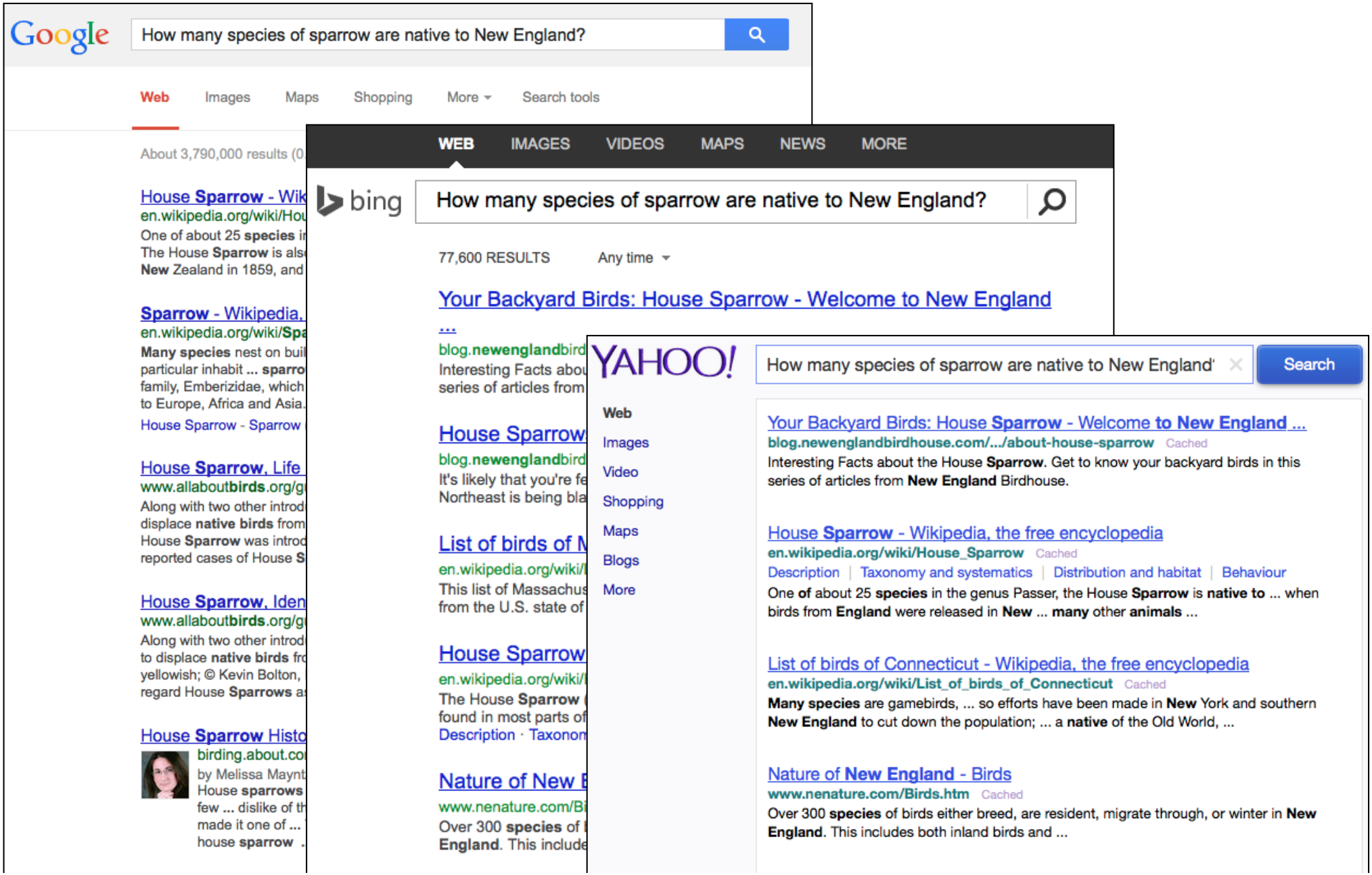
# Information Retrieval
## CS6200

Jesse Anderton
College of Computer and Information Science
Northeastern University

