

# Web Crawling

INF 141: Information Retrieval  
Discussion Session  
Week 4 – Winter 2008

Yasser Ganjisaffar

[yganjisa@ics.uci.edu](mailto:yganjisa@ics.uci.edu)

# Open Source Web Crawlers

**Heritrix**

**Nutch**

**WebSphinx**

**Crawler4j**

# Heritrix

- Extensible, Web-Scale
- Command line tool
- Web-based Management Interface
- Distributed
- Internet Archive's Crawler



# Internet Archive

- dedicated to building and maintaining a free and openly accessible online digital library, including an archive of the Web.

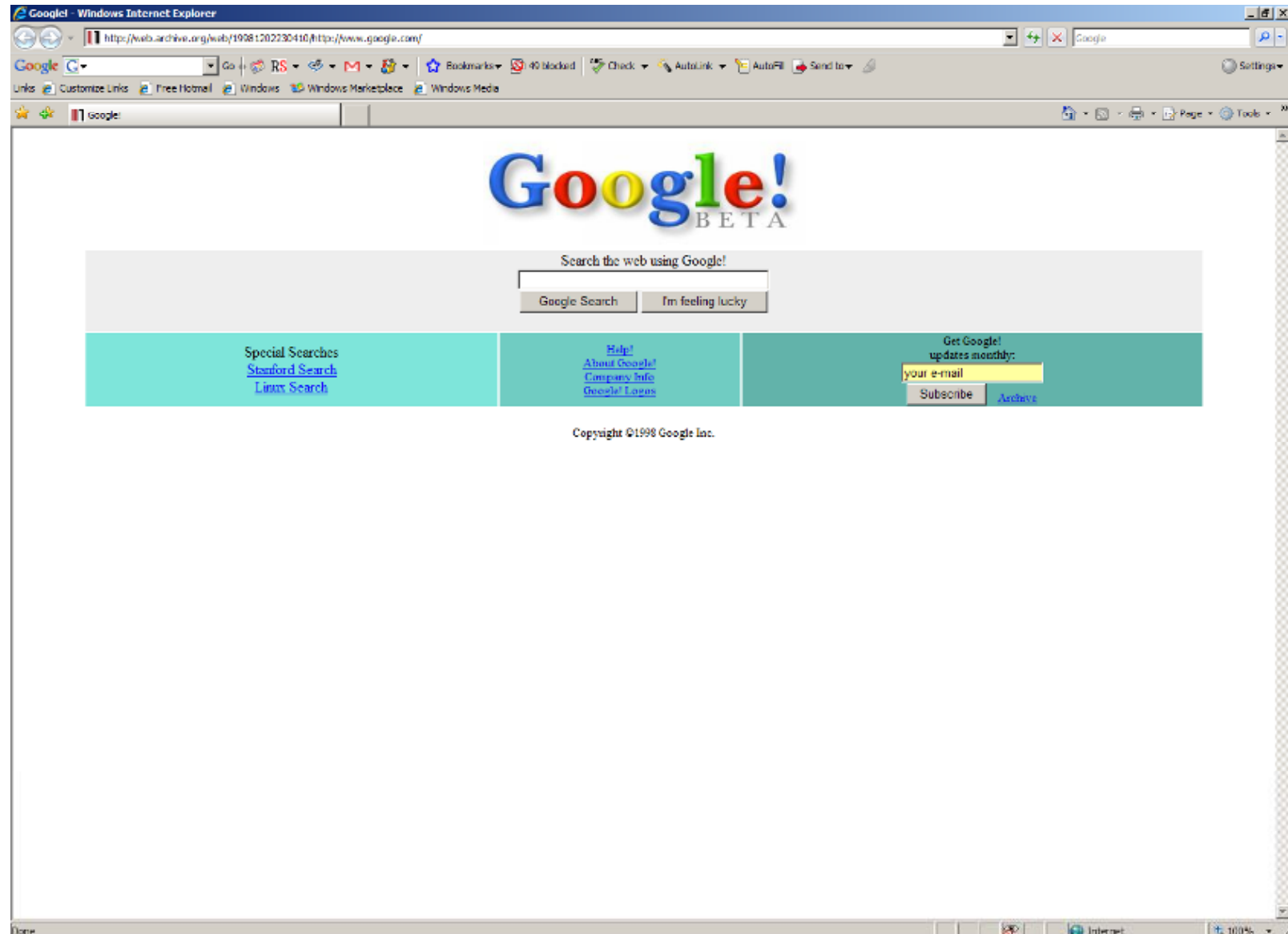


# Internet Archive's WayBack Machine

The screenshot shows a Mozilla Firefox browser window displaying the Internet Archive Wayback Machine search results for the URL <http://google.com>. The search was performed on Jan 01, 1996, to Jul 25, 2008. The results are presented in a table format, showing the number of pages archived for each year and a list of specific dates when the site was updated. The table is as follows:

