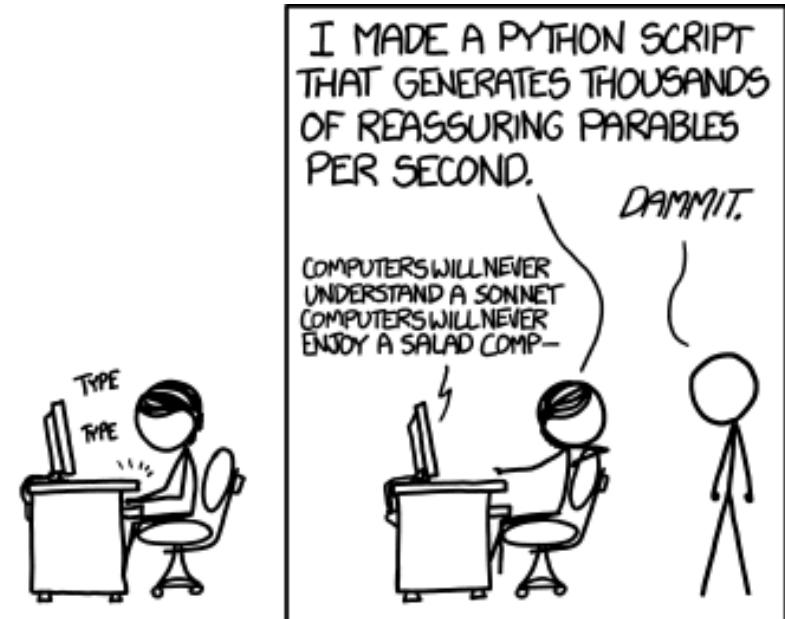


CSE 5525: Speech and Language Processing

Instructor: Wei Xu

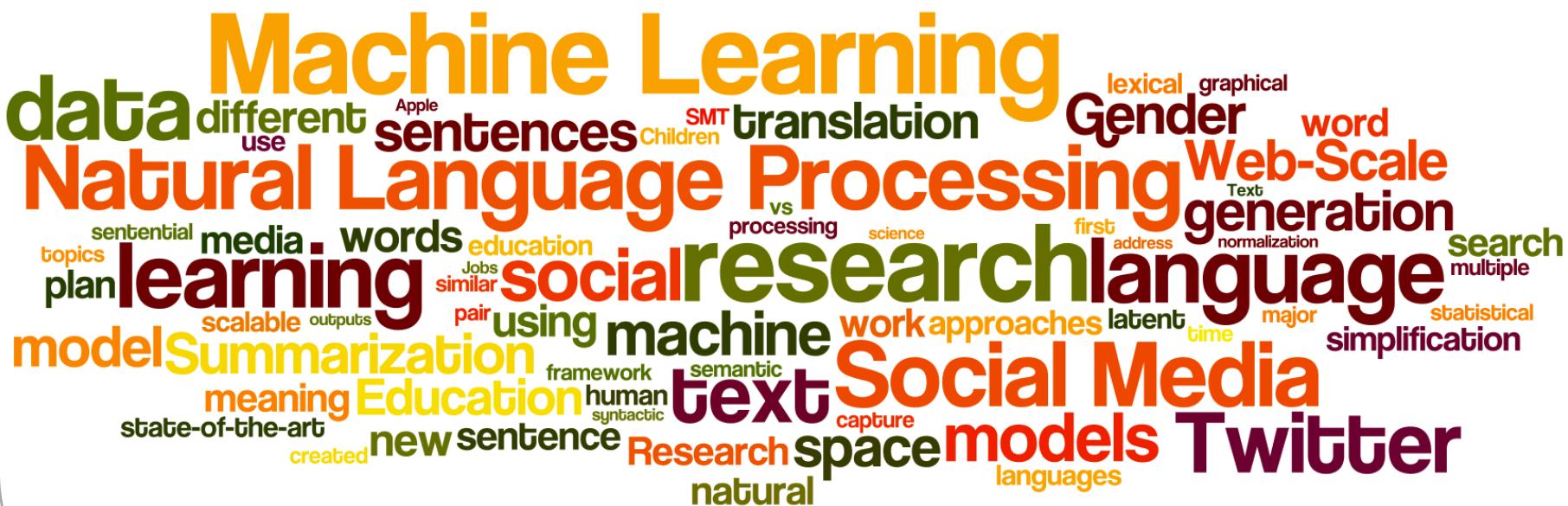




Assistant Professor Wei Xu

Department of Computer Science and Engineering
Ohio State University

[What I Do]



Course Webpage

https://cocoxu.github.io/courses/5525_spring17.html

CSE 5525: Speech and Language Processing

Fundamentals of natural language processing, automatic speech recognition and speech synthesis; lab projects concentrating on building systems to process written and/or spoken language.

Details

Instructor: [Wei Xu](#)

Time: Wednesday and Friday, 12:45PM- 2:05PM (Spring 2017)

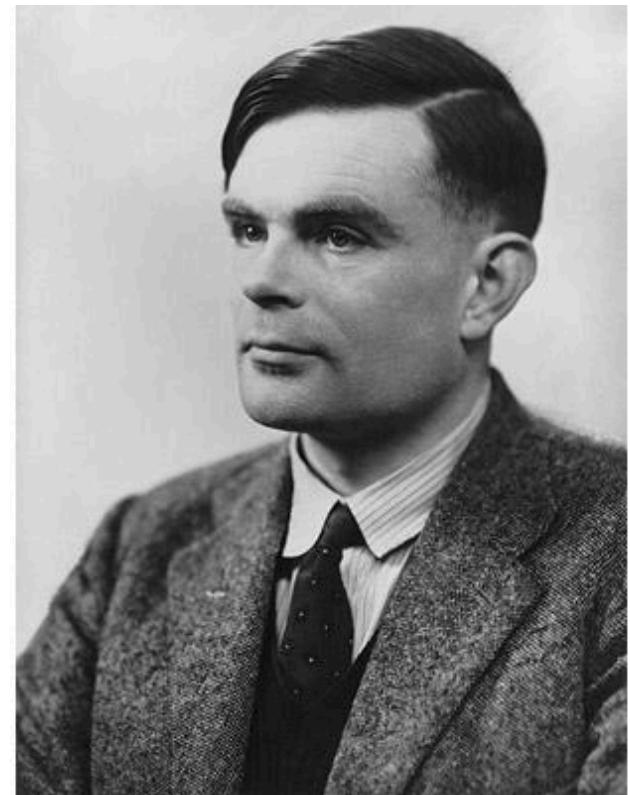
Place: [Bolz Hall 317](#)

