

CS 124/LINGUIST 180

From Languages to Information

DAN JURAFSKY

PROFESSOR OF COMPUTER SCIENCE

PROFESSOR OF LINGUISTICS

STANFORD UNIVERSITY

WINTER 2024

INTRODUCTION AND COURSE OVERVIEW

From Languages to Information

Automatically extracting meaning and structure from:

- Human language (news, social media, etc.)
- Social networks

Interacting with humans via language

- Web search
- Chatbots (= Dialogue Systems)
 - Question Answering
 - Recommendation Systems

Commercial World



OpenAI



Meta

Microsoft[®]



amazon

Social World

Disaster Relief

Help teachers in the classroom

Use data from body-worn camera to
improve Police-Community relations

Train mental health counselors

1. Extracting information from language

Information Retrieval

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

Text-based information retrieval is 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!

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 anfon
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**

Programming Assignment 2: Sentiment

A (fictional) application of sentiment analysis for toxicity intervention that shows how hard it can be!

Emotional Spell-Check

Could we do that tone-check with ChatGPT?

Not yet!

- **Toxicity detection:** detecting hate speech, abuse, harassment, or other kinds of toxic language
 - Widely used in online **content moderation**
- Problem: toxicity classifiers misfire on words for minority identities:
 - The word "blind": (Hutchinson et al., 2020)
 - The word "gay" (Dixon et al. 2018, Oliva et al., 2021)
- Result: censorship of speech by disabled people and others

Could we apply sentiment analysis for humanities or social science questions?

Sentiment in Restaurant Reviews

Dan Jurafsky, Victor Chahuneau, Bryan R. Routledge, and Noah A. Smith. 2014. Narrative framing of consumer sentiment in online restaurant reviews. First Monday 19:4

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 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

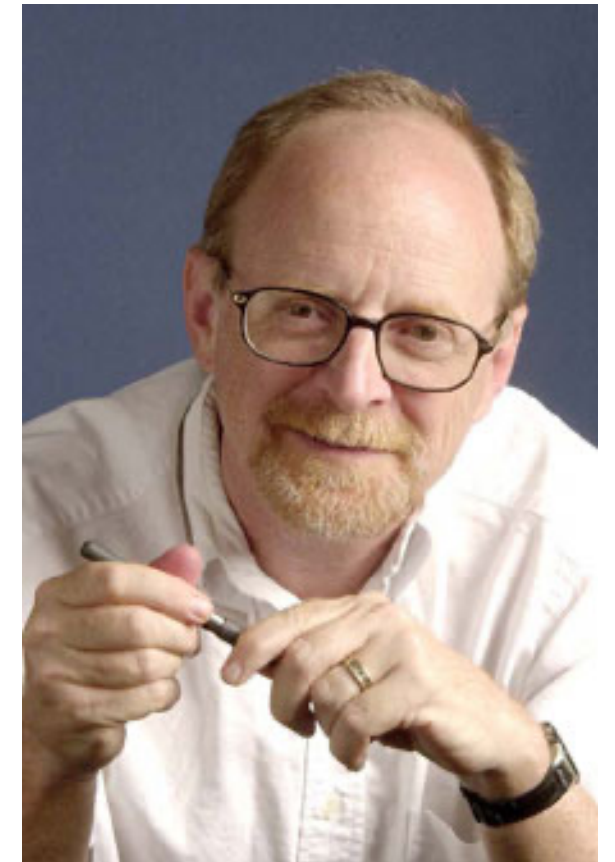
Texts written by **people suffering trauma**

