CSCE 470: Information Storage and Retrieval

XIA (BEN) HU

CSE, TAMU

http://faculty.cs.tamu.edu/xiahu/SP18_CSCE470.html

Point distribution

- Class participation and quizzes - 5%
- Three homework assignments -20%
- Project - 30%
- Three Exams - 45%

- Late penalty, **YES**, increasing *exponentially* wrt the number of days. Late = Original / 2^n , $n > 0$.
- Academic integrity

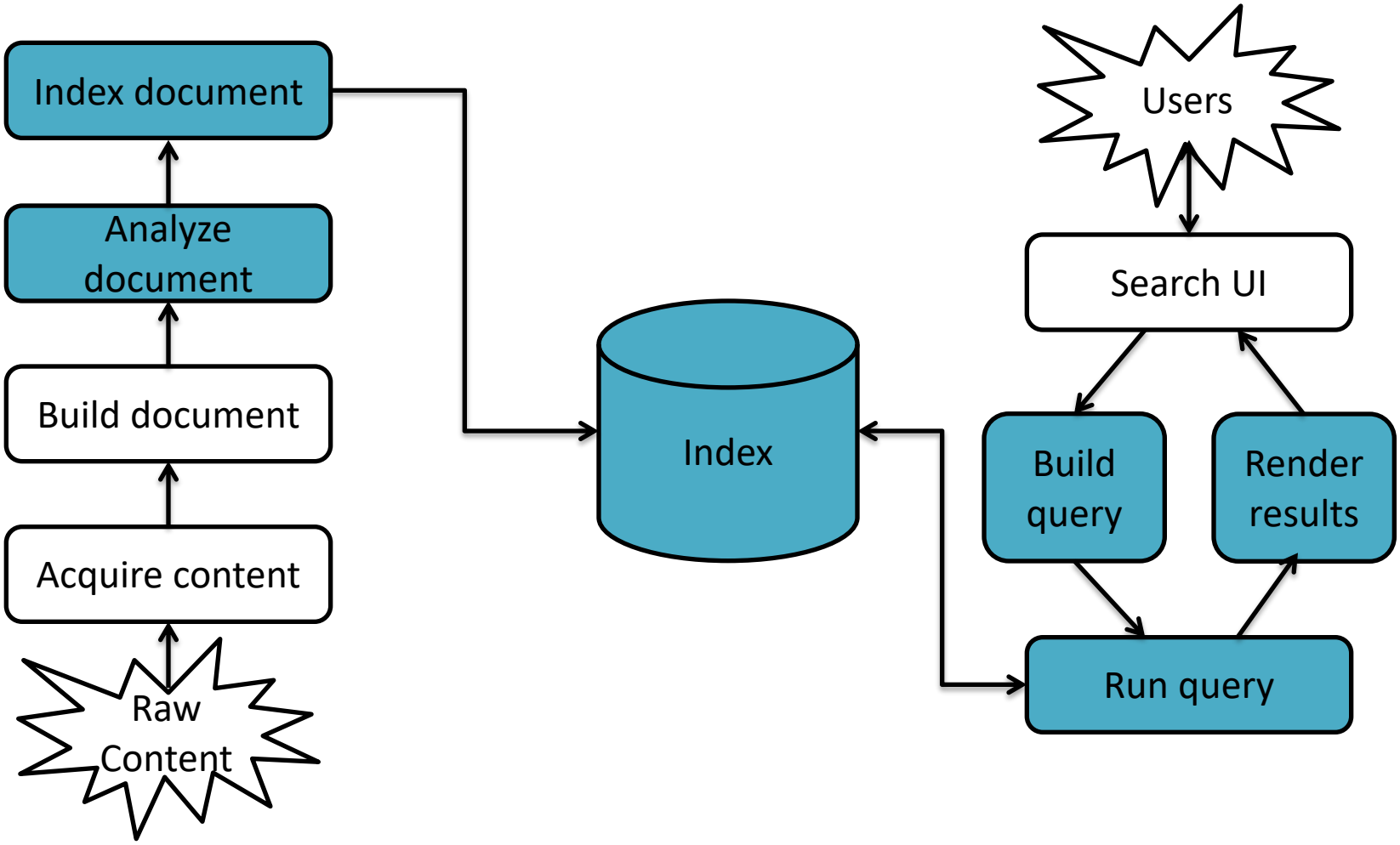
An Aggie does not lie, cheat, or steal, or tolerate those who do.

