## Information Retrieval

INFO 4300 / CS 4300

- Web crawlers
  - Retrieving web pages
  - Crawling the web
    - Desktop crawlers
    - Document feeds
  - File conversion
  - Storing the documents
  - Removing noise

## Desktop Crawls

- Used for desktop search and enterprise search
- Differences from web crawling:
  - Much easier to find the data
  - Responding quickly to updates is more important
  - Must be conservative in terms of disk and CPU usage
  - Many different document formats
  - Data privacy very important

## Document Feeds

- Many documents are ***published***
  - created at a fixed time and rarely updated again
  - e.g., news articles, blog posts, press releases, email
- Published documents from a single source can be ordered in a sequence called a ***document feed***
  - new documents found by examining the end of the feed

## Document Feeds

- Two types:
  - A ***push feed*** alerts the subscriber to new documents
  - A ***pull feed*** requires the subscriber to check periodically for new documents
- Most common format for pull feeds is called ***RSS***
  - Really Simple Syndication, RDF Site Summary, Rich Site Summary, or ...

## RSS Example

```
<?xml version="1.0"?>
<rss version="2.0">
  <channel>
    <title>Search Engine News</title>
    <link>http://www.search-engine-news.org/</link>
    <description>News about search engines.</description>
    <language>en-us</language>
    <pubDate>Tue, 19 Jun 2008 05:17:00 GMT</pubDate>
    <ttl>60</ttl>

    <item>
      <title>Upcoming SIGIR Conference</title>
      <link>http://www.sigir.org/conference</link>
      <description>The annual SIGIR conference is coming!
        Mark your calendars and check for cheap
        flights.</description>
      <pubDate>Tue, 05 Jun 2008 09:50:11 GMT</pubDate>
      <guid>http://search-engine-news.org#500</guid>
    </item>
```

## RSS Example

```
...
   <item>
      <title>New Search Engine Textbook</title>
      <link>http://www.cs.umass.edu/search-book</link>
      <description>A new textbook about search engines
        will be published soon.</description>
      <pubDate>Tue, 05 Jun 2008 09:33:01 GMT</pubDate>
      <guid>http://search-engine-news.org#499</guid>
    </item>
  </channel>
</rss>
```

## RSS Example

```
<?xml version="1.0"?>
<rss version="2.0">
  <channel>
    <title>Search Engine News</title>
    <link>http://www.search-engine-news.org/</link>
    <description>News about search engines.</description>
    <language>en-us</language>
    <pubDate>Tue, 19 Jun 2008 05:17:00 GMT</pubDate>
    <ttl>60</ttl>

    <item>
      <title>Upcoming SIGIR Conference</title>
      <link>http://www.sigir.org/conference</link>
      <description>The annual SIGIR conference is coming!
        Mark your calendars and check for cheap
        flights.</description>
      <pubDate>Tue, 05 Jun 2008 09:50:11 GMT</pubDate>
      <guid>http://search-engine-news.org#500</guid>
    </item>
```

## RSS

- `ttl` tag (time to live)
  - amount of time (in minutes) contents should be cached
- RSS feeds are accessed like web pages
  - using HTTP GET requests to web servers that host them
- Easy for crawlers to parse
- Easy to find new information