Language, Thought and Artificial Intelligence

- Goal: give computers the ability to think (and speak)
- Long history in computer science

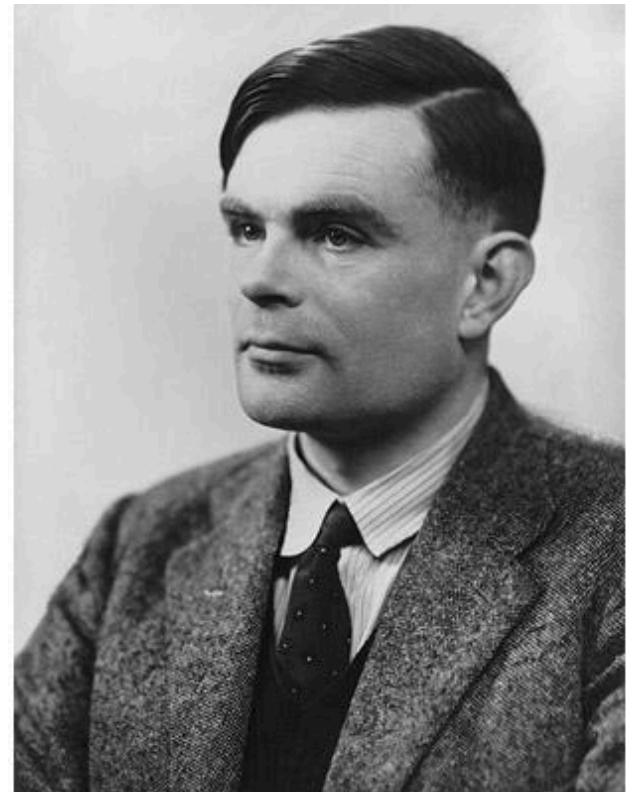
Language, Thought and Artificial Intelligence

- Goal: give computers the ability to think (and speak)
- Long history in computer science
- Turing Test (Alan Turing 1950)



Language, Thought and Artificial Intelligence

- Goal: give computers the ability to think (and speak)
- Long history in computer science
- Turing Test (Alan Turing 1950)
- Loebner Prize
 - Little interest from NLP community...
 - Very simple programs can fool some judges some of the time



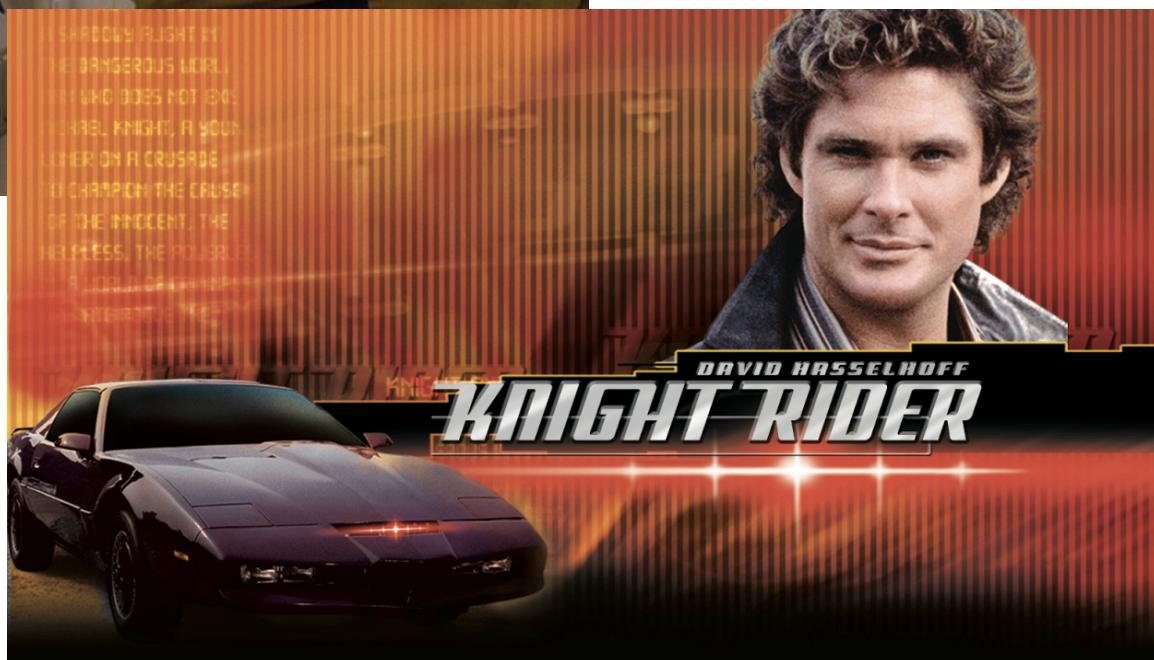
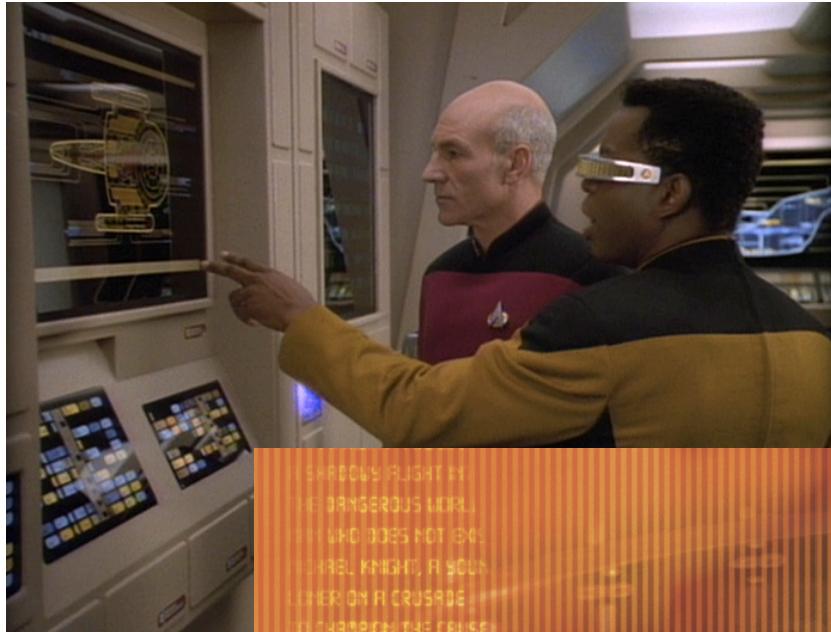
The Chinese Room (Searle 1980)