http://www.tamu.edu/aggiehonor/

Project

- Team Project
  - Real-world application
  - Three checkpoints, including data crawling, search engine, and application
  - Progress report
  - Final report
  - Class presentation and/or demo

CSCE 470 Project – An IR System



Format

- Group Project (2 or 3 students)
  - Task 1: Data crawling (20%)
  - Task 2: Building a search engine (20%)
  - Task 3: An application (20%)
  - Final report (20%)
  - Video demo (20%)

Task 1: Data Crawling

Task 1: Data Crawling

- **Select an open site with text information**
  - Post site's name (e.g., Wikipedia) on ecampus (if you are the first group) under the tag of "project-site-selection"
  - Select the site by replying to the site's name, with names of group members
  - Only 5 groups can crawl the same site (first come, first serve)
  - Please do not post anything unrelated to site selection and group members under the tag "project-site-selection"
  - If you decide to change your site report it again and remove your reply from the old post; otherwise, your first submission is considered as your site
  - Deadline (next Monday, Feb 5th)
- API or Parsing. Many sites have API limits if too many requests are sent to their servers.

How can I find data?

- **Some open source data:**
  - Snopes - https://www.snopes.com/
  - Truth or Fiction - https://www.truthorfiction.com/
  - Factcheck - https://www.factcheck.org/
  - Washington Post's Fact Checker - https://www.washingtonpost.com/news/fact-checker/
  - OpenSecrets - https://www.opensecrets.org/
  - The Sunlight Foundation - http://sunlightfoundation.com/
  - Hoax Slayer - http://www.hoax-slayer.com/
  - Politifact - http://www.politifact.com/

You can use multiple dataset!

Propose a problem you would like to solve

- Example problems:
  - Sentiment analysis for Amazon: Given a movie/product/people, predict whether people are happy or not
  - Recommendation: Given a movie/product/people, recommend related instances
  - Visualization: Given an event, project all of the tweets in a map that we can interact with
  - Any feature in Google/Bing/…

How Can I find a Problem?

- SIGIR Demo Papers
  - http://sigir.org/sigir2014/finaldemos.php
  - http://www.sigir.org/sigir2013/demonstrations.html
- Other conferences
  - CIKM, ECIR

Submit a Proposal

- Write 2 pages. When writing the proposal you should try to answer the following questions:
  - What is the problem you are solving?
  - What data will you use?
  - What work do you plan to do the project?
  - Which algorithms/techniques/models you plan to use/develop? Be as specific as you can!
  - How do you evaluate your method? How will you test it? How will you measure success?
  - What do you expect to submit/accomplish by the end of the semester?

Checklist for Task 1

- Report your site on ecampus, and your group members on ecampus

- A project proposal (for task 3) with no more than two pages on CSNET – Proposal means preliminary
  - UIN-UIN-UIN-proposal.pdf

- Report of Task 1: no more than two pages on CSNET
  - UIN-UIN-UIN-task1.pdf

Submissions for Task 1

- Submit a zip file containing all of the documents you crawled to CSNET
  - Zip file name: UIN-UIN-UIN.zip
  - At least 100k documents
  - If it is full documents, such as Wikipedia, each file only contains one paragraph from the original document. If it is tweet or post (<140 chars), submit the full docs.

- The deadline for this is **Feb 19th, 11:59pm**
- **Class on Feb 19th is saved for project**