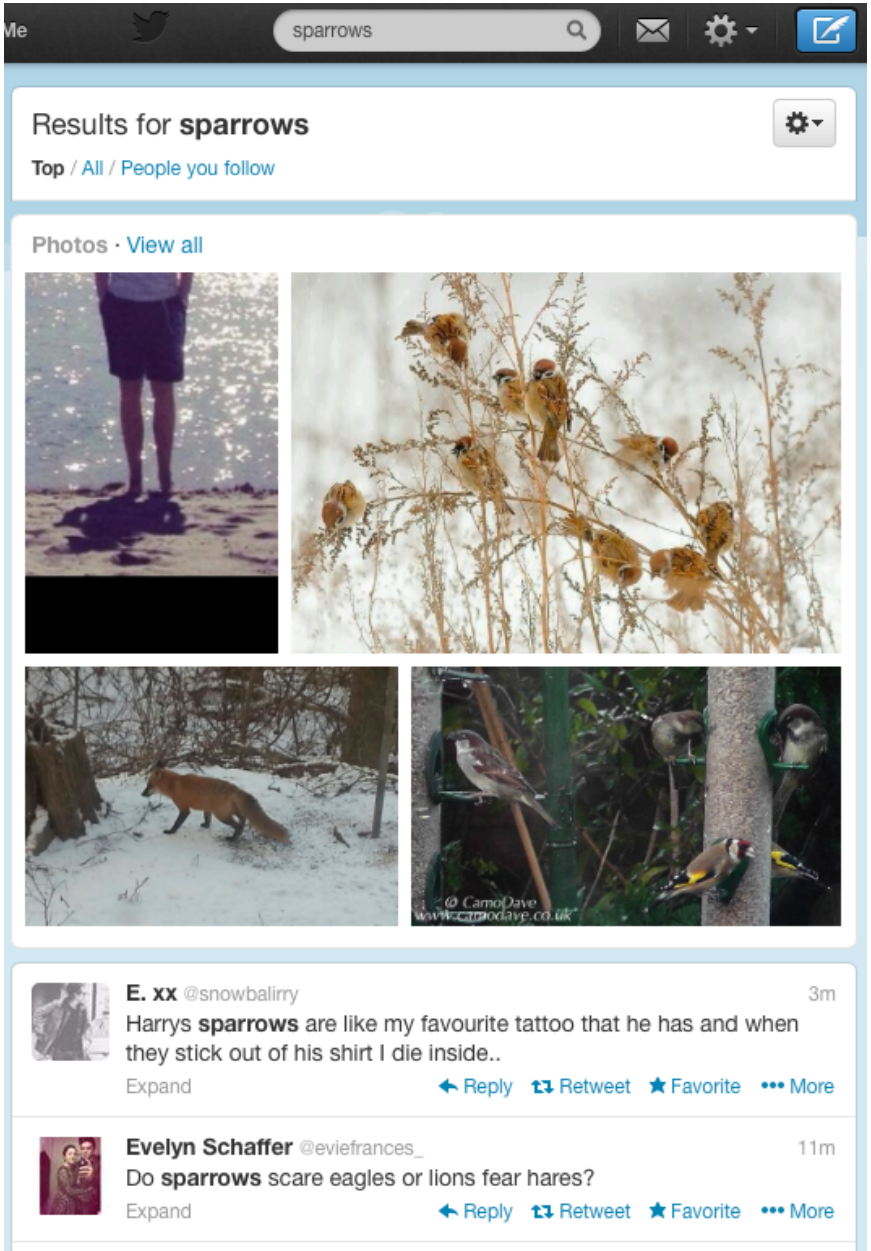
# What is Information Retrieval?

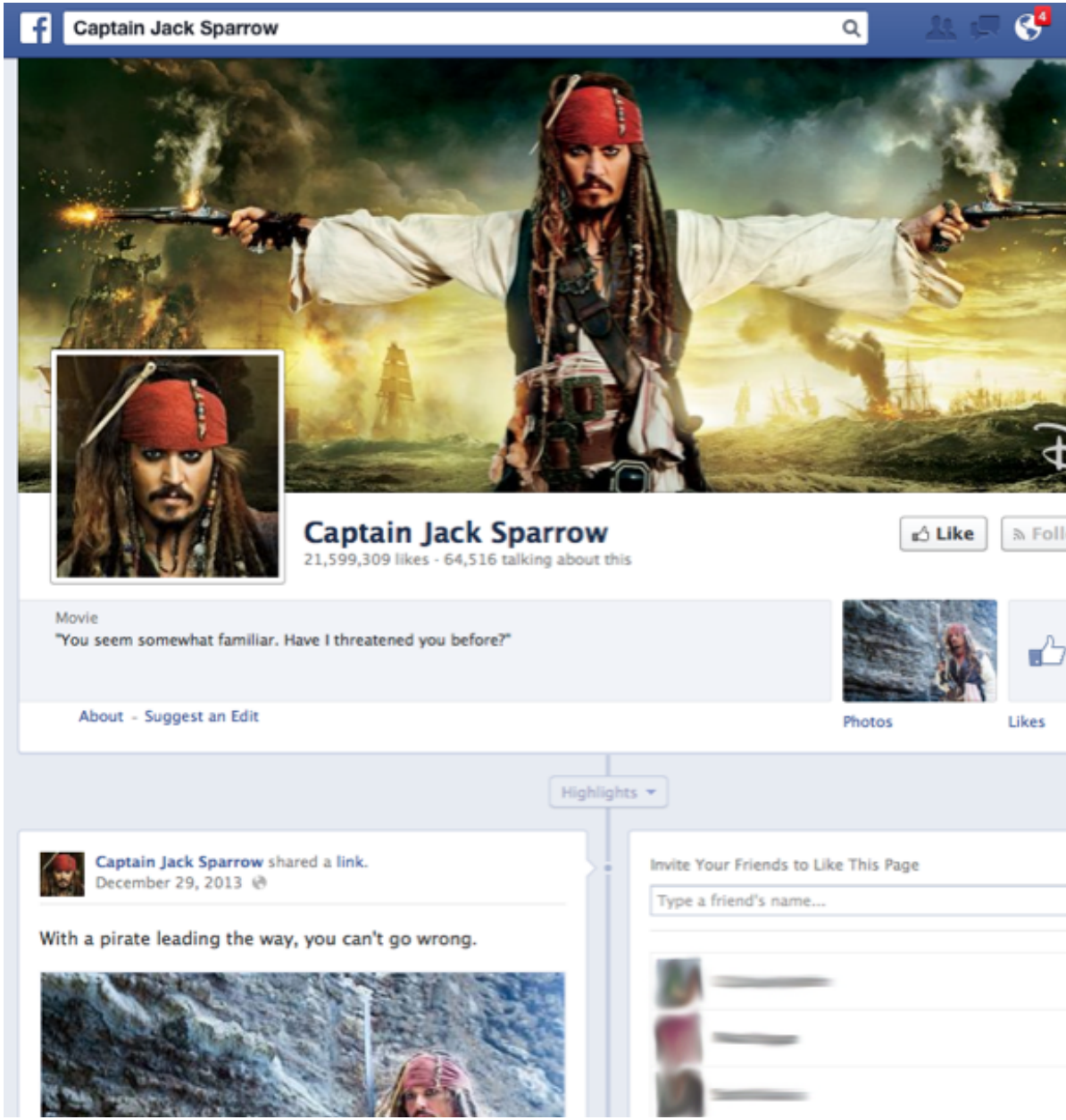
- You have a collection of documents
  - Books, web pages, journal articles, photographs, video clips, tweets, a weather database, …
- You have an *information need*
  - *“How many species of sparrow are native to New England?”*
  - *“Find a new musician I’d enjoy listening to.”*
  - *“Is it cold outside?”*
- You want the documents that best satisfy that need

