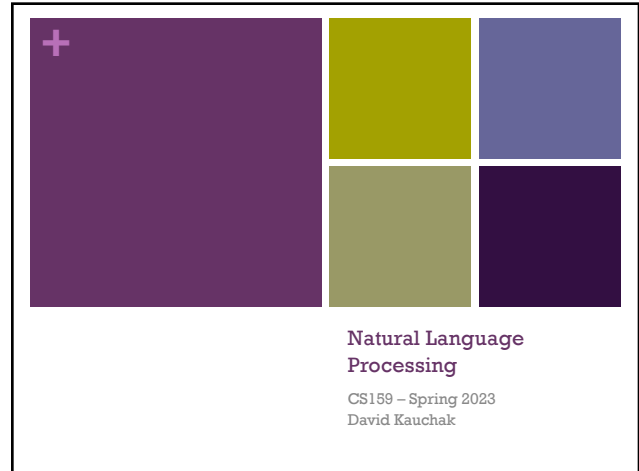


1



2

Administrative stuff

<http://www.cs.pomona.edu/classes/cs159/>

- Office hours, schedule, assigned readings, assignments
- Everything will be posted there

Read the administrative section of the webpage

- ~7 assignments (in a variety of languages)
- 4 quizzes
- final project for the last 3-4 weeks
 - teams of 2-3 people
- Readings

Academic Honesty Collaboration

3

Administrative stuff

Assignment 0 posted already

- Shouldn't take too long
- Due Friday by 5pm

Assignment 1 posted

- Won't cover all material until next Monday
- Due Monday 1/30 at 1pm

4

+ What to expect...

This course will be challenging for many of you

- assignments will be non-trivial
- content can be challenging

But it is a fun field!

We'll cover

- basic linguistics
- probability
- the common problems
- many techniques and algorithms
- common machine learning techniques
- some recent advances in neural networks for language processing
- NLP applications

5

+ Requirements and goals

Requirements

- Competent programmer
 - Some assignments in Java, but I will allow/encourage other languages after the first few assignments
- Comfortable with mathematical thinking
 - We'll use a fair amount of probability, which I will review
 - Other basic concepts, like logs, summation, etc.
- Data structures
 - trees, hashables, etc.

Goals

- Learn the problems and techniques of NLP
- Build real NLP tools
- Understand what the current research problems are in the field

6

+ Getting to know each other

What's your name?

What year?

What has been your favorite CS class?

What's been your least favorite CS class?

Why are you taking this class?

7

+ Videos before class

8

+ What is NLP?

Natural language processing (NLP) is a field of computer science and linguistics concerned with the interactions between computers and human (natural) languages.

- Wikipedia

9

+ What is NLP?

The goal of this new field is to get computers to perform useful tasks involving human language...

- The book

10

+ Where have you seen NLP used?

11

+ Key: Natural text

ALL NATURAL

“A growing number of businesses are making Facebook an indispensable part of hanging out their shingles. Small businesses are using ...”

Natural text is written by people, generally for people

Why do we even care about natural text in computer science?


12

+ Why do we need computers for dealing with natural text?

Google's search knows about over 130 trillion pages

In less than four years, Google's search knowledge of pages have grown by more than 100 trillion new pages.

Barry Schwartz on November 14, 2016 at 9:17 am




Google has updated the [How Search Works](#) page to say they now have knowledge of over 130 trillion pages across the web. The page reads, "Search starts with the web, it's made up of over 130 trillion individual pages and it's constantly growing."

<https://searchengine.land.com/googles-search-indexes-hits-130-trillion-pages-documents-263378>

13


+ Web is just the start...

e-mail




~300 billion e-mails a day

corporate databases




twitter



~500 million tweets a day

Blogs: **~200 million** different blogs



Facebook

14

+ Why is NLP hard?

- Iraqi Head Seeks Arms
- Juvenile Court to Try Shooting Defendant
- Stolen Painting Found by Tree
- Kids Make Nutritious Snacks
- Local HS Dropouts Cut in Half
- Obesity Study Looks for Larger Test Group
- British Left Waffles on Falkland Islands
- Red Tape Holds Up New Bridges
- Hospitals Are Sued by 7 Foot Doctors

15

+ Why is NLP hard?

