Old Dominion University

CS 418/518
Web Programming
Spring 2014

CAPTCHA and Search

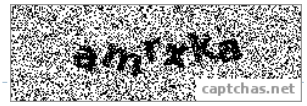
Dr. Michele Weigle

http://www.cs.odu.edu/~mweigle/CS418-S14/

Outline

- CAPTCHA
- Search

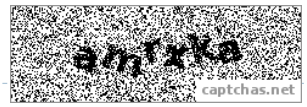
CAPTCHA



- **C**ompletely **A**utomated **P**ublic **T**uring test to tell **C**omputers and **H**umans **A**part
- Challenge-response test provided by server
- User solves problem and is considered (by server) human (and not a machine)
- Goals
  - ensure that interaction is with user
  - prevent spam of all kinds, e.g. mass account creation, posts, etc.

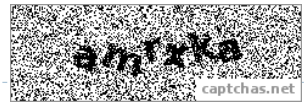
http://www.captcha.net/

CAPTCHA
How to?



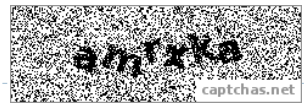
- Distorted text
  - make it hard for Optical Character Recognition (OCR)
  - difficult to distinguish between background and text (color, shape)
  - character overlap
  - out of alignment

CAPTCHA
Problem solved?



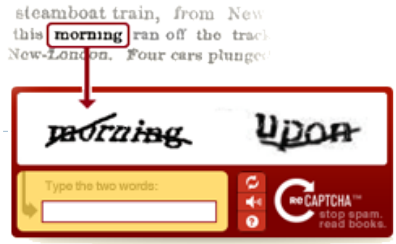
- Vulnerable to relay attacks
  - relay captcha to human when encountered
- Capture and re-use successful session ID
- Dictionary attacks
- "Iron out" images and use ORC, dictionaries

CAPTCHA
Problem solved?



- How about accessibility?
  - Blind users?
    - possible solution: audio stream
    - voice recognition software!
  - Deaf-blind users?
    - ???

reCAPTCHA



- Originates from CMU
  - bought by Google in 2009
- Help needed to digitize books (using OCR)
  - words come from scanned books
- "Wisdom of the Crowds"
  - reCAPTCHA contains
    - 1 term not recognized by OCR
    - 1 term well known
  - Assumption: if user gets known term right, she also gets unknown term right
  - To be confirmed by 2, 3, … others
- Digitization project benefits!

http://www.google.com/recaptcha

video: https://developers.google.com/recaptcha/

reCAPTCHA
Links

- Examples
  - https://www.google.com/webmasters/tools/submit-url?pli=1
  - https://www.blogger.com/comment.g?blogID=25215770&postID=5975815412653416464
- Top 10 Worst Captchas
  - http://www.johnmwillis.com/other/top-10-worst-captchas
- Implementations
  - http://captchas.net/
    - https://weiglevm.cs.odu.edu/~mweigle/captcha/query.php
  - http://www.google.com/recaptcha

WARNING:
reCAPTCHA may not work on weiglevm

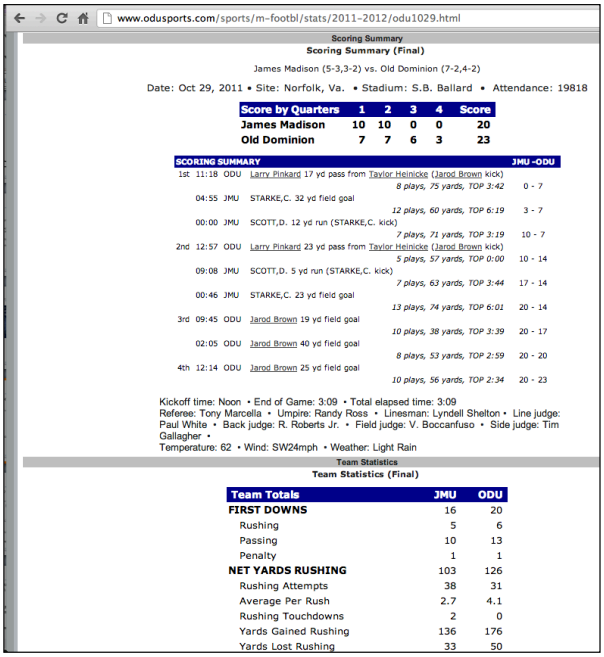
Outline

- CAPTCHA
- Search

Relational Data Model is a Special Case…

```
SELECT ti.name, g.tds, g.passing_yds
     FROM team_info ti, games g
     WHERE ti.name = "Old Dominion"
          AND g.opponent = "James Madison"
          AND g.year = "2011";
```