# Web Search


Google
How many species of sparrow are native to New England?
Web
Images
Maps
Shopping
More
Search tools
About 3,790,000 results (0
House Sparrow - Wik
en.wikipedia.org/wiki/Hou
One of about 25 species i
The House Sparrow is als
New Zealand in 1859, and
Sparrow - Wikipedia,
en.wikipedia.org/wiki/Spa
Many species nest on buil
particular inhabit ... sparro
family, Emberizidae, which
to Europe, Africa and Asia.
House Sparrow - Sparrow
House Sparrow, Life
www.allaboutbirds.org/g
Along with two other introd
displace native birds from
House Sparrow was introc
reported cases of House S
House Sparrow, Iden
www.allaboutbirds.org/g
Along with two other introd
to displace native birds fr
yellowish; © Kevin Bolton,
regard House Sparrows a
House Sparrow Histo
birding.about.co
by Melissa Maynt
House sparrows
few ... dislike of th
made it one of ...
house sparrow
WEB
IMAGES
VIDEOS
MAPS
NEWS
MORE
bing
How many species of sparrow are native to New England?
77,600 RESULTS
Any time
Your Backyard Birds: House Sparrow - Welcome to New England
...
blog.newenglandbird
Interesting Facts abou
series of articles from
House Sparrow
blog.newenglandbird
It's likely that you're fe
Northeast is being bla
List of birds of M
en.wikipedia.org/wiki/
This list of Massachus
from the U.S. state of
House Sparrow
en.wikipedia.org/wiki/
The House Sparrow
found in most parts of
Description · Taxonom
Nature of New E
www.nenature.com/Bi
Over 300 species of
England. This include
YAHOO!
How many species of sparrow are native to New England
Search
Web
Images
Video
Shopping
Maps
Blogs
More
Your Backyard Birds: House Sparrow - Welcome to New England ...
blog.newenglandbirdhouse.com/.../about-house-sparrow Cached
Interesting Facts about the House Sparrow. Get to know your backyard birds in this series of articles from New England Birdhouse.
House Sparrow - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/House_Sparrow Cached
Description | Taxonomy and systematics | Distribution and habitat | Behaviour
One of about 25 species in the genus Passer, the House Sparrow is native to ... when birds from England were released in New ... many other animals ...
List of birds of Connecticut - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/List_of_birds_of_Connecticut Cached
Many species are gamebirds, ... so efforts have been made in New York and southern New England to cut down the population; ... a native of the Old World, ...
Nature of New England - Birds
www.nenature.com/Birds.htm Cached
Over 300 species of birds either breed, are resident, migrate through, or winter in New England. This includes both inland birds and ...

# Site-specific Search



# Product Search

amazon Try Prime | Jesse's Amazon.com | Today's Deals | Gift Cards | Sell | Help

New Year, New You in 2014
› Shop now

Shop by Department | Search | All | sparrows | Go | Hello, Jesse Your Account | Try Prime | 0 Cart | Wish List

**Departments**
**Books**
- Bird Watching
- Outdoors & Nature Reference
- Ornithology
- Nature & Wildlife Photography
- Ecology
- + See more...

**Movies & TV**
- Documentary
- Movies
- DVD
- Drama

**Kindle Store**
- Bird Watching

+ See All 32 Departments

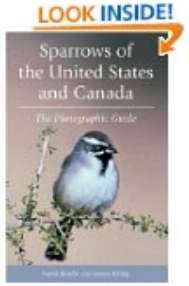
**Eligible for Free Shipping**
- Free Shipping by Amazon

**Book Format**
- Paperback
- Hardcover
- Kindle Edition
- Audible Audio Edition
- Audio CD
- Audio Cassette
- MP3 CD

**Book Language**
- ☐ English
- ☐ French
- ☐ German
- ☐ Spanish

"sparrows"

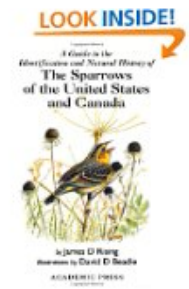
Showing 1 - 16 of 58,641 Results | Choose a Department to enable sortin




**Sparrows of the United States and Canada: The Photographic Guide** by David Beadle and James D. Rising (Nov 30, 2001)

~~$29.95~~ **$19.59** Paperback Prime
Order in the next **4 hours** and get it by **Thursday, Jan 9.**
More Buying Choices - Paperback
**$16.98** new (19 offers)
**$15.60** used (21 offers)

