Introduction to the Course

Sameer Singh and Conal Sathi

BANA 290: ADVANCED DATA ANALYTICS

MACHINE LEARNING FOR TEXT

SPRING 2018

April 3, 2018

Sameer Singh

sameer@uci.edu

Academic Positions

- Assistant Professor at UC Irvine (2016 -)
- Postdoc at University of Washington (2013 -)
- PhD from University of Massachusetts, Amherst (2014)

Research Interests

- Natural Language Processing: information extraction, relation extraction, entity linking and disambiguation, joint modeling
- Machine Learning: interpretable ML, semi-supervised learning, matrix/tensor factorization, probabilistic graphical models

http://sameersingh.org

Conal Sathi

csathi@uci.edu

Education

- M.S. at Stanford University
- studying computer science with the focus on AI
- Published in the intersection of NLP and Graph Mining

Industry Experience

- Hired as the first machine learning engineer at a technology startup
  - focused on e-commerce analytics from email data
- Worked on many real-world applications of ML/NLP
  - from text classification to information extraction to recommendation engines
- Advisor to several technology companies in Northern California
- Many of these companies are looking for data scientists and engineers
  - so reach out to him if you would like career help!

Course Logistics

Meetings

- Room: SB2 117
- Tues 4:00-6:50
- No holidays this quarter (Yay!)

TA

- Yoshitomo Matsubara
- PhD student, Comp Science
- Email: yoshitom@uci.edu
- But, contact us on Piazza

Office Hours (Sameer)

- Room: DBH 4204
- Mostly available Wednesday afternoons (but by appointment only)
- https://calendly.com/sameersingh/office-hours
- TBA for Conal and Yoshi (available on Piazza)

Canvas Course webpage: https://canvas.eee.uci.edu/courses/9097/

Topics (subject to change)

Supervised Learning

- Text Classification: discriminative, generative, libraries
- Feature Engineering: bag of words, lexicons, regular expressions
- Evaluation: Metrics, visualization, debugging errors

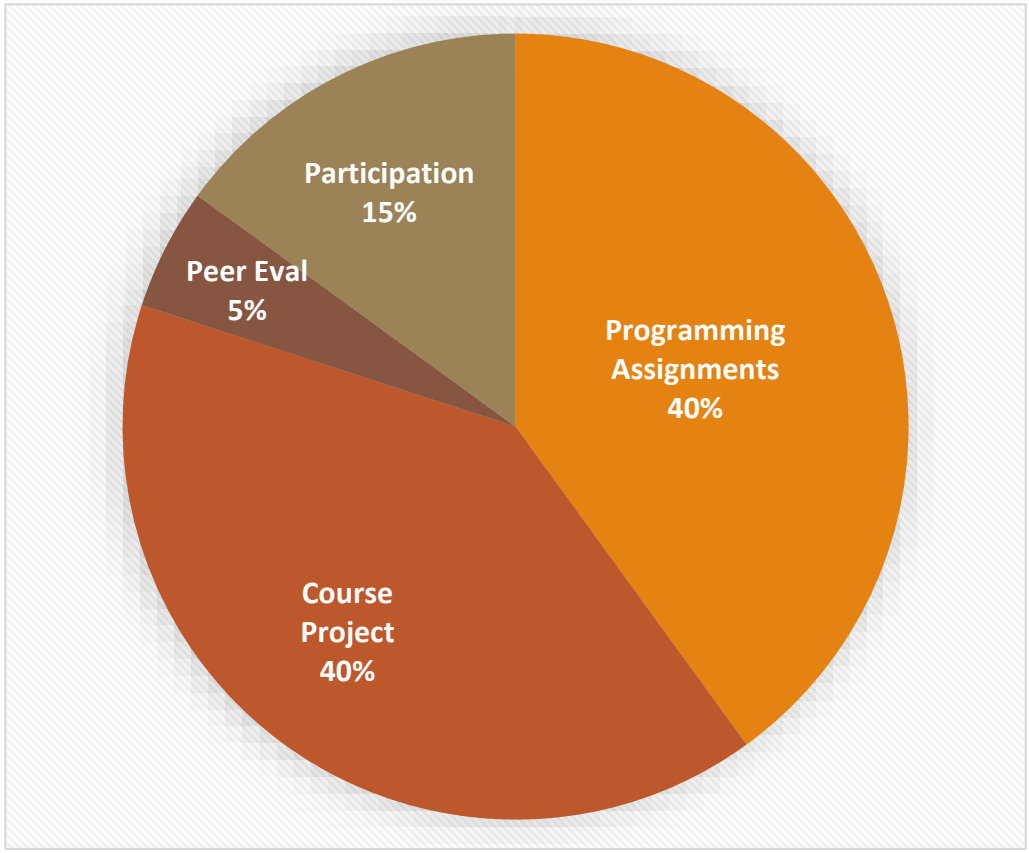
Unsupervised Learning

- Clustering: k-means, DBScan, topic models
- Factorization: latent semantic analysis, word embeddings
- Semi-supervised learning: improving classification performance

Applications and other topics

- Information Extraction: entity tagging, coreference, linking, question answering
- Chat bots and dialog systems, summarization, retrieval

Grading



Submissions

- All submissions through Canvas
- All deadlines are available now
- Will (probably) not be changing..
  - So start planning now

Programming Assignments

4 Programming Assignments

- Every 2 weeks through the quarter

Writing Up (PDF)

- Analysis of your implementation

Source Code (Python)

- Should be pretty straightforward
- Piazza for bugs, weird results, etc.

Late Submissions

- You get five *grace* days
- Full credit when used (no q asked), o.w. 0

Group Projects

- More details in Week 3-4

Groups for the Project

- Team size should be 2, 3 is okay!
  - Try to get diverse team members

Three "submissions"

- Proposal is very short (~1 page), Week 5
- Presentations in Week 9 or 10
- Final report matters the most

Peer Evaluation (individual)

- Everyone will be assigned a few groups
- Comment on their presentation
- Provide suggestions for final report

Participation

In-class participation

- Attend all the classes
  - Submit in-class notebooks
  - Ask questions! Answer them!

Piazza participation

- Propose project ideas
- Ask/answer questions and issues
- Provide feedback to Instructor and TA
- Discuss readings and papers

Evaluations

- Polls/Surveys
- Course Evaluations
  - Mid-Term and Final

Applications of Text Processing

INDUSTRY EXAMPLES

Content summarization

North Italia Claimed

1787 reviews Details

$$ · Pizza, Italian

Ying L.
Seal Beach, CA
44 friends
250 reviews
248 photos
Elite '18

2/18/2018
3 check-ins

We have been here many times before. We came yesterday for dinner around 6pm. We ordered as soon as we sat down cuz we know what we wanted. The waiter came back after the appetizer was served and said he is going on a break, another server will take over. Took at least 40 mins before dinner came out. The second server was no where to be found for us to ask where our food was.

I think she forgot our table and the pasta was cold. I ordered the same Bolognese many times before and the portion was smaller than before. For $19, it should at least come out hot.

We asked for boxes to take the left over home and she didn't even offer the dessert menu...so I paid and left.

Food is good here but the service is horrible.

Hanalee C. and 2 others voted for this review

Useful 3 Funny 2 Cool 2

Lydia T.
San Diego, CA
384 friends
210 reviews
381 photos
Elite '18

3/26/2018
1 check-in

We went family style, nibbled lots of plates and everything was really good. My favorites include: white truffle garlic bread, calamari with arugula and a roasted lemon, mussels (lots of salami atop!), meatballs, **squid ink mafaldine**, cauliflower and egg, margherita pizza. Sprinkled red chili pepper flakes on everything.

Service was attentive, even with it being very busy. We had drinks and dessert and it was a good value. I was pleasantly surprised by how little our tab was.

Soleil S. voted for this review

Useful 1 Funny Cool 1

Steve S.
Orange County, CA
1992 friends
1225 reviews
7906 photos
Elite '18

2/9/2018
1 check-in
Listed in Because I want to be Italian, 2018 Reviews

Fam Bam Sunday Lunch time and I needed a home run. I get this look from wifey when I take her to a new restaurant and the food is just so so. She would say 'Mr. Yelper, 3 stars, why did you take me here, where's the 5 (and then she would laugh)" Ouch! And I've been getting that look way too often lately. Not this time =)

This place is really nice. Vaulted ceiling, modern minimalist decor, a nice big bar on one side and immaculate rows of table on the other.

We got there a little before noon and the place was packed. We luckily snag on of the last few tables available. After us, the line got long (very long) throughout our meal.

