CS 4650/7650: Natural Language Processing

Jacob Eisenstein

Lecture 1: Introduction

August 20, 2013

When and where

- Lecture: ES&T L1105, Tuesday/Thursday 3:05-4:30
- Office hours (Eisenstein): TSRB 228A, Wednesday 11-12
- My email: jacobe@gatech
- TA: Yangfeng Ji. Office hours: Friday 11-12, outside TSRB 228
- Webpage: `https://github.com/jacobeisenstein/gt-nlp-class/`

Prerequisites

- **Officially**: CS 3600, Intro to Artificial Intelligence
- **Unofficially**
  - Basic linear algebra
  - Solid probability and statistics
  - Automata and formal language theory: e.g., finite-state vs context-free languages, etc
  - Ability to analyze and implement dynamic programming algorithms
  - Strong coding ability (Python strongly preferred)
  - Bonuses:
    - Some familiarity with basic machine learning:
      naïve Bayes, logistic regression, perceptron
    - Some familiarity with basic ideas about linguistics

Resources

Readings are are drawn from online resources. They should be completed before the lecture on date assigned.

Resources

Readings are are drawn from online resources. They should be completed before the lecture on date assigned.

Places to find additional material

- Optional textbook: Jurafsky and Martin, **Second Edition**
- Emily Bender, **Linguistic fundamentals for NLP**
- Noah Smith, **Linguistic Structure Prediction**
- Kevin Murphy, **Machine Learning**
- Manning, Raghavan, & Schütze, **Introduction to Information Retrieval**
- Guy Lebanon, **Probability: The Analysis of Data, Vol 1**
- Journals: Computational Linguistics, Journal of Machine Learning Research
- Conferences: ACL, NAACL, EMNLP, EACL, NIPS, ICML, ...

Assignments

Assignments

- Four assigned projects (40%)
- Twelve **short** homework assignments (20%, includes class participation)
- In-class midterm exam (20%)
- Independent project (20%)

Assignments

Assignments

- Four assigned projects (40%)
- Twelve **short** homework assignments (20%, includes class participation)
- In-class midterm exam (20%)
- Independent project (20%)

There will be a lot of reading, a lot of coding, and a lot of math.

Assigned projects

- Through the assigned projects, you will:
  - Build increasingly complex (and practical) NLP systems.
  - Test properties of these systems by performing experiments.
  - Derive properties of NLP systems mathematically.
  - Compete in in-class “bakeoff” competitions.

Assigned projects

- Through the assigned projects, you will:
  - Build increasingly complex (and practical) NLP systems.
  - Test properties of these systems by performing experiments.
  - Derive properties of NLP systems mathematically.
  - Compete in in-class “bakeoff” competitions.
- Grading
  - Assignments are due at the **beginning** of lecture on the date indicated.
  - They’ll be accepted up to 3 days late, with a penalty of 20% per day.
  - You must work alone.
  - I take academic integrity very seriously.
    See `www.honor.gatech.edu` and the online syllabus for more details.
    If you have a question, ask!

Homeworks

- Through the **short** homework assignments you will:
  - Learn to identify linguistic phenomena by labeling real texts.
  - Compare your linguistic analysis with your classmates.
  - Help critique your classmates' independent projects.
  - Not spend more than one hour per assignment.

Homeworks

- Through the **short** homework assignments you will:
  - Learn to identify linguistic phenomena by labeling real texts.
  - Compare your linguistic analysis with your classmates.
  - Help critique your classmates' independent projects.
  - Not spend more than one hour per assignment.
- Grading
  - Submit a PDF online by the beginning of class **and bring a paper copy to class**.
  - Each homework is worth 2 points: one for the writeup and one for the class discussion. If you miss the discussion, you lose the point.
  - There will be twelve homeworks. You may skip two.
  - Homeworks will not be accepted late.
  - You must work alone.

Midterm exam

There will be an in-class midterm on October 8.

