

Introduction to NLP

What is Natural
Language Processing?



Question Answering: IBM's Watson

- Won Jeopardy on February 16, 2011!

WILLIAM WILKINSON'S
“AN ACCOUNT OF THE PRINCIPALITIES OF
WALLACHIA AND MOLDOVIA”
INSPIRED THIS AUTHOR’S
MOST FAMOUS NOVEL



Bram Stoker

Dan Jurafsky



Information Extraction

Subject: **curriculum meeting**

Date: January 15, 2012

To: Dan Jura

Event: Curriculum mtg
Date: Jan-16-2012
Start: 10:00am
End: 11:30am
Where: Gates 159

Hi Dan, we've now scheduled the curriculum meeting.

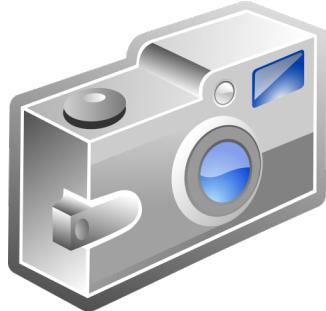
It will be in Gates 159 tomorrow from 10:00-11:30.

-Chris

Create new Calendar entry



Information Extraction & Sentiment Analysis



Size and weight

- ✓ • nice and compact to carry!
- ✓ • since the camera is small and light, I around those heavy, bulky professional cameras
- ✗ • the camera feels flimsy, is plastic and very light in weight you have to be very delicate in the handling of this camera

Attributes:
zoom
affordability
size and weight
flash
ease of use

The screenshot shows a web page with several yellow callout boxes highlighting positive reviews. One box reads: "Nice all around great little camera! My favorite little travel camera over weighted towards action & field shots. The digital image format is decent. I am going to recommend this camera to my wife. 2% off retail until Friday night for new reviews." Another box says: "Great camera! I used this camera for a high quality event during Christmas which is easy to use and I took some great photos! It has a decent zoom, and a decent flash. It does not come with a lens cap which is useful for traveling. This camera does not need you much to adjust anything. Best camera for travel! Very good camera." A third box states: "A really good point-and-shoot camera is a pleasure to use if you just want to take some quick shots. But you need to pay attention to your flash in indoor photography. This camera requires more than 2.4m away with the flash to focus properly (normal flash distance). Go to the camera EX. website. It has some nice video features." A fourth box says: "Great camera & lovely moment. This camera is very nice. It looks good, feels good, very easy to use. I would recommend this camera to anyone who is new to cameras. The only downside is that it is a bit expensive. The price is reasonable though. I think the camera is worth the price. It has some nice features like a flip screen and a good lens." A fifth box reads: "Very nice camera! I was a bit skeptical buying this because it's a bit on the expensive side. But I am so happy with this camera. It has a great lens, and a great flash. And it's very easy to use. I would recommend this camera to anyone who is new to cameras. It's a great camera!"



Machine Translation

- Fully automatic

Enter Source Text:

这 不 过 是 一 个 时 间 的 问 题 .

Translation from Stanford's *Phrasal*:

This is only a matter of time.

- Helping human translators

Enter Source Text:

تعرض الرئيس اللبناني إميل لحود لـ "#حملة عنيفة في مجلس التواب الذي انعقد أمس في جلسة تشريعية علية تحولت إلى "محاكمة" لـ "#رئيس الجمهورية على موقفه من المحكمة الدولية و "الملحاظات" التي ادلّي بها #+ها حول هذا الموضوع.

Translate Clear

Enter Translation:

lebanese |

president
suffered
exposed
president emile
before
presented
offer

Done!



Language Technology

making good progress

mostly solved

Spam detection

Let's go to Agra!



Buy V1AGRA ...



Part-of-speech (POS) tagging

ADJ ADJ NOUN VERB ADV

Colorless green ideas sleep furiously.

Named entity recognition (NER)

PERSON ORG LOC

Einstein met with UN officials in Princeton

Sentiment analysis

Best roast chicken in San Francisco!



The waiter ignored us for 20 minutes.



Coreference resolution

Carter told Mubarak he shouldn't run again.

Word sense disambiguation (WSD)

I need new batteries for my **mouse**.



Parsing



I can see Alcatraz from the window!

Machine translation (MT)

第13届上海国际电影节开幕...