- 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

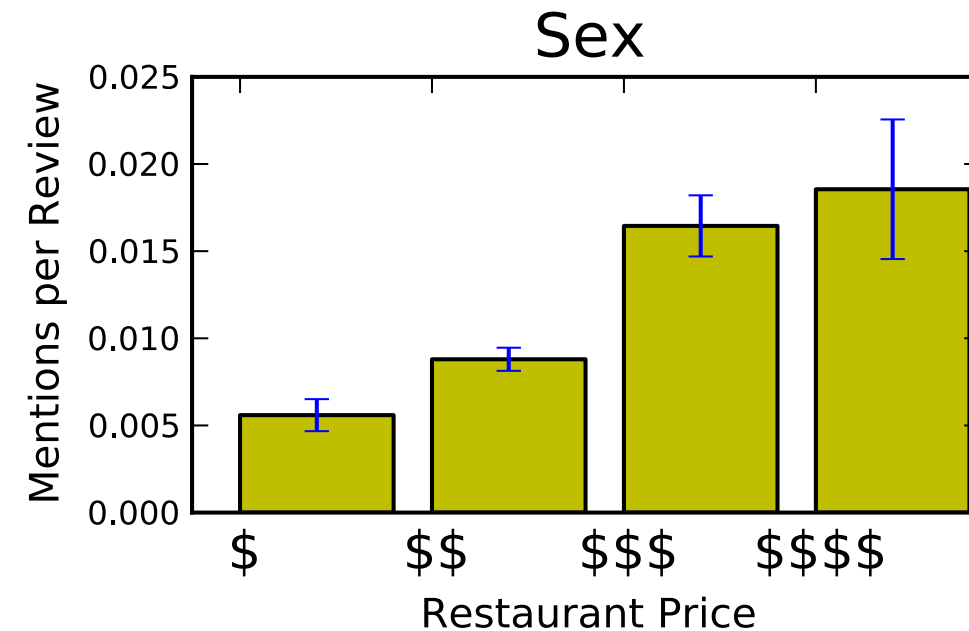
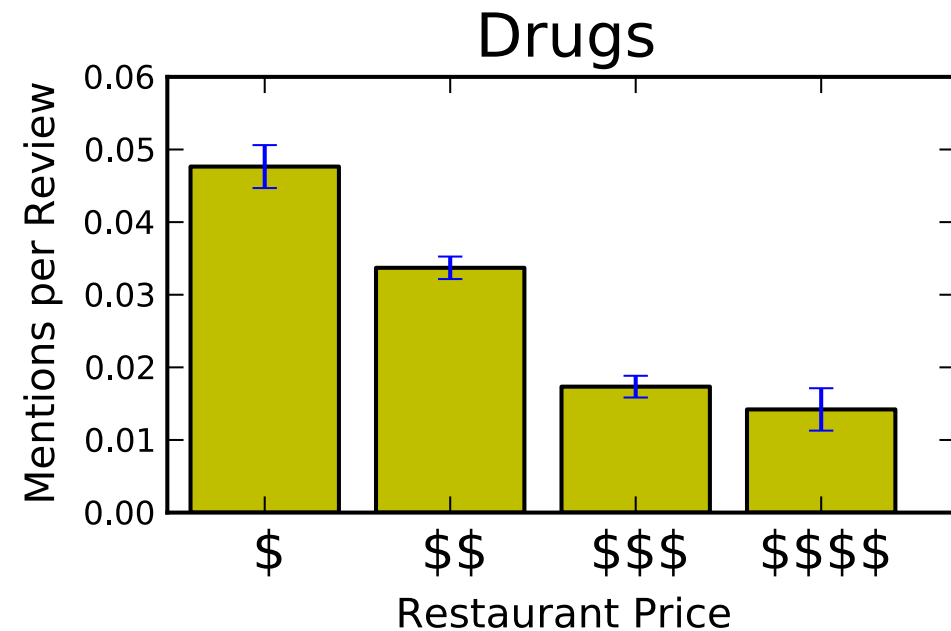
garlic noodles... my *drug* of choice

the fries are *like crack*

orgasmic pastry

sexy food

seductively seared fois gras



NLP can be a tool for cultural analytics!

