

Disaster Relief

Chatbots for Mental Health

Improve Police-Community relations via Body-Cameras
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!*
Text Classification: Disaster Response

Haiti Earthquake 2010
Classifying SMS messages

Mwen thomassin 32 nan pyron
mwen ta renmen jwen yon ti dlo
gras a dieu bo lakay mwen anfom
se sel dlo nou bezwen

I am in Thomassain number 32, in the
area named Pyron. I would like to
have some water. Thank God we are
fine, but we desperately need water.

Programming Assignment 2: Triage!
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**

*Programming Assignment 3: Thumbs up!*
Sentiment Analysis

Emotional Spell-Check
Extracting Social Meaning from Language

**Uncertainty** (students in tutoring)

**Annoyance**
- callers to dialog systems:

**Anger** *(police-community interaction)*

**Deception**

**Emotion**

**Intoxication**

**Flirtation, Romantic interest**
- McFarland, Jurafsky, Ranganath study on heterosexual speed dating
What is the language of hetero flirters?

Women when flirting:
- raise pitch ceiling
- laugh at themselves

Men when flirting:
- raise their pitch floor
- laugh at their date (teasing?)
- 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 at UT Austin
- 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

sex, Drugs, and dessert

orgasmic pastry

sexy food

the fries are like crack

garlic noodles... my drug of choice

seductively seared fois gras

Drugs

Sex

Mentions per Review

Restaurant Price

Mentions per Review

Restaurant Price
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?
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).
Help improve Police-Community Interaction (week 9)

Problems:
• A flood of viral videos show inappropriate officer use of force
• Black Americans especially report more negative interactions with police (Epp et al. 2014)

Could natural language processing help?
• Quantify police-community interactions using body-worn cameras?
• Detect the potential for escalation?
• Help develop officer training?
• And hopefully reduce the chances of violence?
2. Interacting with humans via language
Personal Assistants

Siri

amazon alexa

Facebook M
A personal assistant inside Messenger

Google Now

Hi. I’m Cortana.
Ask me a question!
Question Answering: IBM’s Watson
Recommendation Engines

If you bought…. 

Customers who bought this item also bought

- **First Bite: How We Learn to Eat**
  - By Bee Wilson
  - Paperback
  - $11.37 (prime)

- **The Dorito Effect: The Surprising New Truth About Food and Flavor**
  - By Mark Schatzker
  - Paperback
  - $9.48 (prime)

- **Consider the Fork: A History of How We Cook and Eat**
  - By Bee Wilson
  - Paperback
  - $15.65 (prime)

- **Cuisine and Empire: Cooking in World History (California Studies in...**
  - By Rachel Laudan
  - Paperback
  - $16.20 (prime)
Recommendation Engines: Music

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>Shake it Off</th>
</tr>
</thead>
<tbody>
<tr>
<td>Artist Name</td>
<td>Taylor Swift</td>
</tr>
<tr>
<td>Spotify link</td>
<td>Hate typing? Just drag a spotify track here.</td>
</tr>
</tbody>
</table>

Side B. Track list

- Out of the Woods by Taylor Swift
  - Sorry, Spotify doesn't have that track
  - Delete
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 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!!!!!!

**Aye mate, her duck**
I mate or duck
I’m eight or duck
Eye maid; 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 you've 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
  
  $P(\text{"maison" } \rightarrow \text{"house"})$ high
  
  $P(\text{"L'avocat général" } \rightarrow \text{"the general avocado"})$ low
Models and Tools

Regular Expressions
Edit distance and alignment
Word embeddings
  • vector/neural models of meaning
Language models
Machine Learning classifiers
  • Naïve Bayes
  • Logistic Regression
  • Neural Networks

Network algorithms
  • PageRank
Recommendation algorithms
  • Collaborative filtering
Linguistic tools
  • Sentiment lexicons
Course logistics in brief

Instructor: Dan Jurafsky

TAs: Urvashi Khandelwal (head TA)
    Jennie Chen   Julia Mendelsohn   Charissa Plattner
    Laura Cruz-Albrecht   Matt Mistele   Minh-An Quinn
    Chuma Kabaghe   Vik Pattabi   Sam Redmond

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
Homework
Reading and questions due tomorrow

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

ACTIVITY TODAY
WATCH lecture online tonight!

From (defunct) 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 Quiz, so you can go back to the lecture and retake them.

4. With programming assignments you can see your performance on the training and dev 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 read textbook chapters)
- On average about 90 minutes of video content each week
- Some people watch it speeded up

Some lectures live:
- 7 lectures and 1 group session are required (on final exam, no videos)

In-class group sessions ("active learning")
- Optional but strongly recommended
Logistics More Specifically

Online Video Lectures with embedded quizzes (before class)
  • 20 pages of reading a week

Weekly online Review Quizzes (Tue of following week)

Roughly weekly Python homeworks (Fri of following week)

Final Exam (Tuesday March 19 3:30-6:30)

Class sessions: All encouraged; 7 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 neural word embeddings)
Learning goals

Build a chatbot
Learning goals

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

Win 2019: cs224N Natural Language Processing w/Deep Learning
Win 2019: cs246 Mining Massive Data Sets
Spr 2019: cs222U Natural Language Understanding
Spr 2019: cs276 Information Retrieval and Web Search
Aut 2020: cs224W Analysis of Networks
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