Content summarization

"The **desserts** were really fantastic and my favorite was the tiramisu with the little chocolate spheres which gave it a lovely texture." in 312 reviews

Good For: Dessert

"We started with a yummy glass of their red Zinfandel, which was very yumm, and some **white truffle garlic bread**- my mouth is still salivating." in 200 reviews

$13 White Truffle Garlic Bread

"**Short Rib Radiatori** ($21): tender short ribs w/ radiator shaped pasta w/ parmesan cream, wilted arugula and herbed breadcrumbs." in 159 reviews

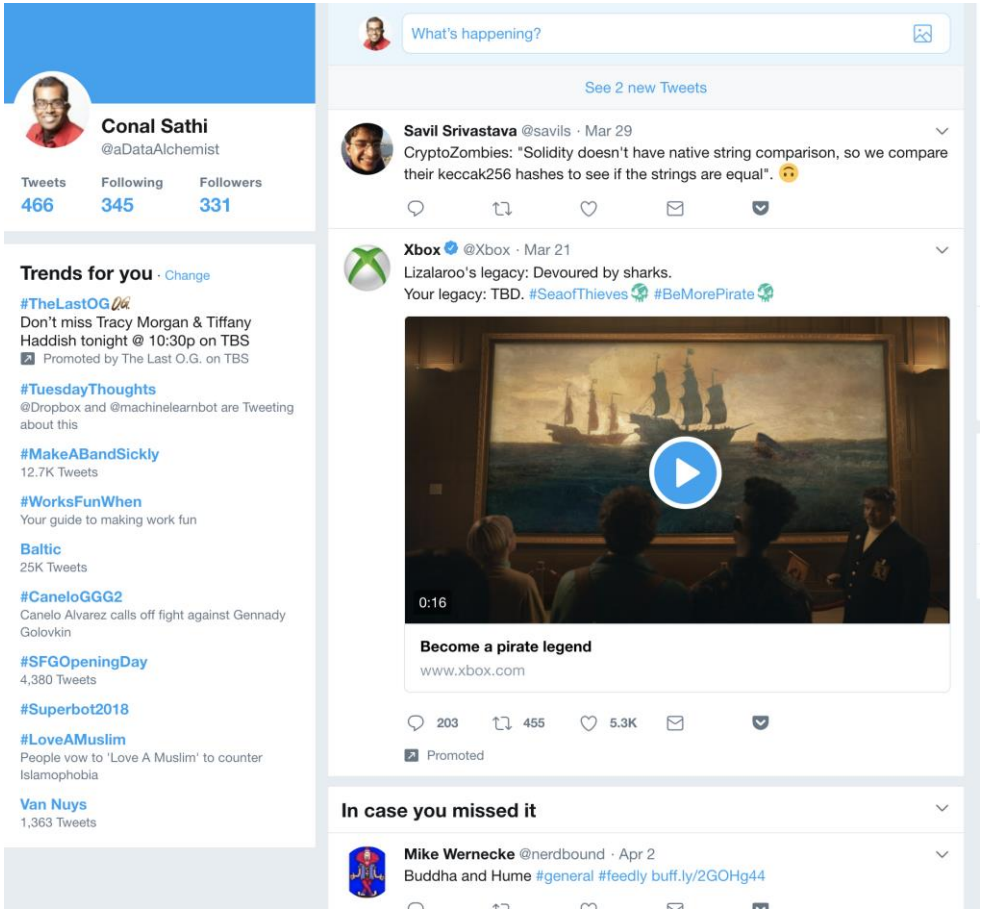
$22 Short Rib Radiatori

Show more review highlights

More business info

Takes Reservations **Yes**
Delivery **No**
Take-out **Yes**
Accepts Credit Cards **Yes**
Accepts Apple Pay **No**
Accepts Android Pay **No**
Good For **Lunch, Dinner, Dessert**
Parking **Valet**
Bike Parking **Yes**
Wheelchair Accessible **Yes**
Good for Kids **Yes**
Good for Groups **Yes**
Attire **Casual**
Ambience **Trendy**
Noise Level **Average**
Alcohol **Full Bar**
Outdoor Seating **Yes**
Wi-Fi **No**
Has TV **Yes**
Waiter Service **Yes**
Caters **Yes**

Content summarization



Content summarization

K nowhere

Tesla stock soars after company releases Q1 production data

Tesla bonds inch up after Q1 production report

Sentiment analysis

- Are your customers happy about your business/product?
- What do they like/dislike?

Sentiment analysis

- Are your customers happy about your business/product?
- What do they like/dislike?
- Examples:
  - Yelp
    - After the latest change in the menu, are people happier or more frustrated?

Sentiment analysis

- Are your customers happy about your business/product?
- What do they like/dislike?
- Examples:
  - Amazon
    - How do people about the product after the purchase is made? What aspects do they like and which aspects do they dislike?

Sentiment analysis

- Are your customers happy about your business/product?
- What do they like/dislike?
- Examples:
  - Twitter
    - After the Super Bowl Ad, are people talking about your brand more positively or more negatively?

Search

Top customer reviews

Amazon Customer

This camera is fantastic!

December 20, 2017

Color: 1080p Black | Verified Purchase

I love this camera! It does everything we wanted it to. Grandpa is in assisted living and we can watch him day and night. We can even talk to him and remind him to stay in bed or tell him that will be late for a visit. Sometimes the sound that I hear on my smartphone is not the greatest quality but what he hears in his room is perfect. I love that we can watch him and I love that the staff know that we can be watching them day or night. One feature that I wish was different is that that app cannot be open with any others on my phone. I've thought about having a separate phone just as the monitor for the camera. It is worth every penny!! I bought my second one just today!!

Comment | 85 people found this helpful. Was this review helpful to you? Yes No Report abuse

M. Y. 'Photographer | Outdoor Enthusiast | Tech Pro' Top Contributor: Photography TOP 1000 REVIEWER

Great product, non-existent Support, and 2 out of 9 cameras failed

August 25, 2017

Color: 1080p White | Verified Purchase

[UPDATE 1/1/2018: I stand by the fact that technical support is POOR. I had never received a reply back to any of my email inquiries. As such, I downgraded the review from 4 to 3 stars.]

I bought 9 of these, 2 of which did not work properly. This then gave me a glimpse of Yi's customer service -- or lack thereof. I ended up returning the 2 defective ones in frustration.

PROS:
- Great design and build quality
- Apps work great, though they are inconsistent across iOS, Android, and Windows, each with their Pros and Cons
- Ability to share camera feed with family
- Ability to record to microSD cards
- 7 days of 6-second motion alerts saved in the Yi Cloud for free
- Motion alerts work great
- Alert regions can be defined (via a rectangle)
- Can schedule when alerts are on/off

Read more

Comment | 87 people found this helpful. Was this review helpful to you? Yes No Report abuse

Daniel A. Gorski

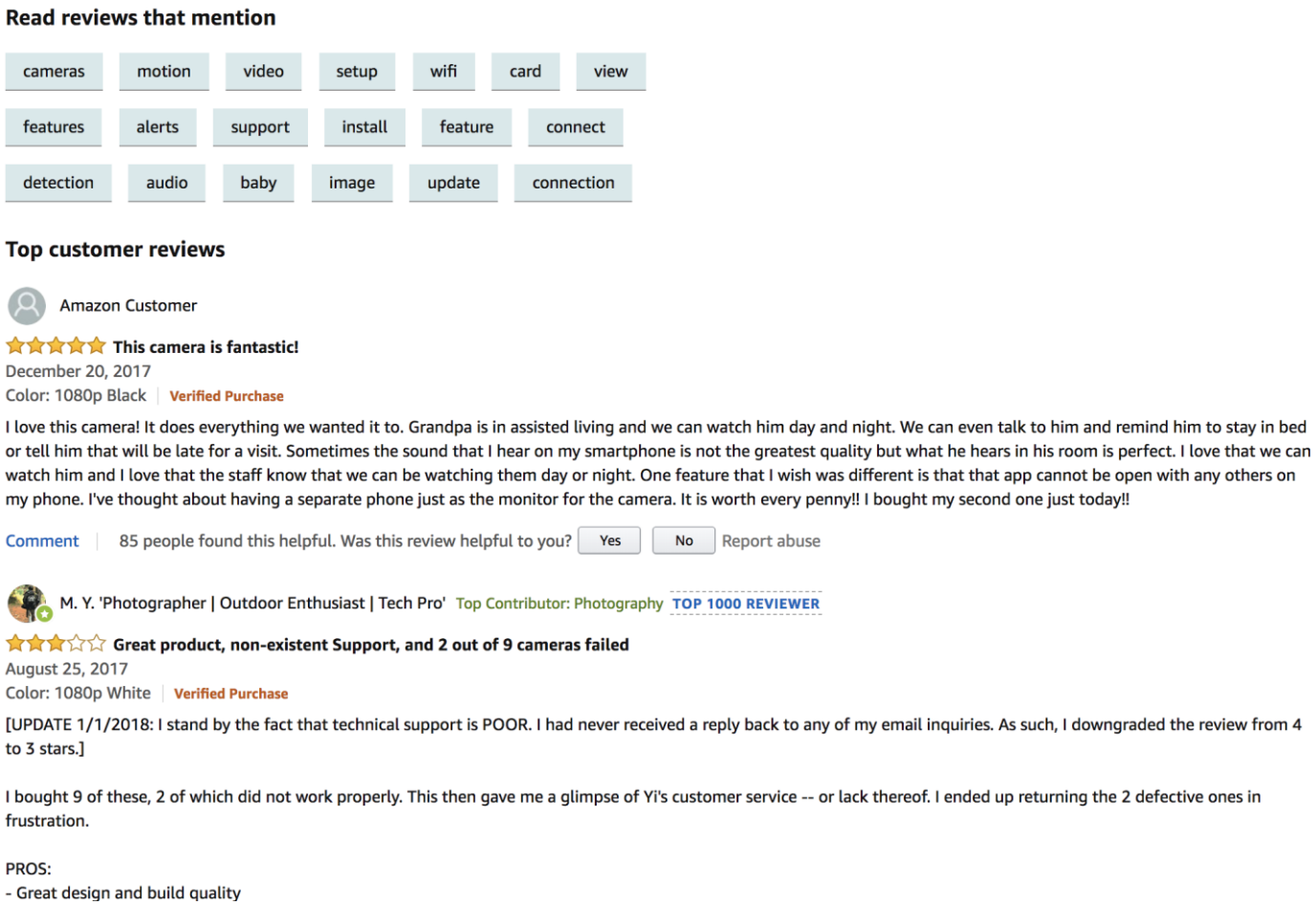
Once Set Up Works Like A Dream

September 6, 2017

Color: 1080p Black | Verified Purchase

Bought two of these cameras, one to keep an eye on my back deck and patio door and the other to keep an eye on my basement. Although it only supports the 2.4 GHz WiFi band (keep this in mind during setup -- if you set up using a 5GHz connection on your smartphone, it will fail. You need to connect your phone specifically to the 2.4GHz band on your router), the video feed is high quality at about 25KB/sec with about a three second delay. You can zoom up to 8x (digital, not optical) to magnify certain spots.