Computational Biology: Comparing Sequences

AGGCTATCACCTGACCTCCAGGCCGATGCCC

TAGCTATCACGACCGCGGGTCGATTGCCCCGAC

–AGGCTATCACCTGACCTCCAGGCCGA––TGCCC––
| | | | | | | | | | | | | | | | | | | | |
TAG–CTATCAC––GACCGC––GGTCGATTGCCCCGAC

Sequence comparison is key to

- Finding genes
- Determining function
- Uncovering evolutionary processes

This is also how simple 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?
- **We'll learn: Network algorithms (Quiz 9)**

Help improve Police-Community Interaction (week 6)

Problems:

- A flood of videos show inappropriate use of force by police
- Black Americans suffer from much more negative interactions with police



Could natural language processing help?

- Quantify police-community interactions using body-worn cameras?
- Help develop officer training?
- Reduce the chances of violence?
- I'll talk about my work with Prof. Jennifer Eberhardt



2. Interacting with humans via language

Personal Assistants

amazon alexa



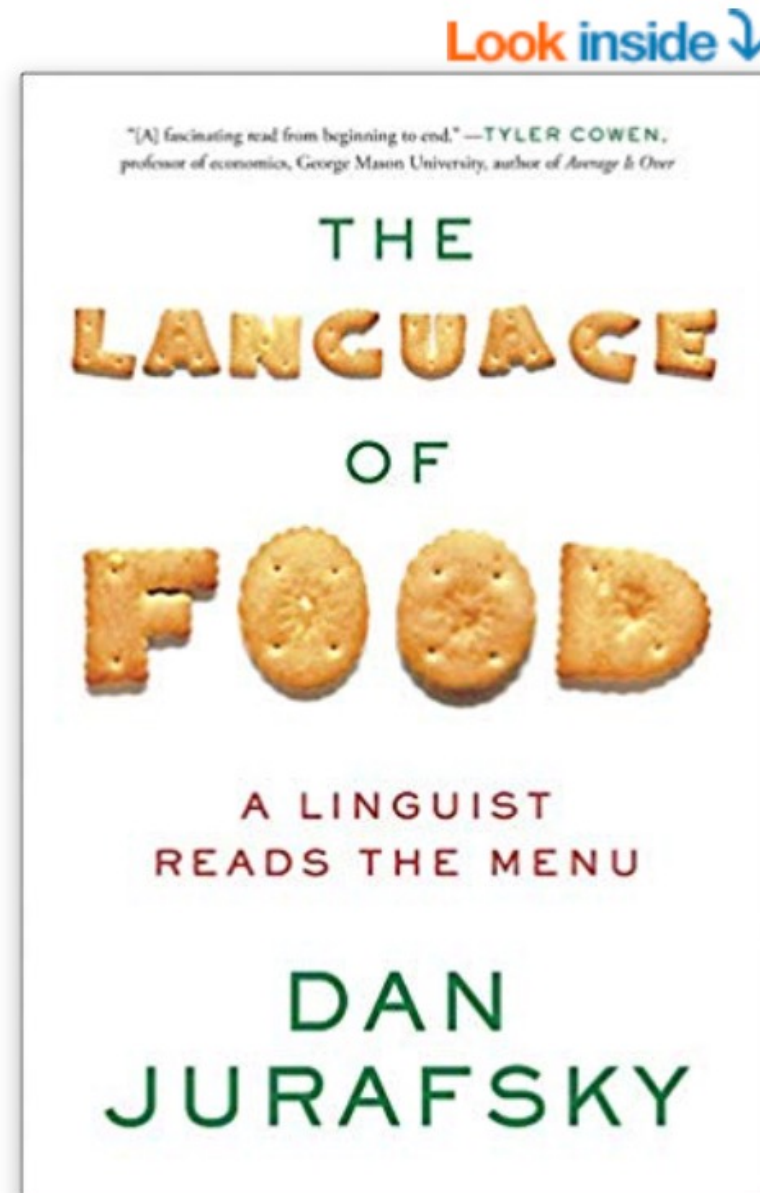
Siri



PA 7 Chatbot!