# Information Retrieval

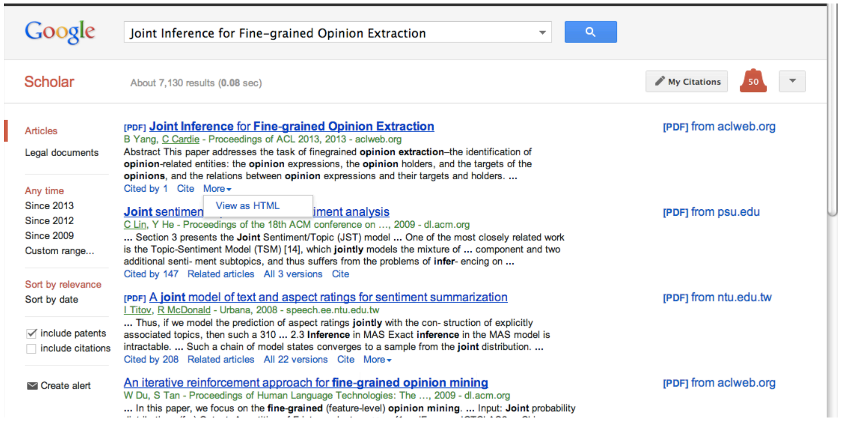
INFO 4300 / CS 4300

- Web crawlers
  - Retrieving web pages
  - Crawling the web
    - » Desktop crawlers
    - » Document feeds
  - File conversion
  - Storing the documents
  - Removing noise

# Conversion

- Text is stored in hundreds of incompatible file formats
  - e.g., raw text, RTF, HTML, XML, Microsoft Word, ODF, PDF
- Other types of files also important
  - e.g., PowerPoint, Excel
- Typically use a conversion tool
  - converts the document content into a tagged text format such as HTML or XML
  - retains some of the important formatting information

# Searching for a .pdf



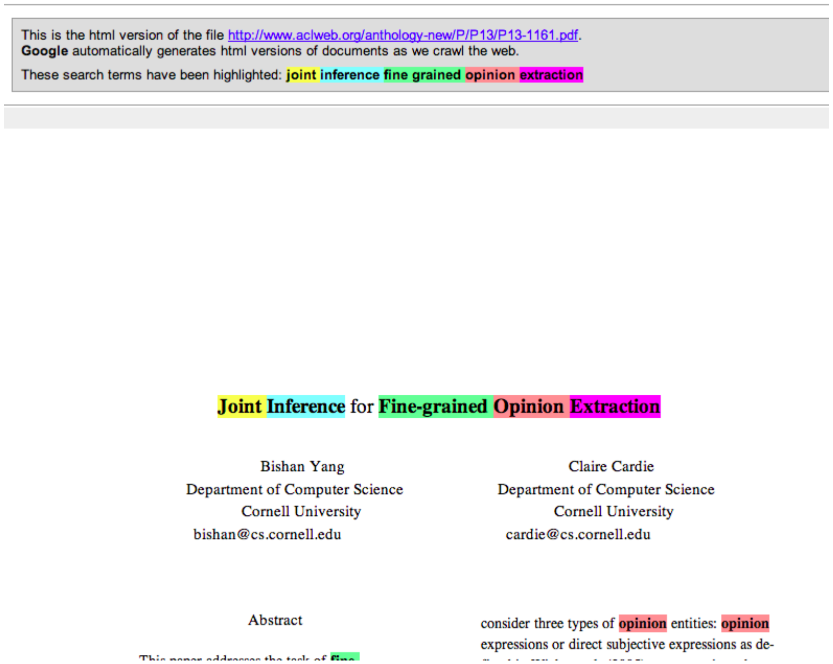
About 7,130 results (**0.08** sec)

[PDF] Joint Inference for Fine-grained Opinion Extraction — [PDF] from aclweb.org
B Yang, C Cardie - Proceedings of ACL 2013, 2013 - aclweb.org
Abstract This paper addresses the task of finegrained **opinion extraction**–the identification of **opinion**-related entities: the **opinion** expressions, the **opinion** holders, and the targets of the **opinions**, and the relations between **opinion** expressions and their targets and holders. ...
Cited by 1 Cite More▾
View as HTML

Joint sentimen iment analysis — [PDF] from psu.edu
C Lin, Y He - Proceedings of the 18th ACM conference on ..., 2009 - dl.acm.org
... Section 3 presents the **Joint** Sentiment/Topic (JST) model ... One of the most closely related work is the Topic-Sentiment Model (TSM) [14], which **jointly** models the mixture of ... component and two additional senti- ment subtopics, and thus suffers from the problems of **infer**- encing on ...
Cited by 147 Related articles All 3 versions Cite