Search



Read reviews that mention

cameras | motion | video | setup | wifi | card | view

features | alerts | support | install | feature | connect

detection | audio | baby | image | update | connection

Top customer reviews

Amazon Customer

This camera is fantastic!

December 20, 2017

Color: 1080p Black | Verified Purchase

I love this camera! It does everything we wanted it to. Grandpa is in assisted living and we can watch him day and night. We can even talk to him and remind him to stay in bed or tell him that will be late for a visit. Sometimes the sound that I hear on my smartphone is not the greatest quality but what he hears in his room is perfect. I love that we can watch him and I love that the staff know that we can be watching them day or night. One feature that I wish was different is that that app cannot be open with any others on my phone. I've thought about having a separate phone just as the monitor for the camera. It is worth every penny!! I bought my second one just today!!

Comment | 85 people found this helpful. Was this review helpful to you? Yes No Report abuse

M. Y. 'Photographer | Outdoor Enthusiast | Tech Pro' Top Contributor: Photography TOP 1000 REVIEWER

Great product, non-existent Support, and 2 out of 9 cameras failed

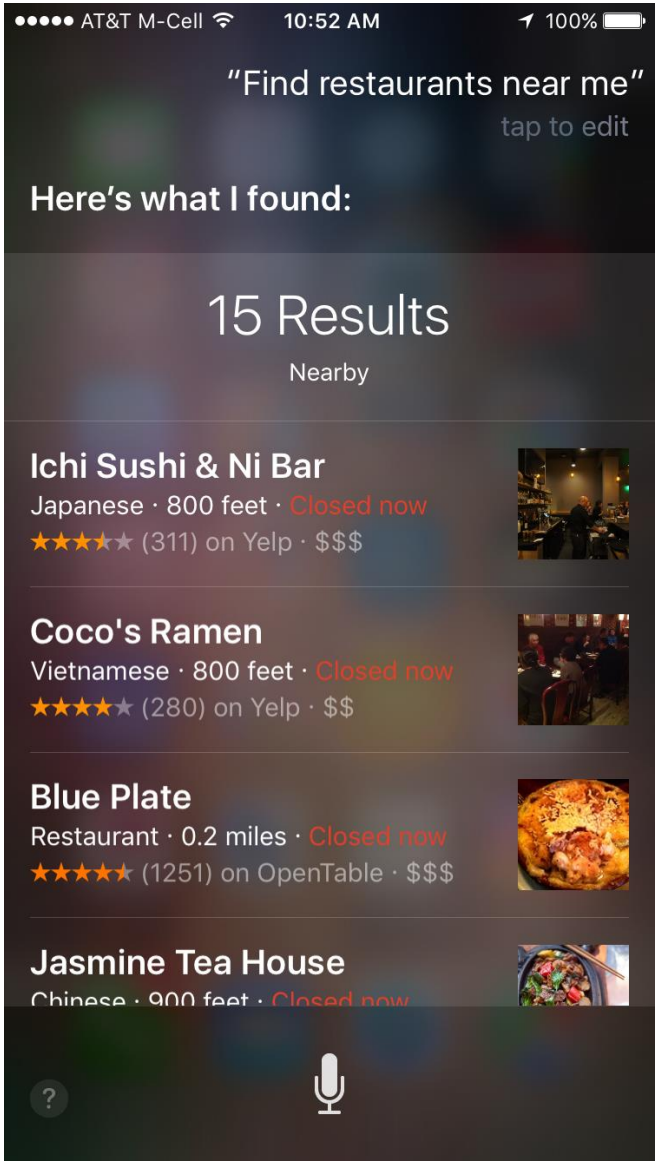
August 25, 2017

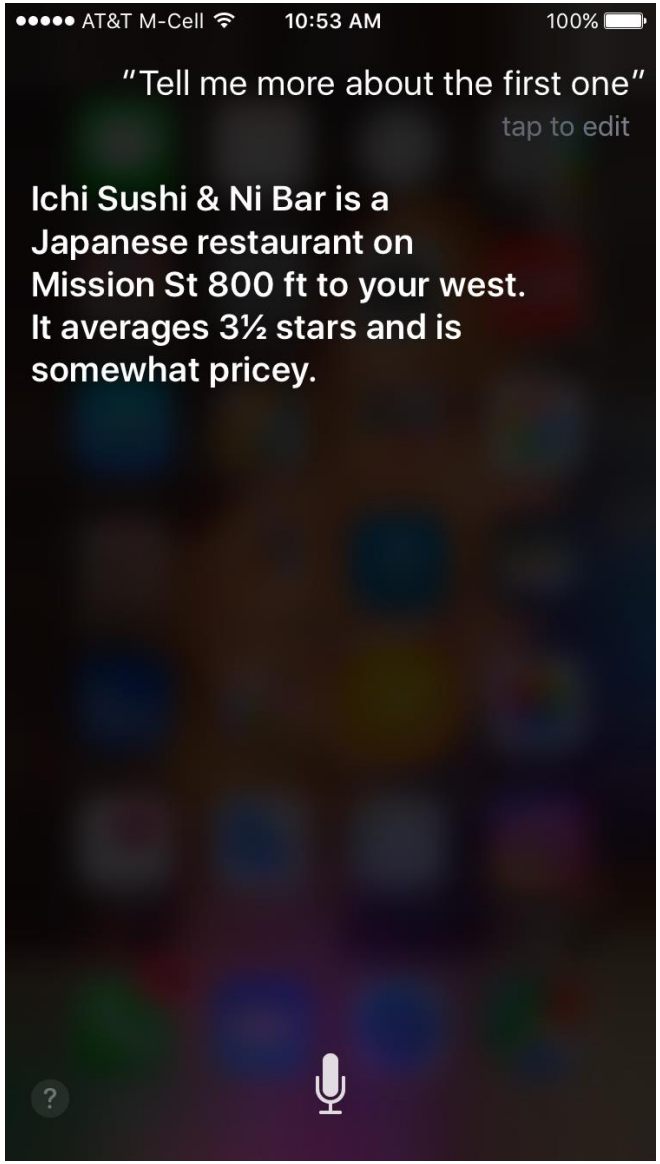
Color: 1080p White | Verified Purchase

[UPDATE 1/1/2018: I stand by the fact that technical support is POOR. I had never received a reply back to any of my email inquiries. As such, I downgraded the review from 4 to 3 stars.]

I bought 9 of these, 2 of which did not work properly. This then gave me a glimpse of Yi's customer service -- or lack thereof. I ended up returning the 2 defective ones in frustration.

PROS:
- Great design and build quality





*Slide comes from Dan Jurafaky

YI

YI Outdoor Security Camera, Cloud Cam Wireless IP Waterproof Night Vision Security Surveillance System - iOS, Android App Available

237 customer reviews | 340 answered questions

#1 Best Seller in Surveillance Camera Lenses

Compare with similar items

					
	This item YI Outdoor Security Camera, Cloud Cam Wireless IP Waterproof Night Vision Security Surveillance System - iOS, Android App Available #1 Best Seller Add to Cart	YI Dome Camera 1080p HD Pan / Tilt / Zoom Wireless IP Security Surveillance System with Night Vision, Remote Monitor with iOS, Android App - Cloud Service Available (Black) #1 Best Seller Add to Cart	YI 1080p Home Camera, Indoor Wireless IP Security Surveillance System with Night Vision for Home / Office / Baby / Pet Monitor with iOS, Android App - Cloud Service Available #1 Best Seller Add to Cart	Zmodo Wireless Security Camera System (2 pack) Smart HD Outdoor WiFi IP Cameras with Night Vision #1 Best Seller Add to Cart	(Pack of 4) Aboom Yi Home Camera Wall Mount ,Customized for YI 1080p/720p Home Camera , 360 degree swivel, Designed for USA Add to Cart
Customer Rating	(237)	(2270)	(2180)	(3221)	(16)
Price	$89.99	$40.59	$44.99	$65.00	$13.99
Shipping	FREE Shipping	FREE Shipping	FREE Shipping	FREE Shipping	FREE Shipping
Sold By	YI Technology	YI Technology	YI Technology	EZoomTek	Aboom

Product description

Search Engine Optimization

- Which keywords should you buy to promote your product/service
- Companies spend a lot of money to buy search terms on Google/Facebook/Mobile

Other applications?

- What other applications have you seen?
- Which ones are you excited about?

