Maciej Medyk – COT6777 – Web Mining

Question 1 – [5.25pt]

What is HTTP? What is HTML? What are their relationships, and how are they related to web mining? [0.75pt]

HTTP is Hyper Text Transfer Protocol which is an application protocol designed to exchange or transfer hypertext. HTML is Hyper Text Markup Language which is a language used for creating web document structure. HTTP protocol is used as web application layer protocol that handles requests and responses done by the server. When response is set and page is found the server will sent files that include HTML or XHTML for client browser to display. Website consists of HTML or XHTML which by tags references various objects like image, JS applet, audio file, etc. HTTP is stateless meaning that server doesn't maintain information about past client requests. Web mining is using both in order to crawl the internet in order to search for content. Web mining program uses HTTP to request pages and receives a response from the server in form of HTML website that it saves and mines for additional links and content sentiment in order to narrow down scope of search and maximize effect of the web mining in order to bring better results.

What is a web crawler? What is the taxonomy of the web crawler? Please list major components of a web crawler. [0.75pt]

Web crawler is a program which systematically browses the web for purpose of indexing and obtaining links from the websites in order map out all of the connections. Web crawlers are used by web search engines and often save a portion of the website for processing later. Crawlers can validate hyperlinks and HTML code or be used for webs crapping. Web crawlers are divided into universal crawlers or preferential crawlers. Universal crawlers crawl though sites without semantic focus while preferential crawlers focus on specific topics and pull website links and content only for the sites of interest. Preferential crawlers are divided into focused crawlers and topical crawlers. Major components of web crawler include page repository, text index, page rank, and text & link analysis.

To develop a large scale universal crawler, the two major issues include "performance" and "policy", please identify at least three components on how to improve the performance and how to design good policy, respectively. [0.75pt]

When developing large scale universal crawler performance of that crawler can be improved if we make sure that page is not downloaded twice. One of the ways is to use HEAD command to return information about the page without downloading a page and comparing it to preexisting downloaded page information. Another way is to compare the link to an array of stored links visited before in order to avoid traveling loops. Second are where performance can be improved is limiting the size of the download from each page to predefined size rather than downloading entire page. Third area to improve performance is to create timeout mechanisms in case the page contains malicious code that hangs the crawler. In case of policy improvements we need to address coverage and freshness. Coverage addresses the new pages that are added while freshness addresses the pages that are recently changed. In both policy that would need to address it would have to record how often on average does the site updates and releases new content. If the site would create more new content on average than other sites then it should be revisited more often. The policy should scan new URL more frequently for the first six months and based on update average it should decide if it should prioritize the URL for more frequent crawling. Third area of policy to cover is should if it should prioritize the content at all as it creates crawler bias and if it does how should that prioritization be defined. If it should be defined by freshness of the page then tracking mechanism needs to be in place as far as how often the site refreshes it pages on average per set period of time.

What is a spider trap (or crawler trap) in web crawling? Please suggest at least three approaches (heuristics) for a web crawler to identify a spider trap. [0.75pt]

Spider traps exist when there are indefinite number sites that are dynamically generated and point to one server and often create loops in which spider is trapped in. Another way the spider can be trapped is when the site directory is indefinitely deep. One way to defend from web crawler traps is to check the length of URL string. If the URL is longer than let's say 128 characters then the link is not pursued. Another way of defending against spider traps is to monitor number of URL per site. If the number of URL per site is very large that may mean that the pages are generated for purposes of trapping the bot. Third defense could be to disable crawling of dynamic pages if bot can detect the presence of dynamic pages by certain attributes.

What is Robots Exclusion Standard? Please use a real-world robots.txt example (e.g., http://www.google.com/robots.txt) to explain the key fields (allow, disallow, sitemap etc.) and the setting of the robots exclusion protocol. [0.75pt]

The Robot Exclusion Standard for web crawlers uses a file called robots.txt which is placed in the root directory of the page. The file specifies which bot can is disallowed from crawling what folders or files by line User-agent. For example if robots.txt says *User-agent: Googlebot* *Disallow: /img/** that means that Google crawler is not allowed to go inside the img folder and crawl anything that is contained there. Another example is if robots.txt contains *User-agent: MSNBot* *Disallow:* that means that MSN/Windows Live bot is allowed everywhere. If there is additional line Crawl-delay: 2 that means that bot needs to slow down in order to abide by the rules. Finally if the robot.txt says *User-agent: ** *Disallow: /* that means that all agents are disallowed from crawling everywhere. There also could be keyword Sitemap which describes all available links of the site and presents the bot with all the links without crawling the site. There is also keyword Allow which is opposite of disallowed which specify what areas bots are fully permitted to.

What is Cloaking and Spamdexing? Please explain how they are used to disguise the webservers and affect a search engine's ranking. [0.75pt]

Cloaking occurs when crawling bot sees substantially different website than what regular user sees. One of the prime examples is serving HTML text website to the search engines while serving Flash rich website to the users. Another area that often goes with Cloaking is Spamdexing where sites are loaded with a lot of specific keywords that are not visible to a user but visible to the crawling bot. So many times the site may have ton of text in its html that is hidden from a user, but visible to the crawling bot that contain only keywords and combinations of certain marketing phrases. Spamdexing also is also known as repetitive mentioning of certain keywords within the text of the website. The clear example would be BMW mentioning phrase like ultimate driving machine around 20 times on just one page. Both Cloaking and Spamdexing are not viewed favorably by the search engine companies who view those practices as dishonest and as an attempt to rig the system for a better page rank score; therefore, most of the search engine companies ban the sites that use those practices.

What is "frontier" in the web crawler? Please suggest two approaches to manage the frontier, and explain the strength and possible weakness of these approaches. [0.75pt]

Frontier is a collection of the collected URLs extracted from downloaded pages stored in the queue. There are two approaches to management of the queue. First one is FIFO (First In First Out) and under this management we see more Breadth For Search type of behavior when crawling sites. This way all the closest links are visited first before children of immediate links are pursued. The strength of this approach is that sites mapped first are most likely very similar to first site. Another strength is that if we are looking for a shortest path between two sites this approach guarantees that. Weakness of this site is that it takes a long while to reach most connected sites on the web. Another way of managing the queue is through LIFO (Last In First Out) approach with we see a behavior of Depth For Search when crawling sites. In this way the crawler will go as deep as it can following just one link from each page until it finds a page with no further links then it traces back to the page that had links before. This approach strength is that that it can find connections to very popular websites faster than FIFO approach as the crawl goes in deep fast and has higher chance to arrive at very popular site faster. Weakness of LIFO approach is that sites it maps many times are loosely connected to the original site.

Question 2 – [6.00pt]

Please download the java file, and use Eclipse or NetBeans to build a web crawling project. Please provide a seed URL and collect at least 50 web pages from the web crawler [1.00pt]