Unstructured Data is More Common…



www.odusports.com/sports/m-footbl/stats/2011-2012/odu1029.html

Scoring Summary

Scoring Summary (Final)

James Madison (5-3,3-2) vs. Old Dominion (7-2,4-2)

Date: Oct 29, 2011 • Site: Norfolk, Va. • Stadium: S.B. Ballard • Attendance: 19818

Score by Quarters	1	2	3	4	Score
James Madison	10	10	0	0	20
Old Dominion	7	7	6	3	23

SCORING SUMMARY				JMU -ODU
1st	11:18	ODU	Larry Pinkard 17 yd pass from Taylor Heinicke (Jarod Brown kick) *8 plays, 75 yards, TOP 3:42*	0 - 7
	04:55	JMU	STARKE,C. 32 yd field goal *12 plays, 60 yards, TOP 6:19*	3 - 7
	00:00	JMU	SCOTT,D. 12 yd run (STARKE,C. kick) *7 plays, 71 yards, TOP 3:19*	10 - 7
2nd	12:57	ODU	Larry Pinkard 23 yd pass from Taylor Heinicke (Jarod Brown kick) *5 plays, 57 yards, TOP 0:00*	10 - 14
	09:08	JMU	SCOTT,D. 5 yd run (STARKE,C. kick) *7 plays, 63 yards, TOP 3:44*	17 - 14
	00:46	JMU	STARKE,C. 23 yd field goal *13 plays, 74 yards, TOP 6:01*	20 - 14
3rd	09:45	ODU	Jarod Brown 19 yd field goal *10 plays, 38 yards, TOP 3:39*	20 - 17
	02:05	ODU	Jarod Brown 40 yd field goal *8 plays, 53 yards, TOP 2:59*	20 - 20
4th	12:14	ODU	Jarod Brown 25 yd field goal *10 plays, 56 yards, TOP 2:34*	20 - 23

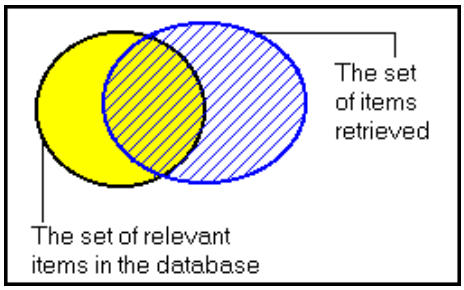
Kickoff time: Noon • End of Game: 3:09 • Total elapsed time: 3:09
Referee: Tony Marcella • Umpire: Randy Ross • Linesman: Lyndell Shelton • Line judge: Paul White • Back judge: R. Roberts Jr. • Field judge: V. Boccanfuso • Side judge: Tim Gallagher •
Temperature: 62 • Wind: SW24mph • Weather: Light Rain

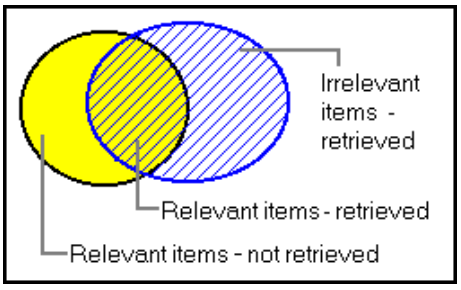
Team Statistics

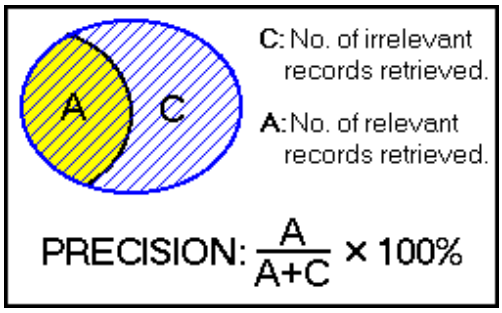
Team Statistics (Final)