Search Results for Jan 01, 1996 - Jul 25, 2008													
1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	
0 pages	0 pages	2 pages	12 pages	73 pages	685 pages	154 pages	61 pages	205 pages	910 pages	348 pages	458 pages	24 pages	
		<a href="#">Nov 11, 1998</a> * <a href="#">Dec 02, 1998</a> *	<a href="#">Jan 17, 1999</a> * <a href="#">Jan 25, 1999</a> * <a href="#">Feb 08, 1999</a> * <a href="#">Apr 22, 1999</a> * <a href="#">Apr 23, 1999</a> * <a href="#">Apr 27, 1999</a> * <a href="#">Apr 28, 1999</a> * <a href="#">May 08, 1999</a> * <a href="#">Oct 01, 1999</a> * <a href="#">Oct 12, 1999</a> * <a href="#">Nov 06, 1999</a> * <a href="#">Nov 29, 1999</a> *	<a href="#">Jan 17, 1999</a> * <a href="#">Jan 25, 1999</a> * <a href="#">Mar 01, 2000</a> * <a href="#">Mar 02, 2000</a> * <a href="#">Mar 03, 2000</a> * <a href="#">Apr 07, 2000</a> * <a href="#">Apr 08, 2000</a> * <a href="#">Apr 09, 2000</a> * <a href="#">May 10, 2000</a> * <a href="#">May 10, 2000</a> * <a href="#">May 10, 2000</a> * <a href="#">May 10, 2000</a> *	<a href="#">Feb 29, 2000</a> * <a href="#">Mar 01, 2000</a> * <a href="#">Mar 02, 2000</a> * <a href="#">Mar 03, 2000</a> * <a href="#">Apr 07, 2000</a> * <a href="#">Apr 08, 2000</a> * <a href="#">Apr 09, 2000</a> * <a href="#">May 10, 2000</a> * <a href="#">May 10, 2000</a> * <a href="#">May 10, 2000</a> * <a href="#">May 10, 2000</a> *	<a href="#">Jan 18, 2001</a> * <a href="#">Jan 19, 2001</a> * <a href="#">Jan 19, 2001</a> * <a href="#">Jan 19, 2001</a> * <a href="#">Jan 19, 2001</a> * <a href="#">Jan 19, 2001</a> * <a href="#">Jan 19, 2001</a> * <a href="#">Jan 19, 2001</a> * <a href="#">Jan 19, 2001</a> * <a href="#">Jan 19, 2001</a> * <a href="#">Jan 19, 2001</a> * <a href="#">Jan 19, 2001</a> *	<a href="#">Jan 23, 2002</a> * <a href="#">Jan 24, 2002</a> * <a href="#">Jan 24, 2002</a> * <a href="#">Feb 06, 2002</a> * <a href="#">Feb 22, 2002</a> * <a href="#">Feb 23, 2002</a> * <a href="#">Mar 31, 2002</a> * <a href="#">Apr 02, 2002</a> * <a href="#">May 23, 2002</a> * <a href="#">May 25, 2002</a> * <a href="#">Jun 02, 2002</a> * <a href="#">Jun 04, 2002</a> * <a href="#">Jun 05, 2002</a> * <a href="#">Jul 02, 2002</a> * <a href="#">Jul 03, 2002</a> * <a href="#">Jul 04, 2002</a> *	<a href="#">Feb 02, 2003</a> * <a href="#">Feb 04, 2003</a> * <a href="#">Feb 05, 2003</a> * <a href="#">Feb 08, 2003</a> * <a href="#">Feb 14, 2003</a> * <a href="#">Feb 15, 2003</a> * <a href="#">Feb 17, 2003</a> * <a href="#">Feb 17, 2003</a> * <a href="#">Mar 24, 2003</a> * <a href="#">Mar 28, 2003</a> * <a href="#">Mar 29, 2003</a> * <a href="#">Apr 02, 2003</a> * <a href="#">Apr 03, 