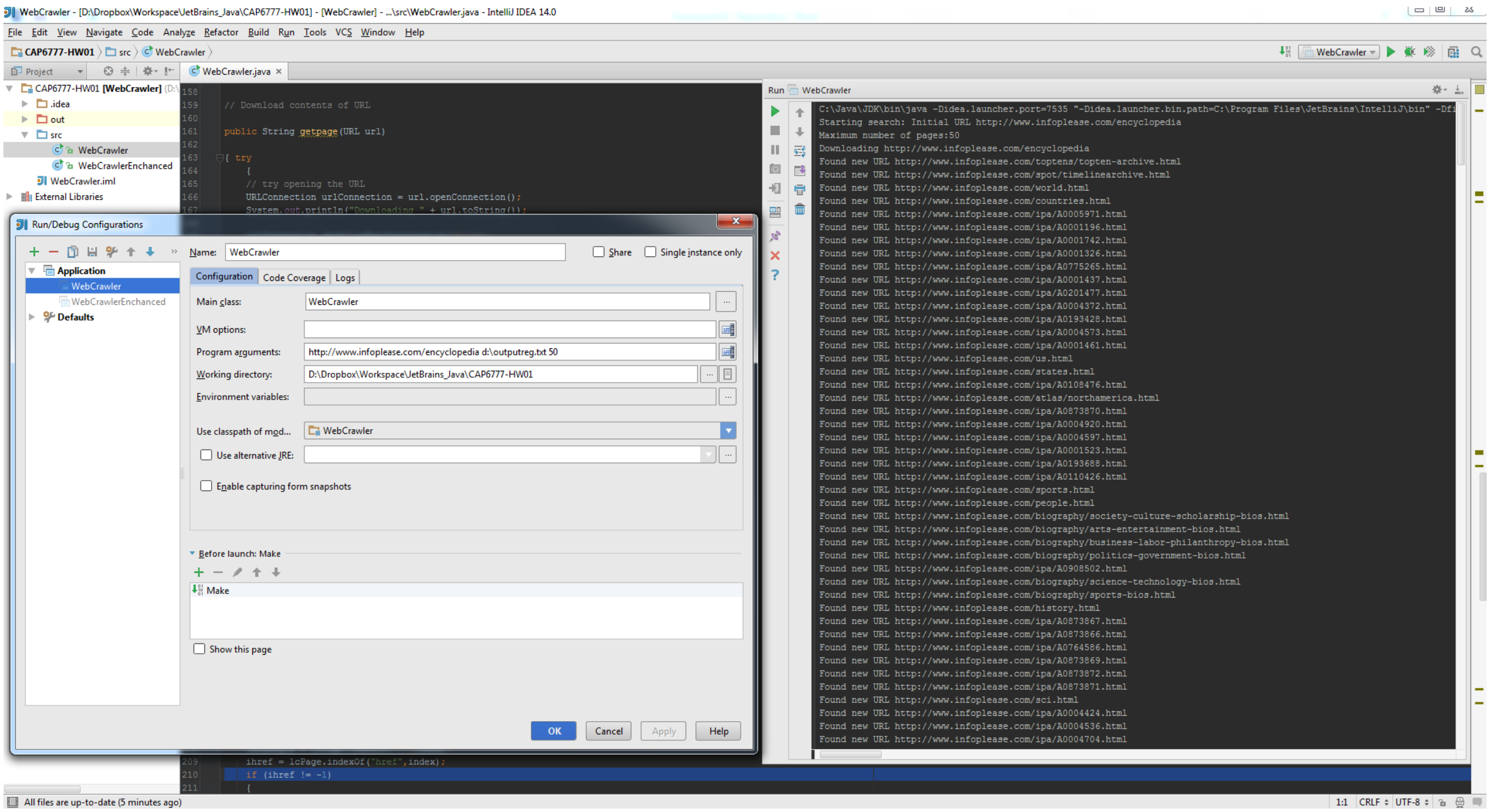
Seed URL is http://www.infoplease.com/encyclopedia

```
Starting search: Initial URL http://www.infoplease.com/encyclopedia
Maximum number of pages:50
Downloading http://www.infoplease.com/encyclopedia
Found new URL http://www.infoplease.com/toptens/topten-archive.html
Found new URL http://www.infoplease.com/spot/timelinearchive.html
Found new URL http://www.infoplease.com/world.html
Found new URL http://www.infoplease.com/countries.html
Found new URL http://www.infoplease.com/ipa/A0005971.html
Found new URL http://www.infoplease.com/ipa/A0001196.html
Found new URL http://www.infoplease.com/ipa/A0001742.html
Found new URL http://www.infoplease.com/ipa/A0001326.html
Found new URL http://www.infoplease.com/ipa/A0775265.html
Found new URL http://www.infoplease.com/ipa/A0001437.html
Found new URL http://www.infoplease.com/ipa/A0201477.html
Found new URL http://www.infoplease.com/ipa/A0004372.html
Found new URL http://www.infoplease.com/ipa/A0193428.html
Found new URL http://www.infoplease.com/ipa/A0004573.html
Found new URL http://www.infoplease.com/ipa/A0001461.html
Found new URL http://www.infoplease.com/us.html
Found new URL http://www.infoplease.com/states.html
Found new URL http://www.infoplease.com/ipa/A0108476.html
Found new URL http://www.infoplease.com/atlas/northamerica.html
Found new URL http://www.infoplease.com/ipa/A0873870.html
Found new URL http://www.infoplease.com/ipa/A0004920.html
Found new URL http://www.infoplease.com/ipa/A0004597.html
Found new URL http://www.infoplease.com/ipa/A0001523.html
Found new URL http://www.infoplease.com/ipa/A0193688.html
Found new URL http://www.infoplease.com/ipa/A0110426.html
Found new URL http://www.infoplease.com/sports.html
Found new URL http://www.infoplease.com/people.html
Found new URL http://www.infoplease.com/biography/society-culture-scholarship-bios.html
Found new URL http://www.infoplease.com/biography/arts-entertainment-bios.html
Found new URL http://www.infoplease.com/biography/business-labor-philanthropy-bios.html
Found new URL http://www.infoplease.com/biography/politics-government-bios.html
Found new URL http://www.infoplease.com/ipa/A0908502.html
Found new URL http://www.infoplease.com/biography/science-technology-bios.html
Found new URL http://www.infoplease.com/biography/sports-bios.html
Found new URL http://www.infoplease.com/history.html
Found new URL http://www.infoplease.com/ipa/A0873867.html
Found new URL http://www.infoplease.com/ipa/A0873866.html
Found new URL http://www.infoplease.com/ipa/A0764586.html
Found new URL http://www.infoplease.com/ipa/A0873869.html
Found new URL http://www.infoplease.com/ipa/A0873872.html
Found new URL http://www.infoplease.com/ipa/A0873871.html
Found new URL http://www.infoplease.com/sci.html
Found new URL http://www.infoplease.com/ipa/A0004424.html
Found new URL http://www.infoplease.com/ipa/A0004536.html
Found new URL http://www.infoplease.com/ipa/A0004704.html
Found new URL http://www.infoplease.com/health.html
Found new URL http://www.infoplease.com/ipa/A0004636.html
Found new URL http://www.infoplease.com/ipa/A0873842.html
Found new URL http://www.infoplease.com/ipa/A0873843.html
Found new URL http://www.infoplease.com/ipa/A0006019.html
Found new URL http://www.infoplease.com/weather.html
Found new URL http://www.infoplease.com/ipa/A0001657.html
Found new URL http://www.infoplease.com/calendar-holidays.html
Found new URL http://www.infoplease.com/ipa/A0875653.html
Found new URL http://www.infoplease.com/ipa/A0875654.html
Found new URL http://www.infoplease.com/ipa/A0875655.html
Found new URL http://www.infoplease.com/spot/archive-society.html
Found new URL http://www.infoplease.com/bus.html
Found new URL http://www.infoplease.com/ipa/A0104511.html
Found new URL http://www.infoplease.com/ipa/A0002080.html
Found new URL http://www.infoplease.com/ipa/A0193934.html
Found new URL http://www.infoplease.com/ipa/A0854971.html
Found new URL http://www.infoplease.com/finance.html
Found new URL http://www.infoplease.com/ipa/A0854972.html
Found new URL http://www.infoplease.com/ipa/A0005920.html
Found new URL http://www.infoplease.com/math/calculator.html
Found new URL http://www.infoplease.com/atlas/calculate-distance.html
Found new URL http://www.infoplease.com/atlas/latitude-longitude.html
Found new URL http://www.infoplease.com/pages/unitconversion.html
Found new URL http://www.infoplease.com/yearbyyear.html
Found new URL http://www.infoplease.com/citing.html
Found new URL http://www.infoplease.com/help.html
Found new URL http://www.infoplease.com/earth-environment.html
Found new URL http://www.infoplease.com/0eaenvatmos.html
Found new URL http://www.infoplease.com/0eaenvbio.html
Found new URL http://www.infoplease.com/0eaenvecol.html
Found new URL http://www.infoplease.com/0eaenvgeo.html
Found new URL http://www.infoplease.com/0eaenvgeol.html
Downloading http://www.infoplease.com/toptens/topten-archive.html
Downloading http://www.infoplease.com/spot/timelinearchive.html
Downloading http://www.infoplease.com/world.html
Found new URL http://www.infoplease.com/cgi-bin/id/A0004372.html
Found new URL http://www.infoplease.com/cgi-bin/id/A0001437.html
Found new URL http://www.infoplease.com/cgi-bin/id/A0005971.html
Downloading http://www.infoplease.com/countries.html
Found new URL http://www.infoplease.com/almanacs.html
Downloading http://www.infoplease.com/ipa/A0005971.html
Found new URL http://www.infoplease.com/news/2016/current-events/index.html
Found new URL http://www.infoplease.com/us/government/presidential-election-campaign-2016.html
Found new URL http://www.infoplease.com/news/2015/current-events/index.html
Found new URL http://www.infoplease.com/news/religious-freedom.html
Found new URL http://www.infoplease.com/world/events/armenian-genocide.html
Found new URL http://www.infoplease.com/world/statistics/migrant-deaths.html
Downloading http://www.infoplease.com/ipa/A0001196.html
Found new URL http://www.infoplease.com/dk/encyclopedia/history.html
Found new URL http://www.infoplease.com/ipa/A0001198.html
Found new URL http://www.infoplease.com/ipa/A0001209.html
Found new URL http://www.infoplease.com/ipa/A0777090.html
Downloading http://www.infoplease.com/ipa/A0001742.html
Found new URL http://www.infoplease.com/ipa/A0873835.html
Found new URL http://www.infoplease.com/ipa/A0873836.html
Downloading http://www.infoplease.com/ipa/A0001326.html
Found new URL http://fun.familyeducation.com/slideshow/historic-sites/61628.html
Found new URL http://www.infoplease.com/world/buildings-structures/7-ancient-wonders.html
Found new URL http://www.infoplease.com/world/buildings-structures/tallest-slideshow.html
Found new URL http://www.infoplease.com/us/slideshow/washington-dc-landmarks.html
Found new URL http://www.infoplease.com/us/slideshow/washington-dc-statues.html
Found new URL http://www.infoplease.com/ipa/A0001327.html
Found new URL http://www.infoplease.com/ipa/A0923082.html
Found new URL http://www.infoplease.com/world/buildings-structures/seven-new-wonders.html
Found new URL http://www.infoplease.com/ipa/A0001328.html
Found new URL http://www.infoplease.com/ipa/A0001338.html
Downloading http://www.infoplease.com/ipa/A0775265.html
Found new URL http://www.infoplease.com/ipa/A0776695.html
Found new URL http://www.infoplease.com/ipa/A0775293.html
Found new URL http://www.infoplease.com/ipa/A0886238.html
Found new URL http://www.infoplease.com/ipa/A0886235.html
Found new URL http://www.infoplease.com/ipa/A0775279.html
Found new URL http://www.infoplease.com/ipa/A0775280.html
Found new URL http://www.infoplease.com/ipa/A0775376.html
Downloading http://www.infoplease.com/ipa/A0001437.html
Found new URL http://www.infoplease.com/ipa/A0001767.html
Found new URL http://www.infoplease.com/world/disasters/fatal-us-mining-accidents.html
Found new URL http://www.infoplease.com/ipa/A0001459.html
Downloading http://www.infoplease.com/ipa/A0201477.html
Found new URL http://www.infoplease.com/ipa/A0772153.html
Found new URL http://www.infoplease.com/ipa/A0772158.html
Found new URL http://www.infoplease.com/ipa/A0772159.html
Found new URL http://www.infoplease.com/ipa/A0772161.html
Found new URL http://www.infoplease.com/spot/quiz/usflag/1.html
Downloading http://www.infoplease.com/ipa/A0004372.html
Found new URL http://www.infoplease.com/ipa/A0873844.html
Found new URL http://www.infoplease.com/ipa/A0873845.html
Found new URL http://www.infoplease.com/ipa/A0873846.html
Found new URL http://www.infoplease.com/ipa/A0873847.html
Found new URL http://www.infoplease.com/ipa/A0873848.html
Found new URL http://www.infoplease.com/ipa/A0873849.html
Downloading http://www.infoplease.com/ipa/A0193428.html
Found new URL http://www.infoplease.com/ipa/A0001299.html
Found new URL http://www.infoplease.com/ipa/A0001300.html
Found new URL http://www.infoplease.com/ipa/A0862135.html
Found new URL http://www.infoplease.com/ipa/A0001295.html
Found new URL http://www.infoplease.com/ipa/A0761157.html
Found new URL http://www.infoplease.com/ipa/A0778427.html
Found new URL http://www.infoplease.com/ipa/A0004582.html
Found new URL http://www.infoplease.com/ipa/A0882520.html
Downloading http://www.infoplease.com/ipa/A0004573.html
Found new URL http://www.infoplease.com/ipa/A0855290.html
Found new URL http://www.infoplease.com/ipa/A0855288.html
Found new URL http://www.infoplease.com/ipa/A0855289.html
Downloading http://www.infoplease.com/ipa/A0001461.html
Found new URL http://www.infoplease.com/society.html
Found new URL http://www.infoplease.com/ipa/A0113529.html
Found new URL http://www.infoplease.com/ipa/A0921143.html
Found new URL http://www.infoplease.com/world/religion/religiously-unaffiliated.html
Found new URL http://www.infoplease.com/world/religion/top-eight-world-religions.html
Found new URL http://www.infoplease.com/world/religion/countries-largest-unaffiliated-population.html
Found new URL http://www.infoplease.com/world/religion/adherents-median-age.html
Found new URL http://www.infoplease.com/ipa/A0922574.html
Found new URL http://www.infoplease.com/ipa/A0881831.html
Found new URL http://www.infoplease.com/world/religion/largest-us-churches.html
Found new URL http://www.infoplease.com/world/religion/largest-us-churches-2010.html
Found new URL http://www.infoplease.com/ipa/A0193644.html
Found new URL http://www.infoplease.com/ipa/A0001474.html
Found new URL http://www.infoplease.com/world/religion/choosing-new-pope.html
Found new URL http://www.infoplease.com/ipa/A0906882.html
Downloading http://www.infoplease.com/us.html
Downloading http://www.infoplease.com/states.html
Downloading http://www.infoplease.com/ipa/A0108476.html
Found new URL http://www.infoplease.com/us/slideshow/san-francisco-landmarks.html
Found new URL http://www.infoplease.com/us/slideshow/chicago-landmarks.html
Found new URL http://www.infoplease.com/us/slideshow/new-orleans.html
Found new URL http://www.infoplease.com/us/slideshow/landmarks-new-york-city.html
Found new URL http://www.infoplease.com/us/slideshow/landmarks-boston.html
Found new URL http://www.infoplease.com/us/slideshow/houston-landmarks.html
Downloading http://www.infoplease.com/atlas/northamerica.html
Found new URL http://www.infoplease.com/atlas/mapindex.html
Found new URL http://www.infoplease.com/atlas/centralamerica.html
Found new URL http://www.infoplease.com/atlas/usa.html
Found new URL http://www.infoplease.com/atlas/country/bermuda.html
Found new URL http://www.infoplease.com/atlas/caribbean.html
Found new URL http://www.infoplease.com/atlas/mexico.html
Downloading http://www.infoplease.com/ipa/A0873870.html
```

```
Found new URL http://www.infoplease.com/ipa/A0902416.html
Found new URL http://www.infoplease.com/us/government/watergate.html
Found new URL http://www.infoplease.com/ipa/A0875904.html
Found new URL http://www.infoplease.com/ipa/A0194019.html
Found new URL http://www.infoplease.com/ipa/A0763770.html
Found new URL http://www.infoplease.com/ipa/A0875838.html
Found new URL http://www.infoplease.com/us/history/race-riots.html
Found new URL http://www.infoplease.com/us/history/girl-scouts.html
Found new URL http://www.infoplease.com/ipa/A0194018.html
Found new URL http://www.infoplease.com/ipa/A0004991.html
Found new URL http://www.infoplease.com/ipa/A0194016.html
Found new URL http://www.infoplease.com/ipa/A0875901.html
Found new URL http://www.infoplease.com/ipa/A0874987.html
Found new URL http://www.infoplease.com/ipa/A0194022.html
Found new URL http://www.infoplease.com/ipa/A0764613.html
Found new URL http://www.infoplease.com/ipa/A0194050.html
Found new URL http://www.infoplease.com/ipa/A0194035.html
Found new URL http://www.infoplease.com/us/history/symbols-united-states.html
Found new URL http://www.infoplease.com/spot/constitutionday.html
Downloading http://www.infoplease.com/ipa/A0004920.html
Found new URL http://www.infoplease.com/ipa/A0004921.html
Found new URL http://www.infoplease.com/ipa/A0110379.html
Found new URL http://www.infoplease.com/ipa/A0922218.html
Found new URL http://www.infoplease.com/ipa/A0110389.html
Found new URL http://www.infoplease.com/ipa/A0113931.html
Found new URL http://www.infoplease.com/familytrends.html
Found new URL http://www.infoplease.com/ipa/A0110390.html
Found new URL http://www.infoplease.com/ipa/A0103715.html
Downloading http://www.infoplease.com/ipa/A0004597.html
Found new URL http://www.infoplease.com/ipa/A0873839.html
Found new URL http://www.infoplease.com/ipa/A0873840.html
Found new URL http://www.infoplease.com/ipa/A0873841.html
Downloading http://www.infoplease.com/ipa/A0001523.html
Found new URL http://www.infoplease.com/ipa/A0763170.html
Found new URL http://www.infoplease.com/ipa/A0922479.html
Found new URL http://www.infoplease.com/business/best-worst-economies-women.html
Found new URL http://www.infoplease.com/us/statistics/leading-occupations-employed-women.html
Found new URL http://www.infoplease.com/us/statistics/leading-occupations-employed-women-2010.html
Found new URL http://www.infoplease.com/business/highest-earnings-women.html
Found new URL http://www.infoplease.com/ipa/A0931343.html
Found new URL http://www.infoplease.com/ipa/A0768502.html
Downloading http://www.infoplease.com/ipa/A0193688.html
Found new URL http://www.infoplease.com/ipa/A0875658.html
Found new URL http://www.infoplease.com/ipa/A0875661.html
Found new URL http://www.infoplease.com/ipa/A0875662.html
Found new URL http://www.infoplease.com/ipa/A0875675.html
Downloading http://www.infoplease.com/ipa/A0110426.html
Found new URL http://www.infoplease.com/ipa/A0110427.html
Found new URL http://www.infoplease.com/ipa/A0771187.html
Found new URL http://www.infoplease.com/ipa/A0110458.html
Found new URL http://www.infoplease.com/ipa/A0878163.html
Found new URL http://www.infoplease.com/ipa/A0110468.html
Found new URL http://www.infoplease.com/ipa/A0110467.html
Found new URL http://www.infoplease.com/ipa/A0193245.html
Found new URL http://www.infoplease.com/ipa/A0768442.html
Found new URL http://www.infoplease.com/ipa/A0771182.html
Found new URL http://www.infoplease.com/us/statistics/postal-workers-bitten-by-dogs-2013.html
Found new URL http://www.infoplease.com/us/statistics/postal-workers-bitten-by-dogs-2012.html
Found new URL http://www.infoplease.com/us/government/postal-worker-dog-bites.html
Downloading http://www.infoplease.com/sports.html
Found new URL http://www.infoplease.com/cgi-bin/id/A0114094.html
Found new URL http://www.infoplease.com/cgi-bin/id/A0003691.html
Found new URL http://www.infoplease.com/cgi-bin/id/A0003203.html
Found new URL http://www.infoplease.com/cgi-bin/id/A0103717.html
Found new URL http://www.infoplease.com/cgi-bin/id/A0002331.html
Downloading http://www.infoplease.com/people.html
Found new URL http://www.infoplease.com/cgi-bin/id/A0880525.html