- Barring a medical emergency, you must take the exam in class on that date.
- The purpose of the midterm is to test your understanding of the concepts covered in class and in the readings.
- The secondary purpose is to encourage you to review those concepts.
- The midterm will include **anything** covered in class and in the readings through October 8. There will not be a practice exam or study guide.

Independent project

- The independent final project will give you an opportunity to delve more deeply into an area of interest.
- You may work in teams of up to three.
- The outcome may be:
  - a piece of software
  - an experimental or theoretical analysis
  - a comprehensive review of some area of research...
- I am happy to discuss possible projects with you.
  Start thinking about this early.

Independent project schedule

- **After the midterm**: start thinking about your project if you haven't already.
- November 12: present your proposal idea in class
- December 3-5: present your project results in class
- December 6: Submit initial project report
- December 13 (est): Submit final project report

In the beginning

While he was inventing the field of AI, Alan Turing asked: “how do we know when we’re done?”

In the beginning

While he was inventing the field of AI, Alan Turing asked: “how do we know when we’re done?”

The Turing Test: Can a computer carry on a conversation so naturally that you can’t distinguish it from a human?

Turing, 1950. “Computing Machinery and Intelligence.” Mind (236): 433-460.

The Turing Test today

http://www.pandorabots.com/pandora/talk?botid=f5d922d97e345aa1
http://www.youtube.com/watch?v=WnzlbyTZsQY

The Turing Test today

http://www.pandorabots.com/pandora/talk?botid=f5d922d97e345aa1

http://www.youtube.com/watch?v=WnzlbyTZsQY

- The best chatbots today avoid deep language understanding and focus on exhaustive string matching.
- In contrast, most of NLP is concerned with building software that understands language on a deeper level.

The Turing Test today

`http://www.pandorabots.com/pandora/talk?botid=f5d922d97e345aa1`

`http://www.youtube.com/watch?v=WnzlbyTZsQY`

- The best chatbots today avoid deep language understanding and focus on exhaustive string matching.
- In contrast, most of NLP is concerned with building software that understands language on a deeper level. **why is this hard?**

Ambiguity

Some real examples:

- Iraqi head seeks arms

Ambiguity

Some real examples:

- Iraqi head seeks arms
- Teacher strikes idle kids

Ambiguity

Some real examples:

- Iraqi head seeks arms
- Teacher strikes idle kids
- Ban on nude dancing on governor's desk

Ambiguity

Some real examples:

- Iraqi head seeks arms
- Teacher strikes idle kids
- Ban on nude dancing on governor's desk
- Stolen painting found by tree

Ambiguity

Some real examples:

- Iraqi head seeks arms
- Teacher strikes idle kids
- Ban on nude dancing on governor's desk
- Stolen painting found by tree

Ambiguity

Some real examples:

- Iraqi head seeks arms
- Teacher strikes idle kids
- Ban on nude dancing on governor's desk
- Stolen painting found by tree

Ambiguity grows with sentence length, sometimes exponentially!

How?

How?

A range of disambiguating cues:

- **The lexicon** (words and their syntactic functions)
  Teacher is almost always a noun, never a verb

How?

A range of disambiguating cues:

- **The lexicon** (words and their syntactic functions)
  Teacher is almost always a noun, never a verb
- **Syntax**
  Sentences rarely have two adjacent verbs (if they are both indicative)
  The sequence ADJECTIVE-VERB-END is unlikely in English

How?

A range of disambiguating cues:

- **The lexicon** (words and their syntactic functions)
  Teacher is almost always a noun, never a verb
- **Syntax**
  Sentences rarely have two adjacent verbs (if they are both indicative)
  The sequence ADJECTIVE-VERB-END is unlikely in English
- **Lexical semantics**
  A ban cannot dance
  A head (the body part) rarely seeks

How?

A range of disambiguating cues:

- **The lexicon** (words and their syntactic functions)
  Teacher is almost always a noun, never a verb
- **Syntax**
  Sentences rarely have two adjacent verbs (if they are both indicative)
  The sequence ADJECTIVE-VERB-END is unlikely in English
- **Lexical semantics**
  A ban cannot dance
  A head (the body part) rarely seeks
- "**Common sense**"
  Teachers aren't supposed to hit kids

How?