2003</a> * <a href="#">Apr 03, 2003</a> * <a href="#">Apr 09, 2003</a> * <a href="#">Apr 21, 2003</a> * <a href="#">Apr 23, 2003</a> *	<a href="#">Jan 03, 2004</a> * <a href="#">Jan 13, 2004</a> * <a href="#">Jan 21, 2004</a> * <a href="#">Jan 26, 2004</a> * <a href="#">Jan 29, 2004</a> * <a href="#">Feb 08, 2004</a> * <a href="#">Feb 11, 2004</a> * <a href="#">Feb 15, 2004</a> * <a href="#">Feb 18, 2004</a> * <a href="#">Feb 25, 2004</a> * <a href="#">Feb 27, 2004</a> * <a href="#">Mar 06, 2004</a> * <a href="#">Mar 25, 2004</a> * <a href="#">Mar 31, 2004</a> * <a href="#">Apr 01, 2004</a> * <a href="#">Apr 03, 2004</a> *	<a href="#">Jan 01, 2005</a> * <a href="#">Jan 01, 2005</a> * <a href="#">Jan 02, 2005</a> * <a href="#">Jan 02, 2005</a> * <a href="#">Jan 03, 2005</a> * <a href="#">Jan 05, 2005</a> * <a href="#">Jan 06, 2005</a> * <a href="#">Jan 06, 2005</a> * <a href="#">Jan 07, 2005</a> * <a href="#">Jan 07, 2005</a> * <a href="#">Jan 07, 2005</a> * <a href="#">Jan 08, 2005</a> * <a href="#">Jan 09, 2005</a> * <a href="#">Jan 09, 2005</a> * <a href="#">Jan 11, 2005</a> * <a href="#">Jan 11, 2005</a> *	<a href="#">Jan 01, 2006</a> * <a href="#">Jan 01, 2006</a> * <a href="#">Jan 01, 2006</a> * <a href="#">Jan 02, 2006</a> * <a href="#">Jan 02, 2006</a> * <a href="#">Jan 02, 2006</a> * <a href="#">Jan 03, 2006</a> * <a href="#">Jan 03, 2006</a> * <a href="#">Jan 03, 2006</a> * <a href="#">Jan 03, 2006</a> * <a href="#">Jan 03, 2006</a> * <a href="#">Jan 04, 2006</a> * <a href="#">Jan 04, 2006</a> * <a href="#">Jan 04, 2006</a> * <a href="#">Jan 04, 2006</a> * <a href="#">Jan 05, 2006</a> * <a href="#">Jan 05, 2006</a> *	<a href="#">Jan 01, 2007</a> * <a href="#">Jan 01, 2007</a> * <a href="#">Jan 02, 2007</a> * <a href="#">Jan 02, 2007</a> * <a href="#">Jan 03, 2007</a> * <a href="#">Jan 03, 2007</a> * <a href="#">Jan 04, 2007</a> * <a href="#">Jan 04, 2007</a> * <a href="#">Jan 05, 2007</a> * <a href="#">Jan 05, 2007</a> * <a href="#">Jan 06, 2007</a> * <a href="#">Jan 06, 2007</a> * <a href="#">Jan 06, 2007</a> * <a href="#">Jan 07, 2007</a> * <a href="#">Jan 07, 2007</a> * <a href="#">Jan 08, 2007</a> *	<a href="#">Jan 02, 2008</a> * <a href="#">Jan 02, 2008</a> * <a href="#">Jan 14, 2008</a> * <a href="#">Jan 16, 2008</a> * <a href="#">Jan 19, 2008</a> * <a href="#">Jan 22, 2008</a> * <a href="#">Jan 28, 2008</a> * <a href="#">Feb 07, 2008</a> * <a href="#">Feb 08, 2008</a> * <a href="#">Feb 09, 2008</a> * <a href="#">Feb 11, 2008</a> * <a href="#">Feb 13, 2008</a> * <a href="#">Feb 15, 2008</a> * <a href="#">Feb 16, 2008</a> * <a href="#">Feb 22, 2008</a> * <a href="#">Feb 27, 2008</a> *