Found new URL http://www.infoplease.com/biography/a-bios.html
Found new URL http://www.infoplease.com/biography/b-bios.html
Found new URL http://www.infoplease.com/biography/c-bios.html
Found new URL http://www.infoplease.com/biography/d-bios.html
Found new URL http://www.infoplease.com/biography/e-bios.html
Found new URL http://www.infoplease.com/biography/f-bios.html
Found new URL http://www.infoplease.com/biography/g-bios.html
Found new URL http://www.infoplease.com/biography/h-bios.html
Found new URL http://www.infoplease.com/biography/i-bios.html
Found new URL http://www.infoplease.com/biography/j-bios.html
Downloading http://www.infoplease.com/biography/society-culture-scholarship-bios.html
Found new URL http://www.infoplease.com/ce6/people/A0813210.html
Found new URL http://www.infoplease.com/ipa/A0908503.html
Found new URL http://www.infoplease.com/biography/activists.html
Found new URL http://www.infoplease.com/biography/anthropologists-archaeologists.html
Found new URL http://www.infoplease.com/biography/economists.html
Found new URL http://www.infoplease.com/biography/educators-scholars.html
Found new URL http://www.infoplease.com/biography/jurists.html
Found new URL http://www.infoplease.com/biography/philosophers.html
Found new URL http://www.infoplease.com/biography/political-scientists.html
Found new URL http://www.infoplease.com/biography/religious-leaders-catholic.html
Found new URL http://www.infoplease.com/biography/religious-leaders-protestant.html
Found new URL http://www.infoplease.com/biography/religious-leaders-jewish.html
Found new URL http://www.infoplease.com/biography/religious-leaders-other.html
Found new URL http://www.infoplease.com/biography/religious-leaders-popes.html
Found new URL http://www.infoplease.com/biography/sociologists.html
Downloading http://www.infoplease.com/biography/arts-entertainment-bios.html
Found new URL http://www.infoplease.com/ce6/people/A0823822.html
Found new URL http://www.infoplease.com/biography/actors.html
Found new URL http://www.infoplease.com/spot/artbio1.html
Found new URL http://www.infoplease.com/biography/architects.html
Found new URL http://www.infoplease.com/biography/comedians.html
Found new URL http://www.infoplease.com/biography/dancers.html
Found new URL http://www.infoplease.com/biography/fashion.html
Found new URL http://www.infoplease.com/biography/filmmakers-directors.html
Found new URL http://www.infoplease.com/biography/journalists.html
Found new URL http://www.infoplease.com/biography/music-rock-pop.html
Found new URL http://www.infoplease.com/biography/music-composers.html
Found new URL http://www.infoplease.com/biography/music-popular.html
Found new URL http://www.infoplease.com/biography/music-concert-instrumentalists.html
Found new URL http://www.infoplease.com/biography/music-country.html
Found new URL http://www.infoplease.com/biography/music-folk-gospel-blues-world.html
Downloading http://www.infoplease.com/biography/business-labor-philanthropy-bios.html
Found new URL http://www.infoplease.com/ce6/people/A0842151.html
Found new URL http://www.infoplease.com/biography/business.html
Found new URL http://www.infoplease.com/biography/labor-leaders.html
Found new URL http://www.infoplease.com/biography/philanthropists.html
Found new URL http://www.infoplease.com/spot/bhmpeople15.html
Found new URL http://www.infoplease.com/spot/asianbios2.html
Found new URL http://www.infoplease.com/spot/hhmbio3.html
Found new URL http://www.infoplease.com/spot/whmbios14.html
Found new URL http://www.infoplease.com/spot/womenceo1.html
Downloading http://www.infoplease.com/biography/politics-government-bios.html
Found new URL http://www.infoplease.com/ce6/people/A0804107.html
Found new URL http://www.infoplease.com/ipa/A0760585.html
Found new URL http://www.infoplease.com/ipa/A0194032.html
Found new URL http://www.infoplease.com/biography/us-supreme-court.html
Found new URL http://www.infoplease.com/ipa/A0108856.html
Found new URL http://www.infoplease.com/ipa/A0108854.html
Found new URL http://www.infoplease.com/biography/us/congress/index.html
Found new URL http://www.infoplease.com/ipa/A0878575.html
Found new URL http://www.infoplease.com/biography/us-military-persons.html
Found new URL http://www.infoplease.com/encyclopedia/0biohist.html
Downloading http://www.infoplease.com/ipa/A0908502.html
Found new URL http://www.infoplease.com/ipa/A0775708.html
Found new URL http://www.infoplease.com/ipka/A0109586.html
Found new URL http://www.infoplease.com/ce6/people/A0821263.html
Found new URL http://www.infoplease.com/ipsa/A0109455.html
Found new URL http://www.infoplease.com/ce6/people/A0820141.html
Found new URL http://www.infoplease.com/ipka/A0761430.html
Found new URL http://www.infoplease.com/ce6/people/A0816541.html
Found new URL http://www.infoplease.com/ce6/people/A0831499.html
Found new URL http://www.infoplease.com/ce6/people/A0826869.html
Found new URL http://www.infoplease.com/ipa/A0758515.html
Found new URL http://www.infoplease.com/ipka/A0109760.html
Found new URL http://www.infoplease.com/ce6/people/A0804198.html
Found new URL http://www.infoplease.com/ce6/people/A0831961.html
Downloading http://www.infoplease.com/biography/science-technology-bios.html
Found new URL http://www.infoplease.com/ce6/people/A0841865.html
Found new URL http://www.infoplease.com/biography/aviators-astronauts.html
Found new URL http://www.infoplease.com/biography/explorers.html
Found new URL http://www.infoplease.com/biography/science-mathematicians.html
Found new URL http://www.infoplease.com/spot/scibio1.html
Found new URL http://www.infoplease.com/spot/scibio2.html
Found new URL http://www.infoplease.com/spot/scibio4.html
Found new URL http://www.infoplease.com/spot/scibio6.html
Found new URL http://www.infoplease.com/spot/asianbios7.html
Found new URL http://www.infoplease.com/spot/hhmbio4.html
Found new URL http://www.infoplease.com/spot/whmbios2.html
Downloading http://www.infoplease.com/biography/sports-bios.html
Found new URL http://www.infoplease.com/ce6/people/A0842106.html
Found new URL http://www.infoplease.com/biography/sports-auto-racing.html
Found new URL http://www.infoplease.com/biography/sports-baseball.html
Found new URL http://www.infoplease.com/biography/sports-basketball.html
Found new URL http://www.infoplease.com/biography/sports-boxing.html
Found new URL http://www.infoplease.com/biography/sports-football.html
Found new URL http://www.infoplease.com/biography/sports-golf.html
Found new URL http://www.infoplease.com/biography/sports-skaters.html
Found new URL http://www.infoplease.com/biography/sports-soccer.html
Found new URL http://www.infoplease.com/biography/sports-tennis.html
Found new URL http://www.infoplease.com/biography/sports-figures.html
Found new URL http://www.infoplease.com/biography/sports-chess.html
Found new URL http://www.infoplease.com/spot/bhmpeople7.html
```

```
Downloading http://www.infoplease.com/history.html
Found new URL http://www.infoplease.com/cgi-bin/id/A0001196.html
Found new URL http://www.infoplease.com/cgi-bin/id/A0873870.html
Found new URL http://www.infoplease.com/cgi-bin/id/A0873871.html
Found new URL http://www.infoplease.com/cgi-bin/id/A0873872.html
Downloading http://www.infoplease.com/ipa/A0873867.html
Found new URL http://www.infoplease.com/ipa/A0194030.html
Found new URL http://www.infoplease.com/ipa/A0101184.html
Found new URL http://www.infoplease.com/us/history/lincoln-resurgence.html
Found new URL http://www.infoplease.com/us/government/presidential-pardons.html
Found new URL http://www.infoplease.com/us/history/president-requirements.html
Found new URL http://www.infoplease.com/us/history/closest-presidential-races.html
Found new URL http://www.infoplease.com/us/slideshow/first-kids.html
Downloading http://www.infoplease.com/ipa/A0873866.html
Found new URL http://www.infoplease.com/us/government/voting-rights.html
Found new URL http://www.infoplease.com/us/government/114-congress.html
Found new URL http://www.infoplease.com/us/government/114-congress-senate.html
Found new URL http://www.infoplease.com/us/government/elections-house-representatives-2014.html
Found new URL http://www.infoplease.com/us/government/113-congress.html
Found new URL http://www.infoplease.com/us/government/2012-elections-senate.html
Found new URL http://www.infoplease.com/us/government/elections-house-representatives-2012.html
Found new URL http://www.infoplease.com/us/government/midterm-elections-senate.html
Found new URL http://www.infoplease.com/us/government/midterm-elections-house-representatives-2010.html
Found new URL http://www.infoplease.com/us/government/112-congress.html
Downloading http://www.infoplease.com/ipa/A0764586.html
Found new URL http://www.infoplease.com/news/2014/midterm-elections.html
Found new URL http://www.infoplease.com/us/government/presidential-election-campaign-2012.html
Found new URL http://www.infoplease.com/us/government/milestones-2012-general-election.html
Found new URL http://www.infoplease.com/us/government/election-2012.html
Downloading http://www.infoplease.com/ipa/A0873869.html
Found new URL http://www.infoplease.com/us/supreme-court/supreme-court-members.html
Found new URL http://www.infoplease.com/ipa/A0101281.html
Found new URL http://www.infoplease.com/us/government/chief-justices-supreme-court.html
Found new URL http://www.infoplease.com/ipa/A0875894.html
Found new URL http://www.infoplease.com/ipa/A0101289.html
Downloading http://www.infoplease.com/ipa/A0873872.html
Found new URL http://www.infoplease.com/ipa/A0101054.html
Found new URL http://www.infoplease.com/ipa/A0194024.html
Found new URL http://www.infoplease.com/ipa/A0194023.html
Found new URL http://www.infoplease.com/ipa/A0875896.html
Found new URL http://www.infoplease.com/ipa/A0194027.html
Found new URL http://www.infoplease.com/ipa/A0194025.html
Found new URL http://www.infoplease.com/ipa/A0194026.html
Found new URL http://www.infoplease.com/ipa/A0908224.html
Found new URL http://fun.familyeducation.com/slideshow/monuments/61486.html
Found new URL http://www.infoplease.com/us/landmarks-timeline.html
Found new URL http://www.infoplease.com/us/most-visited-landmarks.html
Downloading http://www.infoplease.com/ipa/A0873871.html
Found new URL http://www.infoplease.com/ipa/A0101029.html
Found new URL http://www.infoplease.com/ipa/A0101022.html
Found new URL http://www.infoplease.com/ipa/A0101025.html
Found new URL http://www.infoplease.com/ipa/A0101031.html
Found new URL http://www.infoplease.com/ipa/A0194062.html
Found new URL http://www.infoplease.com/ipa/A0101053.html
Found new URL http://www.infoplease.com/ipa/A0194034.html
Found new URL http://www.infoplease.com/ipa/A0194020.html
Found new URL http://www.infoplease.com/ipa/A0194015.html
Found new URL http://www.infoplease.com/ipa/A0194021.html
Found new URL http://www.infoplease.com/ipa/A0930913.html
Found new URL http://www.infoplease.com/ipa/A0932561.html
Found new URL http://www.infoplease.com/us/history/most-important-speeches.html
Downloading http://www.infoplease.com/sci.html
Found new URL http://www.infoplease.com/cgi-bin/id/A0004424.html
Found new URL http://www.infoplease.com/cgi-bin/id/A0001346.html
Found new URL http://www.infoplease.com/cgi-bin/id/A0004536.html
Found new URL http://www.infoplease.com/cgi-bin/id/A0001816.html
Downloading http://www.infoplease.com/ipa/A0004424.html
Found new URL http://www.infoplease.com/ipa/A0873818.html
Found new URL http://www.infoplease.com/ipa/A0873819.html
Found new URL http://www.infoplease.com/ipa/A0873821.html
Found new URL http://www.infoplease.com/ipa/A0001346.html
Found new URL http://www.infoplease.com/science/astronomy/nasa-turns-50.html
Downloading http://www.infoplease.com/ipa/A0004536.html
Found new URL http://www.infoplease.com/ipa/A0004537.html
Found new URL http://www.infoplease.com/ipa/A0872854.html
Found new URL http://www.infoplease.com/ipa/A0778412.html
Found new URL http://www.infoplease.com/ipa/A0192891.html
Found new URL http://www.infoplease.com/ipa/A0004549.html
Found new URL http://www.infoplease.com/ipa/A0004556.html
Found new URL http://www.infoplease.com/ipa/A0004551.html
Downloading http://www.infoplease.com/ipa/A0004704.html
Found new URL http://www.infoplease.com/ipa/A0873828.html
Found new URL http://www.infoplease.com/ipa/A0873829.html
Found new URL http://www.infoplease.com/ipa/A0873830.html
Found new URL http://www.infoplease.com/cig/science-fair-projects/kind-trash-bag-breaks-down-fastest.html
Downloading http://www.infoplease.com/health.html
Found new URL http://www.infoplease.com/ipa/A0001182.html
Found new URL http://www.infoplease.com/news/2014/health-care-update.html
Found new URL http://www.infoplease.com/science/health/americans-without-health-insurance-by-state.html
Downloading http://www.infoplease.com/ipa/A0004636.html
Found new URL http://www.infoplease.com/quizzes/great-inventions-minds/1.html
Found new URL http://www.infoplease.com/ipa/A0004637.html
Found new URL http://www.infoplease.com/ipa/A0193133.html
Found new URL http://www.infoplease.com/ipa/A0004638.html
Found new URL http://www.infoplease.com/ipa/A0883926.html
Downloading http://www.infoplease.com/ipa/A0873842.html
Found new URL http://www.infoplease.com/ipa/A0001816.html
Found new URL http://www.infoplease.com/ipa/A0193009.html
Found new URL http://www.infoplease.com/dk/science/encyclopedia/human-body.html
Found new URL http://www.infoplease.com/ipa/A0779260.html
Found new URL http://www.infoplease.com/ipa/A0001819.html
Found new URL http://www.infoplease.com/ipa/A0193003.html
Found new URL http://www.infoplease.com/ipa/A0193002.html
Found new URL http://www.infoplease.com/ipa/A0762176.html
Found new URL http://www.infoplease.com/ipa/A0873492.html
Found new URL http://www.infoplease.com/ipa/A0762178.html
Found new URL http://www.infoplease.com/ipa/A0770763.html
Found new URL http://www.infoplease.com/dk/science/encyclopedia/plants.html
Found new URL http://www.infoplease.com/ipa/A0932480.html
Found new URL http://www.infoplease.com/ipa/A0932544.html
Found new URL http://www.infoplease.com/ipa/A0932600.html
Found new URL http://www.infoplease.com/ipa/A0932504.html
Found new URL http://www.infoplease.com/ipa/A0932475.html
Found new URL http://www.infoplease.com/ipa/A0932663.html
Found new URL http://www.infoplease.com/cig/biology/chemistry-biology.html
Found new URL http://www.infoplease.com/cig/biology/cell-theory-form-function.html
Downloading http://www.infoplease.com/ipa/A0873843.html
Found new URL http://www.infoplease.com/ipa/A0880382.html
Found new URL http://www.infoplease.com/ipa/A0001817.html
Found new URL http://www.infoplease.com/ipa/A0905215.html
Found new URL http://www.infoplease.com/ipa/A0001826.html
Found new URL http://www.infoplease.com/dk/science/encyclopedia/forces-and-energy.html
Found new URL http://www.infoplease.com/dk/science/encyclopedia/matter-and-materials.html
Found new URL http://www.infoplease.com/dk/science/encyclopedia/electricity-magnetism.html
Found new URL http://www.infoplease.com/ipa/A0762175.html
Found new URL http://www.infoplease.com/ipa/A0001822.html
Found new URL http://www.infoplease.com/ipa/A0905226.html
Found new URL http://www.infoplease.com/science/physics/large-hadron-collider.html
Found new URL http://www.infoplease.com/science/physics/higgs-boson.html
Found new URL http://www.infoplease.com/science/physics/large-hadron-collider-fiction-fact.html
Found new URL http://www.infoplease.com/cig/science-fair-projects/salt-sugar-dissolves-faster-different-liquids.html
Found new URL http://www.infoplease.com/cig/science-fair-projects/all-pennies-created-equal.html
Found new URL http://www.infoplease.com/cig/science-fair-projects/metal-corrodes-fastest.html
Found new URL http://www.infoplease.com/cig/science-fair-projects/matter-much-air-basketball.html
Search complete.
```