Recommendation Engines: The Good

If you bought....



Customers who bought this item also bought



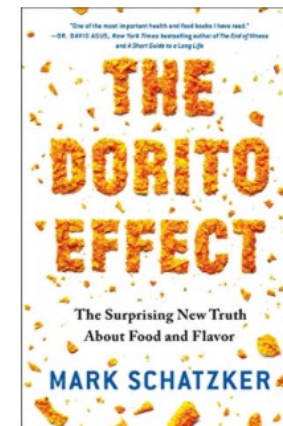
First Bite: How We Learn to Eat

› Bee Wilson

★★★★★ 46

Paperback

\$11.37 ✓prime



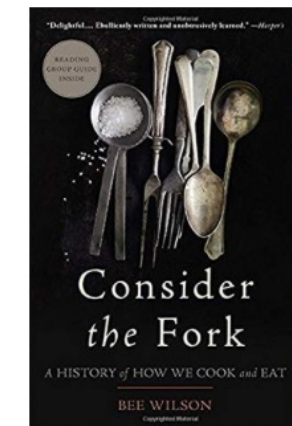
The Dorito Effect: The Surprising New Truth About Food and Flavor

› Mark Schatzker

★★★★★ 193

Paperback

\$9.48 ✓prime



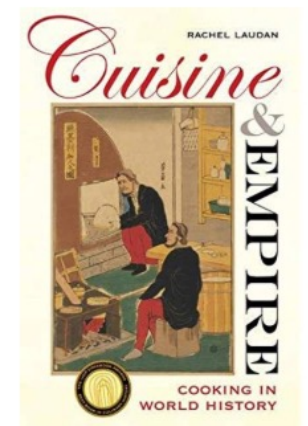
Consider the Fork: A History of How We Cook and Eat

› Bee Wilson

★★★★★ 253

Paperback

\$15.65 ✓prime



Cuisine and Empire: Cooking in World History (California Studies in...)

› Rachel Laudan

★★★★★ 35

Paperback

\$16.20 ✓prime

PA 7 and Quiz 8

The dark side: YouTube Radicalization



ChatGPT (or LLAMA2-CHAT or Claude, etc): Large language models for chat

Act-as prompting and GAtt

I like anything to do with architecture. If it's relevant, suggest something related.



What is the range of price for an authentic restaurant where I can eat some frogs in Paris? Answer in one sentence only.



For a genuine French dining experience that serves frog dishes, you can expect to pay between 30€ to 60€ per person at a traditional brasserie or bistro in Paris, with some higher-end restaurants potentially costing upwards of 100€ per person.



And besides eating frogs, what do you recommend me to do in this city? Don't mention more than 3 things



Sure, here are three things to do in Paris beyond eating frogs:

1. Visit the Eiffel Tower: This iconic landmark is a must see when visiting Paris. You can take the elevator to the top for breathtaking views of the city.
1. Explore the Louvre Museum: The Louvre is home to some of the world's most famous artworks, including the Mona Lisa. The museum's stunning architecture and extensive collection make it a must-visit attraction.
2. Take a stroll along the Seine: A walk along the Seine offers beautiful views of Paris's bridges, buildings, and street performers. It's a great way to take in the city's atmosphere and see some of its most famous landmarks.