Challenges for NLP

BANA 290: ADVANCED DATA ANALYTICS

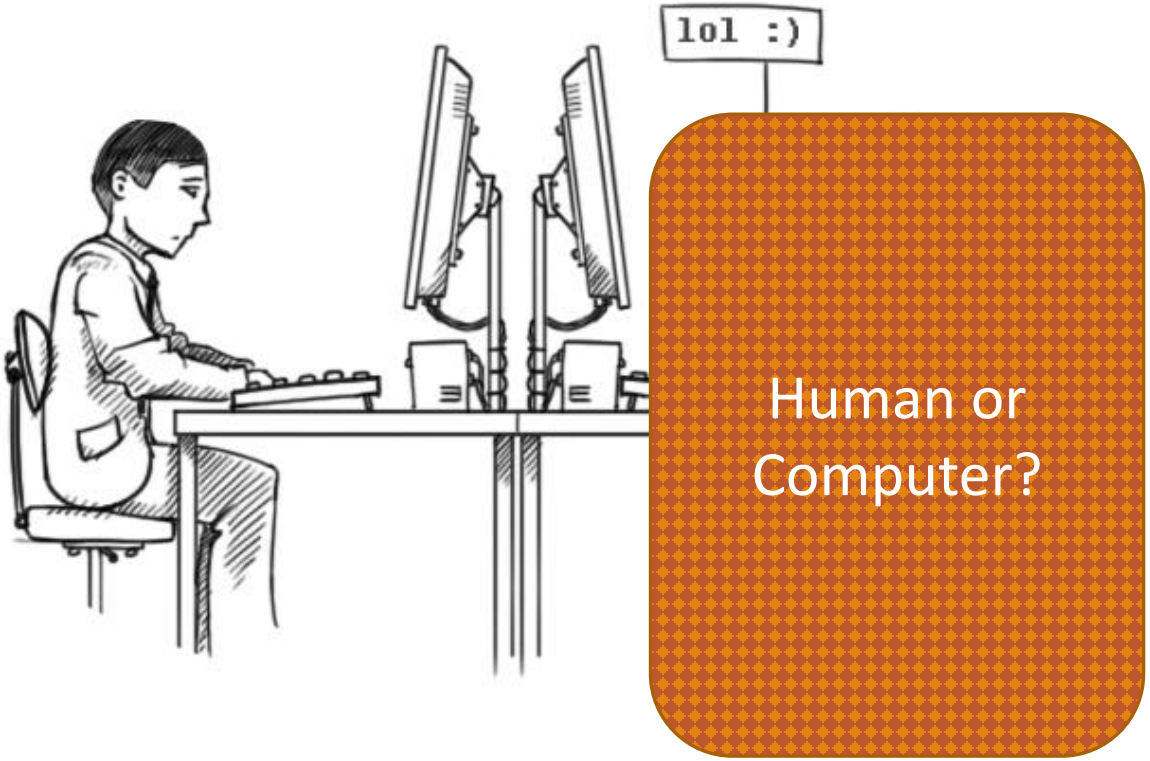
MACHINE LEARNING FOR TEXT

SPRING 2018

April 3, 2018

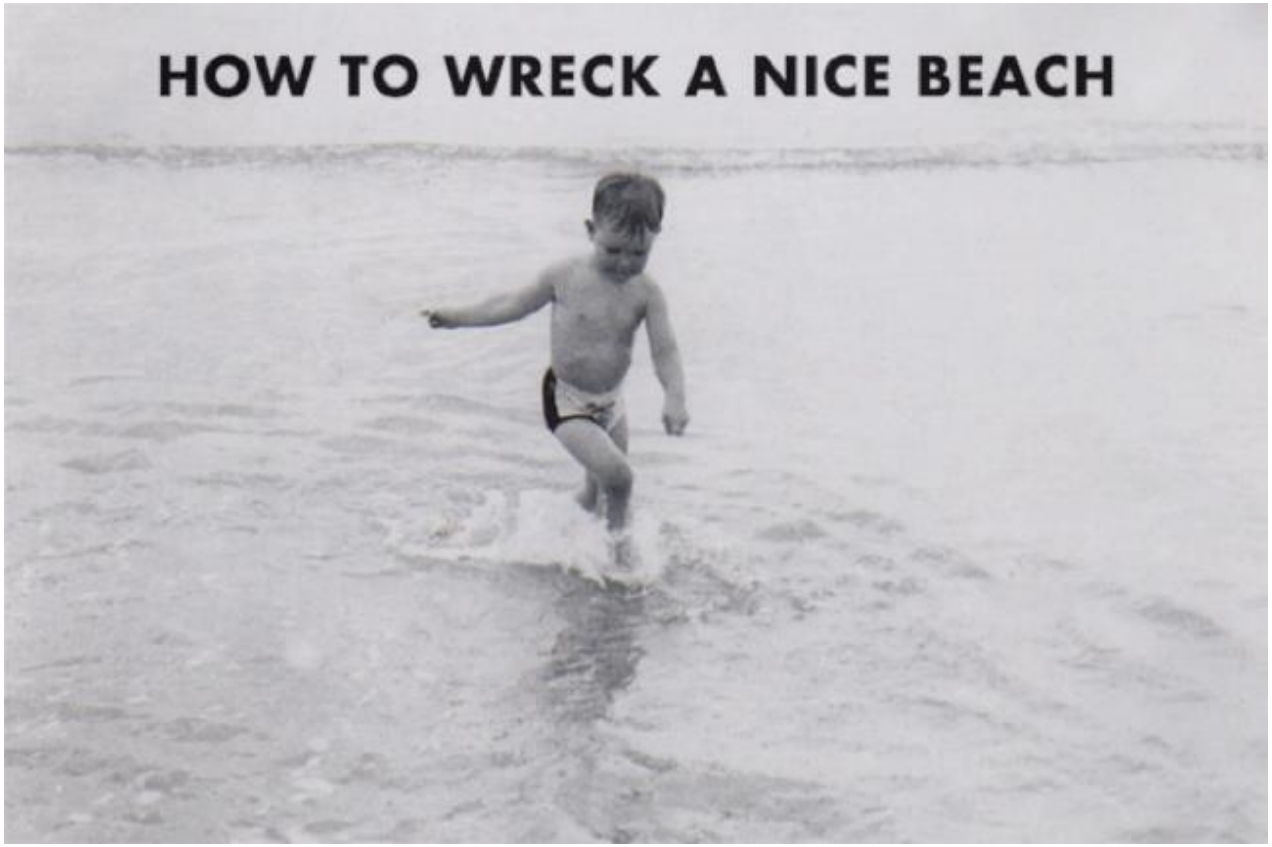
Based on slides from Nathan Schneider, Mohit Bansal, Sebastian Riedel, Yejin Choi, and everyone else they copied from.

Turing's test for Artificial Intelligence



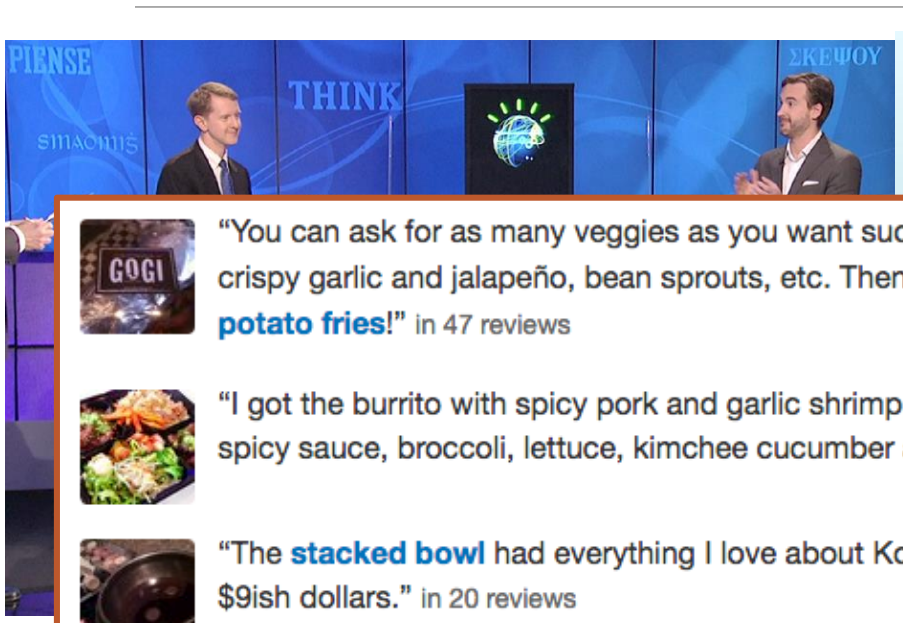
“Deep” understanding

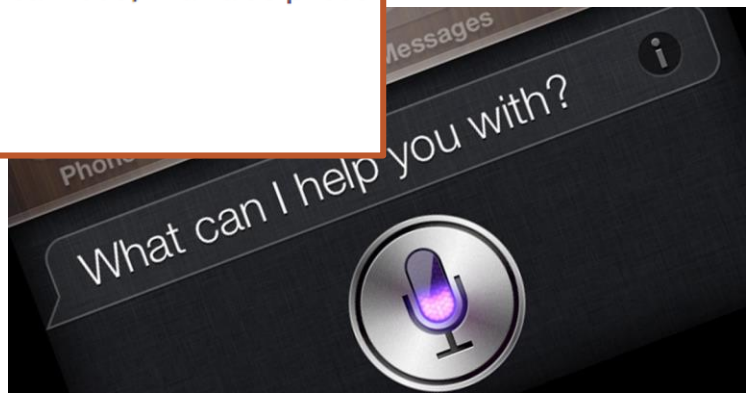
~~Speech Recognition~~



~~Cognitive Sciences/ Psycho-linguistics~~

Lots of Existing Applications





But a long long way to go…

Why isn’t NLP solved yet?