User: Where is Escape Room playing in the Claremont Area?

System: Escape Room is playing at the Edwards in La Verne.

User: When is **it** playing **there**?

System: It's playing at 2pm, 5pm and 8pm

User: I'd like 1 adult and 2 children for **the first show**. How much would **that** cost?

16

+ Why is NLP hard?

Natural language:

- is highly ambiguous at many different levels
- is complex and contains subtle use of context to convey meaning
- is probabilistic?
- involves reasoning about the world
- is highly social
- is a key part in how people interact

However, some NLP problems can be surprisingly easy

17

+ Different levels of NLP

pragmatics/discourse: how does the context affect the interpretation?

semantics: what does it mean?

syntax: phrases, how do words interact

words: morphology, classes of words

18

+ NLP problems and applications

What are some places where you have seen NLP used?

What are NLP problems?

19

+ NLP problems and applications

Lots of problems of varying difficulty

Easier

Word segmentation: where are the words?

**I would've liked Dr. Dave to finish early.
But he didn't.**

20

+ NLP problems and applications

Lots of problems of varying difficulty

Easier

Word segmentation: where are the words?

再往远些看，随着汉字识别和语音识别技术的发展，中文计算机用户将跨越语言差异的鸿沟，在录入上走向中西文求同的道路。

再往远些看，随着汉字识别和语音识别技术的发展，中文计算机用户将跨越语言差异的鸿沟，在录入上走向中西文求同的道路。

21

+ NLP problems and applications

Lots of problems of varying difficulty

Easier

- Speech segmentation
- Sentence splitting (aka sentence breaking, sentence boundary disambiguation)

I would've liked Dr. Dave to finish early. But he didn't.
- Language identification

Soy un maestro con queso.

22

+ NLP problems and applications

Easier continued

- truecasing

i would've liked dr. dave to finish early, but he didn't.
- spell checking

Identifying misspellings is challenging especially in the dessert.
- OCR


4



23

+ NLP problems and applications

Moderately difficult

- morphological analysis/stemming

smarter
smarter
smartly
smartest
smart → smart
- speech recognition
 
- text classification

SPAM  sentiment analysis 

24

+ NLP problems and applications

Moderately difficult continued

- text segmentation: break up text by topics
- part of speech tagging (and inducing word classes)
- parsing

```

    graph TD
      S --> NP1[NP]
      S --> VP[VP]
      NP1 --> PRP[PRP]
      PRP --> I[I]
      VP --> V[V]
      V --> eat[eat]
      VP --> NP2[NP]
      NP2 --> N1[N]
      N1 --> sushi[sushi]
      NP2 --> PP[PP]
      PP --> IN[IN]
      IN --> with[with]
      PP --> N2[N]
      N2 --> tuna[tuna]
  
```

I eat sushi with tuna

25

+ NLP problems and applications

Moderately difficult continued

- word sense disambiguation

As he walked along the side of the stream, he spotted some money by the bank. The money had gotten muddy from being so close to the water.

- grammar correction

We am good at grammar.

- speech synthesis

26

+ NLP problems and applications

Hard (many of these contain many smaller problems)

Machine translation

美国关岛国际机场及其办公室均接获一名自称沙地阿拉伯富商拉登等发出的电子邮件，威胁将会向机场等公众地方发动生化袭击，关岛保持高度戒备。

→

The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport.

27

+ NLP problems and applications

Information extraction

IBM hired Fred Smith as president.

person	company	position
Fred Smith	IBM	president

28

+ NLP problems and applications

Summarization

A company that acts as a middle man between content companies and Internet service providers is accusing Comcast Corp., the nation's largest broadband provider, of anti-competitive behavior. (article 8) Comcast Corp. and NBC Universal made new promises to the Federal Communications Commission that the companies hope will help get the regulatory agency to approve the proposed deal between the media giants. (article 6) At issue is the cable operator's decision to offer the Tennis Channel on a specialty tier of sports networks as opposed to its widely distributed basic tier. (article 9) The quality of television news could deteriorate further under a Comcast-controlled NBC Universal, the Writers Guild of America East warned Wednesday in letters to key Washington officials overseeing the government's review of the proposed merger. (article 5) With regulatory approval still weeks if not months away, Comcast and NBC Universal have extended the term of their merger agreement to March of next year. (article 4) Democrat Michael Capps fears the joint venture would put too much control of content into the hands of a company that also controls how consumers access the Internet and television. (article 3) Susan Fox talked on Wednesday with two senior staff members of FCC Commissioner Meredith Attwell Baker (article 1)

