Prof: Dan Weld
- Most lectures, concepts, perspective.

TA: Jessica Leung
- Project details

Expectations:
- Project (multiple parts, on time!)
- Reading (papers, web - no formal text)
- Class participation / development

Caveat: Life on the cutting edge

My Background
- Research on Intelligent Internet Systems [1991-
  - Internet Softbot
    - Discover Award Finalist '95
  - Webcrawler
    - By Brian Pinkerton
  - Metacrawler & Shopbot
    - Basis for Netbot Inc.
  - Mulder
    - First automated WWW question answerer
  - KnowItAll
    - Massive, autonomous information extraction
  - Intelligence in Wikipedia Project

Background Continued
- Co-founded
  - Netbot (Jango)
  - AdRelevance
  - Nimble Technology
  - Asta Networks

Leaves of absence
- VP Engineering at Netbot
- Venture Partner w/ Madrona Venture Group.

Incredible shortage of software engineers!
Dearth of training

Your Background?
- Classes?
  - 444, 446, 451, 461, 473, 490H
- Concepts?
  - Threads, race condition, deadlock
  - Naive Bayes classifier
  - Hybrid hash join algorithm
  - Precision, recall
- Programming Background?
  - Ruby, .NET, XML, admin own webserver
454 Topics

- Information Retrieval
- Search Engines
  - Crawling, Indexing, Query Processing, Ranking
  - Pagerank, Interfaces
- Text Categorization & Clustering
- Information Extraction
  - Machine Learning
- Internet Advertising
- Security, Cryptography, Malware
- Social Networks
- Temporal Web
- Special Topics

Course Outcomes

- After this course, you should know:
  - How search engines work
  - How to build information extraction systems
  - How to ensure a web site scales
  - How Amazon generates personalized recommendations
  - Cryptography fundamentals
  - Other cool stuff
- Focus: search! (why?)

Why Search?

- A billion or so searches per day...
- Boost to productivity
  - Intellectual & economic
- Search is (still) 'hot'
  - Google, Amazon, Ebay, Farecast
  - Search for/in books, products, music, people, ...
- Fascinating research problem.
- You can learn to be a something of a search expert in one quarter!

What is "Information Extraction"

As a task: Filling slots in a database from sub-segments of text.

Why Information Extraction

- Next-Generation Search
  - People
    - Zoominfo
  - Flipdog
  - Intelius
  - Research Papers
    - Citeseer
  - Google scholar
  - Product search
  - Question Answering

What is "Information Extraction"

As a task: Filling slots in a database from sub-segments of text.

October 14, 2002, 4:00 a.m. PT
For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels—the coveted code behind the Windows operating system—to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That’s a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying...
Example

...Continued

...Continued Some More

CiteSeer vs. Scholar

Grading
- 85% Project (Staged in Parts)
  • Part artifact
  • Part writeup
    - Clear and concise explanation / justification
    - Experimentation
  • Part presentation
- 15% Class participation

Capstone Projects
- Done in Group
  - Why?
- Topics
  - Roll your own
  - Or see me
Start with Concrete Problem

- Text Classification
- Corpus of Wikipedia pages
  - E.g., scientist, writer, author, university
- You’ll use machine learning to construct
  - Program which outputs the ‘type’ of the page
- Details online
  - Done in pairs
  - Due 10/13

Project Possibilities

- Extract Facts from Wikipedia
  - Or recipes, or ...
- Build Ontology of Products & Attributes
- Mine product reviews for attribute valence
- Or suggest something different

Teams & ideas settled by 10/13

Last Quarter’s Projects

- Craigslist++
- University Search
- Twitter Feedrank
- Apartment Listing & Aggregation
- Webcam Identification & Search
- Trail / Hike Search
- Seattle Event Finder
- Automatic Stock Investor

Traditional, Supervised I.E.

Raw Data

Labeled Training Data

Learning Algorithm

Extractor

Kylin: Self-Supervised Information Extraction from Wikipedia

From infoboxes to a training set

Clearfield County was created in 1804 from parts of Huntingdon and Lycoming Counties but was administered as part of Centre County until 1812. Its county seat is Clearfield.

2,972 km² (1,147 mi²) of it is land and 17 km² (7 mi²) of it (0.56%) is water.

As of 2005, the population density was 28.2/km².
What This Course Is Not

• We won’t:
  - Teach you how to be a web master
  - Teach all the latest x-buzzwords in technology
    - XML/SOAP/WSDL
    - (okay, may be a little).
  - Teach web/javascript/java/jdbc... programming

Warning

• No textbook
• Large project component
• Poorly documented, unstable systems
• Field changes quickly
  - Each year is essentially a new course
• Need students to help debug class!