Original WebCrawler crawling http://www.infoplease.com/encyclopedia for 50 links (see Appendix A for results via Dropbox Link)
Link : https://www.dropbox.com/s/tqk2azqxojnx7b1/Appendix%20A%20-%20output-regular-www.infoplease.com-50.txt?dl=0


WebCrawler - [D:\Dropbox\Workspace\JetBrains_Java\CAP6777-HW01] - [WebCrawler] - ...\src\WebCrawler.java - IntelliJ IDEA 14.0
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
CAP6777-HW01 src WebCrawler
WebCrawler
Project
WebCrawler.java
CAP6777-HW01 [WebCrawler] (D:\
.idea
out
src
WebCrawler
WebCrawlerEnchanced
WebCrawler.iml
External Libraries
Run/Debug Configurations
Application
WebCrawler
WebCrawlerEnchanced
Defaults
Configuration
Code Coverage
Logs
Run
WebCrawler
OK
Cancel
Apply
Help
All files are up-to-date (5 minutes ago)
1:1 CRLF UTF-8

```
// Download contents of URL

public String getpage(URL url)

{ try
    {
    // try opening the URL
    URLConnection urlConnection = url.openConnection();
    System.out.println("Downloading " + url.toString());
```

```
    ihref = lcPage.indexOf("href",index);
    if (ihref != -1)
    {
```