[PDF] A joint model of text and aspect ratings for sentiment summarization — [PDF] from ntu.edu.tw
I Titov, R McDonald - Urbana, 2008 - speech.ee.ntu.edu.tw
... Thus, if we model the prediction of aspect ratings **jointly** with the con- struction of explicitly associated topics, then such a 310 ... 2.3 **Inference** in MAS Exact **inference** in the MAS model is intractable. ... Such a chain of model states converges to a sample from the **joint** distribution. ...
Cited by 208 Related articles All 22 versions Cite More▾

An iterative reinforcement approach for fine-grained opinion mining — [PDF] from aclweb.org
W Du, S Tan - Proceedings of Human Language Technologies: The ..., 2009 - dl.acm.org
... In this paper, we focus on the **fine-grained** (feature-level) **opinion mining**. ... Input: **Joint** probability



This is the html version of the file http://www.aclweb.org/anthology-new/P/P13/P13-1161.pdf.
**Google** automatically generates html versions of documents as we crawl the web.

These search terms have been highlighted: **joint** **inference** **fine grained** **opinion** **extraction**

**Joint Inference** for **Fine-grained Opinion Extraction**

Bishan Yang
Department of Computer Science
Cornell University
bishan@cs.cornell.edu

Claire Cardie
Department of Computer Science
Cornell University
cardie@cs.cornell.edu

Abstract

This paper addresses the task of **fine**

consider three types of **opinion** entities: **opinion** expressions or direct subjective expressions as de-

## Character Encoding

- A character encoding is a mapping between bits and glyphs
  - i.e., getting from bits in a file to characters on a screen
  - Can be a major source of incompatibility
- ASCII is basic character encoding scheme for English (since 1963)
  - encodes 128 letters, numbers, special characters, and control characters in 7 bits, extended with an extra bit for storage in bytes

## Character Encoding

- Other languages can have many more glyphs
  - e.g., Chinese has more than 40,000 characters, with over 3,000 in common use
- Many languages have multiple encoding schemes
  - e.g., CJK (Chinese-Japanese-Korean) family of East Asian languages, Hindi, Arabic
  - must specify encoding
  - can't have multiple languages in one file
- Unicode developed to address encoding problems

## Unicode

- Single mapping from numbers to glyphs that attempts to include all glyphs in common use in all known languages
- Unicode is a mapping between numbers and glyphs
  - does not uniquely specify bits to glyph mapping!
  - e.g., UTF-8, UTF-16, UTF-32

## Unicode

- Proliferation of encodings comes from a need for compatibility and to save space
  - UTF-8 uses one byte for English (ASCII), as many as 4 bytes for some traditional Chinese characters
  - variable length encoding, more difficult to do string operations, e.g. find the 10th character
  - UTF-32 uses 4 bytes for every character
- Many applications use UTF-32 for internal text encoding (fast random lookup) and UTF-8 for disk storage (less space)

## Unicode

| Decimal | Hexadecimal | Encoding | | | |
|---|---|---|---|---|---|
| 0–127 | 0–7F | 0xxxxxxx | | | |
| 128–2047 | 80–7FF | 110xxxxx | 10xxxxxx | | |
| 2048–55295 | 800–D7FF | 1110xxxx | 10xxxxxx | 10xxxxxx | |
| 55296–57343 | D800–DFFF | Undefined | | | |
| 57344–65535 | E000–FFFF | 1110xxxx | 10xxxxxx | 10xxxxxx | |
| 65536–1114111 | 10000–10FFFF | 11110xxx | 10xxxxxx | 10xxxxxx | 10xxxxxx |

- e.g., Greek letter pi ($\pi$) is Unicode symbol number 960
- In binary, 00000011 11000000 (3C0 in hexadecimal)
- Final encoding is **110**01111 **10**000000 (CF80 in hexadecimal)