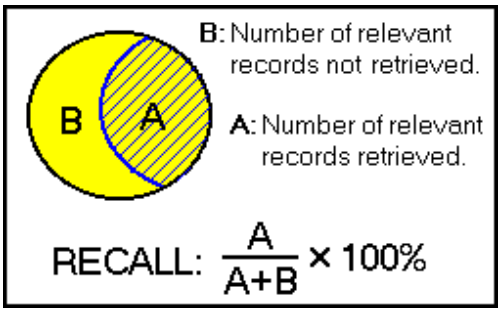
Team Totals	JMU	ODU
FIRST DOWNS	16	20
Rushing	5	6
Passing	10	13
Penalty	1	1
NET YARDS RUSHING	103	126
Rushing Attempts	38	31
Average Per Rush	2.7	4.1
Rushing Touchdowns	2	0
Yards Gained Rushing	136	176
Yards Lost Rushing	33	50

Precision and Recall





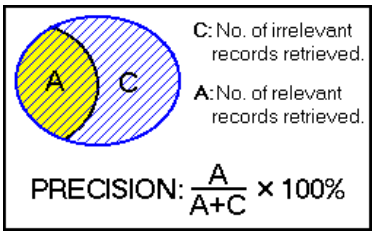


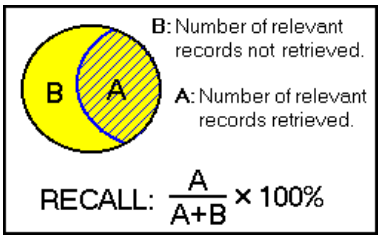


how much extra stuff did you get? ***how much did you miss?***

source: http://www.hsl.creighton.edu/hsl/Searching/Recall-Precision.html

Precision and Recall





10 documents in the index are relevant
search returns 20 documents
5 of which are relevant

$$P = \frac{5}{(5+15)} = 0.25$$

1 out of 4 retrieved documents are relevant

$$R = \frac{5}{(5+5)} = 0.5$$

half of the relevant documents were retrieved

source: http://www.hsl.creighton.edu/hsl/Searching/Recall-Precision.html

Precision and Recall

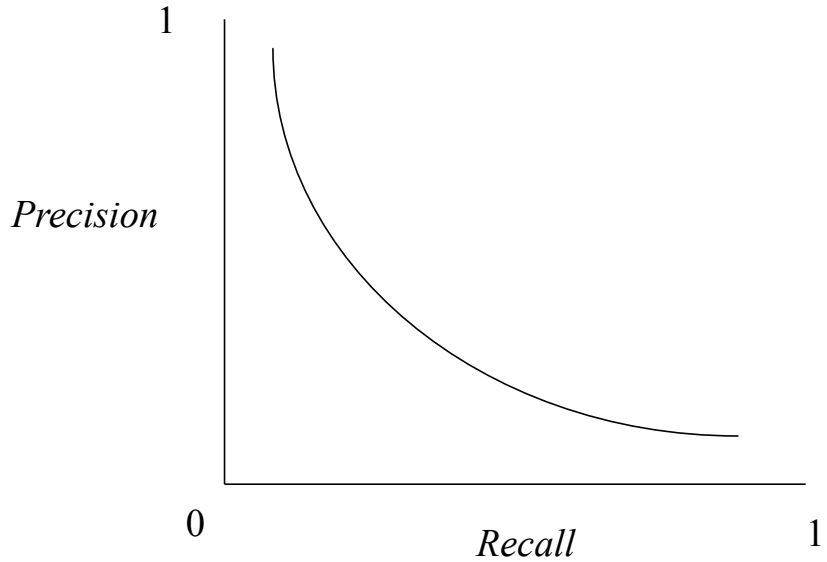


figure 1.2 in FBY

Why Isn't Recall Always 100%?

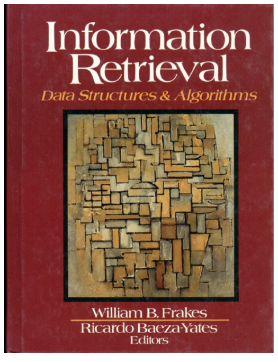
Louisiana State University and Agricultural and Mechanical College?

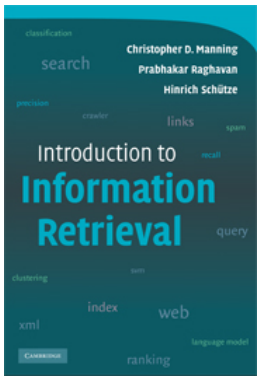
Louisiana State A&M?