★★★★★ (24)
FREE Shipping on orders over $35
Sell this back for an Amazon.com Gift Card
**Books:** See all 9,568 items




**A Guide to the Identification and Natural History of the Sparrows of the United States and Canada (Natural World...** by James D. Rising (Sep 4, 1996)

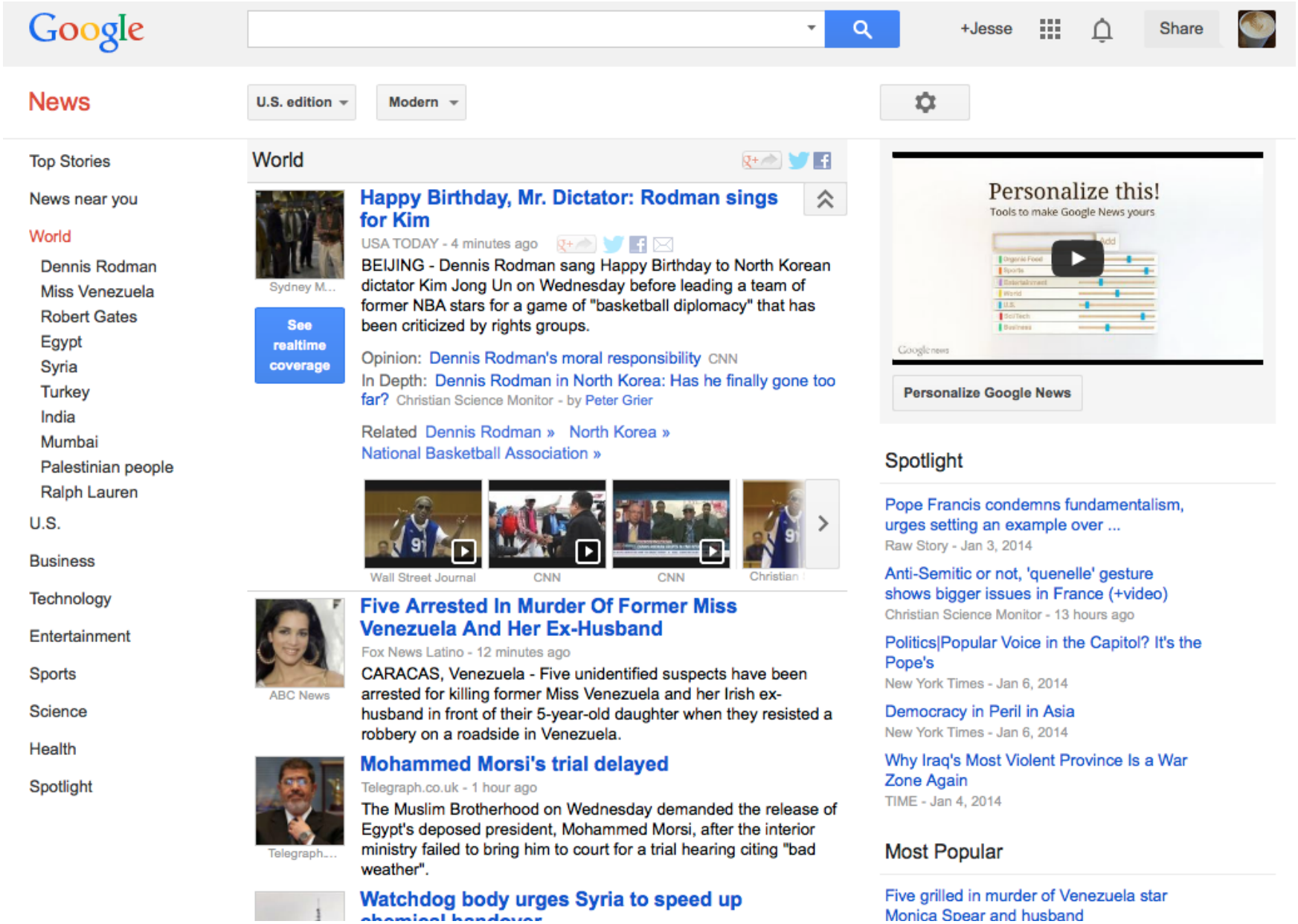
**$28.18** new (19 offers)
**$29.97** used (23 offers)
**$30.00** collectible (1 offer)

★★★★½ (3)
Sell this back for an Amazon.com Gift Card
Other Formats: Hardcover
**Books:** See all 9,568 items

