HW 9: Make your own web page

Processing 1.5: File Export. 2.0 javascript mode

- Makes an index.html!
Everything that humanity knows is now online!!

- Newspapers
- Scientific Articles
- Books
- Encyclopedias (Wikipedia)
- Dictionaries and Thesauri (Wordnet, Framenet, Sentiwordnet, Freebase)
- Penn Treebank: a million words of parsed and semantically labelled news, books etc.
- Plus current and historic (10 years) opinions, reactions, emotions (opinion mining)

How do we find things we want to find?

Why the WWW is so brilliant

- WWW = The Servers + The Data
- All computers use one standard protocol (http) meaning that every computer in every country where people all speak different languages can communicate
- Publishing and accessing information is completely decentralized – no one limits what you put out or search for
- Critical mass of data => Too much personal Data?

Seeing Other People’s Digital Info

Message: WWW exploits one protocol, neutralizing differences at endpoints; the Internet’s universal medium lets us look at other people’s digital info.

Encoding Information: There’s more!

- Bits and bytes encode the information, but that’s not all
- Tags encode format and some structure in word processors
- Tags encode format and some structure in HTML
- Tags are one form of meta-data
- Meta-data is information about information
There are great resources out there

HTML and the Web

- The Web uses http:// protocol
- It's asking for a Web page, which usually means a page expressed in hyper-text markup language, or HTML
  - Hyper-text refers to text containing LINKS that allow you to leave the linear stream of text, see something else, and return to the place you left
  - Markup language is a notation to describe how a published document is supposed to look: what kinds of fonts, text color, headings, images, etc.

Basics of HTML #1

- Rule 1: Content is given directly; anything that is not content is given inside of tags
- Rule 2: Tags made of < and > and used this way:

<table>
<thead>
<tr>
<th>Attribute&amp;Value</th>
<th>Start</th>
<th>Content</th>
<th>End</th>
<th>Tag</th>
</tr>
</thead>
</table>

It produces: **This is paragraph.**

- Rule 3: Tags must be paired or "self terminated"
Basics of HTML #2

- Rule 4: An HTML file has this structure:

```html
<html>
  <head><title>Name of Page</title></head>
  <body>
    Actual HTML page description goes here
  </body>
</html>
```

- Rule 5: Tags must be properly nested
- Rule 6: White space is mostly ignored
- Rule 7: Attributes (style="color:red") preceded by space, name not quoted, value quoted

Example: myfirst.html

```html
<html>
  <head>
    <title>Making a Bake</title>
  </head>
  <body>
    <h1>Making a Bake</h1>
    <img src="cooking-ewan-isabel.jpg" alt="Kids Cooking" width="80%" />
    <p>Where do we get to eat? (?)</p>
  </body>
</html>
```

Example

- Write HTML in text editor: notepad++ or TextWrangler
- The file extension is .html; show it in Firefox or your browser

More on HTML (including good tutorials) at

- http://www.w3schools.com/html/default.asp

HTML Cheat Sheet: In Resources on Course page

- To put in an image (.gif, .jpg, .png), use 1 tag
  `<img src="cooking-ewan-isabel.jpg" alt="Kids Cooking" />`

- To put in a link, use 2 tags
  `<a href="http://users.soee.ucsc.edu/~maw">Prof. Walker's</a>`

- Hyper-text reference – the link
  - Anchor

- More on HTML (including good tutorials) at
  - http://www.w3schools.com/html/default.asp

HW 9: Make your own web page

Example: myfirst.html

- Making a Bake

  - Where do we get to eat? (?)

  ```html
  <html>
    <head>
      <title>Making a Bake</title>
    </head>
    <body>
      <h1>Making a Bake</h1>
      <img src="cooking-ewan-isabel.jpg" alt="Kids Cooking" />
      <p>Where do we get to eat? (?)</p>
    </body>
  </html>
  ```
Tagging and Meta Data

Meta-Data Describes Data

- Meta-data is data about data ... a description of what the data is
  - Knowing what the data is, as in the OED, allows us to process it better for users
  - Here's an example: Search OED for def of "binary"
    - Without meta-data, get 8,311 hits ... which one is the definition?
    - With meta-data, get each definition in order ... how?