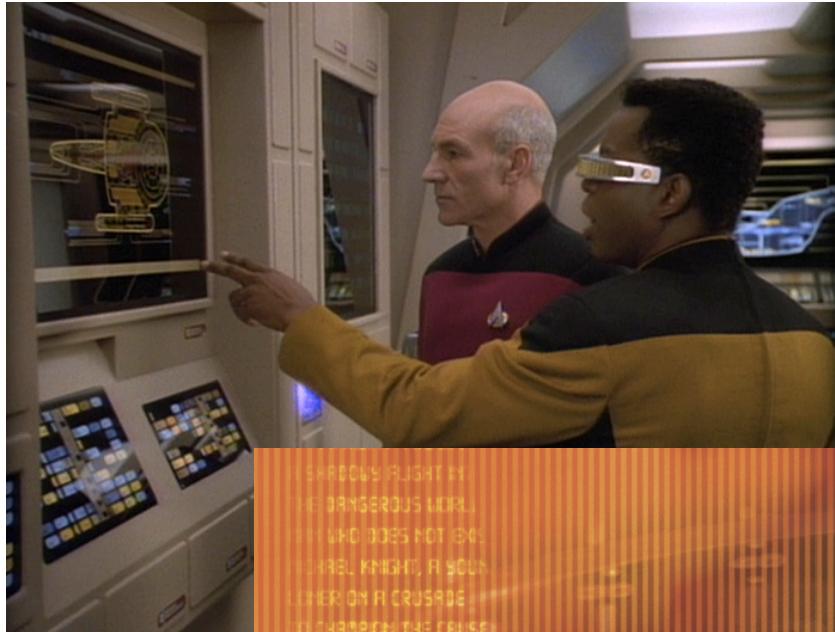
NLP in Science Fiction



NLP in Science Fiction



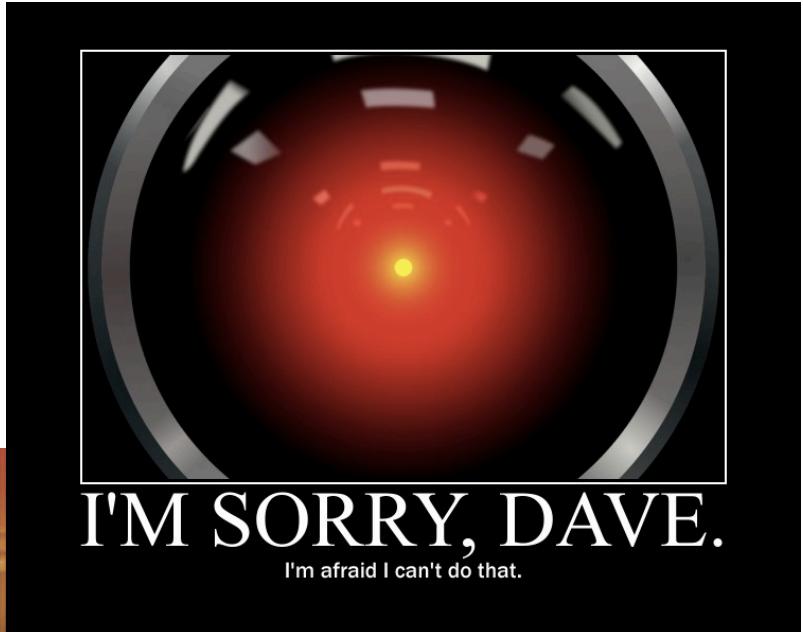
NLP in Science Fiction



IT'S A DANGEROUS WORLD.
ONE WHO DOES NOT DARE
TO FIGHT KNIGHTS, IN SOON
LITER ON A CRUSADE
TO CHAMPION THE CAUSE
OF THE INNOCENT, THE
WEAK, THE POWERLESS,
IS A LONELY GHOST IN
THE DARKNESS.