Task 2: Search Engine

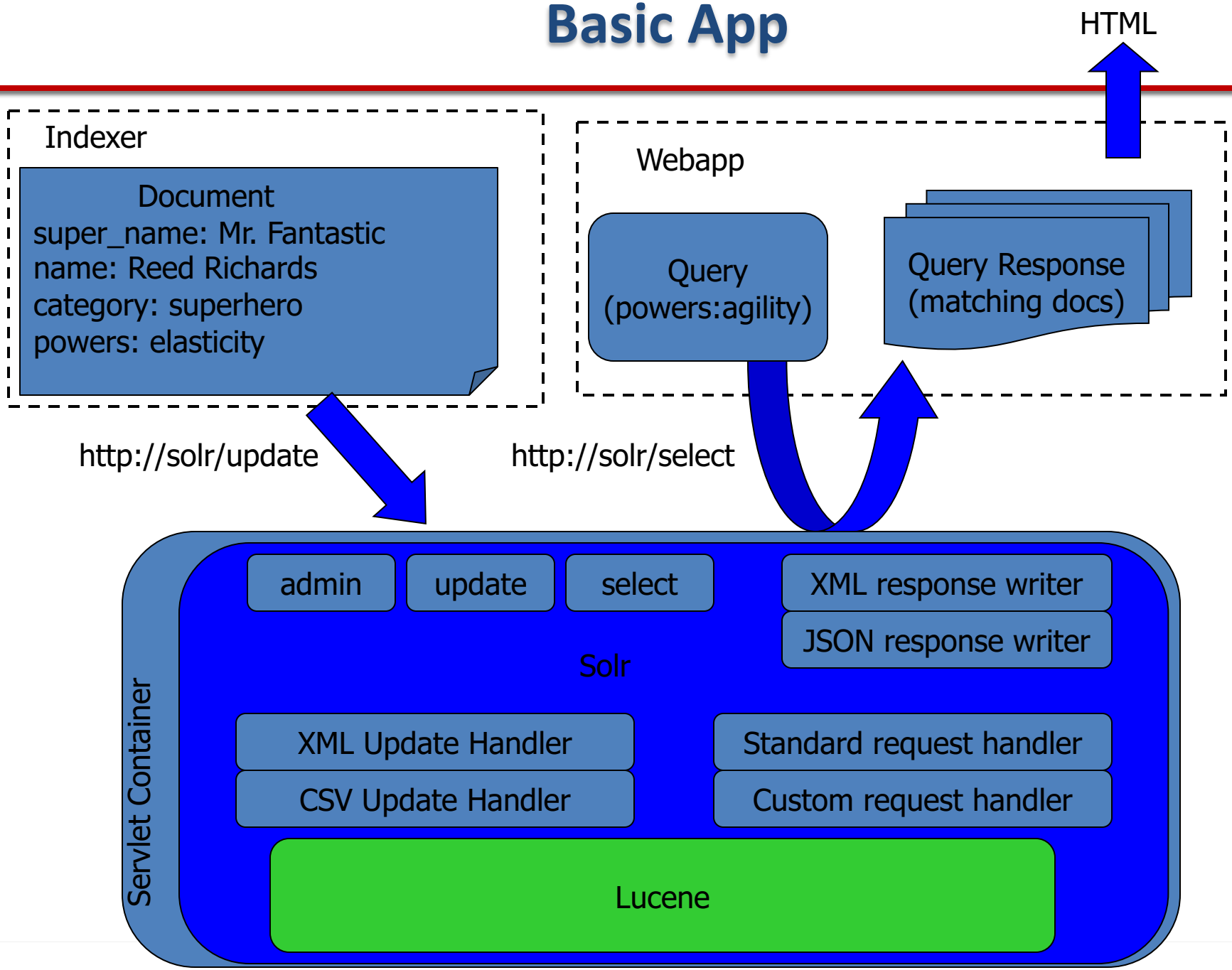
What is Lucene

- High performance, scalable, full-text search **library**
- Focus: Indexing + Searching Documents
  - "Document" is just a list of name+value pairs
- No crawlers or document parsing
- Flexible Text Analysis (tokenizers + token filters)
- 100% Java, no dependencies, no config files

What is Solr

- A full text search server based on Lucene
- XML/HTTP, JSON Interfaces
- Faceted Search (category counting)
- Flexible data schema to define types and fields
- Hit Highlighting
- Configurable Advanced Caching
- Index Replication
- Extensible Open Architecture, Plugins
- Web Administration Interface
- Written in Java5, deployable as a WAR

Basic App


HTML
Indexer
Document
super_name: Mr. Fantastic
name: Reed Richards
category: superhero
powers: elasticity
Webapp
Query
(powers:agility)
Query Response
(matching docs)
http://solr/update
http://solr/select
Servlet Container
admin
update
select
XML response writer
JSON response writer
Solr
XML Update Handler
CSV Update Handler
Standard request handler
Custom request handler
Lucene

Installation & Basic Operations

1. Download
   Solr: http://www.apache.org/dyn/closer.lua/lucene/solr/7.2.1
2. Go to the downloaded folder: solr-7.2.1/bin
3. Terminal: ./solr start
4. browser: http://localhost:8983/solr
5. Create a core
   1. Use web interface
   2. Use command line: (in bin folder) ./solr create -c xing
6. Post you files (documents) to server: ./post -c xing ../example/exampledocs/books.csv
7. Search on server:
   1. http://localhost:8983/solr/xing/select?q=id:0553579908
   2. http://localhost:8983/solr/xing/select?q=cat:book
   3. Question: how do you search name = A Game of Thrones?