## Information Retrieval

INFO 4300 / CS 4300

- Web crawlers
  - Retrieving web pages
  - Crawling the web
    - » Desktop crawlers
    - » Document feeds
  - File conversion
  - ➡ Storing the documents
  - Removing noise

## Storing the Documents

- Many reasons to store converted document text
  - saves crawling time when page is not updated
  - provides efficient access to text for snippet generation, information extraction, etc.
- Database systems can provide document storage for some applications
  - web search engines use customized document storage systems

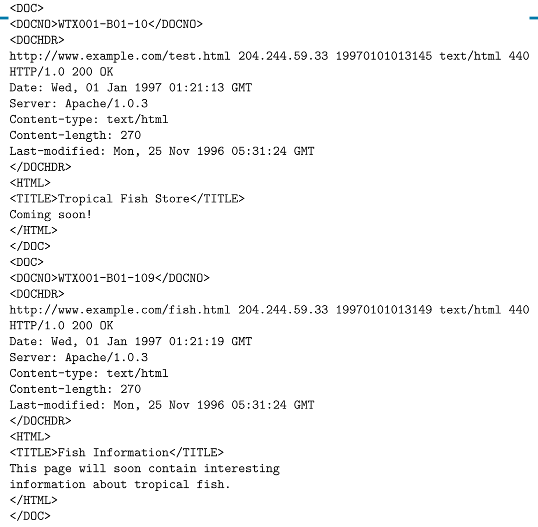
## Storing the Documents

- Requirements for document storage system:
  - Fast random access
    - » request the content of a document based on its URL
    - » hash function based on URL is typical
  - Compression and large files
    - » reducing storage requirements and efficient access
  - Update
    - » handling large volumes of new and modified documents
    - » adding new anchor text

# Large Files

- Store many documents in large files, rather than each document in a file
  - avoids overhead in opening and closing files
  - reduces seek time relative to read time
- Compound documents formats
  - used to store multiple documents in a file
  - e.g., TREC Web

# TREC Web Format



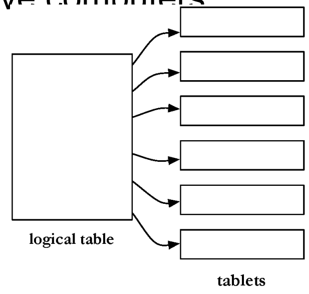
```
<DOC>
<DOCNO>WTX001-B01-10</DOCNO>
<DOCHDR>
http://www.example.com/test.html 204.244.59.33 19970101013145 text/html 440
HTTP/1.0 200 OK
Date: Wed, 01 Jan 1997 01:21:13 GMT
Server: Apache/1.0.3
Content-type: text/html
Content-length: 270
Last-modified: Mon, 25 Nov 1996 05:31:24 GMT
</DOCHDR>
<HTML>
<TITLE>Tropical Fish Store</TITLE>
Coming soon!
</HTML>
</DOC>
<DOC>
<DOCNO>WTX001-B01-109</DOCNO>
<DOCHDR>
http://www.example.com/fish.html 204.244.59.33 19970101013149 text/html 440
HTTP/1.0 200 OK
Date: Wed, 01 Jan 1997 01:21:19 GMT
Server: Apache/1.0.3
Content-type: text/html
Content-length: 270
Last-modified: Mon, 25 Nov 1996 05:31:24 GMT
</DOCHDR>
<HTML>
<TITLE>Fish Information</TITLE>
This page will soon contain interesting
information about tropical fish.
</HTML>
</DOC>
```

# Compression

- Text is highly redundant (or predictable)
- Compression techniques exploit this redundancy to make files smaller without losing any of the content
- Compression of indexes covered later
- Popular algorithms can compress HTML and XML text by 80%
  - e.g., DEFLATE (zip, gzip) and LZW (UNIX compress, PDF)
  - may compress large files in blocks to make access faster