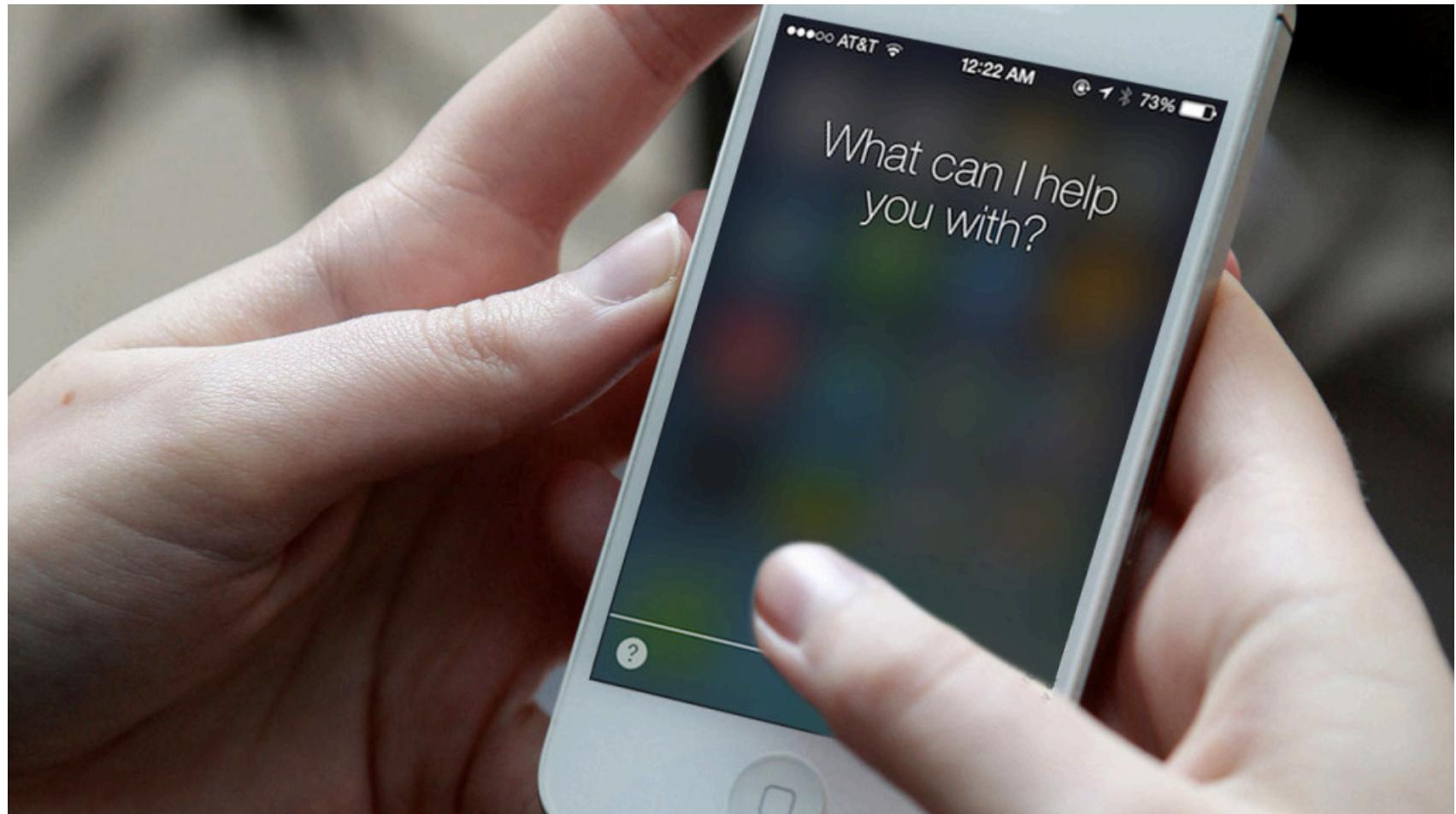
DAVID HASSELHOFF
KNIGHT RIDER



Today's NLP Applications

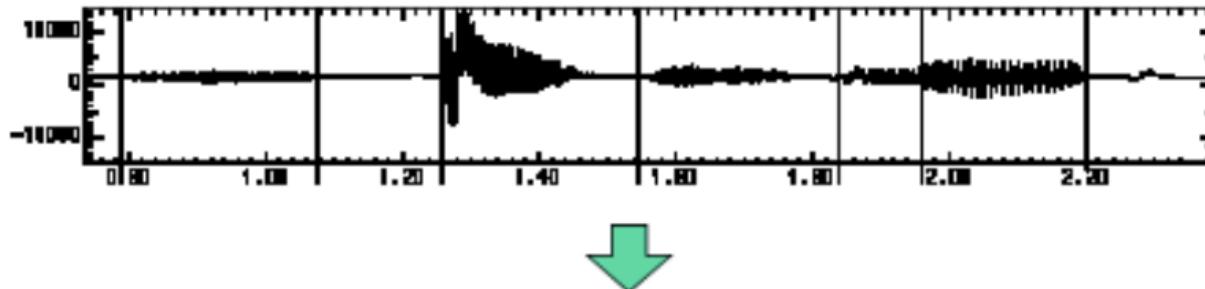
- Speech Interfaces
- Machine Translation
- Search Engines
- Information Extraction
- Summarization
- ...

Speech Interfaces



Speech Interfaces

- Automatic Speech Recognition
 - Audio In, text out
 - SOTA: 0.3% error for digit strings, 5% dictation
50%+ for TV



“Speech Lab”

Speech Interfaces

- Text to Speech
 - Text in Audio Out



Machine Translation



1



Translate

Turn off instant translation



English Spanish French Chinese - detected ▾



English Spanish Arabic ▾

Translate

你喜欢看电影吗?

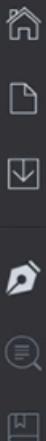


8/5000

do you like watching movies?



Error Correction



My First Car

For years I have been driving an old used car with a lot of mileage, and I hate it. It gets me where I need to go, but I'm tired of fixing leaks and broken parts all the time. Its annoying every times I need to take it to mechanic. Even when they takes care of everything, I know I'll just end up going back there in a few weeks. I have finally decided that I am not going to do it anymore. I have decided to buy a new car! Unfortunately, I have a problem. I have no idea what car to get. Do I want something big? Do I want something stylish? Something economical? I have so many choices that I don't even know where to begin.

I am not sure if I will be able to make a decision on my own. I don't have not a lot of money, either, so I probably don't have many options.

After I did some research, I knew that I would need some expert advice. Eventually, I went to a local dealership to check out some new models. I talked to the saleswoman and listened at she carefully. Her honesty and professionalism were really impressive. She had a lot of vary helpful suggestions and showed myself some safe, affordable choices. After a long discussion I finally decided which one I wanted.



grammarly

Possibly confused word: *its*

every times → *every time*

a mechanic or *the mechanic*

takes → *take*

not

did → *had done*

listened at → *listened to*

she → *her*

really

Natural Language Understanding

- Convert words into semantic representation that can be used to query a database

Where can I find Indian food in Clintonville?

cuisine location



1. Mughal Darbar



42 reviews

\$\$ - Indian

University District

2321 N High St
Columbus, OH 43202
(614) 477-6065

This restaurant takes reservations

[Find a Table](#)

This restaurant accepts pickup orders

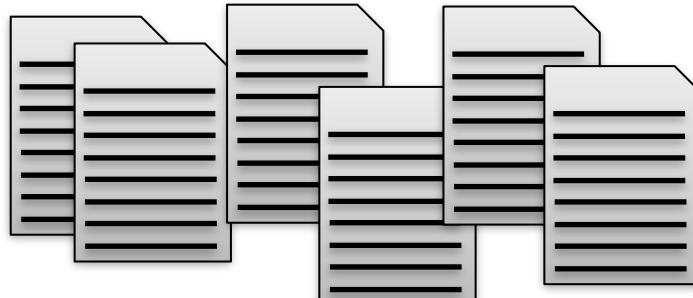
[Start Order](#)

Question Answering



Summarization

User reviews:



"**Ash** always makes visitors feel like her personal guests and never disappoints as a hostess." in 3 reviews



"Great customer service and **delicious food.**" in 2 reviews



"I did opt to eat on their **patio** during the nice summer weather and I would have to say that it was pleasant." in 7 reviews

Outdoor Seating: Yes

Information Extraction

*“Yess! Yess! Its official Nintendo announced today
that they Will release the Nintendo 3DS in north
America march 27 for \$250”*

Information Extraction

*“Yess! Yess! Its official **Nintendo** announced today
that they Will release the **Nintendo 3DS** in north
America march 27 for \$250”*

Information Extraction

*“Yess! Yess! Its official **Nintendo** announced today
that they Will release the **Nintendo 3DS** in north
America march 27 for \$250”*