- Writing down all of these constraints and preferences in a single coherent representation is hard.
- Many (most?) sentences won't satisfy all constraints.
  How to decide which ones can be safely ignored?

How?

- Writing down all of these constraints and preferences in a single coherent representation is hard.
- Many (most?) sentences won't satisfy all constraints.
  How to decide which ones can be safely ignored?
- The answer is data, and probability.

NLP today

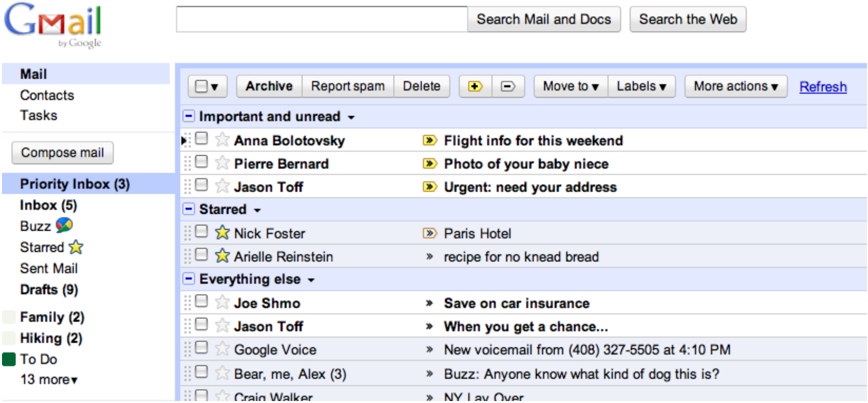
NLP research may still be as ambitious as the Turing test.

NLP today

NLP research may still be as ambitious as the Turing test.
But it may also be very down-to-earth...

- Finding the price of products on the web
- Analyzing reading level or authorship
- Detecting sentiment about products, stocks, or world leaders
- Extracting facts or relations from documents

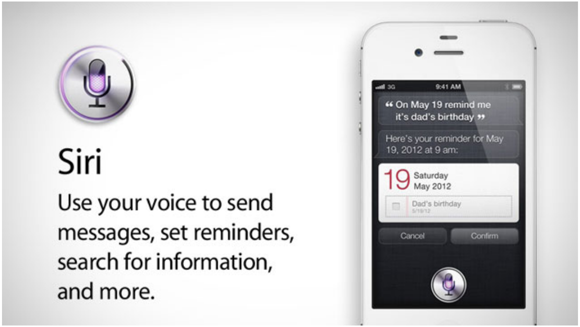
Application: document classification



Important and unread	
Anna Bolotovsky	**Flight info for this weekend**
Pierre Bernard	**Photo of your baby niece**
Jason Toff	**Urgent: need your address**
Starred	
Nick Foster	Paris Hotel
Arielle Reinstein	recipe for no knead bread
Everything else	
Joe Shmo	**Save on car insurance**
Jason Toff	**When you get a chance...**
Google Voice	New voicemail from (408) 327-5505 at 4:10 PM
Bear, me, Alex (3)	Buzz: Anyone know what kind of dog this is?
Craig Walker	NY Lay Over

Email is reliably separated into priority, regular, and spam.

Application: speech and dialogue



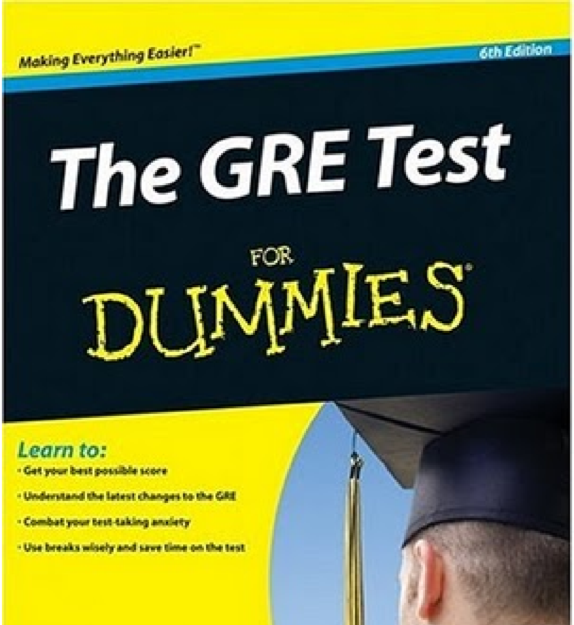
Siri

Use your voice to send messages, set reminders, search for information, and more.