Name: WebCrawler   □ Share   □ Single instance only

Main class: WebCrawler

VM options:

Program arguments: http://www.infoplease.com/encyclopedia d:\outputreg.txt 50

Working directory: D:\Dropbox\Workspace\JetBrains_Java\CAP6777-HW01

Environment variables:

Use classpath of mod...: WebCrawler

□ Use alternative JRE:

□ Enable capturing form snapshots

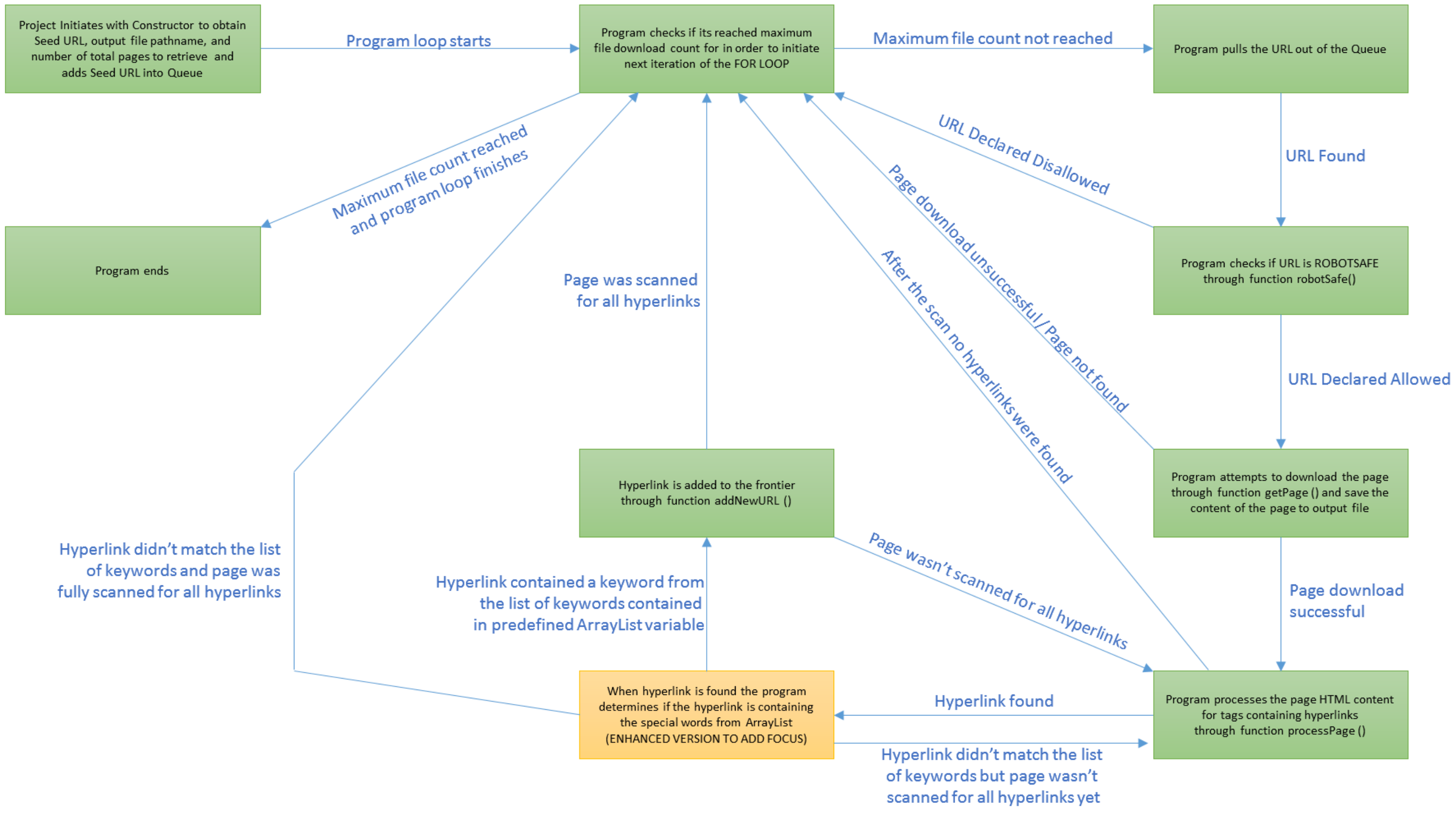
Before launch: Make

Make

□ Show this page

```
C:\Java\JDK\bin\java -Didea.launcher.port=7535 "-Didea.launcher.bin.path=C:\Program Files\JetBrains\IntelliJ\bin" -Dfi
Starting search: Initial URL http://www.infoplease.com/encyclopedia
Maximum number of pages:50
Downloading http://www.infoplease.com/encyclopedia
Found new URL http://www.infoplease.com/toptens/topten-archive.html
Found new URL http://www.infoplease.com/spot/timelinearchive.html
Found new URL http://www.infoplease.com/world.html
Found new URL http://www.infoplease.com/countries.html
Found new URL http://www.infoplease.com/ipa/A0005971.html
Found new URL http://www.infoplease.com/ipa/A0001196.html
Found new URL http://www.infoplease.com/ipa/A0001742.html
Found new URL http://www.infoplease.com/ipa/A0001326.html
Found new URL http://www.infoplease.com/ipa/A0775265.html
Found new URL http://www.infoplease.com/ipa/A0001437.html
Found new URL http://www.infoplease.com/ipa/A0201477.html
Found new URL http://www.infoplease.com/ipa/A0004372.html
Found new URL http://www.infoplease.com/ipa/A0193428.html
Found new URL http://www.infoplease.com/ipa/A0004573.html
Found new URL http://www.infoplease.com/ipa/A0001461.html
Found new URL http://www.infoplease.com/us.html
Found new URL http://www.infoplease.com/states.html
Found new URL http://www.infoplease.com/ipa/A0108476.html
Found new URL http://www.infoplease.com/atlas/northamerica.html
Found new URL http://www.infoplease.com/ipa/A0873870.html
Found new URL http://www.infoplease.com/ipa/A0004920.html
Found new URL http://www.infoplease.com/ipa/A0004597.html
Found new URL http://www.infoplease.com/ipa/A0001523.html
Found new URL http://www.infoplease.com/ipa/A0193688.html
Found new URL http://www.infoplease.com/ipa/A0110426.html
Found new URL http://www.infoplease.com/sports.html
Found new URL http://www.infoplease.com/people.html
Found new URL http://www.infoplease.com/biography/society-culture-scholarship-bios.html
Found new URL http://www.infoplease.com/biography/arts-entertainment-bios.html
Found new URL http://www.infoplease.com/biography/business-labor-philanthropy-bios.html
Found new URL http://www.infoplease.com/biography/politics-government-bios.html
Found new URL http://www.infoplease.com/ipa/A0908502.html
Found new URL http://www.infoplease.com/biography/science-technology-bios.html
Found new URL http://www.infoplease.com/biography/sports-bios.html
Found new URL http://www.infoplease.com/history.html
Found new URL http://www.infoplease.com/ipa/A0873867.html
Found new URL http://www.infoplease.com/ipa/A0873866.html
Found new URL http://www.infoplease.com/ipa/A0764586.html
Found new URL http://www.infoplease.com/ipa/A0873869.html
Found new URL http://www.infoplease.com/ipa/A0873872.html
Found new URL http://www.infoplease.com/ipa/A0873871.html
Found new URL http://www.infoplease.com/sci.html
Found new URL http://www.infoplease.com/ipa/A0004424.html
Found new URL http://www.infoplease.com/ipa/A0004536.html
Found new URL http://www.infoplease.com/ipa/A0004704.html
```

Please draw the flow chart (or the pseudo code) of your preferential web crawler design [1.00pt], and explain how does your approach/design make web crawling focusing on special topics [1.00pt]

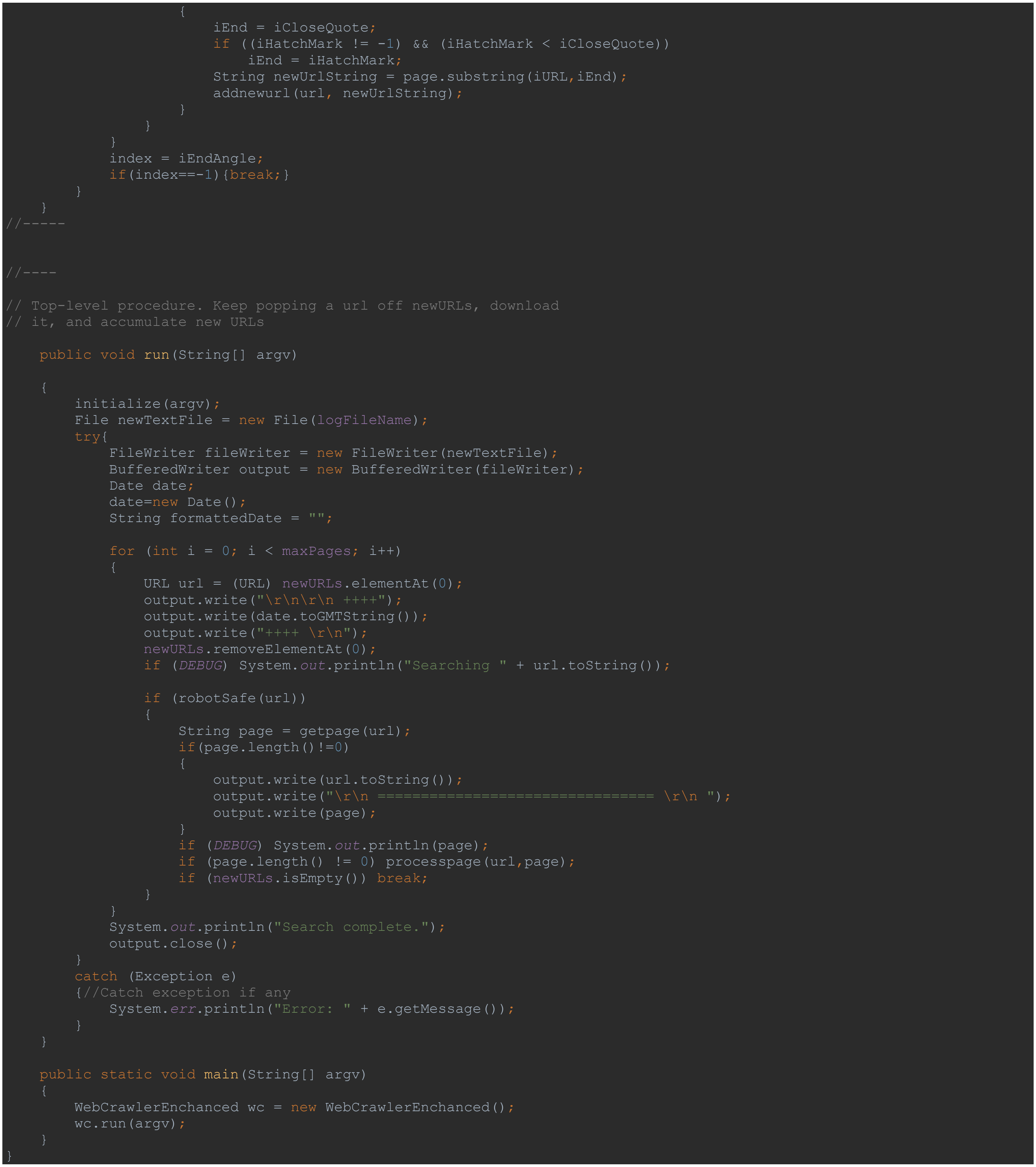

Project Initiates with Constructor to obtain Seed URL, output file pathname, and number of total pages to retrieve and adds Seed URL into Queue
Program loop starts
Program checks if its reached maximum file download count for in order to initiate next iteration of the FOR LOOP
Maximum file count not reached
Program pulls the URL out of the Queue
URL Found
Program checks if URL is ROBOTSAFE through function robotSafe()
URL Declared Disallowed
URL Declared Allowed
Program attempts to download the page through function getPage () and save the content of the page to output file
Page download unsuccessful / Page not found
Page download successful
Program processes the page HTML content for tags containing hyperlinks through function processPage ()
After the scan no hyperlinks were found
Hyperlink found
When hyperlink is found the program determines if the hyperlink is containing the special words from ArrayList (ENHANCED VERSION TO ADD FOCUS)
Hyperlink didn't match the list of keywords but page wasn't scanned for all hyperlinks yet
Page wasn't scanned for all hyperlinks
Hyperlink contained a keyword from the list of keywords contained in predefined ArrayList variable
Hyperlink is added to the frontier through function addNewURL ()
Page was scanned for all hyperlinks
Hyperlink didn't match the list of keywords and page was fully scanned for all hyperlinks
Maximum file count reached and program loop finishes
Program ends