Three main challenges

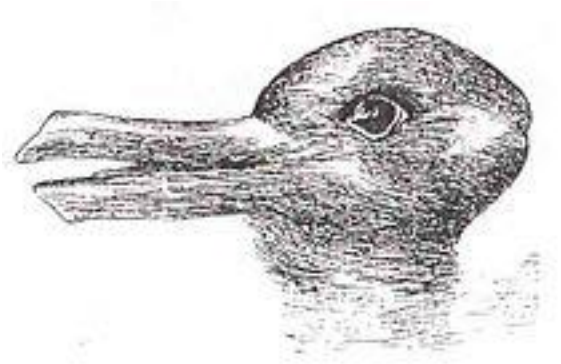
- Ambiguity
- Sparsity
- Variation

Three main challenges

Ambiguity

Sparsity

Variation

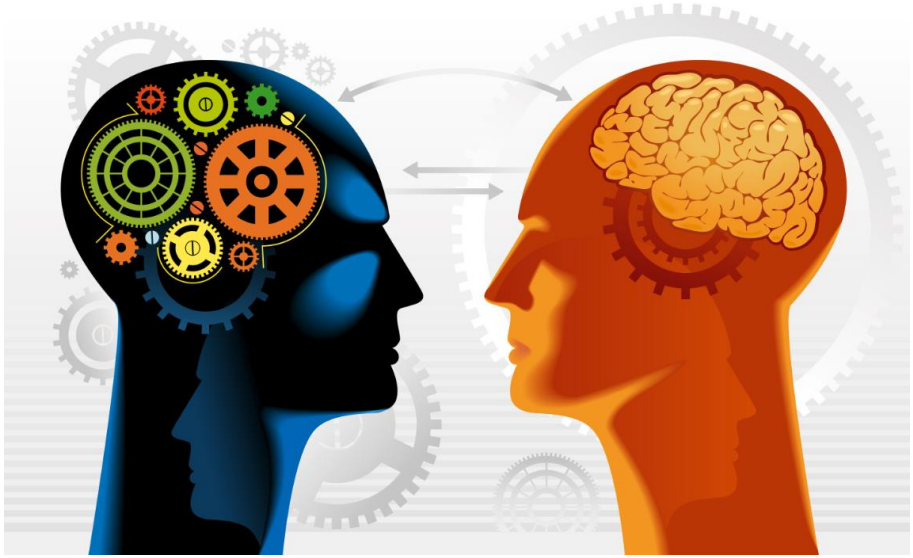


Language is Ambiguous

One tries to be as informative as one possibly can,
and gives as much information as is needed, **and no more.**

- Grice's Maxim of Quantity

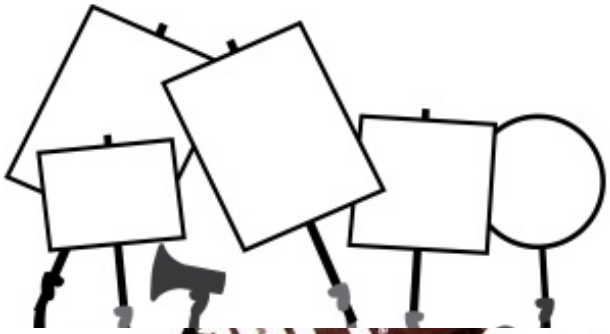
Corollary: The more you know, the less you need.



Computers "know" very little.

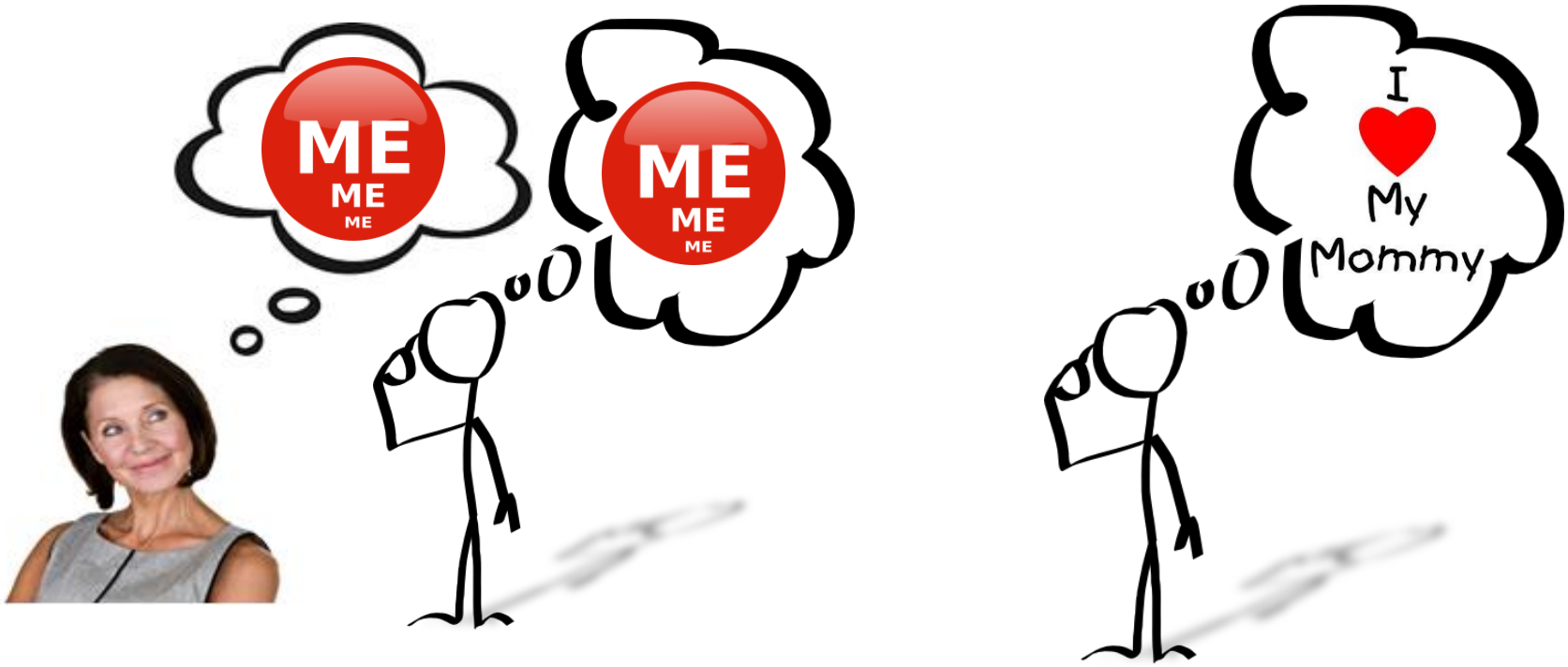
Words have many meanings

Hershey's Bars Protest



Words have many meanings

He knows you like your mother.



Attachment Ambiguities

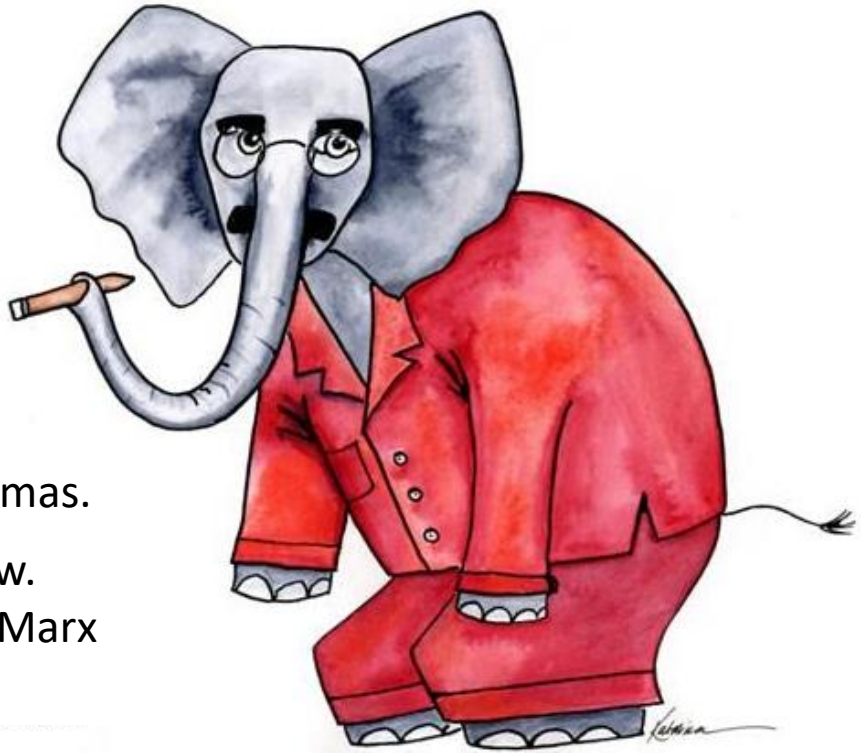
Stolen painting found by tree.

Attachment Ambiguities

One morning I shot an elephant in my pajamas.

How he got into my pajamas I'll never know.

- Groucho Marx



Attachment Ambiguities

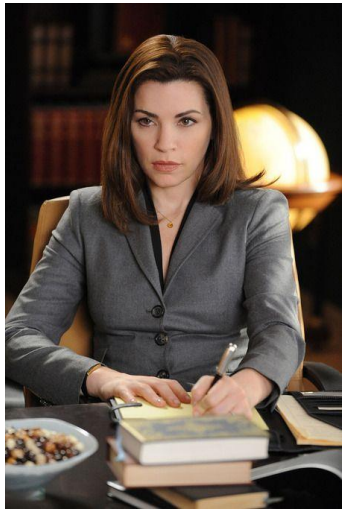
She saw the man with the telescope.