Logistics: Instructor

Instructor: Dan Jurafsky (he/him)

Professor in CS and Linguistics

My office hours:

- Tuesday after class 4:30-5:45
- Margaret Jacks Hall 117

*How to pronounce my name



Picture by Ross Petukhov

Course Staff



Dan Jurafsky
Professor



Deveshi Buch
Head TA



Veronica Rivera
Embedded Ethics Postdoc



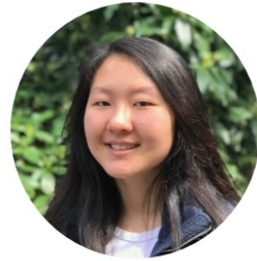
Mo Akintan
TA



Pranav Gurusankar
TA



Natasha Kacharia
TA



Hanna Lee
TA



Amelia Leon
TA



Jasper McAvity
TA



Anwesha Mukherjee
TA



Fahad Nabi
TA



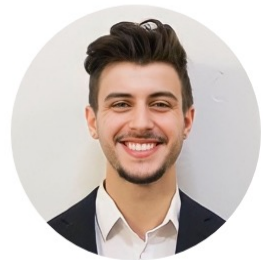
Tolu Oyeniyi
TA



Uma Phatak
Embedded Ethics TA



Michael Ryan
TA



Francis Santiago
TA



Jingwen Wu
TA



Jack Xiao
TA

Let's think about the language modeling task.

Why is it so remarkable?

What makes language interpretation hard?

Ambiguity

Resolving ambiguity is hard

Ambiguity

There are at least half a dozen meanings of this sentence:

The chef made her duck

Go here and type (and vote for) some definitions

<https://pollev.com/danjurafsky451>

Ambiguity

The chef made her duck

The cook cooked waterfowl for a different woman X (person using "she/her" pronouns) to eat

The cook cooked waterfowl belonging to X

The cook cooked waterfowl belonging to the cook

The cook created the (plaster?) waterfowl that X owns

The cook caused X to quickly lower X's head or body

The cook uncovered the true identity of the cook's spy waterfowl

The cook waved their magic wand and turned X into undifferentiated waterfowl

The chef made her duck

The chef caused X to quickly lower her head or body

Part of speech: “duck” can be a Noun or Verb

The chef cooked waterfowl for X (or belonging to X)

Part of speech:

“her” is possessive pronoun (“of her”)

“her” is dative pronoun (“for her”)

The chef cooked waterfowl belonging to the chef (vs to X)

Coreference

“her” can refer to X or to the Chef

The chef made the (plaster) duck statue X (or the chef) owns

Word Meaning : “make” can mean “create” or “cook”

More difficulties:
Non-standard language,
emojis, hashtags, names



chowdownwithchan #crab and #pork #xiaolongbao at @dintaifungusa... where else? 😂👩 Note the cute little crab indicator in the 2nd pic 🦀💕💕