Indexing Data

HTTP POST to http://localhost:8983/solr/update

```
<add><doc>
  <field name="id">05991</field>
  <field name="name"></field>
  <field name="supername">SpiPeter Parkerder-Man<
  <field name="category">superhero</field>
  <field name="powers">agility</field>
  <field name="powers">spider-sense</field>
</doc></add>
```

Data upload methods

```
URL=http://localhost:8983/solr/update/csv
```

- HTTP POST body (curl, HttpClient, etc)

```
curl $URL -H 'Content-type:text/plain;
  charset=utf-8' --data-binary @info.csv
```

- Multi-part file upload (browsers)
- Request parameter

```
?stream.body='Cyclops, Scott Summers,…'
```

- Streaming from URL (must enable)

```
?stream.url=file://data/info.csv
```

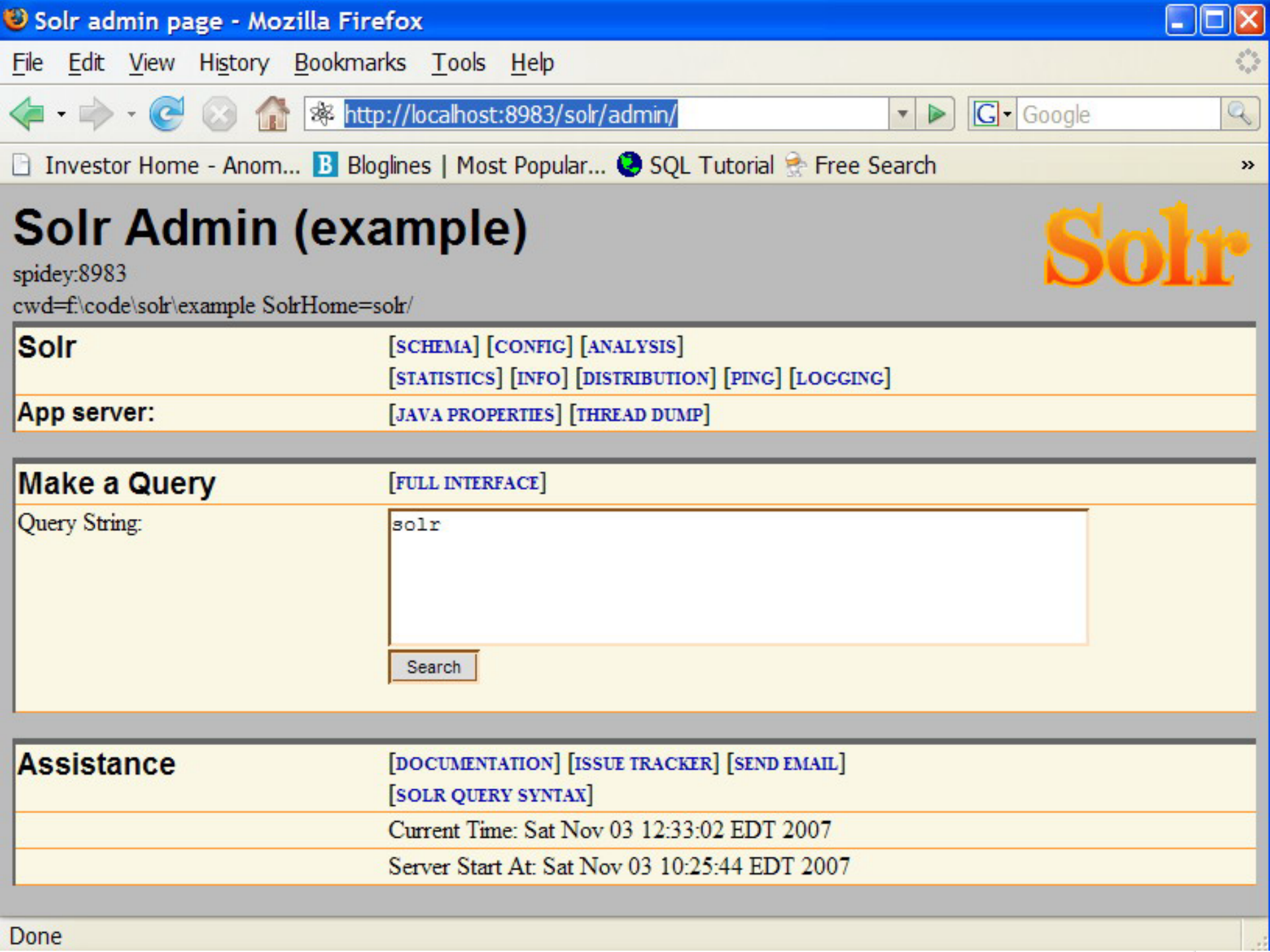
Indexing with SolrJ

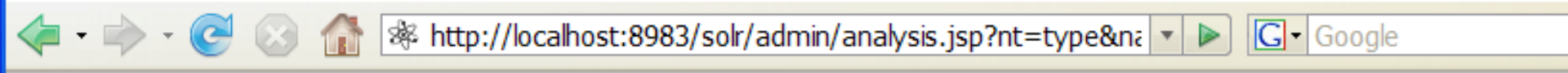
```
// Solr's Java Client API… remote or embedded/local!
SolrServer server = new
    CommonsHttpSolrServer("http://localhost:8983/solr
    ");

SolrInputDocument doc = new SolrInputDocument();
doc.addField("supername","Daredevil");
doc.addField("name","Matt Murdock");
doc.addField("category","superhero");

server.add(doc);
server.commit();
```

Searching

http://localhost:8983/solr/select?q=powers:agility&start=0&rows=2&fl=supername,category

```
<response>
 <result numFound="427" start="0">
  <doc>
   <str name="supername">Spider-Man</str>
   <str name="category">superhero</str>
  </doc>
  <doc>
   <str name="supername">Msytique</str>
   <str name="category">supervillain</str>
  </doc>
 </result>
</response>
```


Solr admin page - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost:8983/solr/admin/
Google
Investor Home - Anom... Bloglines | Most Popular... SQL Tutorial Free Search
Solr Admin (example)
spidey:8983
cwd=f:\code\solr\example SolrHome=solr/
Solr
Solr
[SCHEMA] [CONFIG] [ANALYSIS]
[STATISTICS] [INFO] [DISTRIBUTION] [PING] [LOGGING]
App server:
[JAVA PROPERTIES] [THREAD DUMP]
Make a Query
[FULL INTERFACE]
Query String:
solr
Search
Assistance
[DOCUMENTATION] [ISSUE TRACKER] [SEND EMAIL]
[SOLR QUERY SYNTAX]
Current Time: Sat Nov 03 12:33:02 EDT 2007
Server Start At: Sat Nov 03 10:25:44 EDT 2007
