Welcome

NLP ML Web
Fall 2013
Andrew Rosenberg
TA/Grader: David Guy Brizan
Willie Sutton Jr.

Why do you rob banks?
Willie Sutton Jr.

Why do you rob banks?

Because that’s where the money is.
• > 4.5 billion webpages indexed
  http://www.worldwidewebsize.com/

• 400 million tweets per day
  http://articles.washingtonpost.com/2013-03-21/business/37889387_1_tweets-jack-dorsey-twitter

• 30 billion things shared per day
  http://blog.kissmetrics.com/facebook-statistics/

• > 4 million wikipedia pages

- Natural Language Processing has been using “big data” forever\(^1\).
- Machine Learning exists\(^2\) because of “big data”.

\(^1\) since the mid/late 1980s
\(^2\) in its current form
Moore’s Law

http://en.wikipedia.org/wiki/Moore’s_law
Today

• Course Overview
• Obligatory Syllabus Review
• Term Project Description
• Learning from Data
• Natural Language Processing
• Overview of Web Technology
Course Goals

• Provide an introduction to each of NLP, ML, The Web.

• Plus a look at topics that may not be covered elsewhere.

• No prior knowledge is *required.

• Demonstrate their interaction
Getting to know you

- Name
- Major: Linguistics, Computer Science, Electrical Engineering, other?
- CUNY, Columbia?
- Coursework and other background in each of NLP, ML, Web Technologies
- Prior Research and Future Interests
- Name again
Syllabus/Web Page

- Policies
- Slides
- Lecture Notes
- Assignments
There will be programming.

- Python, C++, Java
- “but it works in idle/msvs/eclipse/xcode”
- Compile using make, ant, maven or a simple series of commands and be run from a *nix command line.
  - cygwin, os x, linux flavors.
- Comfort in a unix environment may/will be helpful for
  - running external tools
  - processing text files
Technical Notes

• You are responsible\textsuperscript{1} to figure out how to
  • configure your work environment
  • run third-party tools
  • use external and third-party APIs
  • Google it. StackOverflow. Talk with classmates.
  • Make sure you have a functional programming environment \textit{today}.

\textsuperscript{1}\textit{i.e.} these are not appropriate topics for office hours and emails
Today

- Course Overview
- Obligatory Syllabus Review
- Term Project Description
- Learning from Data
- Natural Language Processing
- Overview of Web Technology
Pop Quiz

- one question
- ungraded
- anonymous
- five minutes
Pop Quiz

- Solve for $p$

$$ o = \frac{1 - p}{p} $$
Term Project

• “Cognitive Computing”
  • systems that learn and interact naturally with people to extend what either man or machine could do on their own.

• Collect related data from disparate sources.
  • Linking structured and unstructured data

• Use NLP, ML, & Web.

• Bonus: interact with a user or some external resource to assess, refine and/or improve performance.
Sample Projects

• Question Answering
• Specialized Search
• Wikification
• Meta reviews/Recommendations
• Fraud Detection
• Sentiment Analysis
• Trend Report Generation
• Linking Social Media to other related Media
Project Evaluation

- 15 - Innovative use in each of
  - NLP, ML, Web
- 30 - Technical Merit & Correctness
- 10 - Evaluation
- 10 - Written Report: Clarity & Thoroughness
- 10 - Code Documentation: Clarity & Thoroughness
- 10 - Final Presentation
- 5 - Proposal Presentation
- 5 - Written Proposal: Clarity & Concreteness
- 5 - Instructor’s Discretion for exceptional accomplishment in any of the above.
Today

- Course Overview
- Obligatory Syllabus Review
- Term Project Description
- Learning from Data
- Natural Language Processing
- Overview of Web Technology
Learning from Data

- Data
  - Learning Algorithm
    - ≥
    - Programmer or Expert
  - Behavior
- Data
  - Behavior
Major Tasks

- Regression
  - Predict a numeric value given “other information”
- Classification
  - Predict a categorical value given “other information”
- Clustering
  - Identify groups of similar entities.
- Learning Feature Representations
  - What’s the best way to describe this data?
- Evaluation
Classification
Classification
Classification
Classification
Regression

\[ t \]

\[ x \]

Monday, September 9, 13
Regression
Regression
Clustering
Clustering
Clustering
Mechanisms of Machine Learning

- Statistical Estimation
- Numerical Optimization
- Theoretical Optimization
- “Loss Functions”
- Feature Manipulation
- Similarity Measures
Math in Machine Learning

- Probability
- Statistics
- Calculus
  - Vector Calculus
- Linear Algebra
Machine Learning on the Web

- Pick a web site.
- Does it use machine learning?
- Could it?
Machine Learning on the Web

- Google, Bing, Yahoo
- Search
- Ad Delivery
- “Did you mean?”
- Similar Queries
- Info box.
Machine Learning on the Web

- GMail
- Search
- Ad Delivery
- “Did you mean?”
- Suggested Recipients
- Spam Detection
Machine Learning on the Web

- Twitter
  - Trending
  - Do you also know?
  - Promoted Tweets
Machine Learning on the Web

- Amazon
- Search
- Recommended products
- Amazon Web Services
- Supply Chain prediction
Machine Learning and The Web

• Facebook
  • Ads
  • Which status messages show up in your feed?
  • Which messages get delivered?
• Suggested Friends
Machine Learning and The Web

- YouTube
- Ads
- Transcription
- Similar Videos
Machine Learning and The Web

- Customer Service chat bots.
- Metacritic
- Netflix
- Pandora, Spotify, StumbleUpon

- How many of these use Language?
  - How many use language plus something else?
  - How many use multiple sources?
Today

- Course Overview
- Obligatory Syllabus Review
- Term Project Description
- Learning from Data
- Natural Language Processing
- Overview of Web Technology
Natural Language Processing

• Question Answering
• Machine Translation
• Sentiment Analysis
• Automatic Summarization
• Information Extraction
• Search
• (Spoken) Dialog Systems
NLP Machinery

- Part-of-speech tagging
- Parsing
- Language modeling
- Named-entity recognition
- Coreference Resolution
- Word Sense disambiguation
- Word Representations
That’s not how a human processes language.
Natural Language Processing

That’s not how a human processes language.

...no, it’s not.
Today

- Course Overview
- Obligatory Syllabus Review
- Term Project Description
- Learning from Data
- Natural Language Processing
- Overview of Web Technology
Web Technology

- Data, Data, Data
- Software-as-a-Service
- Hack everything

You

The obsessives

Users

“Content”

“Information”

“The Cloud”

Monday, September 9, 13
Web APIs

- Google
- Twitter
- Facebook
- Mechanical Turk
- Ads
Structured web data

- Wikipedia
- Google N-grams
- Freebase
- Yelp
Without an API

- urllib, wget, curl
- beautiful soup
- Chrome Developer

- Vine (yet...)
- Trip Advisor
- Metacritic
Generating Data

- Each of these allow posting data to the web
  - Twitter API
  - Facebook API
  - Youtube API
  - Blogger API
  - Selenium
Amazon Web Services

• Hardware-as-a-Service
• EC2
• Hadoop/MapReduce

• Instagram
  • http://instagram-engineering.tumblr.com/post/13649370142/what-powers-instagram-hundreds-of-instances-dozens-of
The End

- Lecture Notes are posted as they are written.
- Homework 0 is assigned.
- Proposal Presentations are in 4 weeks.
- Sign up for mailing lists:
  - [http://nlpatcuny.cs.qc.cuny.edu/mailinglist.html](http://nlpatcuny.cs.qc.cuny.edu/mailinglist.html)