Models and Tools

Regular Expressions

Edit distance and alignment

Neural word embeddings

Supervised classifiers

- Naive Bayes
- Logistic Regression
- Neural Networks

Neural Large Language Models

Network algorithms

- PageRank

Recommendation algorithms

- Collaborative filtering

Linguistic tools

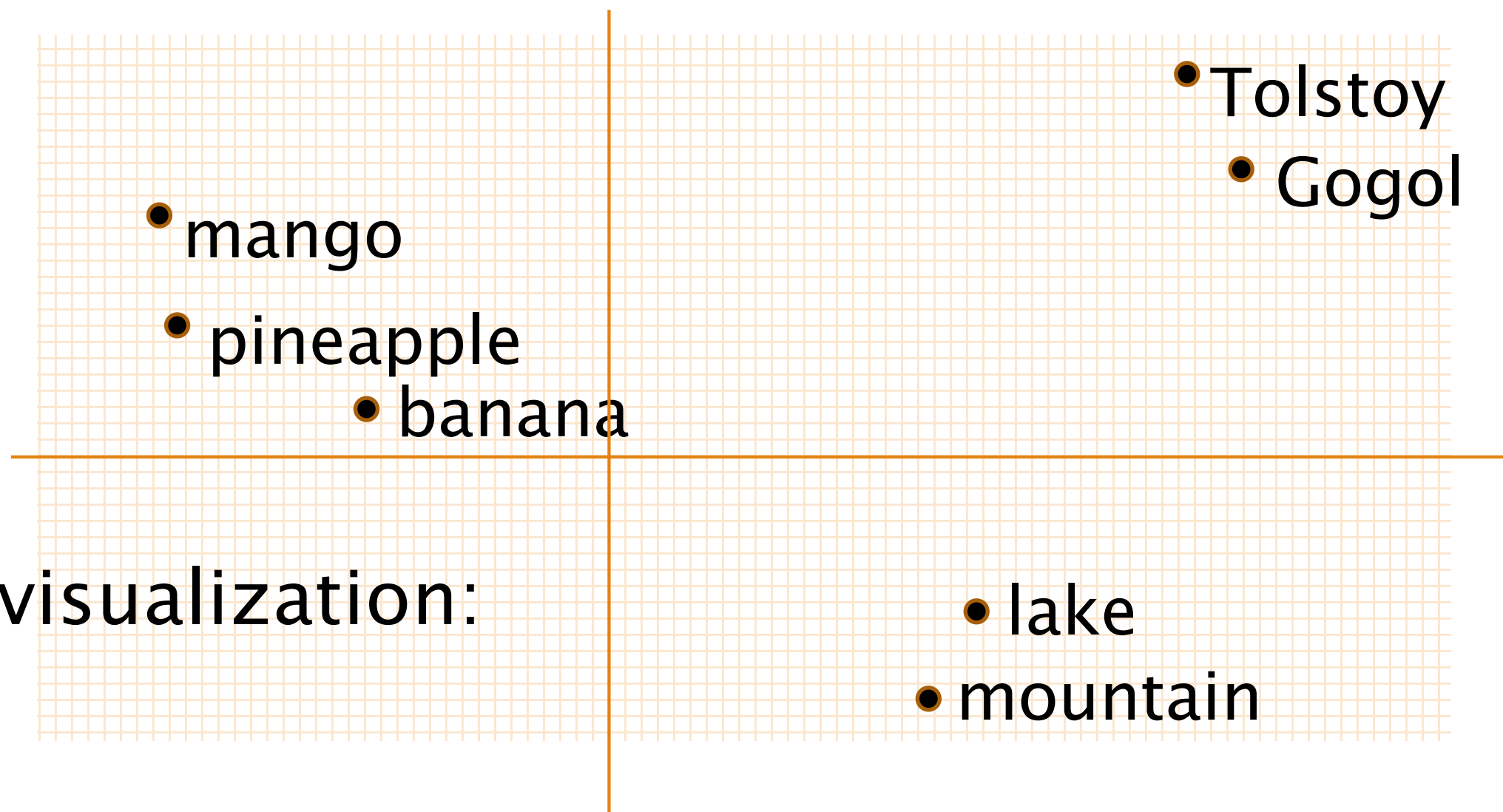
- Sentiment lexicons
- Emotion lexicons

The GUS chatbot architecture
(Siri, Alexa)

Neural chatbots (ChatGPT)

Core of modern NLP: Neural "word embeddings"

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



A 2-D visualization:

Problem: Embeddings reflect cultural bias!

Bolukbasi, Tolga, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. 2016. "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings." In *NeurIPS 2016*, pp. 4349-4357.

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!

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



The Flipped Classroom

Teacher's Role: Guide on the Side



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 **infinite** 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 might be doing wrong on the test set!

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

- Middendorf and Kalish, 1995, Johnstone and Percival 1976, Burns 1985

“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”
- You must do PA7 in groups of 4
- We encourage pair programming on PA1-6

cs124: Flipped classroom

1. Prerecorded video lectures on video:

- About ~90 minutes/week of video lectures
- Some people watch it speeded up

2. Live sessions: (none are recorded)

- 4 required in-person lectures
- 5 required in-class labs(“active learning”)
 - Lab #1 (Unix text tools) next Tuesday is **required in person**
 - Labs #2, #3, #4 are required but attendance extra credit (you can do at home).
 - Lab #5 March 5 (Git and PA7) is **required in person**