And so on…

- Enraged Cow Injures Farmer with Ax
- Ban on Nude Dancing on Governor's Desk
- Teacher Strikes Idle Kids
- Hospitals Are Sued by 7 Foot Doctors
- Iraqi Head Seeks Arms
- Kids Make Nutritious Snacks
- Local HS Dropouts Cut in Half

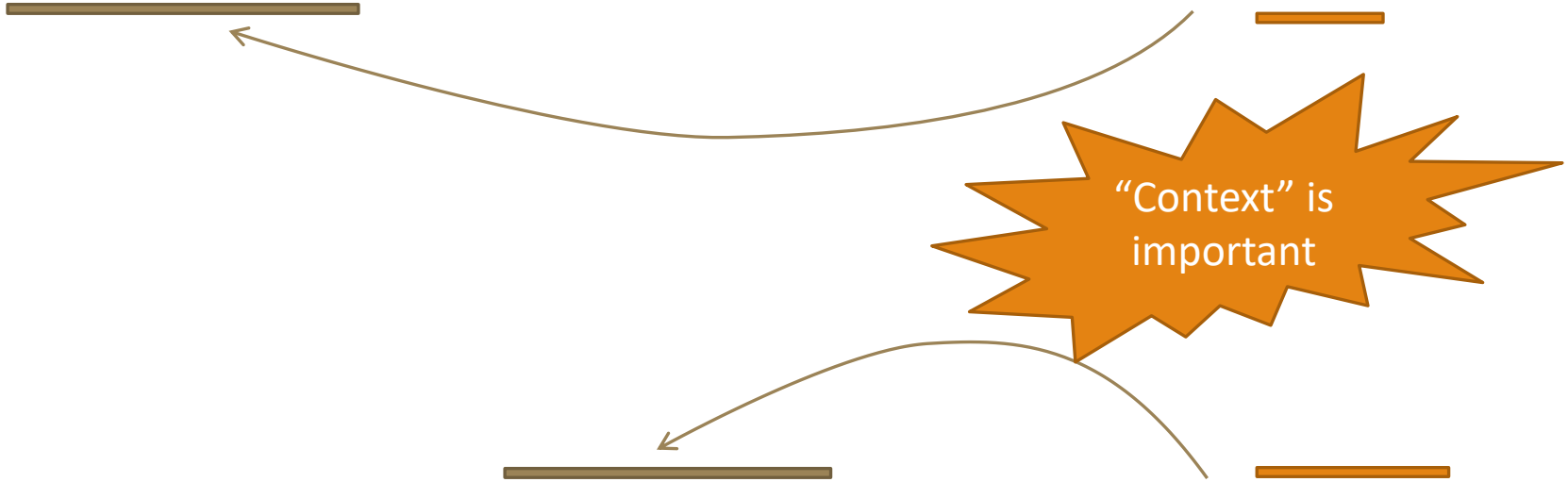
Coreference Ambiguities

My girlfriend and I met my lawyer for a drink, but she became ill and had to leave.



Coreference Ambiguities

The city councilmen refused the demonstrators a permit because they feared violence.



The city councilmen refused the demonstrators a permit because they advocated violence.

Winograd Schema: An Open Challenge for AI

Coreference Ambiguities

Entity Types and Identities

Types	Identities
• Washington, Georgia, Clinton, Adams • John Deere, Williams, Dow Jones, Thomas Cook • Princeton, Amazon, Kingston	• Same Name: Kevin Smith, Jamaica, Springfield • Multiple "Names": President, Obama, Chief, Bambam,…

Animals with Misleading Names

Electric Eel

Not an eel.

Mountain Goat

Not a goat.

Maned Wolf

Not a wolf.

King Cobra

Not a cobra. Also, snakes are typically self-governing.

Peacock Mantis Shrimp

Not a peacock.
Not a mantis.
Also, not a shrimp.

Horny Toad

Not a toad.
Only thinks of you as a friend.

Mayfly

Active through the spring and summer.

Eastern Kingbird

Found in the West.
Many birds do not recognise its authority.