Give Meta-data Using Tags

- We have seen tags in several forms
  - OED ... custom tags created for dictionary
  - HTML ... the tags "mark up" the content so it can be displayed; it's a simple use of tagging

More kinds of meta data all the time

- 'The semantic web'
- An "ontology" i.e. "taxonomy" of the kinds of things there are in the world
  - People, place, thing, animal, organization, country
- New tags within web pages
- Could be put there by people
- Could be programmatically identified by text processing algorithms (like what Watson Jeopardy uses)
What are you supposed to learn?

- HTML lets you programmatically indicate how a particular content should be displayed.
- It can be served up by any HTTP server anywhere in the world.
- Typically uses UTF-8 encoding to guarantee being able to be shown.
- Now, how do search engines use HTML info to do search?

How many different pages?

- Searching for “travel” in Google today yielded “About 3,770,000,000 results (0.33 seconds)”
- At least 40-50 billion pages (www.worldwidewebsize.com)

Important questions: What you should learn

- How can a search engine respond so fast?
- Does it find every relevant link?
- How does a search engine decide what gets listed first?
- If you try another search engine will you get the same result? If so, which is right? Which is better? Which is more authoritative?
- Are sponsored links better than “organic” links? Is the advertising necessary?
- What is the role of government? What should it be?

Indexing Everything: H.G. Wells in 1937

“There is no practical obstacle whatever now to the creation of an efficient index to all human knowledge, ideas and achievements, to the creation, that is, of a complete planetary memory for all mankind… The whole human memory can be, and probably in a short time will be, made accessible to every individual… This is no remote dream, no fantasy.”
As we may think: Vannevar Bush 1940’s

“The difficulty seems to be, not so much that we publish unduly … but rather that publication has been extended far beyond our present ability to make real use of the record. The summation of human experiences is being expanded at a prodigious rate, and the means we use for threading through the consequent maze to the momentarily important item is the same as was used in the days of square-rigged ships. … Our ineptitude in getting at the record is largely caused by the artificiality of systems of indexing.”

In the beginning

- Early catalogues both on the Internet (even before the WWW) and in print.
- “The Whole Internet Catalog” (1994)
- Yahoo was like a library index compiled by human editors.
- Search engines were invented in early 90s with the growing popularity of the web.

Sample Page: Agriculture

- Look up by topic, like a library index
- Where to go

The Library vs the Information Bazaar

- “Yellow pages”, directories, and catalogues
- The “Web” is not hierarchical
  - no structure like a library
- Catalogues are out - search engines are in.
- Why?

How many hosts are out there?

Big data is here
The Library vs the Information Bazaar

- “Yellow pages”, directories, and catalogues
- The “Web” is not hierarchical
  - no structure like a library
- Catalogues are out - search engines are in.
- But - search engines control what you see

“The search tools that help us find needles in the digital haystack have become the lenses through which we view the digital landscape. Businesses and governments use them to distort our picture of reality.”

Blown to Bits pg 110

Web search: It Matters How It Works

1. Gather information.
2. Keep copies.
3. Build an index.
4. Understand the query.
5. Determine the relevance of each possible result to the query.
6. Determine the ranking of the relevant results.
7. Present the results.

Page Rank Algorithms

- The “crown jewels” of search engines lie in their page rank algorithms.
- Factors include:
  - keywords in heading or titles
  - keyword only in the body text
  - site is “trustworthy”
  - links on this page are to relevant pages
  - links to this page are relevant
  - age of the page
  - quality of the text (e.g. absence of misspellings)

Its all free?? : Well no. Who Pays for What?

- Users could pay a subscription fee (early AOL and CompuServe)
- Web sites could pay for being indexed.
- The government could pay (taxes?).
- Advertisers could pay.

- And it matters who pays cause it affects how it works
Web search: It Matters How It Works

1. Gather information.
2. Keep copies.
3. Build an index.
4. Understand the query.
5. Determine the relevance of each possible result to the query.
6. Determine the ranking of the relevant results.
7. Present the results.