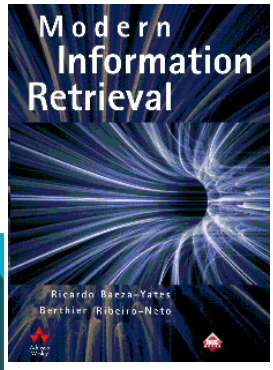
Louisiana State University?

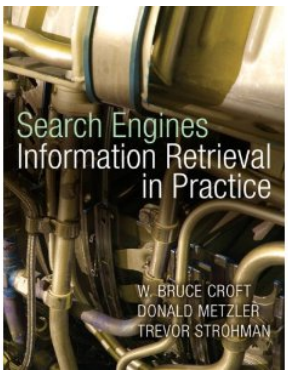
LSU?

Precision and Recall - Literature









Dr. Nelson is teaching IR this semester

Search Example

- Create and populate table for ODU football articles from odusports.com
  - http://www.odusports.com/ViewArticle.dbml?&DB_OEM_ID=31100&ATCLID=208423738
- Fields
  - id
  - title
  - body
  - date
  - url

http://weiglevm.cs.odu.edu/~mweigle/textbook/search.htm

Example 1

LIKE and REGEXP

- We can search rows with the "LIKE" (or "REGEXP") operator
  - http://dev.mysql.com/doc/refman/5.5/en/pattern-matching.html
  - for tables of any size, this will be *s-l-o-w*
- LIKE
  - simple regular expression matching
- REGEXP
  - extended regular expression matching

Example 2

LIKE and REGEXP

- A REGEXP pattern match succeeds if the pattern matches *anywhere* in the value being tested.
- This differs from a LIKE pattern match, which succeeds only if the pattern matches the *entire value*.

http://dev.mysql.com/doc/refman/5.5/en/fulltext-search.html

Full-Text Search – The Better Way

- MATCH()…AGAINST()
  - performs a natural language search over index
- Index = set of one or more columns of the same table
  - column must have type FULLTEXT
- MATCH()
  - takes a comma-separated list that names the columns to be searched
- AGAINST()
  - takes a string to search for
- If used in WHERE clause, results returned in order of relevance score
  - relevance: similarity between search string and index row

FULLTEXT

```
CREATE TABLE odu_football (
 id INT UNSIGNED AUTO_INCREMENT NOT NULL PRIMARY KEY,
 title VARCHAR(200),
 body TEXT,
 date DATE,
 url VARCHAR (200),
 FULLTEXT (title, body))
```

- Can only create FULLTEXT on CHAR, VARCHAR or TEXT columns
- "title" and "body" still available as regular columns
- If you want to search *only* on "title", you need to create a separate index

Example 3

FULLTEXT

- ALTER TABLE to create index
  - also need to alter table to use ENGINE=MYISAM
- Searches
  - Tyree (as in RB Tyree Lee)
  - playoffs
  - Monarchs

Stopwords

- Why no results for "Monarchs"?
- If a word appears in > 50% of the rows then the word is considered a "stop word" and is not matched (unless you are in Boolean mode)
  - this makes sense for large collections (the word is not a good discriminator of records), but can lead to unexpected results for small collections

Stopwords

- Stopwords exist in *stoplists* or *negative dictionaries*
- Idea: remove low semantic content
  - index should only have "important stuff"
- What not to index is domain dependent, but often includes:
  - "small" words: *a, and, the, but, of, an, very,* etc.
  - NASA ADS example
    - http://adsabs.harvard.edu/abs_doc/stopwords.html
  - MySQL full-text index
    - http://dev.mysql.com/doc/refman/5.0/en/fulltext-stopwords.html

Stopwords

- Punctuation, numbers often stripped or treated as stopwords
  - precision suffers on searches for:
    - NASA TM-3389
    - F-15
    - X.500
    - .NET
    - Tree::Suffix
- MySQL also treats words < 4 characters as stopwords
  - too bad for: "Liu", "ORF", "DEA", etc.

