# CSE 591
# Sematic Web Mining

A demo on regular expression, web scraping, and web crawling

BY

Sultan Alzahrani

# We will go over:

- Regular Expressions
- Websites scraper (java+jsoup)
- Crawler jsoup + database to store html files….

# Regular Expression

- Regular expression or rational expression.
- Sequence of defined **string** that describe a search **pattern**.

**String**:

- metacharacters with special meaning.
- Regular character with its literal meaning

- Example:
  - Seriali[sz]e : match either Serialize or Serialise
  - Regular expression that match any number:
    - ^[+-]?(\d+\.?\d*|\.\d+)([eE][+-]?\d+)?$

# Regular Expression

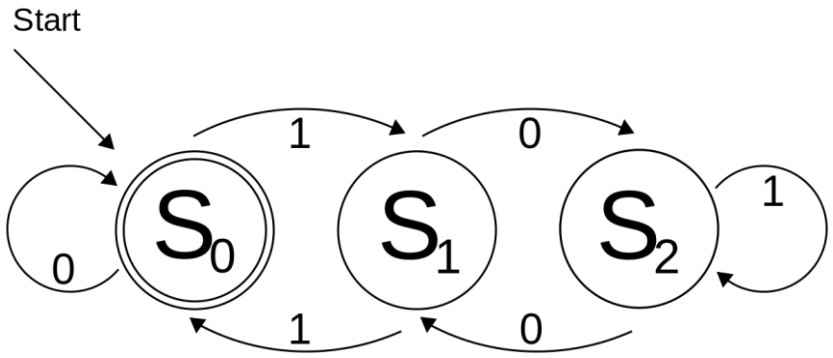
- To match a date in dd-mm-yyyy format
  - ^(0[1-9]|[12][0-9]|3[01])[-\/\.](0[1-9]|1[012])[-\/\.](19|20)\d\d$
  - To match a date in mm-dd-yyyy format
  - ^(0[1-9]|1[012])[-\/\.](0[1-9]|[12][0-9]|3[01])[-\/\.](19|20)\d\d$
  - **What the next regular means?**
    - **(\w+)(\d+)((\1+)(\2+))  A(nt|pple)\1A(?:nt|pple)\1**

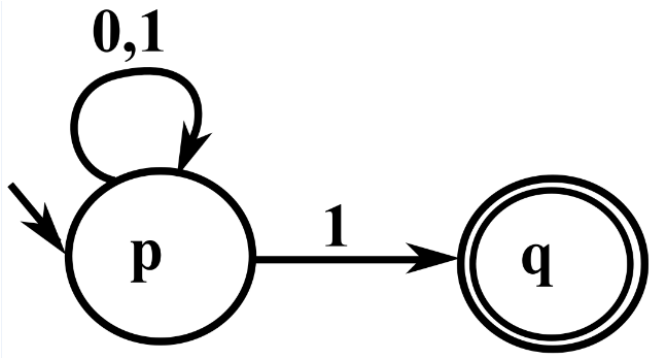
# Regular Expression

- It is supported by many programming languages:
  - Perl
  - Javascript
  - Ruby
  - Java
  - Python
  - C++
  - More….
- But each engine may differ on how to handle and interpret a regular expression.

# Regular Expression

- Regular expression can be used to check for pattern match or to extract information.
- Two kinds of regular expression engines (Deterministic Finite Autonomota) DFA and (Nondeterministic Finite Autonomota) NFA



# Regular Expression

- engines
  - NFA
    - Supports lazy quantifier e.g. sample match: \w{2,4}? ab in **ab**cd
    - Backreferences **A(nt|pple)\1**