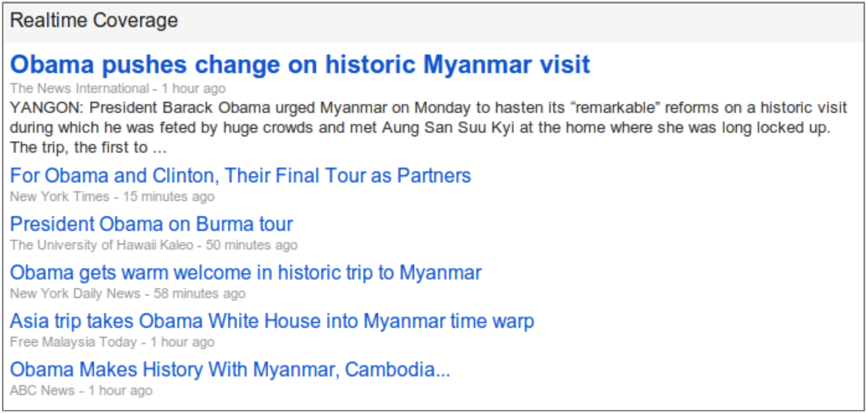
“On May 19 remind me it's dad's birthday”

Here's your reminder for May 19, 2012 at 9 am:

Application: discourse analysis


Making Everything Easier!™
6th Edition
The GRE Test
FOR
DUMMIES®
Learn to:
• Get your best possible score
• Understand the latest changes to the GRE
• Combat your test-taking anxiety
• Use breaks wisely and save time on the test

Application: content and subjectivity analysis



Realtime Coverage

Obama pushes change on historic Myanmar visit

The News International - 1 hour ago

YANGON: President Barack Obama urged Myanmar on Monday to hasten its “remarkable” reforms on a historic visit during which he was feted by huge crowds and met Aung San Suu Kyi at the home where she was long locked up. The trip, the first to ...

For Obama and Clinton, Their Final Tour as Partners

New York Times - 15 minutes ago

President Obama on Burma tour

The University of Hawaii Kaleo - 50 minutes ago

Obama gets warm welcome in historic trip to Myanmar

New York Daily News - 58 minutes ago

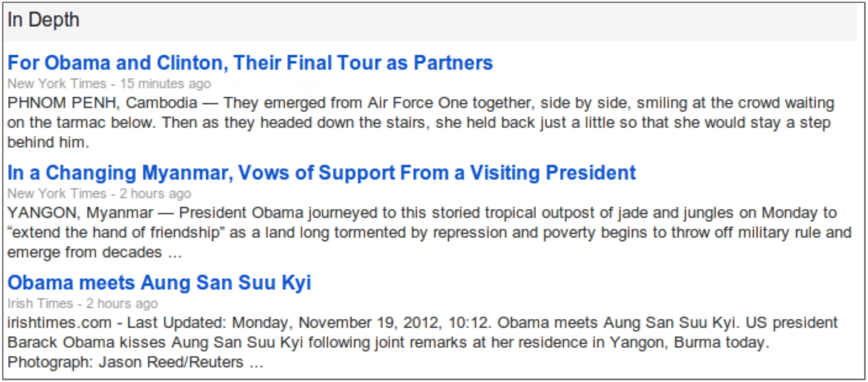
Asia trip takes Obama White House into Myanmar time warp

Free Malaysia Today - 1 hour ago

Obama Makes History With Myanmar, Cambodia...