Search Engines

No one controls what’s published on the WWW ... it is totally decentralized
To find out, search engines crawl Web

- Two parts
  - Crawler visits Web pages building an index of the content (stored in a database)
  - Query processor checks user requests against the index, reports on known pages [You use this!]

Only a fraction of the Web’s content is crawled

1. Gather Information

- Spiders or web crawlers wander the web building indices
- Estimates range from .02% to 3% of information is indexed
- How often does a page get visited?
  - some frequently (daily whitehouse.gov), others rarely
  - Crawler keeps track of which pages change frequently
- How does the crawler find its way and not go in circles?
  - Logins keep bots/crawlers out.

How to crawl the Web:

- Begin with some Web sites, entered “manually”
- Select page not yet crawled; look at its HTML
  - For each keyword, associate it with this page’s URL
  - Harvest words from URL and inside <title> tags …
  - For every link tag on the page, associate the URL with the words inside of the anchor text, that is,
- Save all links and add to list to be crawled

2. Keep Copies

- Spider downloads the page as part of the “visit” in order to create the index.
- Search engine may “cache” the copy.
- Is this legal? What about copyright?
- But wait, browsing requires copying as well.

“(AFP) – Sep 15, 2011”

NEW YORK — Google and publishers told a US judge Thursday they are close to settling a lawsuit over the Internet giant’s controversial book-scanning project…

Net Result From Crawling A Page

- After crawling a page like users.soe.ucsc.edu/~maw
  - the crawler will associate many terms with the URL: reinforcement learning,
  - Terms from URL and anchor are more important in describing the page
  - Copies are cached (things never go away)
3. Build an Index

- list of terms and for each term a list of where it appeared
- more than just the terms
  - terms in bigger font might be more important
  - terms in the title might be more important
- must be very fast to lookup
- could be millions of entries (not just words, but names, special numbers, etc.) requiring Gigabytes of memory
- must fit in the computer’s memory

4. Understand the Query

- Steps 1-3 happen in “the background”
- Not much “understanding” in today’s search engines but that could change soon (QA lecture 3/12).
- Advanced search engine features help

49’ers beat Ravens vs. Ravens beat 49’ers

What about a business called “THE”?

Boolean Queries

Search Engine words are independent

- Words don’t have to occur together
- Use Boolean queries and quotes
- Logical Operators: AND, OR, NOT
  - monet AND water AND lilies
  - "van gogh" OR gauguin
  - vermeer AND girl AND NOT pearl

Make A Query: http://www.google.ca/advanced_search

- When Google gets the query
  - it “ands” the two lists together, finding URLs that are on both lists
  - it counts them up, records time, shows 10 hits

5. Determine Relevance

- “Recall” == what percentage of relevant documents are returned by the search?
- Simple relevance calculation
  - count the number of times each search word appears in the document, add them all up
  - Long documents get higher scores.
  - Uninteresting words like “the” contribute to the score.
  - All word occurrences are not equal (title words should count more).
  - Relevance is an IR term. Page Rank a Search idea
6. Determine Ranking: Page Rank

- Which of the relevant documents should be displayed first?
- Simple solution - put one with highest relevance score first.
- What if many have the same score?
- Are ones with the highest relevance score really the most important? What about the source of the document (e.g. NY Times vs some random blog post).

Page Rank Algorithms

- The “crown jewels” of search engines lie in their page rank algorithms.
- Factors include:
  - keywords in heading or titles (THE META DATA)
  - keyword only in the body text
  - site is “trustworthy”
  - links on this page are to relevant pages
  - links to this page are relevant
  - age of the page
  - quality of the text (e.g. absence of misspellings)

Page Rank

- You want the most likely hits ... how does Google show you what you want?
- Page Rank – a mechanism to estimate the importance of a page; pages are listed by page rank, highest to lowest.

Google Short Index & Long Index

- Known as the short barrels and the long barrels...
- Short index:
  - store the words in link texts that point to a page (inbound links!)
  - the words in a page’s title, and one or two other special things.
- The link text words are attributed to the target page, and not to the page that the link is on.
  - In other words, if my page links to your page, using the link text “Miami hotels”, then the words “Miami” and “hotels” are stored in the short index as though they appeared in your page, but they belong to my page. If 100 pages link to your page, using those same words as link text, then your page will have a lot of entries in the short index for those particular words.
- The long index is used to store all the other words on a page – its actual content.