The program was enhanced with the semantic recognition of URLs contained within the page and if the URL string contains at least one of the keywords that are contained insite the ArrayList<String> called specialWords then that URL is added to the frontier. URLs that do not meet this requirement are not added to the frontier; therefore, are never downloaded and are essentially eliminated from further investigation allowing the web crawler to focus only on the URL that are semantically relevant. In order to do this I had to create a static ArrayList<String> called specialWords that was preloaded with 18 words. Then in addNewURL function I added a for-loop that goes through each word in ArrayList and compares it with URL for content. If at least one word is found in the URL that matches the word in specialWords then the URL is added to the frontier. If the URL does not contain any of special words then URL is ignored and program continues to analyze next URL.

Please turn in the revised source code of your program [1.00pt]

```
import java.text.*;
import java.util.*;
import java.net.*;
import java.io.*;

public class WebCrawlerEnchanced {
    public static final int SEARCH_LIMIT = 200;  // Absolute max pages
    public static final Boolean DEBUG = false;
    public static final String DISALLOW = "Disallow:";
    public static final int MAXSIZE = 200000; // Max size of file

    public static ArrayList<String> specialWords = new ArrayList<String>(Arrays.asList("sport","football","baseball",
                                                   "basketball","soccer","boxing","bowling","archery","golf","hockey",
                                                   "rugby","fishing","hunting","racing","boating","sailing","cycling",
                                                   "tennis"));


    // URLs to be searched
    Vector newURLs;
    // Known URLs
    Hashtable knownURLs;
    // max number of pages to download
    int maxPages;

    // log file name;
    String logFileName;


// initializes data structures.  argv is the command line arguments.

    public void initialize(String[] argv) {
        URL url;
        knownURLs = new Hashtable();
        newURLs = new Vector();
        try { url = new URL(argv[0]); }
        catch (MalformedURLException e)
        {
            System.out.println("Invalid starting URL " + argv[0]);
            return;
        }
        knownURLs.put(url,new Integer(1));
        newURLs.addElement(url);
        System.out.println("Starting search: Initial URL " + url.toString());
        maxPages = SEARCH_LIMIT;
        logFileName=new String(argv[1]);
        if (argv.length > 2)
        {
            int iPages = Integer.parseInt(argv[2]);
            if (iPages < maxPages) maxPages = iPages;
        }
        System.out.println("Maximum number of pages:" + maxPages);

/*Behind a firewall set your proxy and port here!
*/
        Properties props= new Properties(System.getProperties());
        props.put("http.proxySet", "true");
        props.put("http.proxyHost", "webcache-cup");
        props.put("http.proxyPort", "8080");

        Properties newprops = new Properties(props);
        System.setProperties(newprops);
/**/
```

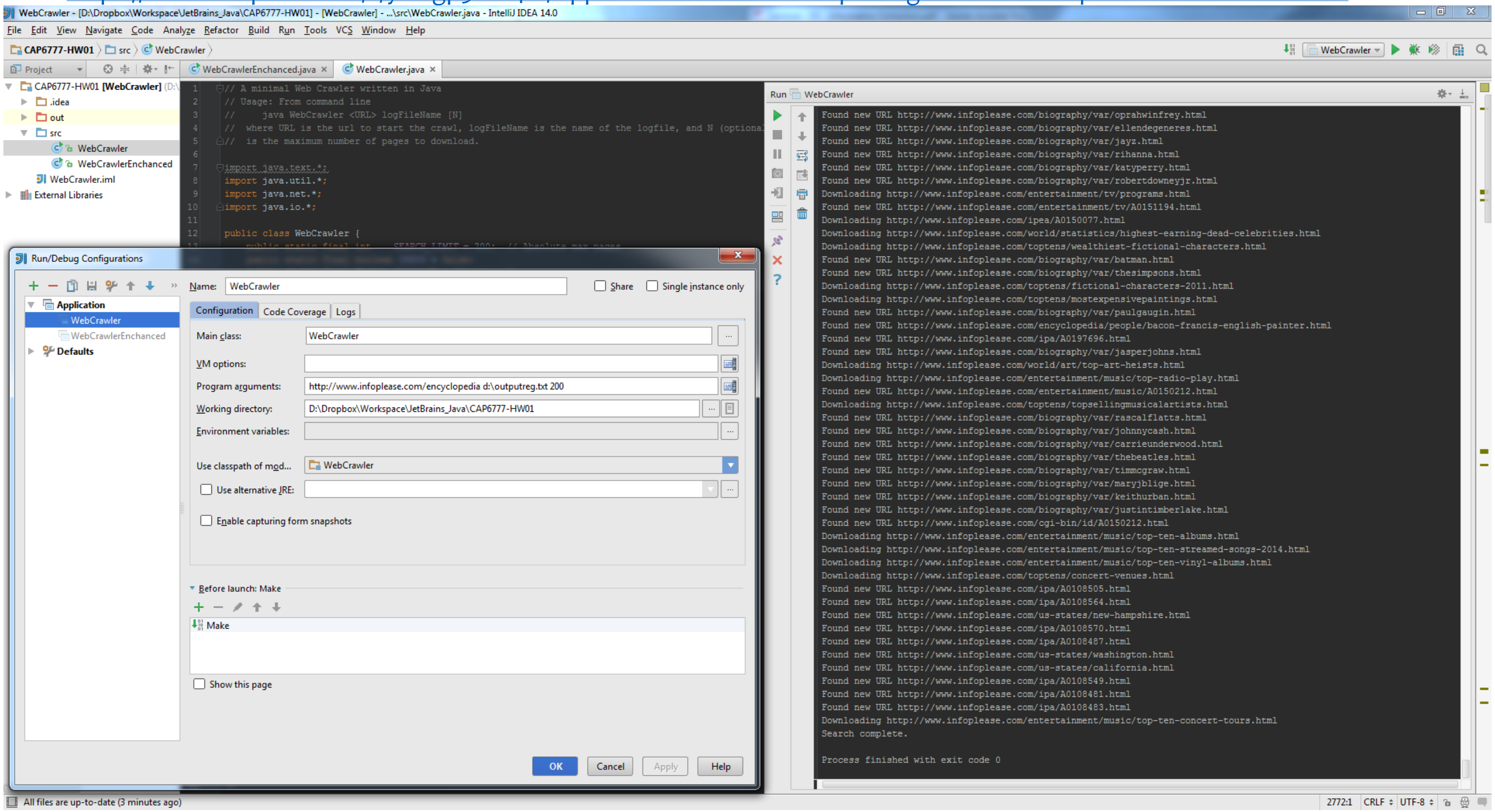
```
    }

// Check that the robot exclusion protocol does not disallow
// downloading url.

    public boolean robotSafe(URL url) {
        String strHost = url.getHost();

        // form URL of the robots.txt file
        String strRobot = "http://" + strHost + "/robots.txt";
        URL urlRobot;
        try { urlRobot = new URL(strRobot);}
        catch (MalformedURLException e)
        {
            // something weird is happening, so don't trust it
            return false;
        }

        if (DEBUG) System.out.println("Checking robot protocol " +
                urlRobot.toString());
        String strCommands;
        try
        {
            InputStream urlRobotStream = urlRobot.openStream();

            // read in entire file
            byte b[] = new byte[1000];
            int numRead = urlRobotStream.read(b);
            strCommands = new String(b, 0, numRead);
            while (numRead != -1)
            {
                numRead = urlRobotStream.read(b);
                if (numRead != -1)
                {
                    String newCommands = new String(b, 0, numRead);
                    strCommands += newCommands;
                }
            }
            urlRobotStream.close();
        } catch (IOException e)
        {
            // if there is no robots.txt file, it is OK to search
            return true;
        }
        if (DEBUG) System.out.println(strCommands);

        // assume that this robots.txt refers to us and
        // search for "Disallow:" commands.
        String strURL = url.getFile();
        int index = 0;
        while ((index = strCommands.indexOf(DISALLOW, index)) != -1)
        {
            index += DISALLOW.length();
            String strPath = strCommands.substring(index);
            StringTokenizer st = new StringTokenizer(strPath);

            if (!st.hasMoreTokens())
                break;

            String strBadPath = st.nextToken();

            // if the URL starts with a disallowed path, it is not safe
            if (strURL.indexOf(strBadPath) == 0)
                return false;
        }

        return true;
    }

// adds new URL to the queue. Accept only new URL's that end in
// htm or html. oldURL is the context, newURLString is the link
// (either an absolute or a relative URL).

    public void addnewurl(URL oldURL, String newUrlString)

    { URL url;
        if (DEBUG) System.out.println("URL String " + newUrlString);
        try
        { url = new URL(oldURL,newUrlString);
            if (!knownURLs.containsKey(url))
            {
                String filename =  url.getFile();
                int iSuffix = filename.lastIndexOf("htm");
                if ((iSuffix == filename.length() - 3) ||
                        (iSuffix == filename.length() - 4))
                {
                    knownURLs.put(url,new Integer(1));

                    for(int k = 0; k < specialWords.size(); k++)
                    {
                        if (url.toString().contains(specialWords.get(k)))
                        {
                            newURLs.addElement(url);
                            System.out.println("Found new URL " + url.toString());
                            break;
                        }
                    }

                }
            }
        }
        catch (MalformedURLException e) { return; }
    }


// Download contents of URL

    public String getpage(URL url)

    { try
    {
        // try opening the URL
        URLConnection urlConnection = url.openConnection();
        System.out.println("Downloading " + url.toString());

        urlConnection.setAllowUserInteraction(false);

        InputStream urlStream = url.openStream();
        // search the input stream for links
        // first, read in the entire URL
        byte b[] = new byte[1000];
        int numRead = urlStream.read(b);
        String content = new String(b, 0, numRead);
        while ((numRead != -1) && (content.length() < MAXSIZE))
        {
            numRead = urlStream.read(b);
            if (numRead != -1)
            {
                String newContent = new String(b, 0, numRead);
                content += newContent;
            }
        }
        return content;

    }
    catch (IOException e)
    {
        System.out.println("ERROR: couldn't open URL ");
        return "";
    }
    }

// Go through page finding links to URLs.  A link is signalled
// by <a href=" ...   It ends with a close angle bracket, preceded
// by a close quote, possibly preceded by a hatch mark (marking a
// fragment, an internal page marker)

    public void processpage(URL url, String page)

    {
        String lcPage = page.toLowerCase(); // Page in lower case
        int index = 0; // position in page
        int iEndAngle, ihref, iURL, iCloseQuote, iHatchMark, iEnd;
        while ((index = lcPage.indexOf("<a",index)) != -1)
        {
            iEndAngle = lcPage.indexOf(">",index);
            ihref = lcPage.indexOf("href",index);
            if (ihref != -1)
            {
                iURL = lcPage.indexOf("\"", ihref) + 1;
                if ((iURL != -1) && (iEndAngle != -1) && (iURL < iEndAngle))
                { iCloseQuote = lcPage.indexOf("\"",iURL);
                    iHatchMark = lcPage.indexOf("#", iURL);
                    if ((iCloseQuote != -1) && (iCloseQuote < iEndAngle))
```



```
                    {
                        iEnd = iCloseQuote;
                        if ((iHatchMark != -1) && (iHatchMark < iCloseQuote))
                            iEnd = iHatchMark;
                        String newUrlString = page.substring(iURL,iEnd);
                        addnewurl(url, newUrlString);
                    }
                }
            }
            index = iEndAngle;
            if(index==-1){break;}
        }
    }
//-----


//----

// Top-level procedure. Keep popping a url off newURLs, download
// it, and accumulate new URLs

    public void run(String[] argv)

    {
        initialize(argv);
        File newTextFile = new File(logFileName);
        try{
            FileWriter fileWriter = new FileWriter(newTextFile);
            BufferedWriter output = new BufferedWriter(fileWriter);
            Date date;
            date=new Date();
            String formattedDate = "";

            for (int i = 0; i < maxPages; i++)
            {
                URL url = (URL) newURLs.elementAt(0);
                output.write("\r\n\r\n ++++");
                output.write(date.toGMTString());
                output.write("++++ \r\n");
                newURLs.removeElementAt(0);
                if (DEBUG) System.out.println("Searching " + url.toString());

                if (robotSafe(url))
                {
                    String page = getpage(url);
                    if(page.length()!=0)
                    {
                        output.write(url.toString());
                        output.write("\r\n ================================ \r\n ");
                        output.write(page);
                    }
                    if (DEBUG) System.out.println(page);
                    if (page.length() != 0) processpage(url,page);
                    if (newURLs.isEmpty()) break;
                }
            }
            System.out.println("Search complete.");
            output.close();
        }
        catch (Exception e)
        {//Catch exception if any
            System.err.println("Error: " + e.getMessage());
        }
    }

    public static void main(String[] argv)
    {
        WebCrawlerEnchanced wc = new WebCrawlerEnchanced();
        wc.run(argv);
    }
}
```