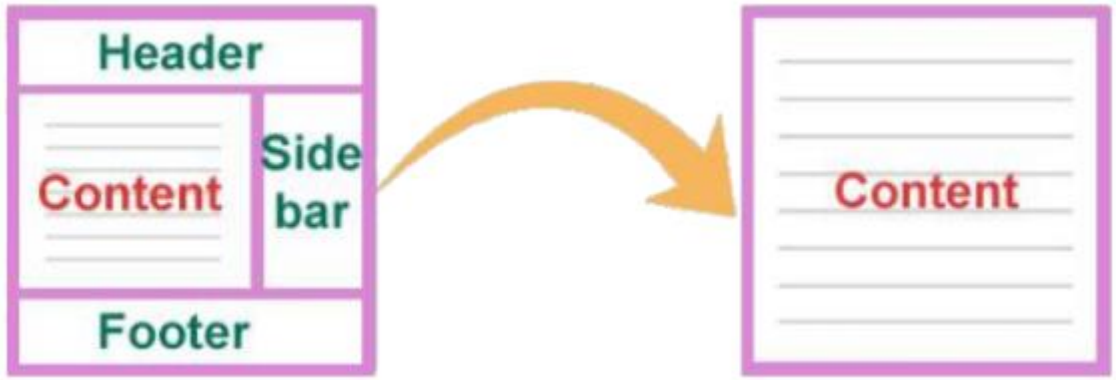
# Examples and References

- Regular Expression for address list: java & python.

- References:
  - http://www.rexegg.com/regex-quickstart.html
  - (Java) http://docs.oracle.com/javase/7/docs/api/java/util/regex/Pattern.html
  - (Python) https://docs.python.org/2/howto/regex.html
  - (The complete tutorial) http://www.princeton.edu/~mlovett/reference/Regular-Expressions.pdf

# Extracting textual data from webpages (Boilerpipe)

- Excluding out markup boilerplate
- URL: https://boilerpipe-web.appspot.com/
- The boilerpipe library provides algorithms to detect and remove the surplus "clutter" (boilerplate, templates) around the main textual content of a web page.



# Boilerpipe Cont.

- **Benefits:**
  - Much smarter than the regular expression.
  - Provides several extraction methods.
  - Returns text in a variety of formats.
  - Helps to avoid manual process of finding content pattern from the source site.
  - Helps to remove boilerplates like headers, footers, menus and advertisements.
- **Extraction Methods**
  - **ArticleExtractor:** A full-text extractor which is specialized on extracting articles. It is having higher accuracy than DefaultExtractor.
  - **DefaultExtractor:** A full-text extractor, but not as good as ArticleExtractor.
  - **LargestContentExtractor:** Like DefaultExtractor, it keeps the largest content block similar to DefaultExtractor.
  - **KeepEverythingExtractor:** Gets everything. We can use this for extracting the title and description.

# Boilerpipe Cont.

- Example: WebScraper\src\TryBoilerPipe.java
- A similar library “Goose” written in scala.