The 13th Shanghai International Film Festival...

Information extraction (IE)

You're invited to our dinner party, Friday May 27 at 8:30



Party
May 27
add

still really hard

Question answering (QA)

Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

Paraphrase

XYZ acquired ABC yesterday

ABC has been taken over by XYZ

Summarization

The Dow Jones is up



The S&P500 jumped

Housing prices rose

Economy is good

Dialog

Where is Citizen Kane playing in SF?



Castro Theatre at 7:30. Do you want a ticket?





Ambiguity makes NLP hard: “Crash blossoms”



Violinist Linked to JAL Crash Blossoms

Teacher Strikes Idle Kids

Red Tape Holds Up New Bridges

Hospitals Are Sued by 7 Foot Doctors

Juvenile Court to Try Shooting Defendant

Local High School Dropouts Cut in Half

Dan Jurafsky



Ambiguity is pervasive

New York Times headline (17 May 2000)

Fed raises interest rates

Fed raises interest rates

Fed raises interest rates 0.5%



In-video quizzes!

- Most lectures will include a little quiz
 - Just to check basic understanding
 - Simple, multiple-choice.
 - You can retake them if you get them wrong



Why else is natural language understanding difficult?

non-standard English

Great job @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either♥

segmentation issues

the New York-New Haven Railroad
the New York-New Haven Railroad

idioms

dark horse
get cold feet
lose face
throw in the towel

neologisms

unfriend
Retweet
bromance

world knowledge

Mary and Sue are sisters.
Mary and Sue are mothers.

tricky entity names

Where is *A Bug's Life* playing ...
Let It Be was recorded ...
... a mutation on the *for* gene ...

But that's what makes it fun!



Making progress on this problem...

- The task is difficult! What tools do we need?
 - Knowledge about language
 - Knowledge about the world
 - A way to combine knowledge sources
- How we generally do this:
 - probabilistic models built from language data
 - $P(\text{"maison"} \rightarrow \text{"house"})$ **high**
 - $P(\text{"L'avocat général"} \rightarrow \text{"the general avocado"})$ **low**
 - Luckily, rough text features can often do half the job.



This class

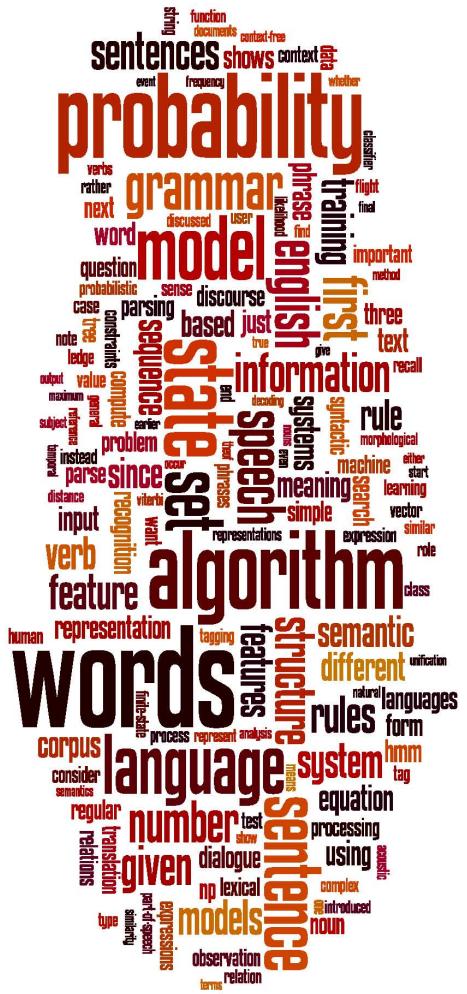
- Teaches key theory and methods for statistical NLP:
 - Viterbi
 - Naïve Bayes, Maxent classifiers
 - N-gram language modeling
 - Statistical Parsing
 - Inverted index, tf-idf, vector models of meaning
- For practical, robust real-world applications
 - Information extraction
 - Spelling correction
 - Information retrieval
 - Sentiment analysis

Dan Jurafsky



Skills you'll need

- Simple linear algebra (vectors, matrices)
- Basic probability theory
- Java or Python programming
 - Weekly programming assignments



Introduction to NLP

What is Natural
Language Processing?