Getting the Rank

Example 4

```
mysql> SELECT id, MATCH(title,body) AGAINST('playoffs') from odu_football;
+----+---------------------------------------+
| id | MATCH(title,body) AGAINST ('playoffs') |
+----+---------------------------------------+
|  1 |                    0.493198305368423 |
|  2 |                                    0 |
|  3 |                                    0 |
|  4 |                                    0 |
|  5 |                       0.552978515625 |
|  6 |                                    0 |
+----+---------------------------------------+
6 rows in set (0.00 sec)
```

Getting the Rank in Order

Example 5

```
mysql> SELECT id, MATCH(title,body) AGAINST('playoffs')
AS score FROM odu_football WHERE MATCH(title,body) AGAINST('playoffs')
ORDER BY score DESC;

+----+-------------------+
| id | score             |
+----+-------------------+
|  5 |    0.552978515625 |
|  1 | 0.493198305368423 |
+----+-------------------+
2 rows in set (0.00 sec)
```


Boolean Mode

Example 6

```
mysql> SELECT id, title FROM odu_football
WHERE MATCH(title,body) AGAINST('+Monarchs' IN BOOLEAN MODE);

+----+-----------------------------------------------------------------------+
| id | title                                                                 |
+----+-----------------------------------------------------------------------+
|  1 | ODU to Host Watch Party Sunday at Sheraton Norfolk Waterside at 1:30pm  |
|  2 | Monarchs Remain No. 4 in FCS Polls                                    |
|  3 | Monarchs Hammer Georgia State, 53-27                                  |
|  4 | Monarchs Win Rain Soaked Oyster Bowl Over Delaware, 31-26             |
+----+-----------------------------------------------------------------------+
4 rows in set (0.00 sec)
```

- Does not use the 50% threshold
- Does use stopwords, length limitation
- Operator list
  - http://dev.mysql.com/doc/refman/5.5/en/fulltext-boolean.html

Blind Query Expansion
(AKA Automatic Relevance Feedback)

- General assumption: user query is insufficient
  - too short
  - too generic
  - too many results
- How does one keep up with LSU's multiple names / nicknames?
  - Tigers, Bayou Bengals, LSU, LSU-A&M, Louisiana State
- Idea:
  - run the search twice
  - 1) run the search with the requested terms
  - 2) the search phrase for the second search is the original search phrase concatenated with the few most highly relevant documents from the first search

Blind Query Expansion
(AKA Automatic Relevance Feedback)

- Use `WITH QUERY EXPANSION`
- Because blind query expansion tends to increase noise significantly by returning non-relevant documents, it is meaningful to use only when a search phrase is rather short.
- https://dev.mysql.com/doc/refman/5.5/en/fulltext-query-expansion.html

Blind Query Expansion
(AKA Automatic Relevance Feedback)

Example 7

```
SELECT title,body FROM odu_football WHERE MATCH(title,body) AGAINST('Tyree' IN BOOLEAN MODE);

+-------------------------------------------+-------------------------------------------+
| title                                     | body
+-------------------------------------------+-------------------------------------------+
| Monarchs Win Rain Soaked Oyster Bowl      | Taylor Heinicke threw for 375 yards and ran
| Over Delaware, 31-26                      | for three touchdowns while Tyree Lee rushed
|                                           | for a career-high 128 yards and a touchdown
|                                           | as No. 6/7 Old Dominion University football
|                                           | defeated No. 20/16 Delaware 31-26 on a windy
|                                           | Saturday afternoon at Foreman Field at S.B.
|                                           | Ballard Stadium.
+-------------------------------------------+-------------------------------------------+
```

```
SELECT title,body FROM odu_football WHERE MATCH(title,body) AGAINST('Tyree' WITH QUERY EXPANSION);
```

adds

```
| RECORD BOOKS SHATTERED: #5 ODU With 64-61   | Heinicke threw for a Division I record 730
| Victory Over #18/19 UNH                     | yards and five touchdowns as Jarod Brown kicked
|                                             | a 25-yard field goal with 41 seconds left and
|                                             | Andre Simmons intercepted a pass to clinch the
|                                             | 64-61 win over #18/19 New Hampshire Saturday
|                                             | afternoon.
+---------------------------------------------+-------------------------------------------+
```


For More Information…

- MySQL documentation:
  - http://dev.mysql.com/doc/refman/5.5/en/fulltext-search.html
- Chapter 13 "Building a Content Management System"
- CS 751/851 "Introduction to Digital Libraries"
  - http://www.cs.odu.edu/~mln/teaching/
  - esp. "Information Retrieval Concepts" lecture
- CS 895 "Web-based Information Retrieval"

MySQL examples in this lecture based on those found at dev.mysql.com
content snippets taken from www.odusports.com

Outline

- CAPTCHA
- Search

Up Next:
Apr 10, 15 - Student Presentations

Apr 17 - Project 4 Status Reports

May 1 (8:30-11:30am) - Project 4 Demos