COMPANY	PRODUCT	DATE	PRICE	REGION

PRODUCT RELEASE

Information Extraction

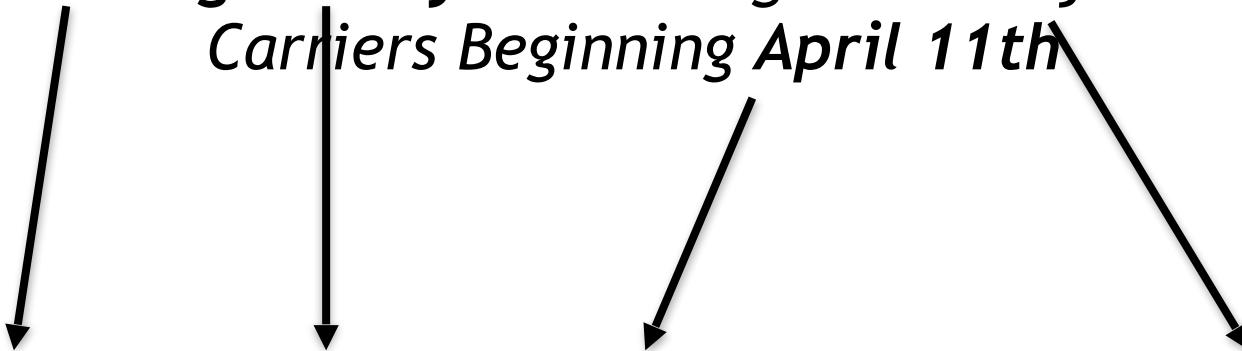
“Yess! Yess! Its official ~~Nintendo~~ announced today that they Will release the ~~Nintendo~~ 3DS in north America march 27 for \$250”

COMPANY	PRODUCT	DATE	PRICE	REGION
Nintendo	3DS	March 27	\$250	North America

PRODUCT RELEASE

Information Extraction

*Samsung Galaxy S5 Coming to All Major U.S.
Carriers Beginning April 11th*



COMPANY	PRODUCT	DATE	PRICE	REGION
Samsung	Galaxy S5	April 11	?	U.S.
Nintendo	3DS	March 27	\$250	North America

PRODUCT RELEASE

Information Extraction

Samsung Galaxy S5 Coming to All Major U.S.

- State of the art is maybe 80%, for single easy fields: 90%+
- Redundancy helps a lot!
- Much of human knowledge is waiting to be harvested from the Web!

COMPANY	PRODUCT	DATE	PRICE	REGION
Samsung	Galaxy S5	April 11	?	U.S.
Nintendo	3DS	March 27	\$250	North America

PRODUCT RELEASE

NLP as a Field

- Tasks (each have their own datasets and evaluations)
 - Syntactic Parsing
 - Summarization
 - Machine Translation
 - Information Extraction
 - ...
- Methods
 - Heuristic / Rule-based Approaches (dominated until the 1990s statistical revolution)
 - Probabilistic Graphical Models
 - Neural Networks (deep learning)
 - ...

Why is NLP hard?

- “Any speaker of any human language can produce and understand an infinite number of sentences.”

(Fromkin, Rodman & Hyams. 2011. An Introcution to Language.)

Why is NLP hard?

- “Any speaker of any human language can produce and understand an infinite number of sentences.”

(Fromkin, Rodman & Hyams. 2011. An Introcution to Language.)

Creativity!

Layers of Linguistic Annotation

UTF8: UN wants Yemen's warring parties to resume talks on Jan 14.

Layers of Linguistic Annotation

Words: UN-wants-Yemen-'s-warring-parties-to-resume-talks-on-Jan-14-.
(Tokenization)

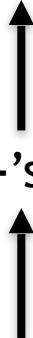


UTF8: UN wants Yemen's warring parties to resume talks on Jan 14.

Layers of Linguistic Annotation

Morphology:
(stemming)

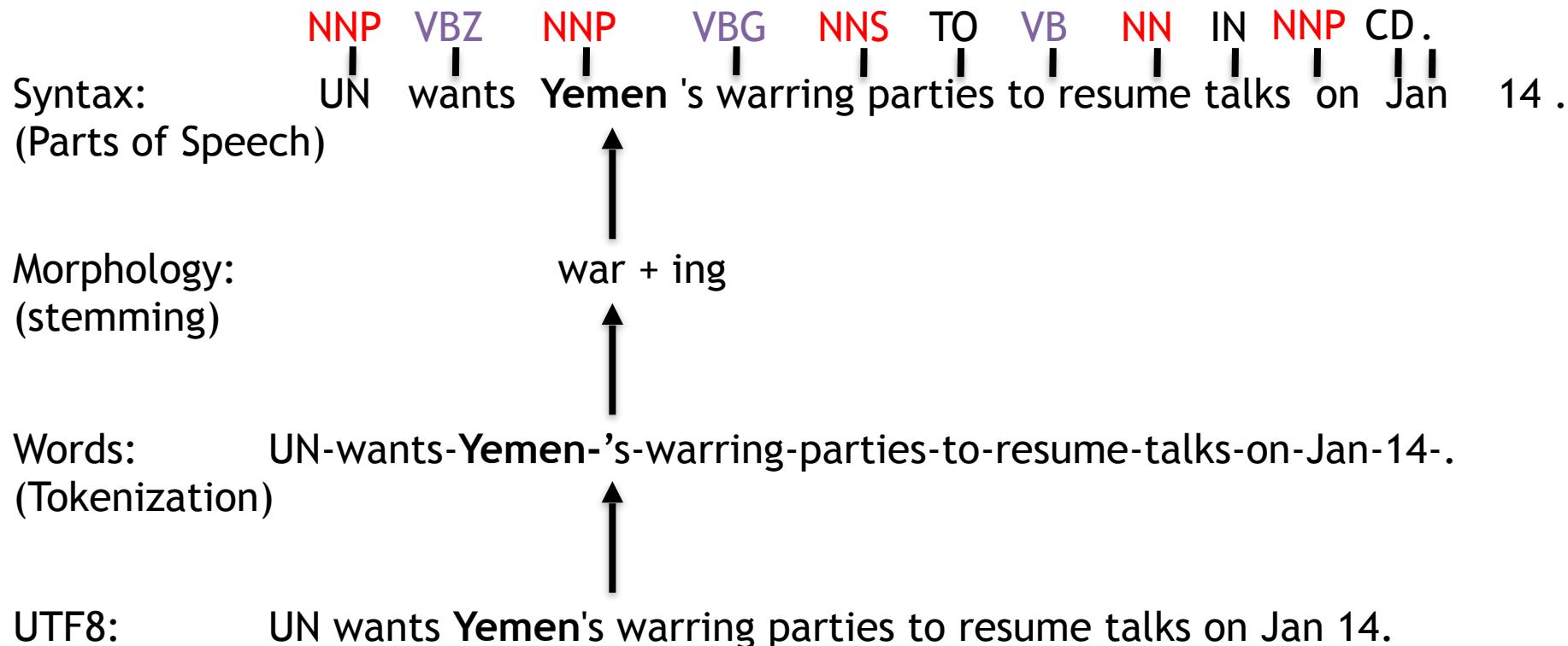
war + ing



Words: UN-wants-Yemen-'s-warring-parties-to-resume-talks-on-Jan-14-.
(Tokenization)

UTF8: UN wants Yemen's warring parties to resume talks on Jan 14.