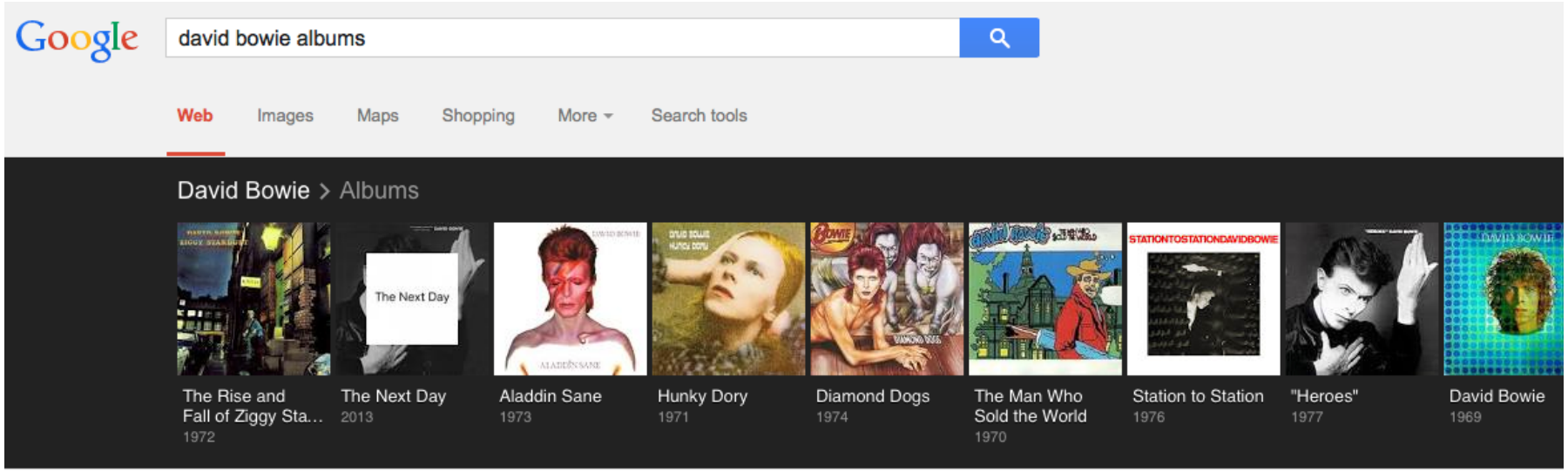

**Sparrows** 1999 Unrated
from **$57.90** DVD

★★★★★ (13)
Runtime: 1 hr 24 mins
Starring: Mary Pickford, Roy Stewart, et al.
Directed by: Tom McNamara and William Beaudine
Trade in this item for an Amazon.com Gift Card
**Movies & TV:** See all 319 items

# But also grouping related documents



# And mining the web for knowledge




Category:**David Bowie albums** - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Category:**David_Bowie_albums**
This category contains **albums** by **David Bowie**. See also categories: **David Bowie** songs, Tin Machine **albums**, and **David Bowie album** covers ...

**David Bowie** discography - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**David_Bowie**_discography
Jump to Studio **albums** - 1972, **The Rise and Fall of Ziggy Stardust and the Spiders from Mars**. Released: 6 June 1972; Label: RCA; Format: LP. 5, 75, 11 ...
The Singles Collection - Best of Bowie - Tin Machine - Tin Machine II

**David Bowie** (1967 **album**) - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**David_Bowie**_(1967_**album**)
**David Bowie** is the debut **album** by British musician **David Bowie**, released in 1967 on Deram Records. Its content bears little overt resemblance to the type of ...

**David Bowie** - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**David_Bowie**
**David-Bowie** Chicago 2002-08-08 photoby Adam-Bielawski-cropped. ... " Starman" and the **album The Rise and Fall of Ziggy Stardust and the Spiders from Mars**. ... Bowie's latest studio **album The Next Day** was released in March 2013. ..... **Hunky Dory** (1971) found Visconti, Bowie's producer and bassist, supplanted in both ...

**David Bowie** Discography--The List of **David Bowie Albums**
www.brianhartzog.com/**david-bowie**/**david-bowie**-discography.htm