# BigTable

- Google's document storage system
  - Customized for storing, finding, and updating web pages
  - Handles large collection sizes using inexpensive computers

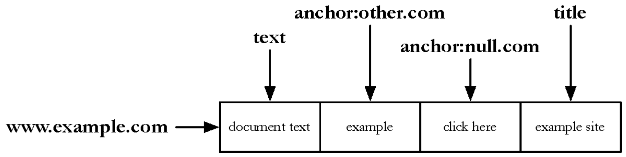


## BigTable

- No query language, no complex queries to optimize
- Only row-level transactions
- Tablets are stored in a replicated file system that is accessible by all BigTable servers
- Any changes to a BigTable tablet are recorded to a transaction log, which is also stored in a shared file system
- If any tablet server crashes, another server can immediately read the tablet data and transaction log from the file system and take over

## BigTable

- Logically organized into rows
- A row stores data for a single web page



- Combination of a row key, a column key, and a timestamp point to a single *cell* in the row

## BigTable

- BigTable can have a huge number of columns per row
  - all rows have the same column groups
  - not all rows have the same columns
  - important for reducing disk reads to access document data
- Rows are partitioned into tablets based on their row keys
  - simplifies determining which server is appropriate

## Detecting Duplicates

- Duplicate and near-duplicate documents occur in many situations
  - Copies, versions, plagiarism, spam, mirror sites
  - 30% of the web pages in a large crawl are exact or near duplicates of pages in the other 70%
- Duplicates consume significant resources during crawling, indexing, and search
  - Little value to most users

## Duplicate Detection

- *Exact* duplicate detection is relatively easy
- *Checksum* techniques
  - A checksum is a value that is computed based on the content of the document
    - e.g., sum of the bytes in the document file

| T | r | o | p | i | c | a | l |  | f | i | s | h | *Sum* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 54 | 72 | 6F | 70 | 69 | 63 | 61 | 6C | 20 | 66 | 69 | 73 | 68 | 508 |

  - Possible for files with different text to have same checksum
- Functions such as a *cyclic redundancy check* (CRC), have been developed that consider the positions of the bytes

## Near-Duplicate Detection

- More challenging task
  - Are web pages with same text context but different advertising or format near-duplicates?
- A near-duplicate document is defined using a threshold value for some similarity measure between pairs of documents
  - e.g., document *D1* is a near-duplicate of document *D2* if more than 90% of the words in the documents are the same

## Near-Duplicate Detection

- *Search*:
  - find near-duplicates of a document *D*
  - $O(N)$ comparisons required
- *Discovery*:
  - find all pairs of near-duplicate documents in the collection
  - $O(N^2)$ comparisons
- IR techniques are effective for search scenario
- For discovery, other techniques used to generate compact representations

## Information Retrieval

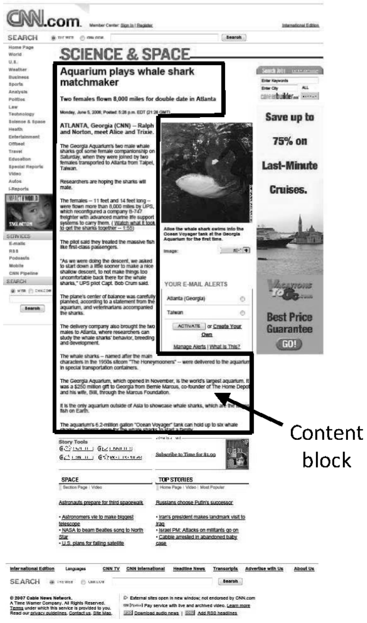
INFO 4300 / CS 4300

- Web crawlers
  - Retrieving web pages
  - Crawling the web
    - Desktop crawlers
    - Document feeds
  - File conversion
  - Storing the documents
  - ➡ Removing noise