Given the same set of seed URLs, please use original Webcrawler.java to collect 200 web pages, and also use your new preferential web crawler to collect 200 web pages. For 200 web pages collected from different web crawlers, please check the percentage of the web pages containing the specific keywords, and report the values in your report [1.00pt].

Original WebCrawler crawling http://www.infoplease.com/encyclopedia for 200 links (see Appendix B for results via Dropbox Link)
Link : https://www.dropbox.com/s/yl6agp5frbi4lri/Appendix%20B%20-%20output-regular-www.infoplease.com-200.txt?dl=0



WebCrawler - [D:\Dropbox\Workspace\JetBrains_Java\CAP6777-HW01] - [WebCrawler] - ...\src\WebCrawler.java - IntelliJ IDEA 14.0

```
// A minimal Web Crawler written in Java
// Usage: From command line
//     java WebCrawler <URL> logFileName [N]
//  where URL is the url to start the crawl, logFileName is the name of the logfile, and N (optiona
//  is the maximum number of pages to download.

import java.text.*;
import java.util.*;
import java.net.*;
import java.io.*;

public class WebCrawler {
```

Run/Debug Configurations

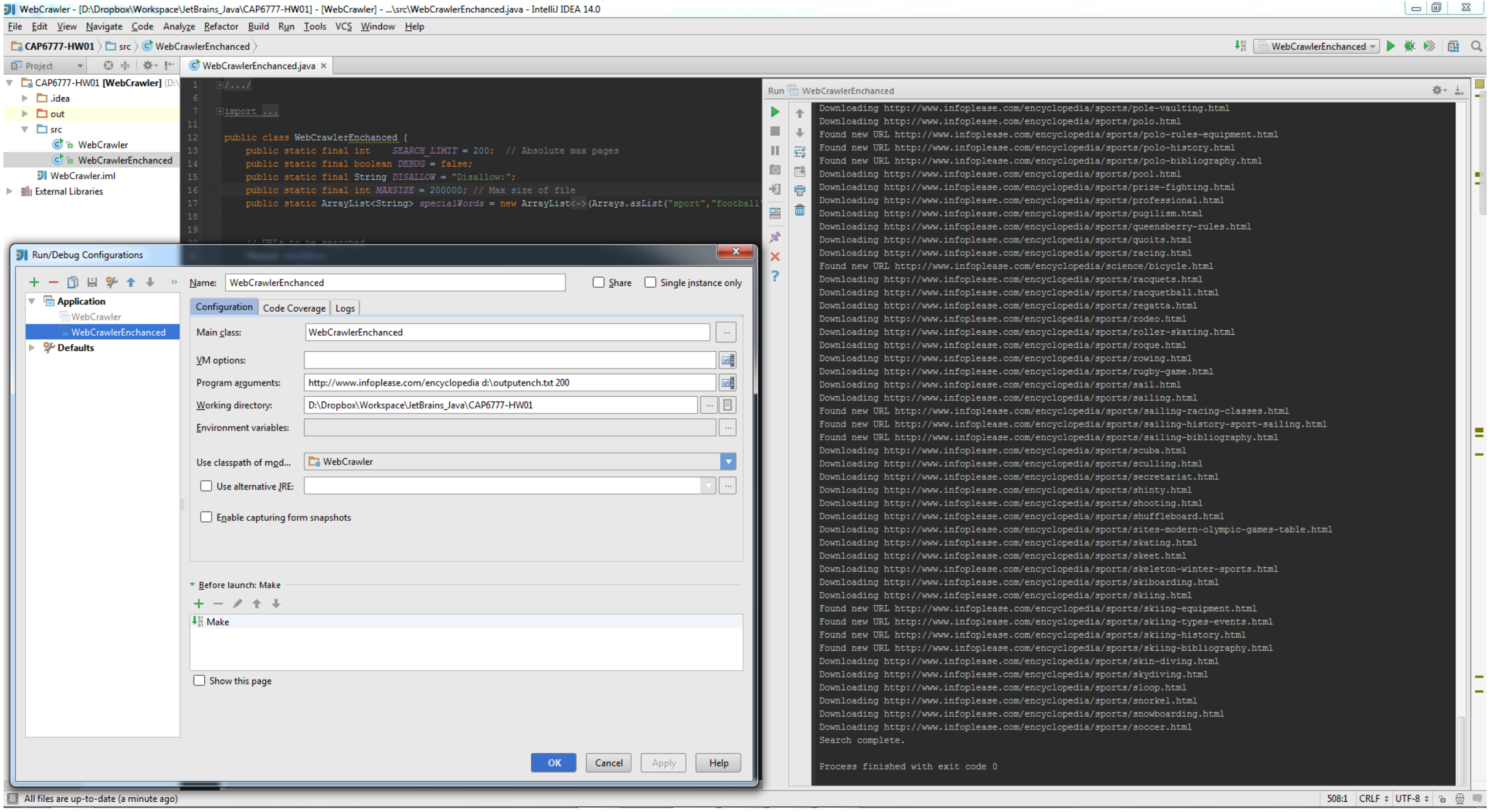
Name: WebCrawler
□ Share □ Single instance only

Main class: WebCrawler
VM options:
Program arguments: http://www.infoplease.com/encyclopedia d:\outputreg.txt 200
Working directory: D:\Dropbox\Workspace\JetBrains_Java\CAP6777-HW01
Environment variables:
Use classpath of mod...: WebCrawler
□ Use alternative JRE:
□ Enable capturing form snapshots
Before launch: Make
□ Show this page

```
Found new URL http://www.infoplease.com/biography/var/oprahwinfrey.html
Found new URL http://www.infoplease.com/biography/var/ellendegeneres.html
Found new URL http://www.infoplease.com/biography/var/jayz.html
Found new URL http://www.infoplease.com/biography/var/rihanna.html
Found new URL http://www.infoplease.com/biography/var/katyperry.html
Found new URL http://www.infoplease.com/biography/var/robertdowneyjr.html
Downloading http://www.infoplease.com/entertainment/tv/programs.html
Found new URL http://www.infoplease.com/entertainment/tv/A0151194.html
Downloading http://www.infoplease.com/ipea/A0150077.html
Downloading http://www.infoplease.com/world/statistics/highest-earning-dead-celebrities.html
Downloading http://www.infoplease.com/toptens/wealthiest-fictional-characters.html
Found new URL http://www.infoplease.com/biography/var/batman.html
Found new URL http://www.infoplease.com/biography/var/thesimpsons.html
Downloading http://www.infoplease.com/toptens/fictional-characters-2011.html
Downloading http://www.infoplease.com/toptens/mostexpensivepaintings.html
Found new URL http://www.infoplease.com/biography/var/paulgaugin.html
Found new URL http://www.infoplease.com/encyclopedia/people/bacon-francis-english-painter.html
Found new URL http://www.infoplease.com/ipa/A0197696.html
Found new URL http://www.infoplease.com/biography/var/jasperjohns.html
Downloading http://www.infoplease.com/world/art/top-art-heists.html
Downloading http://www.infoplease.com/entertainment/music/top-radio-play.html
Found new URL http://www.infoplease.com/entertainment/music/A0150212.html
Downloading http://www.infoplease.com/toptens/topsellingmusicalartists.html
Found new URL http://www.infoplease.com/biography/var/rascalflatts.html
Found new URL http://www.infoplease.com/biography/var/johnnycash.html
Found new URL http://www.infoplease.com/biography/var/carrieunderwood.html
Found new URL http://www.infoplease.com/biography/var/thebeatles.html
Found new URL http://www.infoplease.com/biography/var/timmcgraw.html
Found new URL http://www.infoplease.com/biography/var/maryjblige.html
Found new URL http://www.infoplease.com/biography/var/keithurban.html
Found new URL http://www.infoplease.com/biography/var/justintimberlake.html
Found new URL http://www.infoplease.com/cgi-bin/id/A0150212.html
Downloading http://www.infoplease.com/entertainment/music/top-ten-albums.html
Downloading http://www.infoplease.com/entertainment/music/top-ten-streamed-songs-2014.html
Downloading http://www.infoplease.com/entertainment/music/top-ten-vinyl-albums.html
Downloading http://www.infoplease.com/toptens/concert-venues.html
Found new URL http://www.infoplease.com/ipa/A0108505.html
Found new URL http://www.infoplease.com/ipa/A0108564.html
Found new URL http://www.infoplease.com/us-states/new-hampshire.html
Found new URL http://www.infoplease.com/ipa/A0108570.html
Found new URL http://www.infoplease.com/ipa/A0108487.html
Found new URL http://www.infoplease.com/us-states/washington.html
Found new URL http://www.infoplease.com/us-states/california.html
Found new URL http://www.infoplease.com/ipa/A0108549.html
Found new URL http://www.infoplease.com/ipa/A0108481.html
Found new URL http://www.infoplease.com/ipa/A0108483.html
Downloading http://www.infoplease.com/entertainment/music/top-ten-concert-tours.html
Search complete.

Process finished with exit code 0
```