ABC News - 1 hour ago

Application: content and subjectivity analysis



In Depth

For Obama and Clinton, Their Final Tour as Partners

New York Times - 15 minutes ago

PHNOM PENH, Cambodia — They emerged from Air Force One together, side by side, smiling at the crowd waiting on the tarmac below. Then as they headed down the stairs, she held back just a little so that she would stay a step behind him.

In a Changing Myanmar, Vows of Support From a Visiting President

New York Times - 2 hours ago

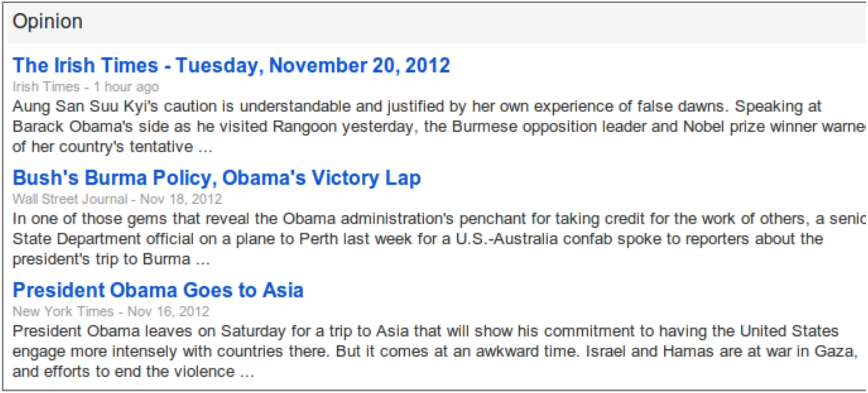
YANGON, Myanmar — President Obama journeyed to this storied tropical outpost of jade and jungles on Monday to "extend the hand of friendship" as a land long tormented by repression and poverty begins to throw off military rule and emerge from decades ...

Obama meets Aung San Suu Kyi

Irish Times - 2 hours ago

irishtimes.com - Last Updated: Monday, November 19, 2012, 10:12. Obama meets Aung San Suu Kyi. US president Barack Obama kisses Aung San Suu Kyi following joint remarks at her residence in Yangon, Burma today. Photograph: Jason Reed/Reuters ...

Application: content and subjectivity analysis



Opinion

The Irish Times - Tuesday, November 20, 2012

Irish Times - 1 hour ago

Aung San Suu Kyi's caution is understandable and justified by her own experience of false dawns. Speaking at Barack Obama's side as he visited Rangoon yesterday, the Burmese opposition leader and Nobel prize winner warne of her country's tentative ...

Bush's Burma Policy, Obama's Victory Lap

Wall Street Journal - Nov 18, 2012

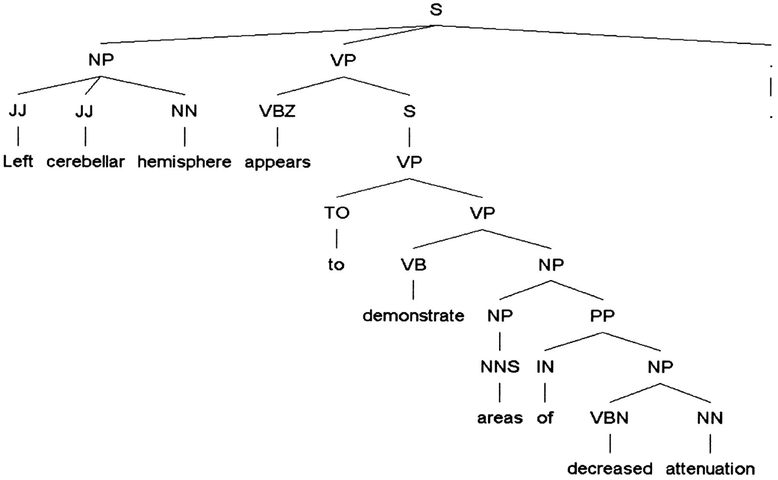
In one of those gems that reveal the Obama administration's penchant for taking credit for the work of others, a senic State Department official on a plane to Perth last week for a U.S.-Australia confab spoke to reporters about the president's trip to Burma ...