# Google - 1998



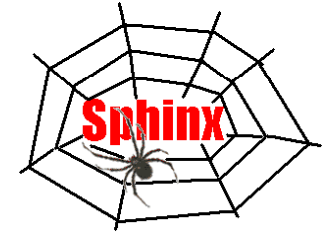
# Nutch



- Apache's Open Source Search Engine
- Distributed
- Tested with 100M Pages

# WebSphinx

- 1998-2002
- Single Machine
- Lots of Problems (Memory leaks, ...)
- Reported to be very slow





# Crawler4j

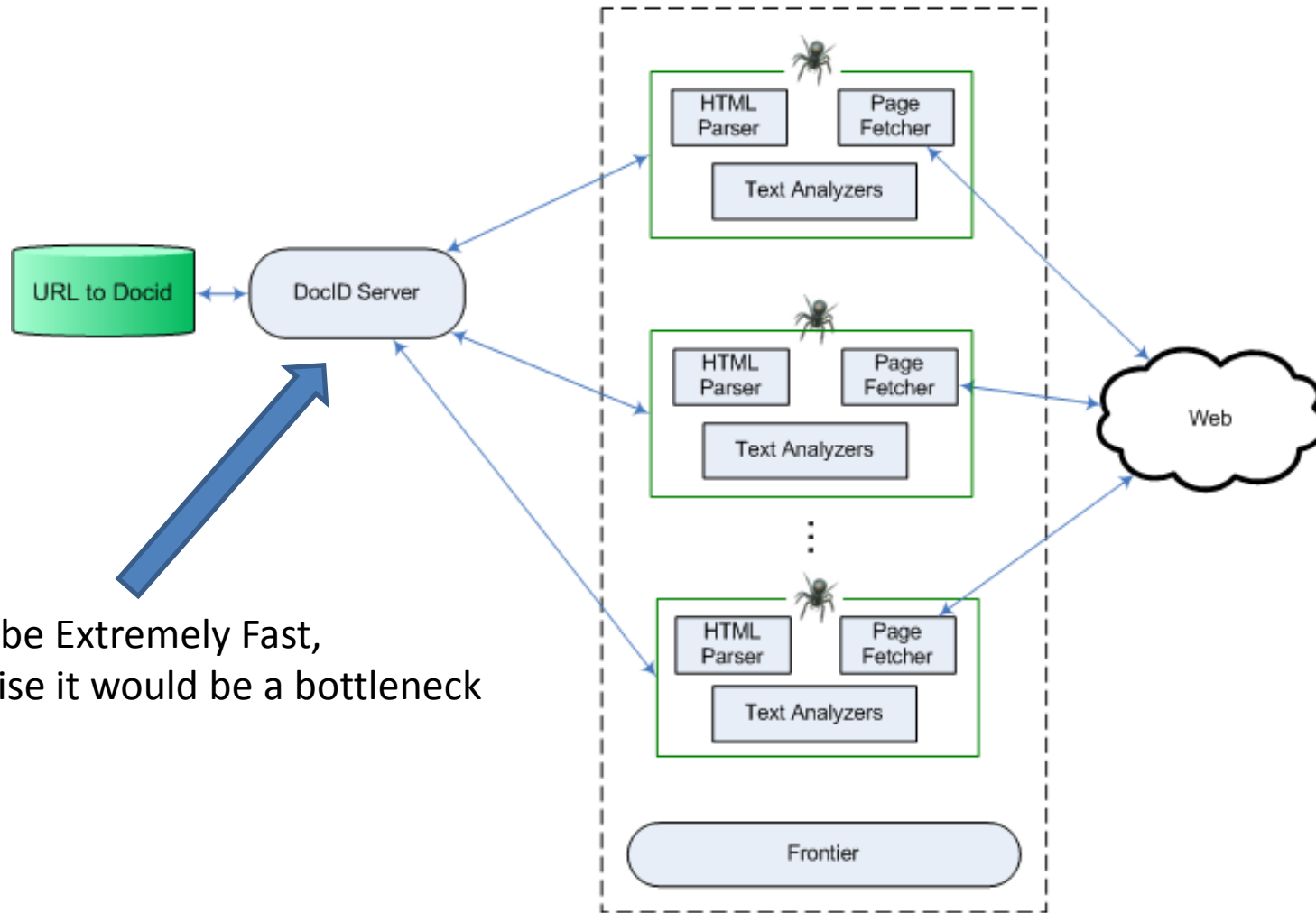
- Single Machine
- Should Easily Scale to 20M Pages
- Very Fast

Crawled and Processed the whole English Wikipedia in 10 hours (including time for extracting palindromes and storing link structure and text of articles).