Layers of Linguistic Annotation



Layers of Linguistic Annotation

Syntax: UN wants Yemen's warring parties to resume talks on Jan .
(Parts of Speech) NNP VBZ NNP VBG NNS TO VB NN IN NNP CD.

Layers of Linguistic Annotation

Syntax:
(Constituents)

```
(ROOT
  (S
    (NP (NNP UN))
    (VP (VBZ wants)
      (S
        (NP
          (NP (NNP Yemen) (POS 's))
          (VBG warring) (NNS parties))
        (VP (TO to)
          (VP (VB resume)
            (NP (NNS talks))
            (PP (IN on)
              (NP (NNP Jan) (CD 14)))))))
      (. .)))
```

Syntax: UN VBZ NNP VBG NNS TO VB NN IN NNP CD.
(Parts of Speech) wants Yemen 's warring parties to resume talks on Jan



Layers of Linguistic Annotation

Semantics/
Discourse:

DiplomaticTalks-Event
Between: Yemen's warring parties
Mediator: UN
Date: 1/14/2016



Syntax:
(Constituents)

```
(ROOT
  (S
    (NP (NNP UN))
    (VP (VBZ wants)
      (S
        (NP
          (NP (NNP Yemen) (POS 's))
          (VBG warring) (NNS parties))
        (VP (TO to)
          (VP (VB resume)
            (NP (NNS talks))
            (PP (IN on)
              (NP (NNP Jan) (CD 14)))))))
      (. .)))
```

Why is NLP hard?

- Why is it easy for computers to parse Python, but not English?

Why is NLP hard?

- Why is it easy for computers to parse Python, but not English?

Ambiguity!

Lexical Ambiguity



(Detroit Free Press, Nov. 12, 1974)

Lexical Ambiguity



Structural Ambiguity

the guardian

[News](#) | [Sport](#) | [Comment](#) | [Culture](#) | [Business](#) | [Money](#) | [Life & style](#) |

[News](#) › [UK news](#) › [Diana, Princess of Wales](#)

Princess Diana dresses to be auctioned

Ten dresses, including gowns designed by Catherine Walker, Bruce Oldfield and Zandra Rhodes, to go under the hammer

Lauren Cochrane

guardian.co.uk, Friday 15 March 2013 17.26 GMT

Structural Ambiguity



Students Cook & Serve Grandparents

On Thursday, September 9, Gorman School hosted the first annual Grandparent's Day.

All Grandparents were invited to a school wide pancake breakfast. Upper grade students served as excellent chefs, as well as taking responsibility for serving the food and the clean up after-

What will you get out of this class?

- Understanding of a major field of Artificial Intelligence
- Basic machine learning skills
- Ideas that you could transform into a startup company or academic research
- Summer interns?

What to Expect

What to Expect

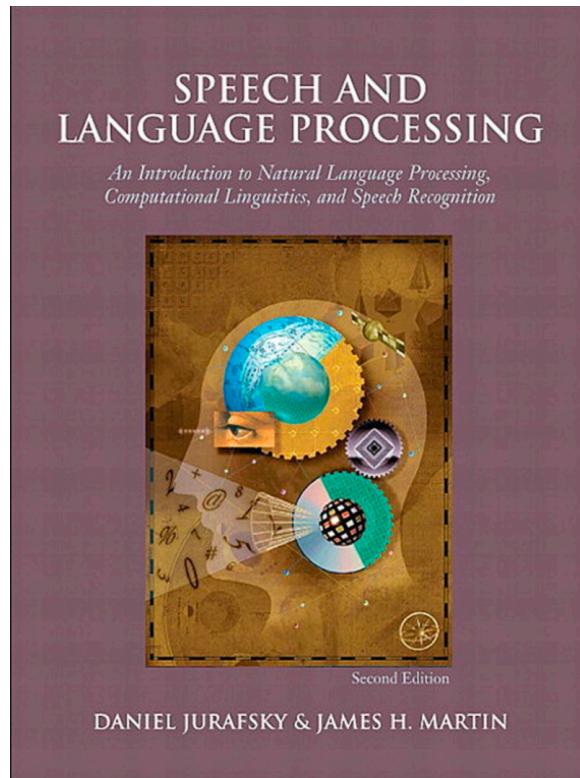
- Lots of math and programming
- A bit of Linguistics
- Computing Resources:
 - Experiments could take hours to run or debug depending the efficiency and quality of your code. We recommend you start **early** on homework assignments and final project.

Prerequisites

- (CSE 3521 or CSE 5521)
and (CSE 5522 or Stat 3460 or Stat 3470)
- Math:
 - Basic Probability
 - Basic Linear Algebra
- Programming
 - Python or ability to learn Python quickly
 - Numpy/Scipy (Python libraries for scientific computing)
 - Linux/Unix (for windows users: <https://www.cygwin.com/>)

Textbook

- Books
 - Jurafsky and Martin (2nd edition and 3rd edition draft)
 - Some other readings as well



Grading

- Participation (5%)
- Homework Assignments (55% individual)
 - one written assignment
 - three programming assignments
 - and market research
- Midterm (20%)
- Final Project (20% group of 3-4)

Homework

- #0 written assignment (simple):
 - Probability & Programming Basics
 - Due next week on January 11th
- Programming assignments (subject to change):
 - #1: Text Classification (Naive Bayes)
 - #2: Tagging (Structured Perceptron)
 - #3: Machine Translation (Word Alignment)

Homework

- Assignments are your individual work.