# Web Scraper

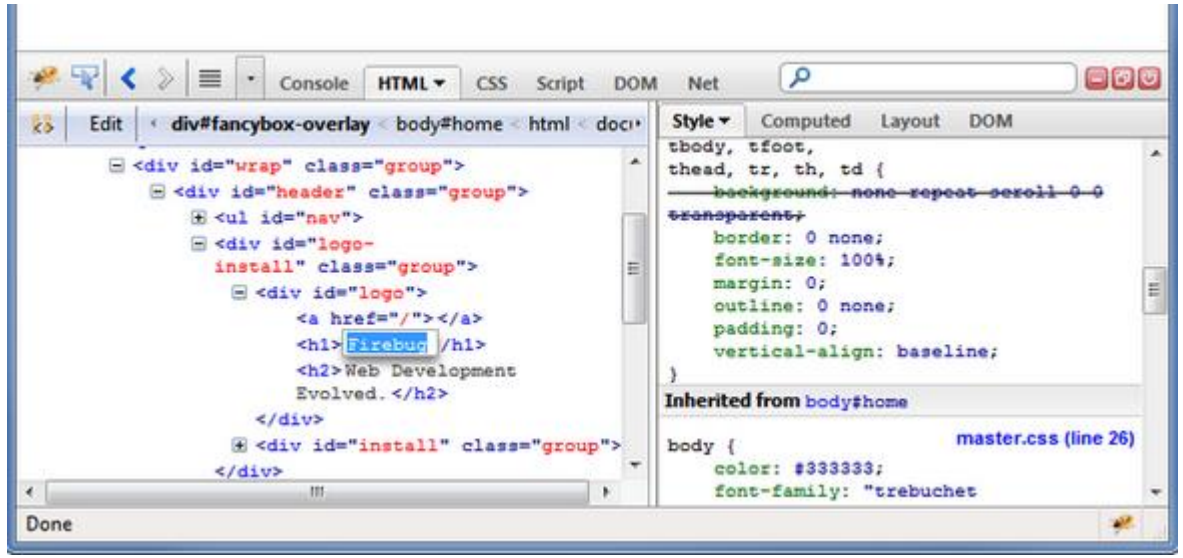
- Extracting information from websites.
- Which directory you are allowed to crawl
  - http://www.asu.edu/robots.txt
- Two main extracting tool:
  - (1) Xpath
    - used to navigate through elements and attributes in an XML document.
    - Syntax (http://www.w3schools.com/xsl/xpath_syntax.asp)
      - Xml: Read xml document and query it
      - Java: http://enira.net/?p=497

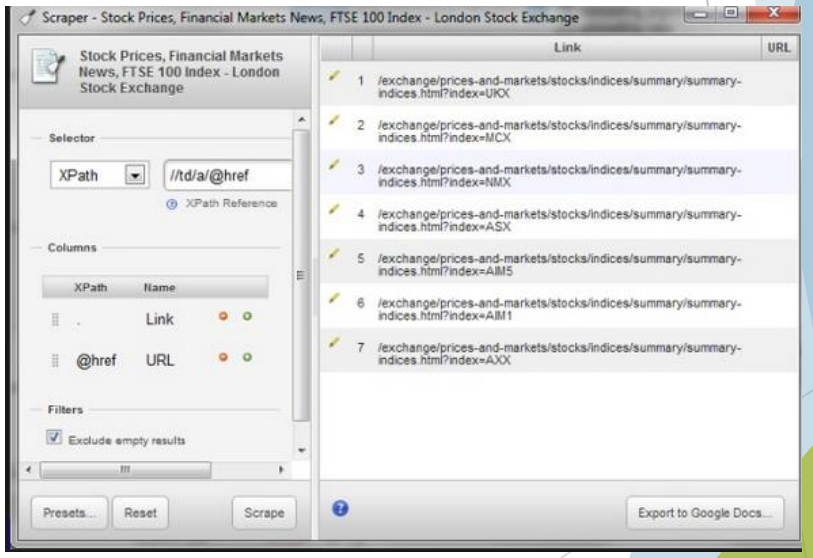
# Web Scraper

- (2) CSSselector
  - Likewise xpath, patterns are used to select the element(s)
  - Widely used with better compatibility and more reliability with inconsistent html pages
  - Syntax: http://www.w3schools.com/cssref/css_selectors.asp
  - Java example (jsoup) and syntax. http://jsoup.org/cookbook/extracting-data/selector-syntax
  - Java libraries: **Selenium** **(better for dynamic contents)and** **Jsoup**
  - Example: **WebScraper/Main.java**
- **For more scraping and interact with website, you may use HmlUnit java library.**
- **R programmers cab rvest package.**

# Internet Explorer plug-in

- Firefox => firebug
- Chrome =>web scraper ad-on



# Web Crawler

- Known as web spider or web robots, bots.
- Web crawlers are mainly used to create a copy of all the visited pages for later processing by a search engine, that will index the downloaded pages to provide fast searches.
- Different media files can be stored to database along with html files.
- Different APIs support target crawling
  - Example: https://www.kimonolabs.com/
- Java Example WebCrawler_v4
- Some websites offers platforms to connect their resources with some secure communication.