# What is a docid?

- A unique sequential integer value that uniquely identifies a Page/URL.
- Why use it?
  - Storing links:
    - “http://www.ics.uci.edu/”-”http://www.ics.uci.edu/about” (53 bytes)
    - 120-123 (8 bytes)
  - ...

# Docid Server



Should be Extremely Fast,  
otherwise it would be a bottleneck

# Docid Server

```
public static synchronized int getDocID(String URL) {  
    if (there is any key-value pair for key = URL) {  
        return value;  
    } else {  
        lastdocid++;  
        put (URL, lastdocid) in storage;  
        return lastdocid;  
    }  
}
```

# Docid Server

- Key-value pairs are stored in a B+-tree data structure.
- Berkeley DB as the storage engine

# Berkeley DB

- Unlike traditional database systems like MySQL and others, Berkeley DB comes in form of a jar file which is linked to the Java program and runs in the process space of the crawlers.
- No need for inter-process communication and waiting for context switch between processes.
- You can think of it as a large HashMap:



# Crawler4j

- It is not polite
  - Does not respects robots.txt limitations
  - Does not limit number of requests sent to a host per second.
    - For example:
      - Wikipedia’s policy does not allow bots to send requests faster than 1 request/second.
      - Crawler4j has a history of sending 200 requests/second
  - Introduces itself as a Firefox agent!
    - Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.9.0.4) Gecko/2008102920 Firefox/3.0.4
    - Compare with Google’s user agent:
      - Mozilla/5.0 (compatible; Googlebot/2.1; http://www.google.com/bot.html)

# Crawler4j

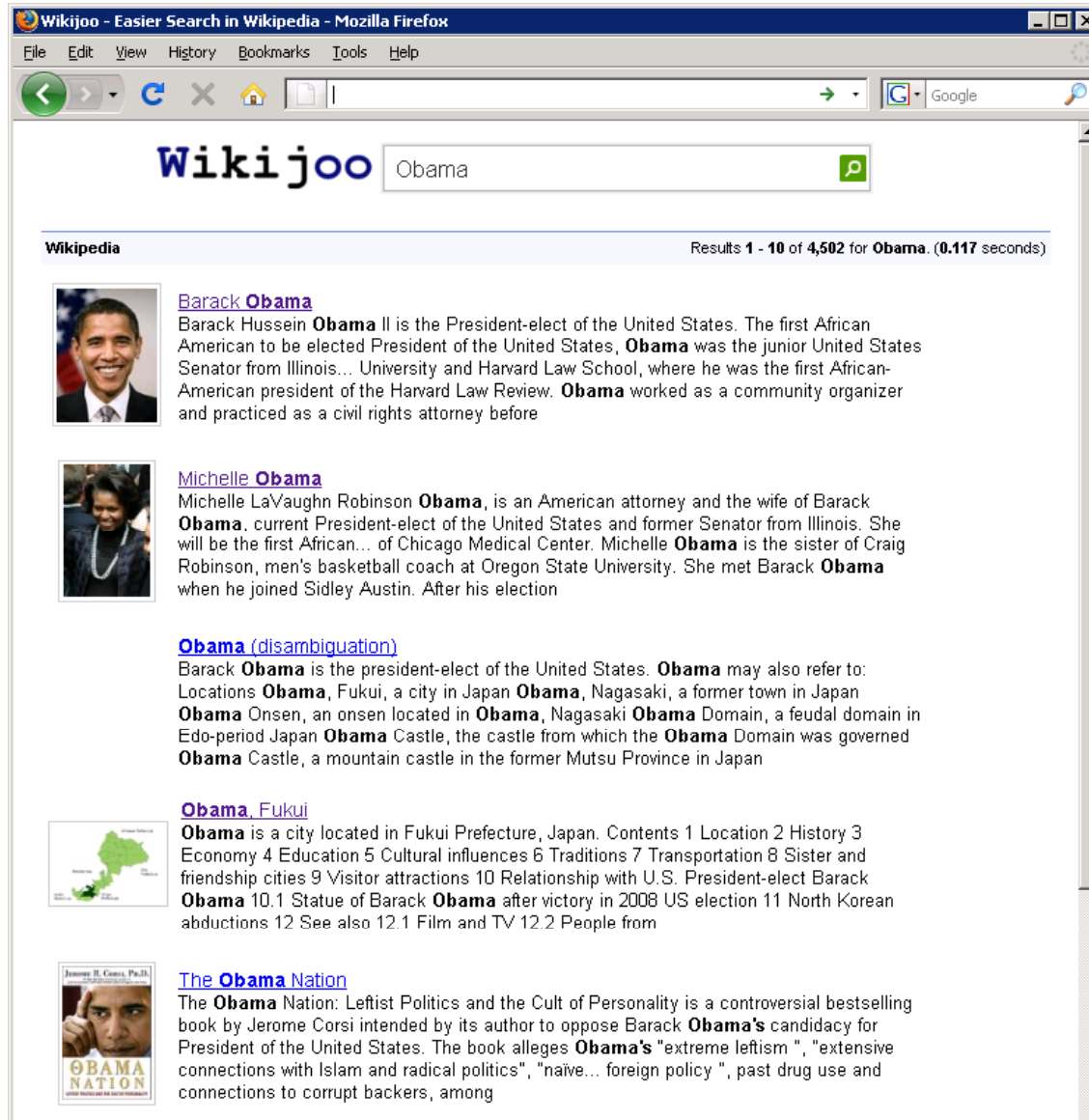
- Only Crawls Textual Content
  - Don't try to download images and other media with it.
- Assumes that page is encoded in UTF-8 format.