The screenshot shows a web browser window with the URL <https://theory.stanford.edu/~aiken/moss/> in the address bar. The page title is "Moss". Below the title, there is a large heading "A System for Detecting Software Plagiarism". Underneath this heading, there is a section titled "UPDATES" containing a list of historical events. The browser interface includes standard navigation buttons (back, forward, search) and a toolbar with various icons.

Moss

A System for Detecting Software Plagiarism

UPDATES

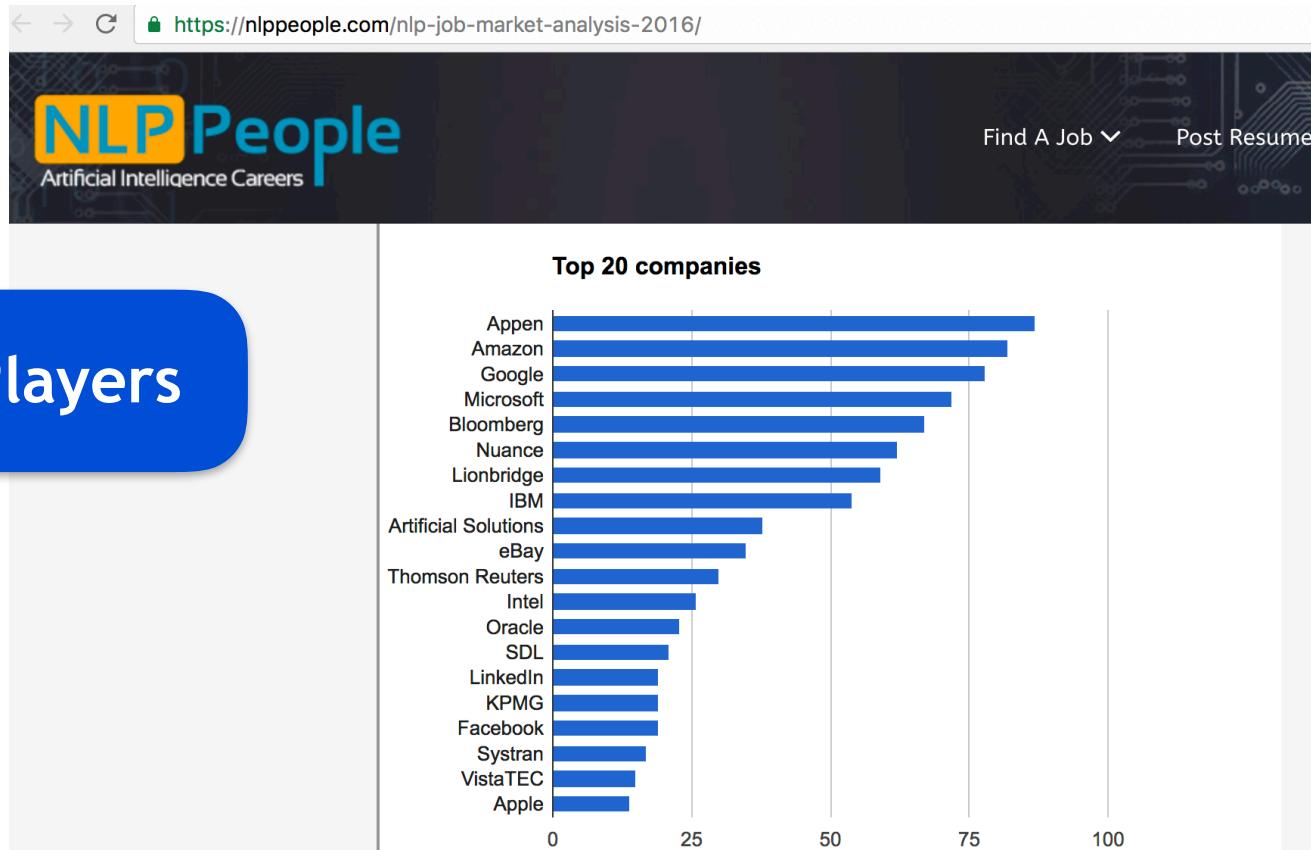
- May 18, 2014 Community contributions (including a Windows submission GUI from Shane May, thanks!) are now in their own section on this page.
- May 14, 2014 And here is a [Java version](#) of the submission script. Thanks to Bjoern Zielke!
- May 2, 2014 Here is a [PHP version](#) of the submission script. Many thanks to Phillip Rehs!
- June 9, 2011 There were two outages over the last couple of days that lasted no more than a hour each (I think). I've made some changes to the disk management software that should prevent these problems from recurring.
- April 29, 2011 There was an outage lasting a few hours today, the first since last summer, but everything is back up.
- August 1, 2010 Everything is back to normal.
- July 27, 2010 The Moss server is back on line. There may be some more tuning and possibly downtime in the coming weeks, but any outages should be brief. New registrations are not yet working, but people with existing accounts can submit jobs.
- July 25, 2010 As many (many!) people have noticed, the Moss server has been down for all of July. Unfortunately the hardware failed while I was away on a trip. I am hopeful it will be back up within a few days.

What is Moss?

Moss (for a Measure Of Software Similarity) is an automatic system for determining the similarity of programs. To date, the main application of Moss has been in detecting plagiarism in programming classes. Since its development in 1994, Moss has been very effective in this role. The algorithm behind moss is a significant improvement over other cheating detection algorithms (at least, over those known to us).

Homework

- Market Research on companies that use speech or language technologies.



Homework

- Market Research on companies that use speech or language technologies.

The screenshot shows a web browser window with the URL <https://angel.co/natural-language-processing>. The page title is "Natural Language Processing Startups". Below the title, it says "\$4.9M AVERAGE VALUATION". There are social media sharing icons for Twitter and Facebook. At the bottom, there are statistics: 596 COMPANIES, 1,936 INVESTORS, 13,910 FOLLOWERS, and 222 JOBS. A blue button on the left says "Startups".