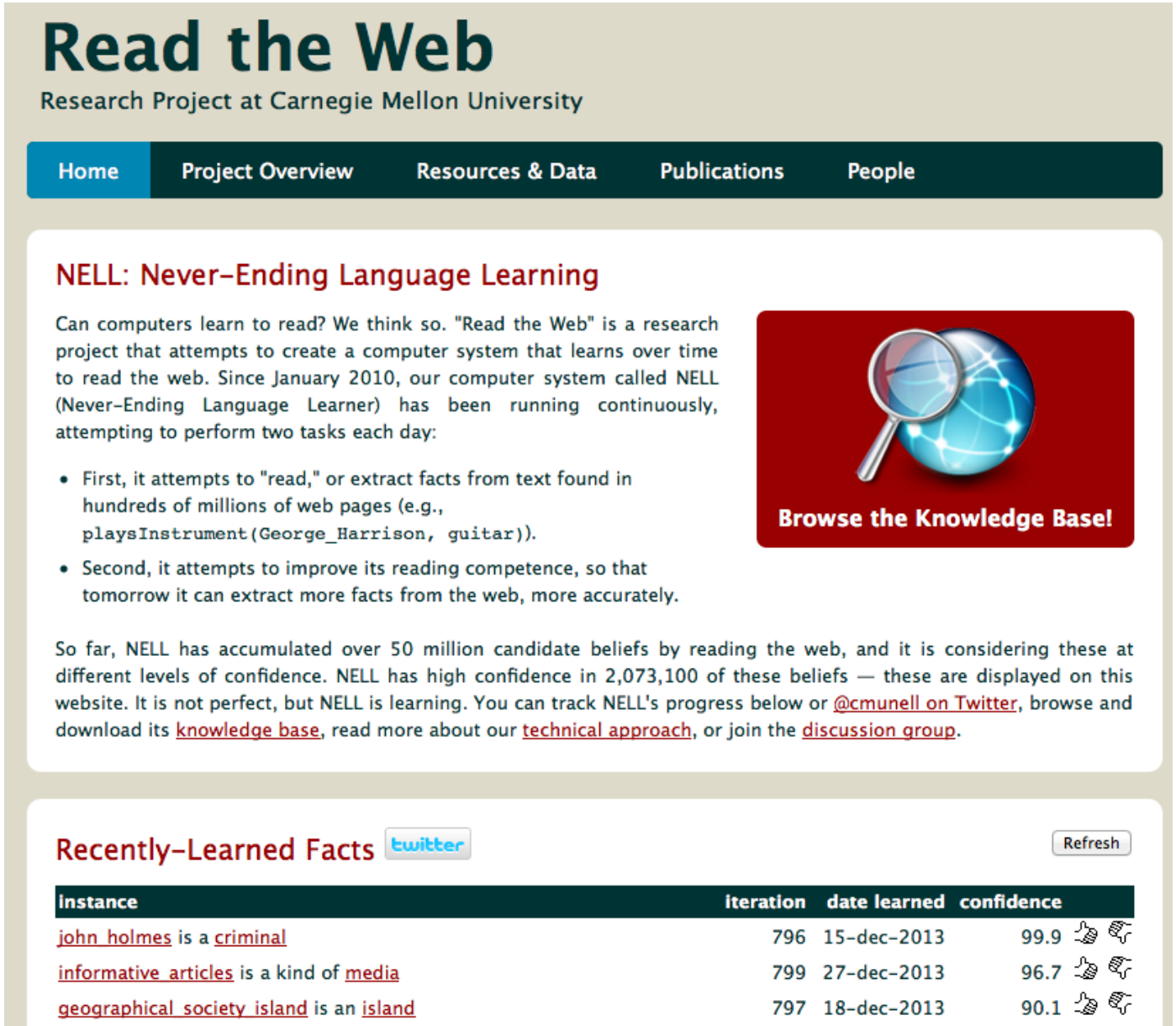

## David Bowie

Musician

David Robert Jones, known by his stage name David Bowie, is an English musician, singer-songwriter, producer, actor and arranger.
Wikipedia

**Born:** January 8, 1947 (age 67), Brixton, London, United Kingdom
**Spouse:** Iman Abdulmajid (m. 1992), Angela Bowie (m. 1970–1980)
**Children:** Alexandria Zahra Jones, Duncan Jones
**Movies:** Labyrinth, The Prestige, The Man Who Fell to Earth, More

# And learning how to read



# Read the Web

Research Project at Carnegie Mellon University

Home | Project Overview | Resources & Data | Publications | People

## NELL: Never-Ending Language Learning

Can computers learn to read? We think so. "Read the Web" is a research project that attempts to create a computer system that learns over time to read the web. Since January 2010, our computer system called NELL (Never-Ending Language Learner) has been running continuously, attempting to perform two tasks each day:

- First, it attempts to "read," or extract facts from text found in hundreds of millions of web pages (e.g., `playsInstrument(George_Harrison, guitar)`).
- Second, it attempts to improve its reading competence, so that tomorrow it can extract more facts from the web, more accurately.

Browse the Knowledge Base!

So far, NELL has accumulated over 50 million candidate beliefs by reading the web, and it is considering these at different levels of confidence. NELL has high confidence in 2,073,100 of these beliefs — these are displayed on this website. It is not perfect, but NELL is learning. You can track NELL's progress below or @cmunell on Twitter, browse and download its knowledge base, read more about our technical approach, or join the discussion group.