birdandmoon.com

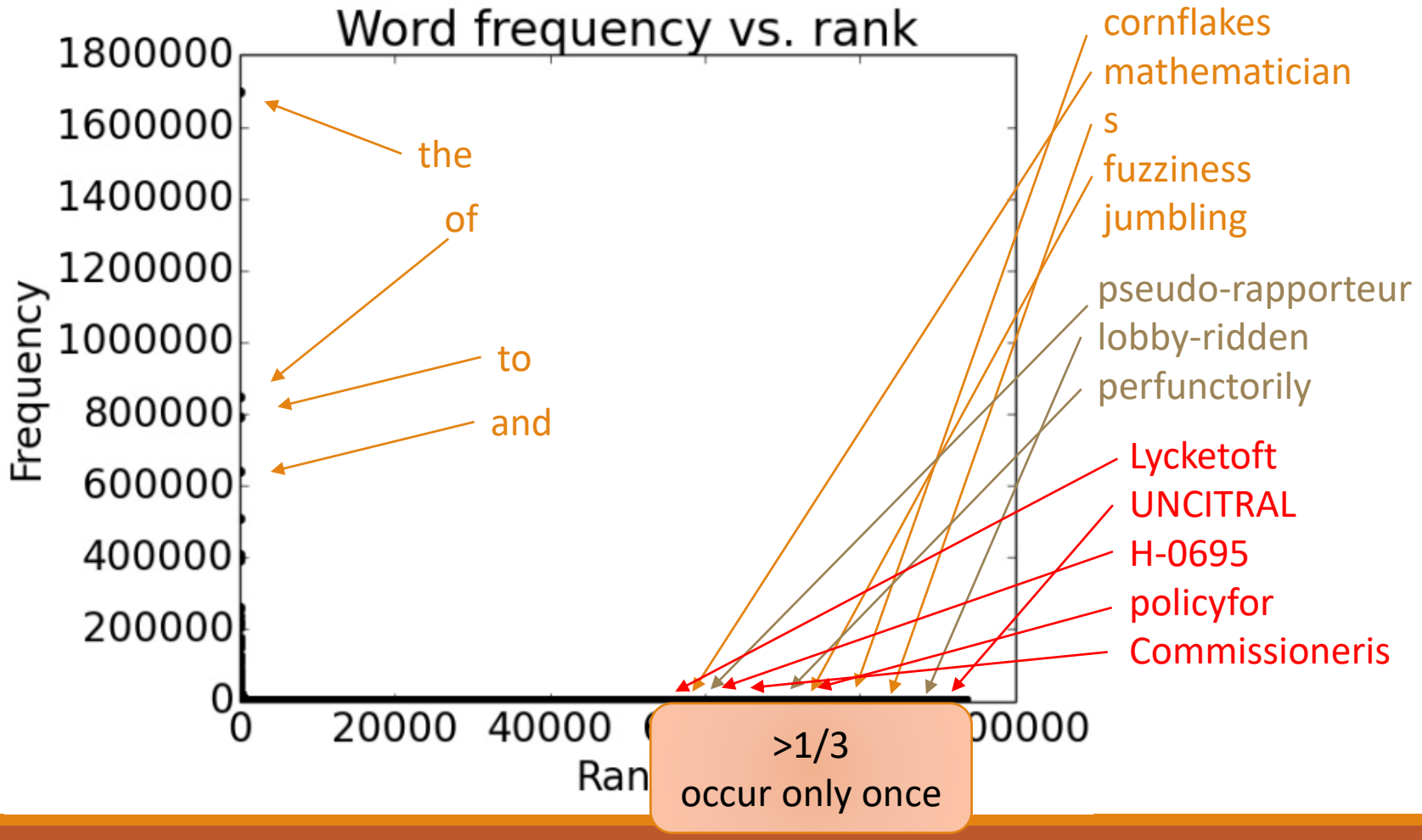
Three main challenges

Ambiguity

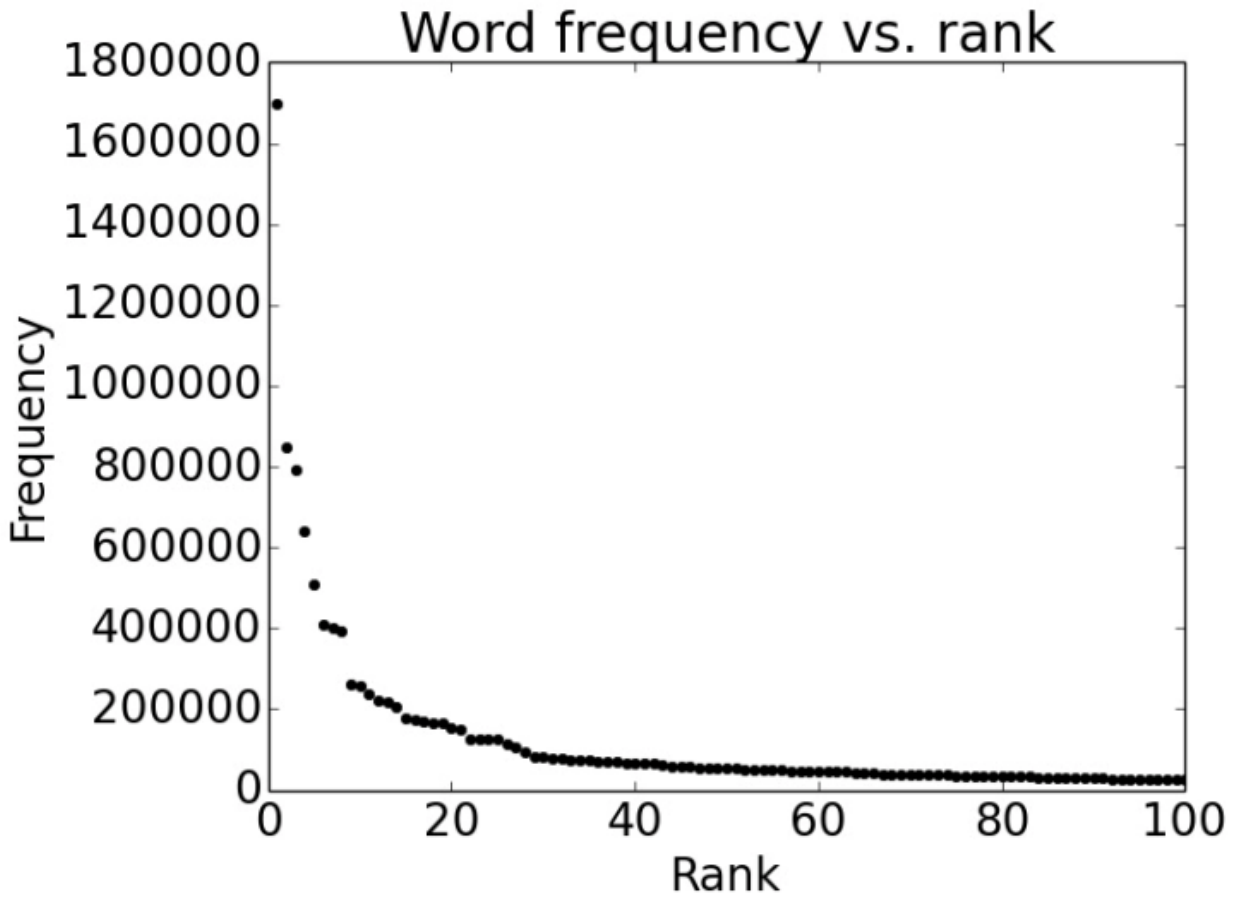
Sparsity

Variation

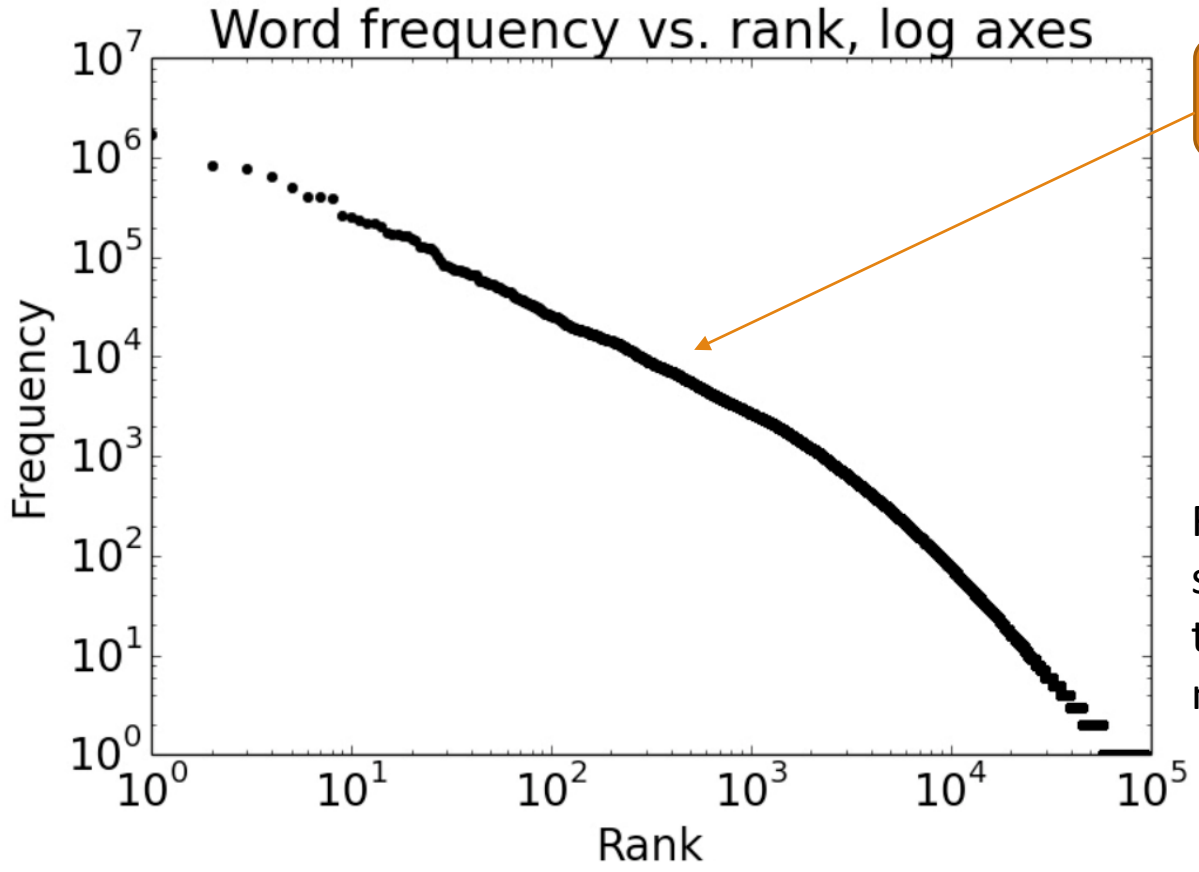
Sparsity of Words



Sparsity of Words



Rescaling the Axes



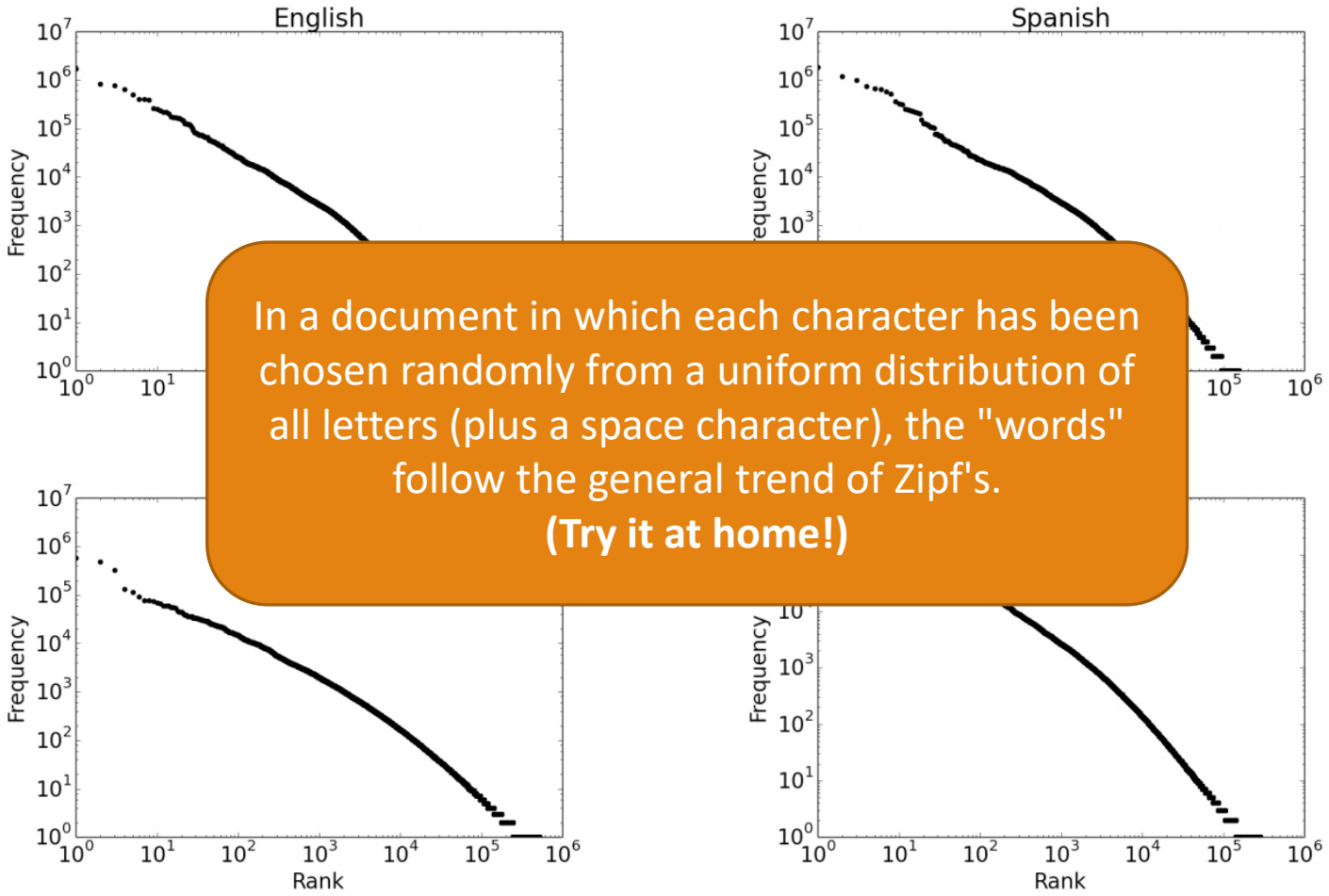
Zipf's Law

$f \times r = k$

$\log f + \log r = \log k$

Regardless of the size of the data, there will be many rare words.

Not unique to English



In a document in which each character has been chosen randomly from a uniform distribution of all letters (plus a space character), the "words" follow the general trend of Zipf's.
(Try it at home!)

Three main challenges

Ambiguity

Sparsity

Variation

Many ways to say something

She gave the book to Tom vs. She gave Tom the book
Some kids popped by vs. A few children visited
Is that window still open? vs Please close the window

Variations in Domains

Its vanished trees, the trees that had made way for Gatsby's house, had once pandered in whispers to the last and greatest of all human dreams; for a transitory enchanted moment man must have held his breath in the presence of this continent, compelled into an aesthetic contemplation he neither understood nor desired, face to face for the last time in history with something commensurate to his capacity for wonder.

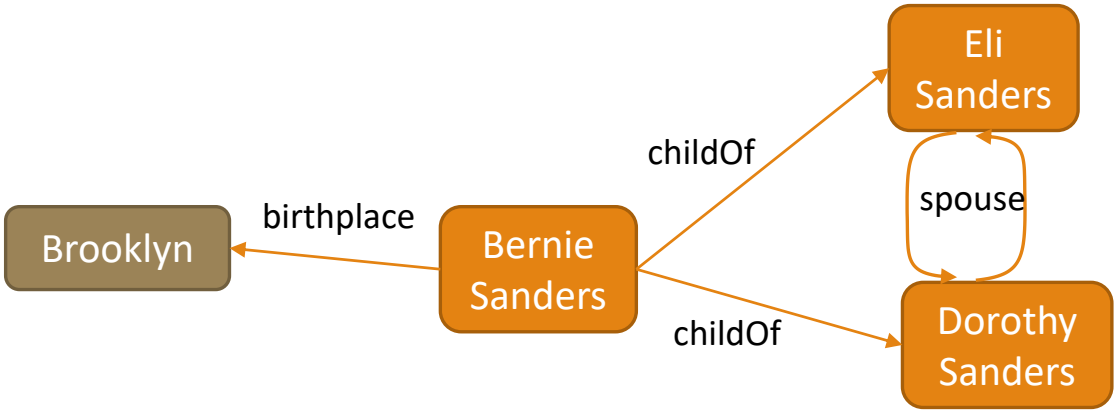
ikr smh he asked fir yo last name so he can add u on fb lolololtw

Tools & Methods

HOW CAN WE GET COMPUTERS TO SOLVE THIS PROBLEM?