Processing a query: short and long indices

- First try to get enough results from the short index.
- If you can’t get enough results, then use the long index to add to what they have.
- It means that, if they can get enough results from the short index – that’s the index that contains words in link texts and page titles – then they don’t even look in the long index where the actual contents of pages are stored. Page content isn’t even considered if they can get enough results from the link texts and titles index – the short index.
- Thus: link texts are very powerful for Google rankings.
  - Much more powerful than page titles, because a page can have the words from only one title in the short index, but it can have the words from a great many link texts in there.
  - Page titles and meta descriptions were the second most powerful ranking factors, because they are stored in the short index.
Google Query Algorithm

- In a first paper Larry Page and Sergey Brin gave their algorithm for processing a Google query

Processing a query: short and long indices

- First try to get enough results from the short index.
- If you can’t get enough results, then use the long index to add to what they have.
  - It means that, if they can get enough results from the short index – that’s the index that contains words in link texts and page titles – then they don’t even look in the long index where the actual contents of pages are stored. Page content isn’t even considered if they can get enough results from the link texts and titles index – the short index.
  - Thus: link texts are very powerful for Google rankings.
  - Much more powerful than page titles, because a page can have the words from only one title in the short index, but it can have the words from a great many link texts in there.
  - Page titles and meta descriptions were the second most powerful ranking factors, because they are stored in the short index.

What does Google use to search?

- An inbound link is simply a hyperlink that can also have an href description: the anchor text
- To put in a link, use 2 tags
  `<a href="http://users.soe.ucsc.edu/~maw">Prof. Walker’s </a>`
  - Hyper-text reference – the link  Anchor  End
- Inbound links are important because of the way that Google stores a page’s data, and the way that they process a search query.

Google’s PageRank Algorithm

- If lots of pages point TO this page, this must be a “more important” page
- Tweaking the page rank algorithm can make or break a small business.
7. Presenting Results

- Mostly just a list.
- Maybe there are better forms.
- Those sponsored links…

Search Engines … A Summary

- A search engine has two parts
  - Crawler, to index the data
  - Query Processor, to answer queries based on index
- In the case of many hits, a query processor must rank the results; page rank does that by
  - "using data differentially" … not all associations are equivalent; anchors and file names count more
  - "noting relationship of pages" … a page is more important if important pages link to it

Google, Bing, Yahoo and other Search Engines Use All of These Ideas
Who Pays for What?

- Users could pay a subscription fee (early AOL and CompuServe)
- Web sites could pay for being indexed.
- The government could pay (taxes?).
- Advertisers could pay.

Placements, Clicks, and Auctions

- Buy higher position in the ranking - FTC said don’t do it without flagging it as such.
- Banner ads displayed when search included certain terms.
  - pay for view or pay for click throughs?
- Companies bid for popular terms.
- Companies exercise editorial power (censorship?) by refusing certain ads.

Manipulating Search Results

- White text on white background with words that will raise your rank.
- Google Bombing - “miserable failure” search in 2000 yielded white house biography of George Bush
- Companies that will help you move up in the ranking with changes to your web site.

Google Bombing

- Companies purchasing inbound links to increase their page rank
- If its not on the first page it might as well not be there
- Google has begun trying to change the page rank algorithm to move away from inbound links.
- However, this was their core competency from 1996-2007, what differentiated them from AltaVista, Yahoo, Netscape etc.

Important questions: What you should learn

- How can a search engine respond so fast?
- Does it find every relevant link?
- How does a search engine decide what gets listed first?
- If you try another search engine will you get the same result? If so, which is right? Which is better? Which is more authoritative?
- Are sponsored links better than “organic” links? Is the advertising necessary?
- What is the role of government? What should it be?

Summary

- Search engines offer unprecedented access to information.
- Search engines place the power to shape what we see into the hands of a few companies.
- Search engines continue to evolve.
  - Recently adding in Google Plus
  - Startups on indexing twitter etc
- Question Answering is the next new thing!