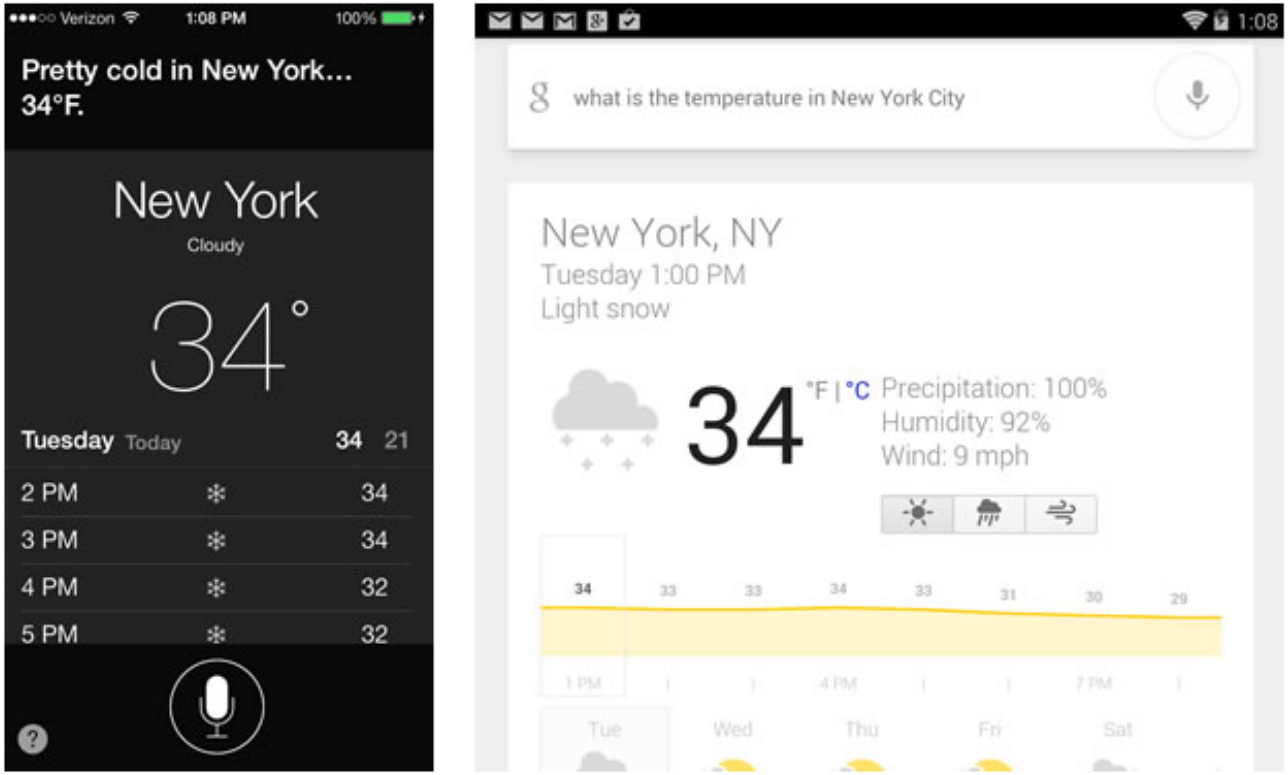
## Recently-Learned Facts

twitter

Refresh

| instance | iteration | date learned | confidence |
|---|---|---|---|
| john_holmes is a criminal | 796 | 15-dec-2013 | 99.9 |
| informative_articles is a kind of media | 799 | 27-dec-2013 | 96.7 |
| geographical_society_island is an island | 797 | 18-dec-2013 | 90.1 |

# And answering everyday questions




Siri and Google Now.
(Credit: Screenshot by Lance Whitney/CNET)

http://news.cnet.com/8301-13579_3-57615135-37/siri-battles-google-now-in-new-contest/

That’s a lot of stuff.

Where do we start?

# Course Goals

- To help you understand the fundamentals of search engines.
  - How to crawl, index, and search documents
  - How to evaluate and compare different search engines
  - How to modify search engines for specific applications
- To provide broad coverage of the major issues in information retrieval
- As time permits, to take a closer look at particular applications of Information Retrieval in industry

# Course Materials

- Suggested books:
  - Search Engines: Information Retrieval in Practice, by Croft, Metzler, and Strohman
  - Introduction to Information Retrieval, by Manning, Raghavan, and Schütze
    - Available for free online!
- Occasional research papers may be suggested for further reading.

# Grading

- If you focus on learning the material, you'll probably get an A
- 40%: 2-3 Homework assignments
  - Some coding, some math, some system design
- 60%: 3 Projects
  - Coding, plus evaluating and explaining your results
  - *A few* of you can do your own final project in place of the third project. Come and see me later in the course if you're interested.
- Quizzes
  - Extra credit only. Meant to measure your comprehension and my lecturing.
  - Probably posted on Piazza.

# Late Policy

- Assignments are due by 10pm on the announced due date (generally the day before a lecture)
- You may turn in one assignment up to four days late without asking in advance or providing a reason.
- After your first late assignment, you will be penalized by 20% per day late. If you feel you have a good reason to submit an assignment late, please talk to me me *in advance*.
- I will be showing correct answers a week after the due date, so I will not accept any assignments after that.

# Collaborating

- What do you do if you need help?
  - Post a question on Piazza
  - Come to office hours, or ask for an appointment
  - Talk to your friends, and report in your assignment who you spoke with
- You are responsible for writing and understanding everything you submit
  - Don't prioritize getting a grade over understanding the material. We are looking for cheaters, both manually and using plagiarism detection software.
  - If you copy another student's work, or if another student copies yours, expect to be caught, to receive zero credit for the assignment, and to be reported to the university.
  - But if you are having a problem finishing an assignment, please come talk to me. I want to help you.

# Contacting Us

- Instructor: Jesse Anderton
  - jesse@ccs.neu.edu
  - Office Hours: Thursdays, 10am-12pm, 472 WVH
- TA: Maryam Bashir
  - maryam@ccs.neu.edu
  - Office Hours: Tuesdays, 10:00am-12:00pm 472 WVH
- TA: Ting Chen
  - tingchen@ccs.neu.edu
  - Office Hours: Mondays, 2:30-4:30pm 472 WVH
- Course website: http://www.ccs.neu.edu/course/cs6200s14/
- Piazza: https://piazza.com/ccs.neu.edu/spring2014/cs6200

# Course Topics

- Architecture of a search engine
- Data acquisition
- Text representation
- Information extraction
- Indexing
- Query processing
- Ranking
- Evaluation
- Classification and clustering
- Social search
- More…

# A brief history of IR

## Let's start with Vannevar Bush, in the aftermath of WWII

This has not been a scientist's war; it has been a war in which all have had a part. The scientists, burying their old professional competition in the demand of a common cause, have shared greatly and learned much. It has been exhilarating to work in effective partnership. Now, for many, this appears to be approaching an end. What are the scientists to do next?

There is a growing mountain of research. But there is increased evidence that we are being bogged down today as specialization extends. The investigator is staggered by the findings and conclusions of thousands of other workers—conclusions which he cannot find time to grasp, much less to remember, as they appear. Yet specialization becomes increasingly necessary for progress, and the effort to bridge between disciplines is correspondingly superficial.

Consider a future device for individual use, which is a sort of mechanized private file and library. It needs a name, and, to coin one at random, "memex" will do. A memex is a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory.