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 Quizzes (due Tue of following week)

7 Python homeworks (due Fri of following week)

One midterm that is online and can be taken during an extended 24-hour period (Feb 29)

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, both by prompting of a large language model, and by building dedicated components

Learning goals

Understand and implement PageRank and other social network functions

Learning goals

Understand large language models and be able to prompt them, reason about what they can do, and and reason about their social implications

Learning goals

Work on computational projects in groups

PA1-6: Pair programming is encouraged

PA7: Must be done in groups of 3-4

This class is the undergrad intro to:

Win 2024: cs224N NLP w/Deep Learning (Hashimoto and Yang)

Spr 2024: cs224N NLP w/Deep Learning (Manning)

Win 2024: cs246 Mining Massive Data Sets (Leskovec)

Spr 2024: cs224C: NLP for Computational Social Science (Yang)

Spr 2024: cs224S Spoken Language Processing (Maas)

Fall 2023: cs329R Race and NLP (Jurafsky + Eberhardt)

Fall 2023: cs224V Conversational Virtual Assistants with Deep Learning (Lam)

Not this year: cs222U Natural Language Understanding (Potts)

Not this year: cs346 Ethical and Social Issues in NLP (Jurafsky)

Not this year: cs 329X: Human-Centered NLP (Yang)

Not this year: cs276 Information Retrieval and Web Search (Manning)

Should I take 124 or 224N?

CS124 is designed for sophomores or juniors

- It's gentle (I explain everything) and broad (covering many topics, not just NLP but also recommendation engines, IR, social networks)
- Mastery learning, quizzes, programming assignments with starter code and scaffolding.
- No research project, but a fun chatbot final homework

CS224N and 224U are deeper, more focused, grad courses

- They assume you are familiar with machine learning and will jump right into optimization
- More details of implementation and you will learn about and implement more advanced things
- 124 stops right before Transformers and attention; 224N will teach transformers and attention
- Learning via research: novel research projects as a large component

CS224?? (Human Centered NLP) and CS346 (Social and Ethical Issues in NLP) require 224N or 224U

CS224C: more applied focus, applying NLP to social science: (NLP for Computational Social Science)

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

Syllabus

cs124.stanford.edu

Where do I find all the programming assignments and quizzes and readings?

Everything is on the webpage cs124.stanford.edu

Except the videos which are on Canvas Modules!

In other words:

- Lectures slides: webpage
- Lab instructions: webpage
- Tutorial information: webpage
- Programming assignments: webpage (points to git where they live)
- Weekly quizzes: webpage (points to gradescope where they live)
- Practice midterm: webpage
- Midterm: gradescope
- Videos: canvas

Coming up this week: Thursday

Optional tutorial on jupyter notebooks and PA0, getting ready for PA1

Come to class **with your laptops** and we'll go through PA0 together!

This tutorial will be led by amazing TA Jasper McAvity!!! But I and many other CAs will be there!

Action Items Before Thursdays class!

- 1) Read the syllabus webpage at cs124.stanford.edu
- 2) Look at PA0 (you can find it from the webpage)
- 3) Watch Canvas Videos on "PA0 Mac Setup" (or "PA0 Windows Setup"), also pointed to by webpage

Coming up next
week (Tuesday)

"Unix for poets":

grep

sort

Key UNIX tools for
dealing with text files and
regular expressions.



Action Items before next Tuesday's class!

1) Watch the "week 1" videos on Canvas by Sunday (since the quiz is also due Tuesday)

3) Download this file to your laptop

http://cs124.stanford.edu/nyt_200811.txt

4) 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 pa0 Windows video about how to download and install Ubuntu (it's pointed to from the website)
- Watch Chris Gregg's excellent UNIX videos here: Logging in, first 7 File System, and first 8 useful commands

<https://web.stanford.edu/class/archive/cs/cs107/cs107.1186/unixref/>

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