Natural Language Processing Startups

\$4.9M AVERAGE VALUATION



596 COMPANIES 1,936 INVESTORS 13,910 FOLLOWERS 222 JOBS

◀ All Markets

PARENTS

Deep Information Technology

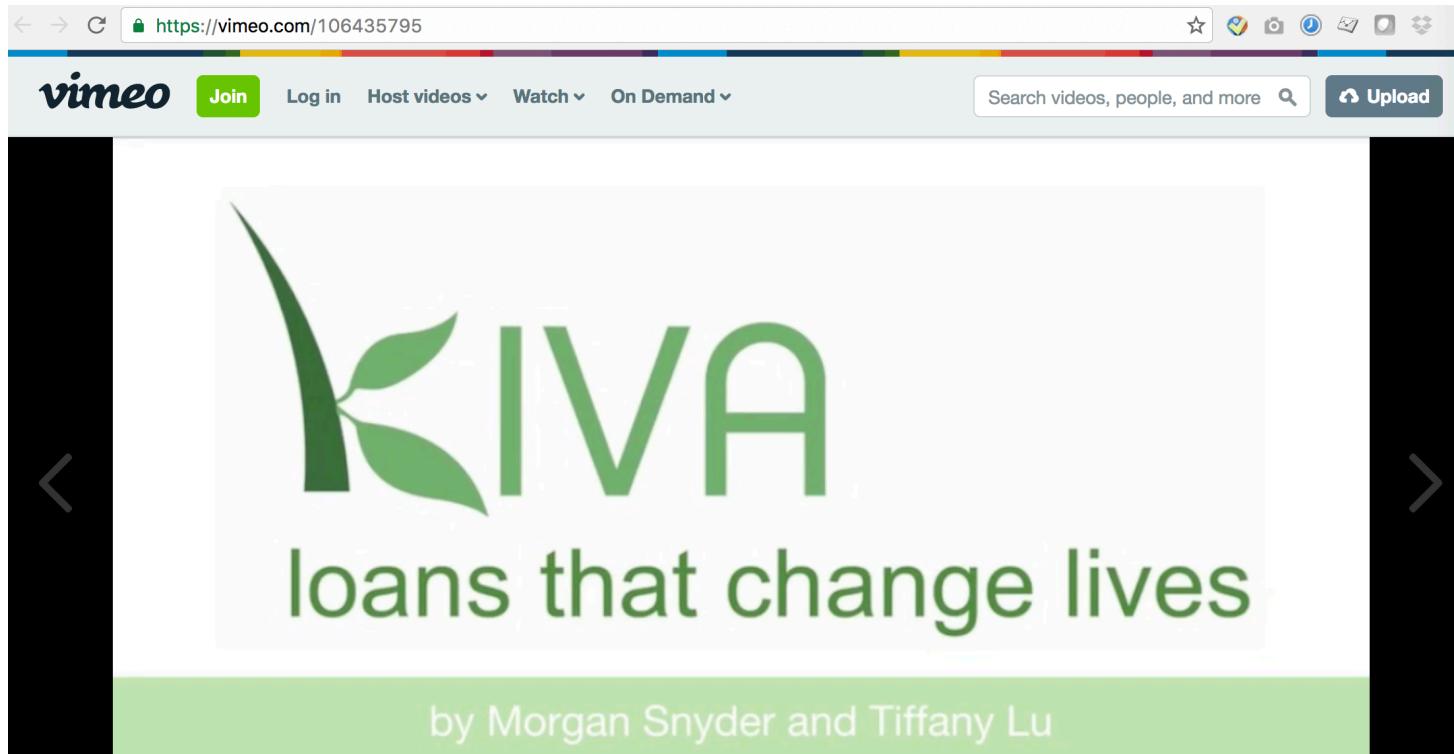
SIBLINGS

Algorithms

Company	Joined	Followers	Signal
Edgecase (formerly Compare Metrics) White-Label Product Discovery Plat... Austin · Semantic Search	Jul '11	202	
Vurb Mobile Search, Messaging, and Co... San Francisco · Semantic Search	Oct '11	1009	
Yummly The smart and personal platform for ...	Nov '11	407	

Market Research

- a short ~5 minute video presentation about a company (due on February 3rd)



Midterm

- In class, the week before spring break
- but, not final exam



Final Project (group of 3-4)

- Step 1: form a team and brainstorm 3 ideas
- Step 2: select an idea (talk to the instructor during office hours) and collect your data
- Step 3: conduct analyses of your data
- Step 4: develop a working system
- Step 5: write a 2-page report and give an in-class presentation (last class of the semester))

QA and Course Resources

- Carmen (homework submission & grades)
- Piazza (discussion and announcements)

The screenshot shows the Piazza platform interface for the course SP17 CSE 5525 - Spch & Lang Proc (10019). The left sidebar contains links for Spring 2017, Home, Announcements, Assignments, Discussions, Grades, People, Pages, Files, Syllabus, Modules, Conferences, Collaborations, Chat, and Piazza. The main content area has a header with tabs for Q & A, Resources, Statistics, and Manage Class. It features a search bar, a pinned post button, and a search input field. Below this is a feed of posts. A yellow-highlighted post titled "Welcome to Piazza!" provides an introduction to the platform. To the right, a "note" section titled "Welcome to Piazza!" explains the platform's purpose and usage.

Welcome to Piazza!

Piazza is a Q&A platform designed to get you great answers from classmates and a list of tips you might find handy as you get started:

- 1. Ask questions!**

The best way to get answers is to ask questions! Ask questions on Piazza rather than email because everyone can benefit from the response (and so you can get answers from classmate and instructor).

- 2. Edit questions and answers wiki-style.**

Think of Piazza as a Q&A wiki for your class. Every question has just a single student answer, which is great for getting answers quickly. You can also edit other students' answers collectively (and a single instructor's answer for instructors).

- 3. Add a followup to comment or ask further questions.**

To comment on or ask further questions about a post, start a followup discussion. This is a great way to keep the conversation going and make sure everyone is up-to-date.

- 4. Go anonymous.**

Shy? No problem. You can always opt to post or edit anonymously.

- 5. Tag your posts.**

It's far more convenient to find all posts about your Homework 3 or Midterm 1 when you've used a keyword. Click a blue tag in a post or the question feed to filter for posts containing that tag.

- 6. Format code and equations.**

Office Hours

- Instructor's Office Hours:
Wednesdays 4:30-5:30pm, Dreese 495
- TA: Wuwei Lan



Guest Lectures and Invited Talks

- January 13th – Zhou Yu (CMU)
- Late Feb or March – Micha Elsner (OSU)
- March 3rd – Ray Mooney (UT Austin)

By next week, please

- Do Homework #0
 - due at the beginning of class on next Wed
 - hand in paper copy
- Watch three example video presentations of market research

Market Research

- [Example Presentation 1](#) (Iceland's Crowdsourced Constitution by Abhishek GAdiraju)
- [Example Presentation 2](#) (Silk Road by Shreshth Khilani)
- [Example Presentation 3](#) (Kiva: Loans That Change Lives by Morgan Snyder and Tiffany Lu)