# Crawler4j

- There is another version of crawler4j which is:
  - Polite
  - Supports all types of content
  - Supports all encodings of text and automatically detects the encoding
  - Not open source ;)

# Crawler4j – Crawler of Wikijoo



The screenshot shows a Mozilla Firefox browser window with the Wikijoo search engine. The search term "Obama" is entered in the search bar. The results page displays several entries:

- Wikipedia** Results 1 - 10 of 4,502 for Obama. (0.117 seconds)
- Barack Obama**: Barack Hussein **Obama** II is the President-elect of the United States. The first African American to be elected President of the United States, **Obama** was the junior United States Senator from Illinois... University and Harvard Law School, where he was the first African-American president of the Harvard Law Review. **Obama** worked as a community organizer and practiced as a civil rights attorney before
- Michelle Obama**: Michelle LaVaughn Robinson **Obama**, is an American attorney and the wife of Barack **Obama**, current President-elect of the United States and former Senator from Illinois. She will be the first African... of Chicago Medical Center. Michelle **Obama** is the sister of Craig Robinson, men's basketball coach at Oregon State University. She met Barack **Obama** when he joined Sidley Austin. After his election
- Obama (disambiguation)**: Barack **Obama** is the president-elect of the United States. **Obama** may also refer to: Locations **Obama**, Fukui, a city in Japan **Obama**, Nagasaki, a former town in Japan **Obama** Onsen, an onsen located in **Obama**, Nagasaki **Obama** Domain, a feudal domain in Edo-period Japan **Obama** Castle, the castle from which the **Obama** Domain was governed **Obama** Castle, a mountain castle in the former Mutsu Province in Japan
- Obama, Fukui**: **Obama** is a city located in Fukui Prefecture, Japan. Contents 1 Location 2 History 3 Economy 4 Education 5 Cultural influences 6 Traditions 7 Transportation 8 Sister and friendship cities 9 Visitor attractions 10 Relationship with U.S. President-elect Barack **Obama** 10.1 Statue of Barack **Obama** after victory in 2008 US election 11 North Korean abductions 12 See also 12.1 Film and TV 12.2 People from
- The Obama Nation**: The **Obama** Nation: Leftist Politics and the Cult of Personality is a controversial bestselling book by Jerome Corsi intended by its author to oppose Barack **Obama's** candidacy for President of the United States. The book alleges **Obama's** "extreme leftism", "extensive connections with Islam and radical politics", "naïve... foreign policy", past drug use and connections to corrupt backers, among