*As We May Think, Vannevar Bush. The Atlantic, Jul. 1, 1945.*

# A brief history of IR

- Vannevar Bush in 1945 imagined a system involving cards and photography.
- Suddenly, computers.
- Search of digital libraries was one of the earliest tasks computers were used for.
- By the 1950s, rudimentary search systems could find documents that contained particular terms.
- Documents were ranked based on how often the specific search terms appeared in them — *term frequency* weighting

# A brief history of IR

- In the 60s, new techniques were developed that treated a document as a *term vector*.
  - Using a “bag of words” model: assuming that the number of occurrences of each term matters but term order does not
  - A query can also be represented as a term vector, and the vectors can be compared to measure similarity between the document and query
- Work also started on clustering documents with similar content
- The concept of *relevance feedback* was introduced: the best few documents are assumed to be matches, and documents which are similar to them are assumed to also be relevant to the original query.
- Some of the first commercial systems appeared in the 60s, sold to companies who wanted to search their private records

# A brief history of IR

- Before the Internet, search was mainly about finding documents in your own collection
- The emphasis was largely on *recall* — making sure you find every relevant document
- Documents were mainly text files, and did not contain references to other documents
- Just after the Internet, this was all changed
  - Collection sizes jumped to billions of documents
  - Documents are structured in networks, providing extra relevance information, and often have other useful metadata (e.g. how many FaceBook likes?)
  - You can't possibly know what's in every document
  - A "document" can be pages long or just 120 characters, or could be an image or video clip, a file download, an abstract fact, or something else entirely
  - You usually care more about *precision* — making sure your first few results are relevant — because people only look at the first few results (except for when they don't…)

# Challenges of IR

- Text documents are generally free-form
  - The metadata is there, but you have to find it
  - Most web pages contain lots of extra content — ads, navigation bars, comments — that might or might not be of interest
  - Spam filtering is hard
- Searching multimedia content has its own challenges
  - What are the features? How do you extract them?

# Challenges of IR

- Running a query is hard
  - You have less than one second to search the full text of billions of documents to find the best ten matches
  - …and the user only gave you two or three words
  - …and one was misspelled, and one was “the”
  - …and maybe throw a good relevant ad in, so you can pay the bills
- Working at web scale means massive distributed systems, sub-linear algorithms, and careful use of heuristics

# Challenges of IR

- Comparing the query text to the document text and determining what is a good match is the core issue of information retrieval
  - Exact matching of words is not enough
  - Many different ways to write the same thing in a “natural language” like English
  - e.g., does a news story containing the text “bank director in Amherst steals funds” match the query “bank scandals in western mass?”
  - Some stories will be better matches than others

# Relevance

- What is relevance?
- Simple (and simplistic) definition: A relevant document contains the information that a person was looking for when they submitted a query to the search engine
- Many factors influence a person's decision about what is relevant: e.g., task, context, novelty, style

# Relevance

- Retrieval models define a particular view of relevance based on some idea of what users want
- Ranking algorithms used in search engines are based on retrieval models
- Most models are based on statistical properties of text rather than deep linguistic analysis
  - i.e., counting simple text features such as words instead of parsing and analyzing the sentences

# Users and Information Needs

- Search evaluation is user-centered
- Keyword queries are often poor descriptions of actual information needs
- Interaction and context are important for understanding user intent
- Query refinement techniques such as query expansion, query suggestion, relevance feedback improve ranking

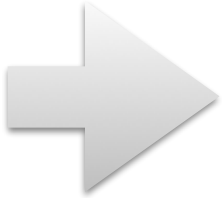
# Research and Industry

- A search engine is the practical application of information retrieval techniques to large scale text collections
- Web search engines are the best-known examples, but there are many others
- Open source search engines are important for research and development
  - e.g., Lucene, Lemur/Indri, Galago
- Researchers are focused on many, but not all, of the tasks that industry search engines care about

# Research and Industry

Research Tasks

- Relevance
  - Effective ranking
- Evaluation
  - Testing and measuring
- Information needs
  - User interaction



Search Engines

- Performance
  - Efficient search and indexing
- Incorporating new data
  - Coverage and freshness
- Scalability
  - Growing with data and users
- Adaptability
  - Tuning for applications
- Specific problems
  - e.g. Spam

# Search Engine Issues

- Performance
- Measuring and improving the efficiency of search
  - e.g., reducing response time, increasing query throughput,  increasing indexing speed
- Indexes are data structures designed to improve search efficiency
  - Designing and implementing them are major issues for search engines

# Search Engine Issues

- Dynamic data
- The “collection” for most real applications is constantly changing in terms of updates, additions, deletions
  - e.g., web pages
- Acquiring or “crawling” the documents is a major task
  - Typical measures are *coverage* (how much has been indexed) and *freshness* (how recently was it indexed)
- Updating the indexes while processing queries is also a design issue

# Search Engine Issues

- Scalability
  - Making everything work with millions of users every day, and many terabytes of documents
  - Distributed processing is essential
- Adaptability
  - Changing and tuning search engine components such as ranking algorithms, indexing strategies, interfaces for different applications

# Search Engine Issues

- Spam
- For web search, spam in all its forms is one of *the* major issues
- Affects the efficiency of search engines and, more seriously, the *effectiveness* of the results
- Proliferation of spam varieties
  - e.g. spamdexing or term spam, link spam, “optimization”
- New subfield called *adversarial IR*, since spammers are “adversaries” with different goals

# Further Reading

- Chapters 1 and 2 of *Search Engines* by Croft, Metzler, and Strohman

- As We May Think, Vannevar Bush, 1941

  http://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/

- The History of Information Retrieval Research, Croft and Sanderson, IEEE Xplore

  http://ieeexplore.ieee.org/xpls/icp.jsp?arnumber=6182576