29

+ NLP problems and applications

Natural language understanding

- Text => semantic representation (e.g. logic, probabilistic relationships)

Information retrieval and question answering

"How many programmers in the child care department make over \$50,000?"

"Who was the fourteenth president?"

"How did he die?"

30

+ NLP problems and applications

Text simplification

Alfonso Perez Munoz, usually referred to as Alfonso, is a former Spanish footballer, in the striker position.

↓

Alfonso Perez is a former Spanish football player.

31

+ Where are we now?

Many of the "easy" and "medium" problems have reasonable solutions

- spell checkers
- sentence splitters
- word segmenters/tokenizers

natral language →

Print Layout View Sec 1 Pages: 1 of 150 Words: 0 of 56282

32

+ Where are we now?

Parsing

Stanford Parser (<http://nlp.stanford.edu:8080/parser/>)

Stanford Parser

Please enter a sentence to be parsed:
My dog also likes eating bananas.

Language: English Sample Sentence Parse

Your query
My dog also likes eating bananas.

Tagging
My/DT dog/NN also/ADV likes/VB eating/VBG bananas/NN ././.

Parse

```

(ROOT
  (S
    (NP (NNP My) (NN dog))
    (VP (VBZ likes)
      (NP (VBN eating) (NN bananas))
      (PUNCT .)
      (PUNCT .)))
  )
  )
  
```

33

+ Where are we now?

Machine translation

How is it?

Soy un maestro con queso.

I am a teacher with cheese.

34

+ Where are we now?

Machine translation

- Getting better every year
- enough to get the jist of most content, but still no where near a human translation
- better for some types of text

<http://translate.google.com>

Many commercial versions...

- systran
- language weaver

35

+ Where are we now?

Information extraction

- Structured documents (very good!)
 - www.dealtime.com
 - www.google.com/shopping
 - AKT technologies
- Lots of these
 - FlipDog
 - WhizBang! Labs
 - ...
 - work fairly well

Text Annotations Annotation Sets Print

The departure of **Mr. Hogan**, who originally moved to **San Francisco** as service director from **United States** in **1992**, surprised aviation analysts, as it was believed that he had been brought into the senior executive team of the airline, as part of the group's management succession planning.

He played a leading role in the strategic planning by the rebranding of the airline as **USA** in preparation for its entry **into** the crowded long haul market with the launch of services from **Los Angeles** to the **US**.

USA has taken on the costs of entry into the **North Atlantic** market of an unfortunate time, as airlines in **North America** and facing the toughest conditions for 20 years with many carriers plunging into loss.

USA, in which **United** of **Germany** and **USA Scandinavian Airlines** each own stakes of **20%**, **USA** offered a **\$2.5 billion** bid in pre-bid talks. **USA** from **\$11.5 billion** **\$16.7 billion** to **\$2.5 billion** on a business that grew **\$100 million** to **\$250 million**.

In the **last months** **USA** is understood that passenger volumes have fallen by around **50 percent**. The share of available seats filled, the load factor, has declined by around two percentage points, but this has been offset by a strong increase in yields, or average fare levels, by more than **50 percent**.

USA move to tighten its management structure follows a warning on **USA** from **USA** that the industry faces challenges in months with a "more difficult" **USA** ahead.

Key annotations

- Location
- Money
- Person

Original markup annotations

- Location
- Money
- Person

36

+ Where are we now?