Done



Solr Admin (example)

spidey:8983

cwd=f:\code\solr\example SolrHome=solr/

Field Analysis

Field type

simple

Field value (Index)

verbose output ☐

highlight matches ☑

The Justice League saved the day by defeating Darkseid

Field value (Query)

verbose output ☐

the defeat of DarkSeid

Analyze

Index Analyzer

The	Justice	League	saved	the	day	by	defeating	Darkseid
Justice	League	saved	day	defeating	Darkseid			
justice	league	saved	day	defeating	darkseid			
justic	leagu	save	day	defeat	darkseid			

Query Analyzer

the	defeat	of	DarkSeid
defeat	DarkSeid		
defeat	darkseid		
defeat	darkseid		

Checklist for Task 2

- Report of Task 2: no more than two pages on CSNET
  - UIN-UIN-UIN-task2.pdf

- Video recording to show your interaction with the search engine
  - A link in a text file (web link or youtube link) on CSNET
  - Audio is not required
  - Show your IDs in the video or captions
  - No more than five minutes
  - UIN-UIN-UIN-solr.txt

- The deadline for this is **March 21st, 11:59pm (reading day)**

Task 3: Application

Checklist for Task 3

- Final report: no more than six pages on CSNET
  - UIN-UIN-UIN-project.pdf
  - Reports for task1 and task2 can be re-used

- Video recording to show your demo
  - A link in a text file (web link or youtube link) on CSNET
  - Audio is required
  - Show your IDs in the video or captions
  - No more than five minutes
  - UIN-UIN-UIN-demo.txt

  - The deadline for this is **April 20th, 11:59pm**

Project Report

- **Introduction/Motivation/Problem Definition (25%)**
  What is it that you are trying to solve/achieve and why does it matter.

- **Prior Work (10%)**
  How does your project relate to previous work. Please give a short summary on each paper you cite and include how it is relevant.

- **Model/Algorithm/Method (20%)**
  This is where you give a detailed description of your primary contribution. It is especially important that this part be clear and well written so that we can fully understand what you did.

- **Results and findings (25%)**

- **Style and writing (20%)**
  Overall writing, grammar, organization and neatness.