Ancient History

• Pre-history: Dewey Decimal system
  - Bizarre medieval rituals performed by hand

• 1960: Ted Nelson → Xanadu
  - Hypertext vision of WWW
  - Why did it fail?
  - Focus on copyright issues
  - Still a thorny problem
  - Focus on stable, bidirectional links
  - "Trying to fix HTML is like trying to graft arms and legs onto hamburger" -- Ted Nelson

1961 Kleinrock paper on packet switching
Contrast with phone lines - circuit switched.

Paleolithic Era

1965 Gordon Moore proposes law
1966 Design of ARPAnet
1968 Doug Engelbart:
The first WIMP

1969 First ARPAnet message
  UCLA -> SRI
1970 ARPAnet spans country, has 5 nodes
1971 ARPAnet has 15 nodes
1972 First email programs, FTP spec

The Personal Computer Era

1974 Intel launches 8080;
  TCP design
1975 Gates/Allen write Basic - Altair 8800
1976 Jobs/Wozniak form Apple Computer
  111 hosts on ARPAnet
1979 Visicalc
1981 Microsoft has 40 employees; IBM PC
1984 Launch of Macintosh
1986 Microsoft goes public

Internet Ramps Up

1983 ARPAnet uses TCP/IP, Design of DNS
  1000 hosts on ARPAnet
1985 Symbolic.com first registered domain name
1989 100,000 hosts on Internet
1990 Cisco Systems goes public
  Tim Berners-Lee creates WWW at CERN
Web Search Pre-History

- 1950s: "Information Retrieval" (IR) term coined
- 1960s-70s: SMART system, vector space model, Gerald Salton (Cornell) father of IR
- 1980s: Proprietary document DBs
  - (Lexis-Nexis, Medline)
- 1990: Archie (index file names, anon. ftp)
- 1991: Gopher (menus, links to servers)
- 1992: Veronica (index of menu items on gophers)
- 1993: Jughead (keyword + boolean search)
  - Rapid evolution, but what is missing?

Modern History of Search

- 1993: WWW Wanderer (first crawler)
- 1994: WebCrawler, Lycos (1st widely-used SEs)
  - WebCrawler was a UW class project by Brian Pinkerton
- 1994: Yahoo directory (Stanford; founded ‘95)
  - Amazon founded
  - Netscape founded (90% mkt share → 1%)
- 1995: Ebay
  - MetaCrawler (1st major meta-SE)
  - UW Master’s thesis by Erik Selberg

Discovery of the Biz Model

1996: Flash by Macromedia
  - later acquired by Adobe
1997: goto.com
  - "sponsored links" pay-per-click
  - AskJeeves
    - manually-powered question answering
  - Netbot
    - comparison-shopping search
1998: Open directory launched
  - Google, pagerank algorithm
  - Paypal founded

Turn of the Millennium

- 1999: becomes dominant browser
  - Napster starts operation
  - Search Engines → portals (Yahoo, Excite)
    - "Search is a commodity"
- 2000: Flipdog
  - Commercial information extraction
- 2001: Bittorrent protocol (soon 35% of internet)
  - Ascendance of Google
    - "Search is nirvana"
- 2002: IE peaks at 90% market share

Approaching the Present

- 2003: Skype released
- 2004: Facebook founded
  - Social news (Digg)
- 2005: Youtube founded
  - 9.5 B videos shown per month
  - 33 months after founding!
- 2006: Twitter founded
- 2007: Google Streetview
  - Apple iPhone
- 2009: Facebook 200M users

Future of the Net

- Domination of Mobile Devices (cellphone, etc)
- Link-Spamming (Arms race to bias SE ranking)
- Local Search, Digital Earth
- Image & Video search
- Social news (Digg / Twitter)
- Crowd Sourcing
- What else?
**Mechanical Turk**

Built in 1770 by Wolfgang von Kempelen

---

**Observations**

- Internet/Web *evolved* - it wasn’t created
- Scalability beats structure
  - search engines over directories
  - Web over hypertext
- “We are 10 seconds from the Big Bang”
  - John Doerr

---

**Adoption**

Facilitating Innovation

- Newer technologies taking hold at double or triple previous rates
- The pace of innovation is increasing

---

**Accelerating**

Years to Reach 50MM Users:
- Radio: 38
- TV: 13
- Cable: 10
- Internet: 5

---

**And now?**

**Number of Facebook Users**

- 2004: 1,000,000
- 2008: 500,000,000
For Next Time

- **Add yourself to mailing list**
  - We'll send out a key email tomorrow
  - Be sure to get it!

- **Think about ps1**
  - Form a group of 2 people

- **Think about project**
  - Form a group of 4 people

### 33 months after founding

<table>
<thead>
<tr>
<th>Property</th>
<th>Videos Viewed (MM)</th>
<th>Share (%) of Videos</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Internet</td>
<td>9,491</td>
<td>100.0%</td>
</tr>
<tr>
<td>Google Sites</td>
<td>2,966</td>
<td>31.3%</td>
</tr>
<tr>
<td>Fox Interactive Media</td>
<td>4,9</td>
<td>4.4%</td>
</tr>
<tr>
<td>Yahoo Sites</td>
<td>328</td>
<td>3.5%</td>
</tr>
<tr>
<td>Viacom Digital</td>
<td>245</td>
<td>2.6%</td>
</tr>
<tr>
<td>Time Warner Network</td>
<td>184</td>
<td>1.9%</td>
</tr>
<tr>
<td>Microsoft Sites</td>
<td>181</td>
<td>1.9%</td>
</tr>
<tr>
<td>Disney Online</td>
<td>96</td>
<td>1.0%</td>
</tr>
<tr>
<td>ABC.com</td>
<td>88</td>
<td>0.9%</td>
</tr>
<tr>
<td>ESPN</td>
<td>87</td>
<td>0.9%</td>
</tr>
<tr>
<td>Break</td>
<td>47</td>
<td>0.5%</td>
</tr>
</tbody>
</table>