# Your job?

```
public class MyCrawler extends WebCrawler {  
  
    Pattern filters = Pattern.compile(".*(\\. (css|js|bmp|gif|jpe?g"  
        + "|png|tiff?|mid|mp2|mp3|mp4" + "|wav|avi|mov|mpe|ram|m4v|pdf"  
        + "|rm|smil|wmv|swf|wma|zip|rar|gz))$");  
  
    public MyCrawler() {  
    }  
  
    public boolean shouldVisit(WebURL url) {  
        String href = url.getURL().toLowerCase();  
        if (filters.matcher(href).matches()) {  
            return false;  
        }  
        if (href.startsWith("http://ics.uci.edu/")) {  
            return true;  
        }  
        return false;  
    }  
  
    public void visit(Page page) {  
        int docid = page.getWebURL().getDocid();  
        String url = page.getWebURL().getURL();  
        String title = page.getTitle();  
        String text = page.getText();  
        String html = page.getHTML();  
        ArrayList<WebURL> links = page.getURLs();  
    }  
}
```

# Your job?

```
public class Controller {  
    public static void main(String[] args) throws Exception {  
        CrawlController controller = new CrawlController("/extra/grad_space/yganjisa/crawl");  
        controller.addSeed("http://ics.uci.edu/");  
        controller.start(MyCrawler.class, 10);  
    }  
}
```

# Crawler4j Objects

- Page
  - String html: getHTML()
  - String text: getText()
  - String title: getTitle()
  - WebURL url: getWebURL()
  - ArrayList<WebURL> urls: getURLs()
- WebURL
  - String url: getURL()
  - int docid: getDocid()

# Assignment 03

- Feel free to use any crawler you like.
- You might want to filter pages which are not in the main namespace:
  - [http://en.wikipedia.org/wiki/\*\*Wikipedia\*\*:Searching](http://en.wikipedia.org/wiki/Wikipedia:Searching)
  - [http://en.wikipedia.org/wiki/\*\*Category\*\*:Linguistics](http://en.wikipedia.org/wiki/Category:Linguistics)
  - [http://en.wikipedia.org/wiki/\*\*Talk\*\*:Main\\_Page](http://en.wikipedia.org/wiki/Talk:Main_Page)
  - [http://en.wikipedia.org/wiki/\*\*Special\*\*:Random](http://en.wikipedia.org/wiki/Special:Random)
  - Image:, File:, Help:, Media:, ...
- Set Maximum heap size for java:
  - `java -Xmx1024M -cp .:crawler4j.jar ir.assignment03.Controller`

# Assignment 03

- Dump your partial results
  - For example, after processing each 5000 page write the results in a text file.
- Use *nohup* command on remote machines.
  - `nohup java -Xmx1024M -cp .:crawler4j.jar ir.assignment03.Controller`

# Shell Scripts

run.sh:

```
#!/bin/bash
cp="."
for f in $(ls lib/*); do
    cp=$cp:$f
done
java -Xmx2048M -classpath $cp ir.assignment03.Controller
```

Available online: <http://www.ics.uci.edu/~yganjisa/TA/>



# Shell Scripts

## run-nohup.sh:

```
#!/bin/bash
cp="."
for f in $(ls lib/*); do
    cp=$cp:$f
done
nohup java -Xmx2048M -classpath $cp ir.assignment03.Controller > crawl-log.txt
```

## Check logs:

```
tail -f crawl-log.txt
```

Available online: <http://www.ics.uci.edu/~yganjisa/TA/>

# Tips

- Do not print on screen frequently.
- Your seeds should be accepted in the shouldVisit() function.
  - Pattern: `http://en.wikipedia.org/wiki/*`
  - Bad seed:
    - `http://en.wikipedia.org/`
  - Good seed:
    - `http://en.wikipedia.org/wiki/Main_Page`

# Openlab

- From a Linux/MAC machine:
  - Connecting to shell:
    - `ssh myicsid@openlab.ics.uci.edu`
  - File transfer:
    - scp or other tools
- From a Windows machine:
  - Connecting to shell:
    - Download putty.exe
      - <http://www.chiark.greenend.org.uk/~sgtatham/putty/download.html>
  - File transfer:
    - WinSCP (<http://winscp.net/>)

**QUESTIONS?**