CMU's NELL (Never Ending Language Learner)

<http://rtw.ml.cmu.edu/rtw/>

Recently-Learned Facts buolker

Instance

red_harvester_ant is an invertebrate

crockpot_mushroom_chicken is a type of meat

football is a hobby

colgate_palmolive is a magazine

kirkwood is a city

ken_takahashi plays the sport baseball

andy_warhol is a visual artist in the field of printmaking

john_hayward held the position of vice_admiral

tom_osborne works for nebraska

in is a company headquartered in the city nashville

37

+ Where are we now?

Why do people do this?

: David.last_name@pomona.edu

38

+ Where are we now?

Information retrieval/query answering

How are search engines?

What are/aren't the good at?

How do they work?

39

+ Where are we now?

Information retrieval/query answering

search engines:

- pretty good for some things

who was the fifteenth president of the united states

About 928,000 results (0.17 seconds) [Advanced search](#)

[James Buchanan - Fast Facts - Fifteenth President James Buchanan](#)

James Buchanan, Fifteenth President of the United States. Credit: Library of Congress, Prints and Photographs Division. LC-61-92101-9228 D&C ...
americanhistory.about.com/od/.../a/jf/j_buchanan.htm - Cached - Similar

- does mostly pattern matching and ranking
- no deep understanding
- still requires user to "find" the answer

40

+ Where are we now?

Question answering

- wolfram alpha

WolframAlpha[™] computational knowledge engine

who is the fifteenth president of the united states?

Input interpretation:
United States President 15th

Result:
James Buchanan

41

+ Where are we now?

Question answering: wolfram alpha

WolframAlpha[™] computational knowledge engine

what is the most popular car color in the united states?

WolframAlpha doesn't understand your query
Showing instead result for query: the united states

Other queries to try: popular

Assuming "united states" is a country (Use as a music work instead)

Input interpretation:
United States

full name	United States of America
alternate names	America US USA U.S. U.S.A.
internet code	.us

42

+ Question answering

The Science Behind an Answer

Watson performs so fast that it can rival the greatest human contestants in understanding a Jeopardy! clue and arriving at a single, precise answer. The significance of this accomplishment can be difficult to comprehend.

Watch the video to see how the computing system designed to play Jeopardy! works.

The first person mentioned by name in "The Man in the Iron Mask" is the hero of a previous book by the same author.

Possible Answers

- bathe
- battle
- be
- beam
- bear
- beat
- become
- beg

43

+ Where are we now?

Question answering

- Many others systems
 - TREC question answering competition
 - language computer corp
 - answerbus
 - ...

44

+ Where are we now?

Summarization

NewsBlaster

(Columbia) <http://newsblaster.cs.columbia.edu/>

Ukraine crisis: Russian forces reportedly seen in rebel cities Donetsk and Luhansk

Summary from multiple countries, from articles in English

Pro-Russian rebels softened their demand for full independence Monday, saying they would respect Ukraine's sovereignty in exchange for autonomy a shift that reflects Moscow's desire to strike a deal at a new round of peace talks. [\(article 17\)](#) The Ukrainian government, whose forces have been fighting the pro-Russian rebel groups for months, has said any aid needs its approval and has to go through the Red Cross or the United Nations. [\(article 19\)](#) Diplomats at a new round of talks on easing the crisis in Ukraine must push for an immediate, unconditional cease-fire between Ukrainian government troops and pro-Russian separatists, Russia's foreign minister said Monday. [\(article 2\)](#) The talks later Monday in Minsk, the Belarusian capital, come as Ukrainian troops are facing a resurgent rebel force. [\(article 3\)](#) NATO leaders this week will be asked to approve the creation of a high-readiness force and the stockpiling of military equipment and supplies in Eastern Europe to help protect member nations there against potential Russian aggression, the alliance's secretary general said Monday. [\(article 11\)](#) On Monday, the rebels pushed Ukrainian government forces from an airport near Luhansk, the second-largest rebel-held city, the latest in a series of military gains. [\(article 11\)](#)

45

+ Where are we now?

Voice recognition

- pretty good, particularly with speaker training
 - Apple OS/Siri
 - Android/Google
 - Alexa, Google Assistant, etc...
 - IBM ViaVoice
 - Dragon Naturally Speaking

Speech generation

- The systems can generate the words, but getting the subtle nuances right is still tricky
 - Apple OS
 - <http://translate.google.com>

46