President Obama Goes to Asia

New York Times - Nov 16, 2012

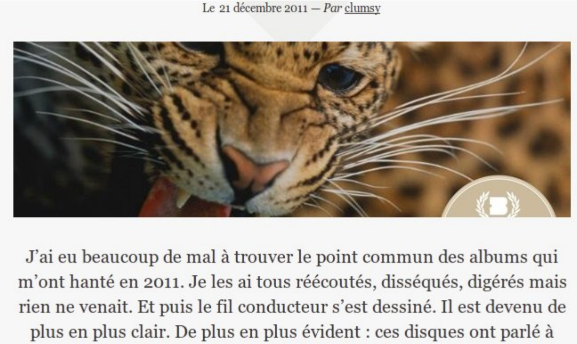
President Obama leaves on Saturday for a trip to Asia that will show his commitment to having the United States engage more intensely with countries there. But it comes at an awkward time. Israel and Hamas are at war in Gaza, and efforts to end the violence ...

Syntactic analysis today



Modern syntactic parsers get 90% accuracy on English newstext.

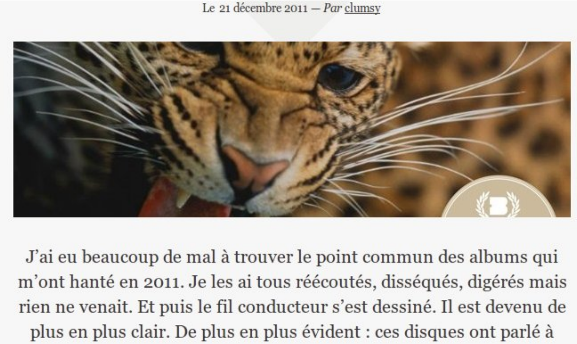
Machine translation today



Le 21 décembre 2011 — *Par* clumsy

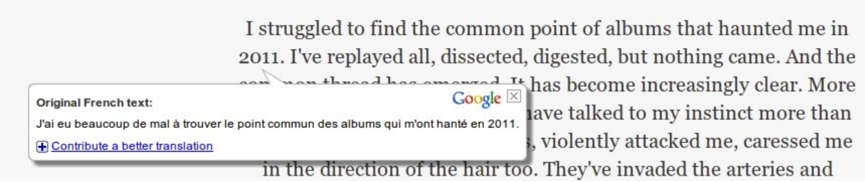
J’ai eu beaucoup de mal à trouver le point commun des albums qui m’ont hanté en 2011. Je les ai tous réécoutés, disséqués, digérés mais rien ne venait. Et puis le fil conducteur s’est dessiné. Il est devenu de plus en plus clair. De plus en plus évident : ces disques ont parlé à

Machine translation today



Le 21 décembre 2011 — *Par* clumsy

J'ai eu beaucoup de mal à trouver le point commun des albums qui m'ont hanté en 2011. Je les ai tous réécoutés, disséqués, digérés mais rien ne venait. Et puis le fil conducteur s'est dessiné. Il est devenu de plus en plus clair. De plus en plus évident : ces disques ont parlé à



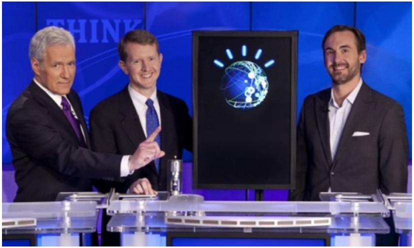
I struggled to find the common point of albums that haunted me in 2011. I've replayed all, dissected, digested, but nothing came. And the common thread has emerged. It has become increasingly clear. More ... ave talked to my instinct more than ... s, violently attacked me, caressed me in the direction of the hair too. They've invaded the arteries and

Original French text:

J'ai eu beaucoup de mal à trouver le point commun des albums qui m'ont hanté en 2011.