Corpus

Entity resolution,
Entity linking,
Relation extraction…



Document

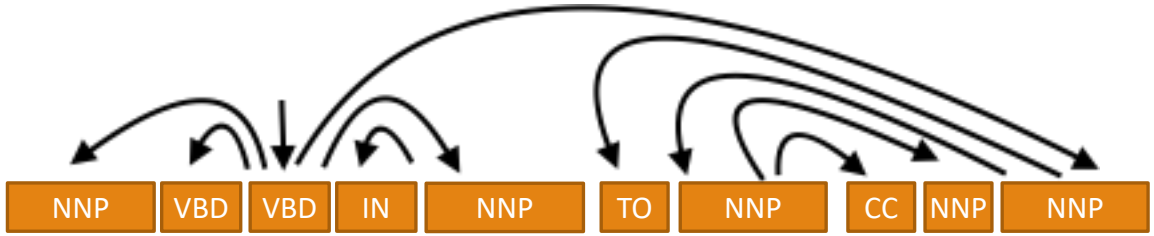
Discourse analysis,
Coreference,
Sentiment analysis...

Bernie..
Bernie Sanders...
Mrs. Sanders..
the city
.. his mother ..
his father
he
Eli

Person — Sanders
was born in
Location — Brooklyn
, to
Person — Dorothy
and
Person — Eli Sanders.

Sentence

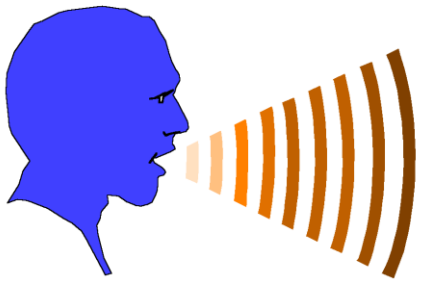
Dependency Parsing,
Part of speech tagging,
Named entity recognition…



Sanders was born in Brooklyn, to Dorothy and Eli Sanders.

Two Different Approaches

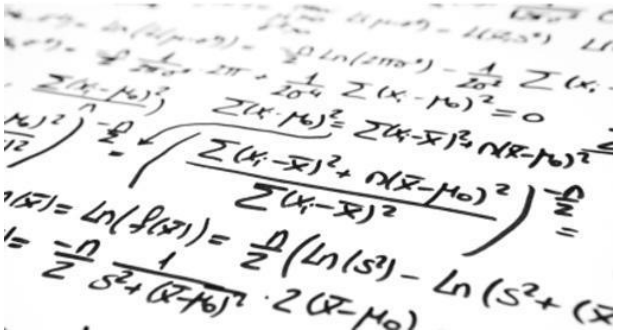
DIRECTLY USE LINGUISTICS



Expensive, time-consuming...

… but also, incomplete!

MACHINE LEARNING!



Automatically learn from data!

… if the right data exists

"Every time I fire a linguist, my accuracy goes up."
- Frederick Jelinek

Example: Machine Translation

From https://medium.com/@ageitgey/machine-learning-is-fun-part-5-language-translation-with-deep-learning-and-the-magic-of-sequences-2ace0acca0aa

Example: Machine Translation

Quiero ir a la playa más bonita.

Step 1: Break into Chunks

Example: Machine Translation

Step 2: Translations for each chunk

Example: Machine Translation

Step 3: Generate all possible sequences

Quiero ir a la playa más bonita.

In same order

I love | to leave | at | the seaside | more tidy.
I mean | to be on | to | the open space | most lovely.
I like | to be |on | per the seaside | more lovely.
I mean | to go | to | the open space | most tidy.

In different order

I try | to run | at | the prettiest | open space.
I want | to run | per | the more tidy | open space.
I mean | to forget | at | the tidiest | beach.
I try | to go | per | the more tidy | seaside.

Step 4: Find the most human sounding one

I try | to leave | per | the most lovely | open space.

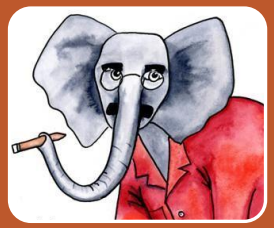
I want | to go | to | the prettiest | beach.

I want to go to the prettiest beach.

In summary…

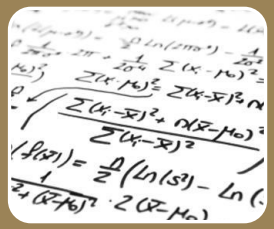
Language to Knowledge

- Lots of applications…
- Made a lot of progress, but not done



It's quite difficult

- Varied, sparse, and lots of ambiguities
- Context really matters



Machine Learning!

- With enough data and math, we can do it
- The future looks really exciting for NLP

Regular Expressions!

RULE-BASED APPROACH FOR DEALING WITH TEXT


WHENEVER I LEARN A NEW SKILL I CONCOCT ELABORATE FANTASY SCENARIOS WHERE IT LETS ME SAVE THE DAY.
OH NO! THE KILLER MUST HAVE FOLLOWED HER ON VACATION!
BUT TO FIND THEM WE'D HAVE TO SEARCH THROUGH 200 MB OF EMAILS LOOKING FOR SOMETHING FORMATTED LIKE AN ADDRESS!
IT'S HOPELESS!
EVERYBODY STAND BACK.
I KNOW REGULAR EXPRESSIONS.
PERL!
TAP TAP

What are regular expressions?

- Rules/patterns that can be applied to text
- It can be used for matching:
  - Is this text a phone number or not?
- It can also be used for search:
  - Give me all the phone numbers in this text document
- Why have rule-based solutions when we have machine learning?

Disjunctions and Ranges

- “or” operations
  - E.g. r'cat|dog'
    - Matches “cat” and matches “dog”
- Can also give a range
  - E.g.
    - r'[0-9]' matches any digit
    - r'[A-Z]' matches any capital letter

Optionals, Wildcards, Repeats

- Optionals
  - E.g. r'cats?' match ‘cat’ and ‘cats’
  - E.g. r'fox(es)?' match ‘fox’ and ‘foxes’
- Wildcards
  - E.g. r'c.t' matches ‘cut’ and ‘cat’
- * is for zero or more occurrences
  - E.g. r'oo*h!' matches “oh!”, “ooh!”, “oooh!”, etc
- + is for 1 or more occurrences
  - E.g. r'o+h!' matches “oh!”, “ooh!”, “oooh!”, etc
- {} for expressing count
  - E.g. r’[0-9]{3}’ matches “123”, “345”, “789”, etc.
  - E.g. r’o{2,4}h!’ matches “ooh!”, “oooh!”, “ooooh!”, but not “oh!” Or “ooooooooh!”

In-class exercise

WRITE AND TEST YOUR OWN REGULAR EXPRESSIONS

More tools for regular expressions

- Anchors ‘^’ and ‘$’ to signify the beginning and end of the text
  - E.g. r'^the' matches ‘the’ but not ‘other’
  - E.g. r'the$' matches ‘the’ but not ‘theology’
- Negative disjunction
  - E.g. r'the[^m]' matches “then” but not “them”
  - E.g. r’the[^a-z]’ matches “the” but not “them” or “then"

In-class exercise

WRITE AND TEST YOUR OWN REGULAR EXPRESSIONS

Tokenization

- Process of breaking text into “tokens” or “words”
- Why?
  - Words are the basic semantic units
  - By breaking a document into words, you can do search, categorizing documents, etc.
  - You can compare the similarity of two documents even if they’re not identical

Tokenization

- Process of breaking text into “tokens” or “words”
- What regex could you use to do this?

Tokenization

- Process of breaking text into “tokens” or “words”
- What regex could you use to do this?
  - What about r’[A-Za-z]+’ ?

Tokenization

- Process of breaking text into “tokens” or “words”
- What regex could you use to do this?
  - What about r’[A-Za-z]+’ ?
    - Misclassifications:
      - ‘Ph.D’ would get the tokens “Ph” and “D”
      - ‘can’t’ would get the tokens ‘can’ and ’t'

Issues with tokenizing other languages?

- Thoughts?

Word normalization

- Need to normalize terms
  - Otherwise you will miss documents when you do a search engine (as well as in most NLP tasks)
  - E.g. “U.S.A.” and “USA”
  - E.g. “the” and “The”
    - Are there any problems though if we ignore case?

Word normalization

- Need to normalize terms
  - Otherwise you will miss documents when you do a search engine (as well as in most NLP tasks)
  - E.g. “U.S.A.” and “USA”
  - E.g. “the” and “The”
    - Are there any problems though if we ignore case?
      - US vs. us
      - Fed vs. fed
      - Windows vs windows

Other normalization

- Lemmatization
  - Cat vs. cats
  - Fox vs. foxes
  - Excite vs excited
- Stemming is one approach
  - Rule-based approach that just chops off the suffix

Sentence segmentation

- Sentences are the next building blocks
- How do we break into sentences?
  - Use punctuation?

Edit distance

- We can compare the distances between two texts
- Intuitively, “cat” and “cats” should be less distant than “cat” and “dog”
- Applications?

Edit distance

- We can compare the distances between two texts
- Intuitively, “cat” and “cats” should be less distant than “cat” and “dog”
- Applications?
  - Spell corrections
  - Computational biology (aligning different DNA sequences together)
  - Machine translation, information extraction, speech recognition, etc.

Edit distance

- Minimum edit distance between two strings
- Minimum number of editing operations to transform one string to the other
  - Insertion
  - Deletion
  - Substitution

Edit distance

- Minimum edit distance between two strings
- Minimum number of editing operations to transform one string to the other
  - Insertion
  - Deletion
  - Substitution