## Removing Noise

- Many web pages contain text, links, and pictures that are not directly related to the main content of the page
- This additional material is mostly *noise* that could negatively affect the ranking of the page
- Techniques have been developed to detect the content blocks in a web page
  - Non-content material is either ignored or reduced in importance in the indexing process

## Noise Example



CNN.com Member Center: Sign In | Register International Edition

SEARCH ◉ THE WEB ○ CNN.COM Search

Home Page
World
U.S.
Weather
Business
Sports
Analysis
Politics
Law
Technology
Science & Space
Health
Entertainment
Offbeat
Travel
Education
Special Reports
Video
Autos
I-Reports

SERVICES
E-mails
RSS
Podcasts
Mobile
CNN Pipeline

SEARCH ◉ WEB ○ CNN.COM Search

SCIENCE & SPACE

**Aquarium plays whale shark matchmaker**

**Two females flown 8,000 miles for double date in Atlanta**

Monday, June 5, 2006; Posted: 5:28 p.m. EDT (21:28 GMT)

**ATLANTA, Georgia (CNN) -- Ralph and Norton, meet Alice and Trixie.**

The Georgia Aquarium's two male whale sharks got some female companionship on Saturday, when they were joined by two females transported to Atlanta from Taipei, Taiwan.

Researchers are hoping the sharks will mate.

The females -- 11 feet and 14 feet long -- were flown more than 8,000 miles by UPS, which reconfigured a company B-747 freighter with advanced marine life support systems to carry them. ( Watch what it took to get the sharks together -- 1:55)

The pilot said they treated the massive fish like first-class passengers.

"As we were doing the descent, we asked to start down a little sooner to make a nice shallow descent, to not make things too uncomfortable back there for the whale sharks," UPS pilot Capt. Bob Crum said.

The plane's center of balance was carefully planned, according to a statement from the aquarium, and veterinarians accompanied the sharks.

The delivery company also brought the two males to Atlanta, where researchers can study the whale sharks' behavior, breeding and development.

The whale sharks -- named after the main characters in the 1950s sitcom "The Honeymooners" -- were delivered to the aquarium in special transportation containers.

The Georgia Aquarium, which opened in November, is the world's largest aquarium. It was a $250 million gift to Georgia from Bernie Marcus, co-founder of The Home Depot and his wife, Bill, through the Marcus Foundation.

It is the only aquarium outside of Asia to showcase whale sharks, which are the fish on Earth.

The aquarium's 6.2-million gallon "Ocean Voyager" tank can hold up to six whale sharks, so there's room for the whale sharks to start a family.

Alice the whale shark swims into the Ocean Voyager tank at the Georgia Aquarium for the first time.

Image:

YOUR E-MAIL ALERTS
Atlanta (Georgia) ○
Taiwan ○
ACTIVATE or Create Your Own
Manage Alerts | What Is This?

Search Jobs
Enter Keywords
Enter City ALL
careerbuilder.com

Save up to 75% on Last-Minute Cruises.
VacationsToGo.com
Best Price Guarantee
GO!

Story Tools

Subscribe to Time for $1.99

SPACE
Section Page | Video
Astronauts prepare for third spacewalk
- Astronomers vie to make biggest telescope
- NASA to beam Beatles song to North Star
- U.S. plans for falling satellite

TOP STORIES
Home Page | Video | Most Popular
Russians choose Putin's successor
- Iran's president makes landmark visit to Iraq
- Israel PM: Attacks on militants go on
- Cabbie arrested in abandoned baby case

International Edition Languages CNN TV CNN International Headline News Transcripts Advertise with Us About Us

SEARCH ◉ THE WEB ○ CNN.COM Search

© 2007 Cable News Network.
A Time Warner Company. All Rights Reserved.
Terms under which this service is provided to you.
Read our privacy guidelines. Contact us. Site Map.

External sites open in new window; not endorsed by CNN.com
Pay service with live and archived video. Learn more
Download audio news | Add RSS headlines