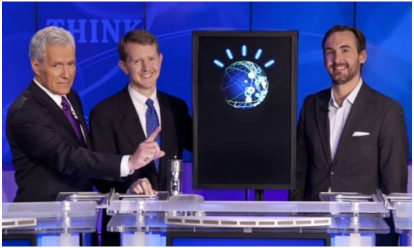
Fast, accurate, and (somewhat) fluent translation for many language pairs

Information extraction today



Watson extracts facts from millions of documents, parses complex questions, and outperforms the best human players.

Information extraction today



Watson extracts facts from millions of documents, parses complex questions, and outperforms the best human players.

This goes way beyond search and string match. An example question:

Wanted for general evilness, last seen at the Tower of Barad-Dur.
It's a giant eye, folks, kinda hard to miss

Data-driven NLP

- All of these success stories result from applying statistical analysis to large amounts of linguistic data.
- This data-driven approach will be the focus of this course.

Corpora

A **corpus** is a collection of text:
often annotated in some way, but sometimes just lots of text.

The development of large corpora made data-driven NLP possible. Some examples:

- Brown corpus:
  1M words of text with part-of-speech tags
- Penn Treebank:
  1M words of text with parse trees
- Europarl:
  1.8M aligned French-English sentence pairs
- Google n-grams:
  1.2B 5-grams and their counts
  (to think about: why is this useful?)

Corpora

A **corpus** is a collection of text:
often annotated in some way, but sometimes just lots of text.

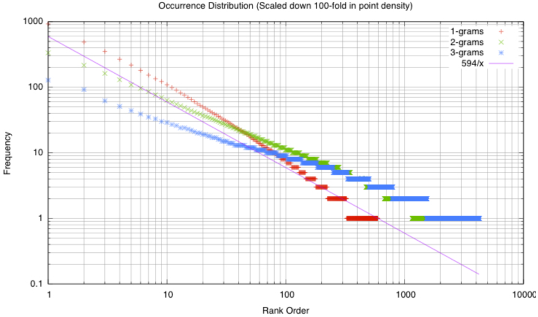
The development of large corpora made data-driven NLP possible. Some examples:

- Brown corpus:
  1M words of text with part-of-speech tags
- Penn Treebank:
  1M words of text with parse trees
- Europarl:
  1.8M aligned French-English sentence pairs
- Google n-grams:
  1.2B 5-grams and their counts
  (to think about: why is this useful?)

How much data is enough?

How much is enough?

Answer: there's no data like more data.



- There is always a "long tail" of rare but important phenomena.
- We are increasingly interested in languages with few resources, and new problems without annotations.
- This means we have to be smart.

The NLP stack

Natural language processing applications are typically built in a **stack**

- From “low-level” phenomena like words and morphemes...
- ... to “high-level” phenomena like semantics and discourse.

This course

Outline of topics

- **Words**: text classification, language models, morphology
- **Sequences**: hidden Markov models, part-of-speech tagging
- **Trees**: context free grammars, parsing
- **General graphs**: semantics and discourse
- **Learning**: unsupervised and semi-supervised methods
- **Applications**: translation, information extraction, dialogue

This course

Outline of topics

- **Words**: text classification, language models, morphology
- **Sequences**: hidden Markov models, part-of-speech tagging
- **Trees**: context free grammars, parsing
- **General graphs**: semantics and discourse
- **Learning**: unsupervised and semi-supervised methods
- **Applications**: translation, information extraction, dialogue

In each section, we will cover linguistic issues, computational representations, and statistical techniques.
You will build software that put these ideas into practice.

Course goals

By the end of the semester, you should have learned:

- What are the range of linguistic phenomena we need to address to build useful language technology.
- How to select linguistic representations that are appropriate for the problem you want to solve.
- How to apply modern machine learning techniques to solve language processing problems.
- What are the existing resources (software and data) that can help.

You will also learn to read current research papers in the field.

For next time

- Homework 1: identify ambiguous sentences in the news
- Read Chapter 1 of **Linguistic Fundamentals for NLP**, if you haven't already.
- Next topic: **supervised learning**. Read:
  - **LxMLS notes** on linear algebra, probability, and classification
  - Optional: survey on word sense disambiguation