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
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 3: Search!*
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 Thomassin 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
A (fictional) application of sentiment analysis for toxicity intervention

Emotional Spell-Check
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 is used for "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*
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?

Peter S. Bearman, James Moody and Katherine Stovel
Chains of affection: The structure of adolescent romantic and sexual networks
Image drawn by Mark Newman
Help improve Police-Community Interaction (week 9)

Problems:
• A flood of viral videos show inappropriate officer use of force
• Black Americans report more negative interactions with police

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

amazon alexa

Siri

Google Now

Facebook M

Hi. I’m Cortana.
Recommendation Engines: The Good

If you bought....

Customers who bought this item also bought

First Bite: How We Learn to Eat
› Bee Wilson
🌟🌟🌟🌟🌟 46
Paperback
$11.37 ✉️prime

The Dorito Effect: The Surprising New Truth About Food and Flavor
› Mark Schatzker
🌟🌟🌟🌟🌟 193
Paperback
$9.48 ✉️prime

Consider the Fork: A History of How We Cook and Eat
› Bee Wilson
🌟🌟🌟🌟🌟 253
Paperback
$15.65 ✉️prime

Cuisine and Empire: Cooking in World History (California Studies in...)
› Rachel Laudan
🌟🌟🌟🌟🌟 35
Paperback
$16.20 ✉️prime
The dark side: YouTube Radicalization

Caleb Cain was a college dropout looking for direction. He turned to YouTube.
Why is language interpretation hard?
Ambiguity

Resolving ambiguity is hard
Ambiguity

There are at least 6 meanings of this sentence:

I made her duck

Go here and type (and vote for) some definitions

https://pollev.com/danjurafsky451
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

**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, emojis, hashtags, names

chowdownwithchan #crab and #pork #xiaolongbao at @dintaifungusa... where else? 😂cry_smiley_head Serum! Note the cute little crab indicator in the 2nd pic 🦀❤️
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:
- Neural and other machine learning models built from language data
Models and Tools

Regular Expressions

Edit distance and alignment

Neural word embeddings

Language models

Machine Learning classifiers
  - Naïve Bayes
  - Logistic Regression
  - Neural Networks

Network algorithms
  - PageRank

Recommendation algorithms
  - Collaborative filtering

Linguistic tools
  - Sentiment lexicons
  - Emotion lexicons

The GUS chatbot architecture (Siri, Alexa, Cortana, etc)
Neural word embeddings

A word's meaning is a point in 300-dimensional space

A 2-D visualization:

- mango
- pineapple
- banana
- lake
- mountain
- Tolstoy
- Gogol
Embeddings are the core of NLP

Core technology for any NLP task (question answering, machine translation, information retrieval, etc)

• Finding synonyms for words
• Deciding if two sentences have similar meaning
How to learn these "embeddings"?

Push words *together in space* they occur *together in text*

Read millions of words.
When you see:

**Banana, mango, or pineapple are all delicious...**

Move *banana* closer to *mango*
Move *banana* further from *Tolstoy*
Problem: Embeddings reflect cultural bias!


Ask “Paris : France :: Tokyo : x”
◦ x = Japan

Ask “father : doctor :: mother : x”
◦ x = nurse

Ask “man : computer programmer :: woman : x”
◦ x = homemaker

What can we do about this problem? Week 5!
Logistics: Instructor

Instructor: Dan Jurafsky (he/him)
Professor in CS and Linguistics
Department Chair of Linguistics
My office hours:

- Tuesday after class 4:00-5:45

*How to pronounce my name
Course logistics

TAs: Dan Iter (head TA)

Jesus Cervantes
Bryan Kim
Grace Lam

Allison Lettiere
Rose Li
Ben Newman

Arjun Sawhney
Lauren Zhu

Time: TuTh 2:30-3:50

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!
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/midterms: 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
- Most important: programming assignments
- Active learning (“constructivism”), learning by doing

**Collaborative learning**: Learn from each other
- Use class time for group problem-solving
- “Small group active learning”
cs124: Semi-flipped classroom

1. **Prerecorded video lectures on video:** I expect you to:
   - Watch video lectures (~90 min/week)
   - Some people watch it speeded up

2. **Live sessions:**
   - 5 lectures, required (either in person or watch recording)
   - 4 In-class group sessions (“active learning”), **not recorded**
     - #1 (this Thursday) is strictly required
     - #2,3,4 are optional but strongly recommended
Logistics More Specifically

Online Video Lectures w/embedded non-graded questions (watch before class)

20 pages of reading a week (up to you when to read)

Weekly online Review Quizzes (Tue of following week)

6 Python homeworks (Fri of following week)

Two midterms during class time (but we will have alternate times on both days)

- May 6
- June 3
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 neural network and train it using stochastic gradient descent
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 2021:  cs224N Natural Language Processing w/Deep Learning
Win 2021:  cs246 Mining Massive Data Sets
Spr 2021:  cs222U Natural Language Understanding
Win 2021:    cs224S Spoken Language Processing
? 2022:    cs346 Ethical and Social Issues in NLP
? 2022:    cs276 Information Retrieval and Web Search
Aut 2022:    cs224W Machine Learning with Graphs
Aut 2022:    cs221 Artificial Intelligence
Should I take 124 or 224X?

CS124 is designed for sophomores or juniors
◦ It's gentle (my goal is to explain everything) and broad (covering many many topics)
◦ Mastery learning, quizzes and programming homeworks

CS224X are deeper, more focused, grad courses
◦ They assume you are familiar with machine learning and will jump right into optimization and do advanced stuff
◦ Learning via research: novel research projects as a large component

(You should of course take all of them!!)
Syllabus

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

Unix for poets

grep

sort
Action Items Before Thursday's class!

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

2) Sign up for piazza there.

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

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

5) If you don't know UNIX yet (haven't had cs107):
   - For people using a Windows 10 machine, if you don't have Ubuntu on your machine:
     - Watch the first 9 minutes of Bryan's lovely pa0 video about how to download and install Ubuntu:
       https://canvas.stanford.edu/courses/136258/modules/items/873662
   - Watch Chris Gregg's excellent UNIX videos here: Logging in, first 7 File System, and first 8 useful commands
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 5pm!

*Just kidding; don't be evil