All files are up-to-date (3 minutes ago)

Enhanced WebCrawler crawling http://www.infoplease.com/encyclopedia for 200 links (see Appendix C for results via Dropbox Link)
Link : https://www.dropbox.com/s/bdnotsu1cv7dnuc/Appendix%20C%20-%20output-enchanced-www.infoplease.com-200.txt?dl=0



WebCrawler - [D:\Dropbox\Workspace\JetBrains_Java\CAP6777-HW01] - [WebCrawler] - ...\src\WebCrawlerEnchanced.java - IntelliJ IDEA 14.0

```
/.../

import ...

public class WebCrawlerEnchanced {
    public static final int    SEARCH_LIMIT = 200;  // Absolute max pages
    public static final boolean DEBUG = false;
    public static final String DISALLOW = "Disallow:";
    public static final int MAXSIZE = 200000; // Max size of file
    public static ArrayList<String> specialWords = new ArrayList<~>(Arrays.asList("sport","football
```

Run/Debug Configurations

Name: WebCrawlerEnchanced
□ Share □ Single instance only

Main class: WebCrawlerEnchanced
VM options:
Program arguments: http://www.infoplease.com/encyclopedia d:\outputench.txt 200
Working directory: D:\Dropbox\Workspace\JetBrains_Java\CAP6777-HW01
Environment variables:
Use classpath of mod...: WebCrawler
□ Use alternative JRE:
□ Enable capturing form snapshots
Before launch: Make
□ Show this page

```
Downloading http://www.infoplease.com/encyclopedia/sports/pole-vaulting.html
Downloading http://www.infoplease.com/encyclopedia/sports/polo.html
Found new URL http://www.infoplease.com/encyclopedia/sports/polo-rules-equipment.html
Found new URL http://www.infoplease.com/encyclopedia/sports/polo-history.html
Found new URL http://www.infoplease.com/encyclopedia/sports/polo-bibliography.html
Downloading http://www.infoplease.com/encyclopedia/sports/pool.html
Downloading http://www.infoplease.com/encyclopedia/sports/prize-fighting.html
Downloading http://www.infoplease.com/encyclopedia/sports/professional.html
Downloading http://www.infoplease.com/encyclopedia/sports/pugilism.html
Downloading http://www.infoplease.com/encyclopedia/sports/queensberry-rules.html
Downloading http://www.infoplease.com/encyclopedia/sports/quoits.html
Downloading http://www.infoplease.com/encyclopedia/sports/racing.html
Found new URL http://www.infoplease.com/encyclopedia/science/bicycle.html
Downloading http://www.infoplease.com/encyclopedia/sports/racquets.html
Downloading http://www.infoplease.com/encyclopedia/sports/racquetball.html
Downloading http://www.infoplease.com/encyclopedia/sports/regatta.html
Downloading http://www.infoplease.com/encyclopedia/sports/rodeo.html
Downloading http://www.infoplease.com/encyclopedia/sports/roller-skating.html
Downloading http://www.infoplease.com/encyclopedia/sports/roque.html
Downloading http://www.infoplease.com/encyclopedia/sports/rowing.html
Downloading http://www.infoplease.com/encyclopedia/sports/rugby-game.html
Downloading http://www.infoplease.com/encyclopedia/sports/sail.html
Downloading http://www.infoplease.com/encyclopedia/sports/sailing.html
Found new URL http://www.infoplease.com/encyclopedia/sports/sailing-racing-classes.html
Found new URL http://www.infoplease.com/encyclopedia/sports/sailing-history-sport-sailing.html
Found new URL http://www.infoplease.com/encyclopedia/sports/sailing-bibliography.html
Downloading http://www.infoplease.com/encyclopedia/sports/scuba.html
Downloading http://www.infoplease.com/encyclopedia/sports/sculling.html
Downloading http://www.infoplease.com/encyclopedia/sports/secretariat.html
Downloading http://www.infoplease.com/encyclopedia/sports/shinty.html
Downloading http://www.infoplease.com/encyclopedia/sports/shooting.html
Downloading http://www.infoplease.com/encyclopedia/sports/shuffleboard.html
Downloading http://www.infoplease.com/encyclopedia/sports/sites-modern-olympic-games-table.html
Downloading http://www.infoplease.com/encyclopedia/sports/skating.html
Downloading http://www.infoplease.com/encyclopedia/sports/skeet.html
Downloading http://www.infoplease.com/encyclopedia/sports/skeleton-winter-sports.html
Downloading http://www.infoplease.com/encyclopedia/sports/skiboarding.html
Downloading http://www.infoplease.com/encyclopedia/sports/skiing.html
Found new URL http://www.infoplease.com/encyclopedia/sports/skiing-equipment.html
Found new URL http://www.infoplease.com/encyclopedia/sports/skiing-types-events.html
Found new URL http://www.infoplease.com/encyclopedia/sports/skiing-history.html
Found new URL http://www.infoplease.com/encyclopedia/sports/skiing-bibliography.html
Downloading http://www.infoplease.com/encyclopedia/sports/skin-diving.html
Downloading http://www.infoplease.com/encyclopedia/sports/skydiving.html
Downloading http://www.infoplease.com/encyclopedia/sports/sloop.html
Downloading http://www.infoplease.com/encyclopedia/sports/snorkel.html
Downloading http://www.infoplease.com/encyclopedia/sports/snowboarding.html
Downloading http://www.infoplease.com/encyclopedia/sports/soccer.html
Search complete.

Process finished with exit code 0
```

All files are up-to-date (a minute ago)

Seed URL: http://www.infoplease.com/encyclopedia			
Keyword	Original	Enhanced	% Original / Enhanced
sport	1337	5625	24%
football	14	118	12%
baseball	16	228	7%
basketball	14	68	21%
soccer	4	67	6%
boxing	6	216	3%
bowling	1	61	2%
archery	0	20	0%
golf	9	70	13%
hockey	10	125	8%
rugby	0	37	0%
fishing	0	28	0%
hunting	0	32	0%
racing	4	166	2%
boating	0	51	0%
sailing	0	38	0%
cycling	0	19	0%
tennis	9	186	5%
Keyword Total Count	1424	7155	20%
Keyword Average	79	398	20%

Please suggest one additional approach (show your design as a flow chart or a diagram) which may help improve the accuracy of preferential web crawler, so the collected documents are closely related to the topics [1.00pt]

At this moment my enhanced version of the program scans the URLs if they are containing specific words. However, that reduces the ability of web crawler to only look at the URL that are specifically named with those words. A major improvement would be to assess the neighboring text of the URL within the page. The text that comes before or after URL is often very telling of the content that URL will link to while that URL may be abbreviated to achieve smaller page download. In this way grabbing several words that come before the anchor tag and couple of words that comes after anchor tag might be much better approach. To illustrate the example

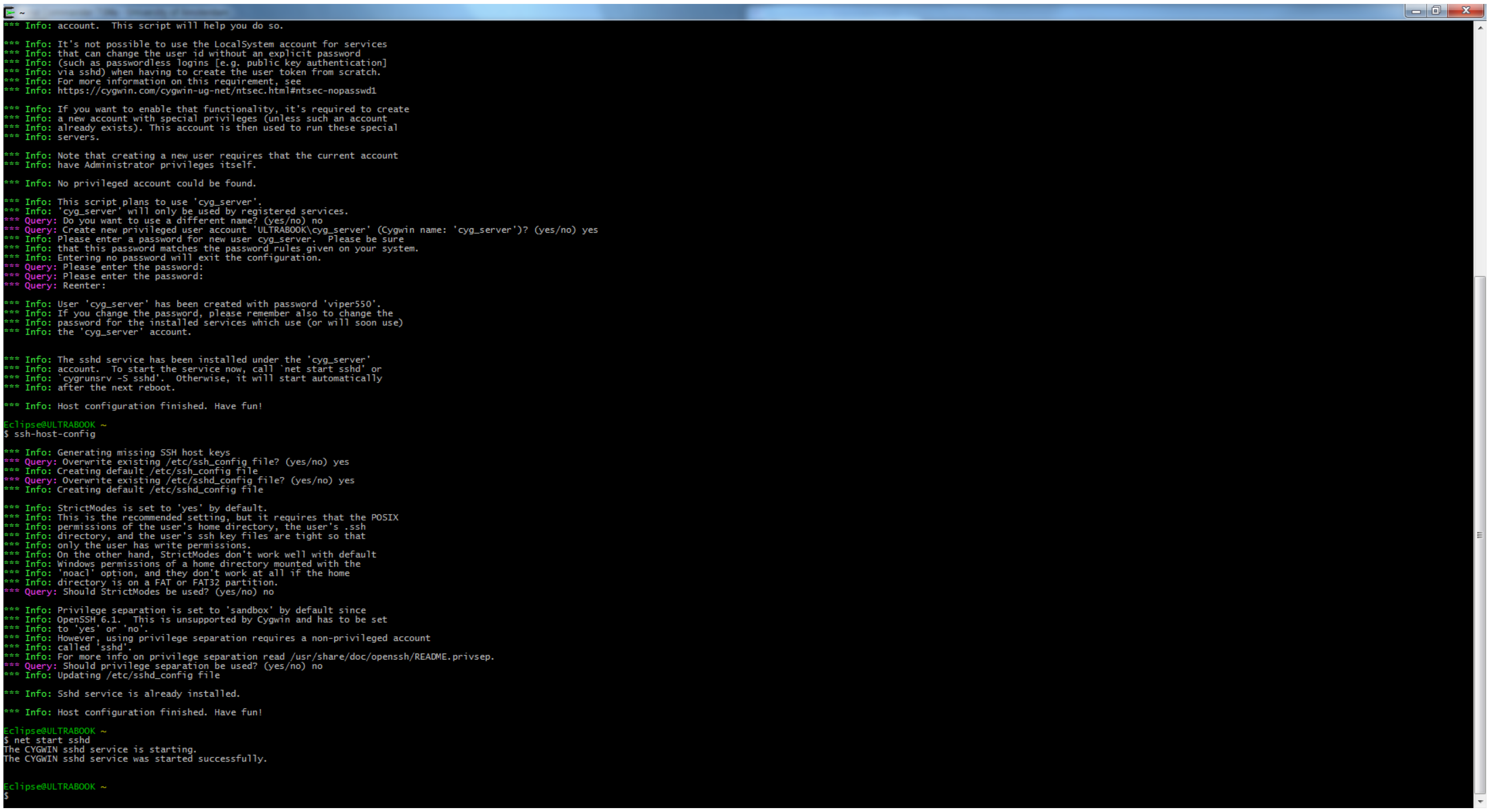
```
<ul>
  <li><a href=”article-2016-05-25.html”>Football season has ended</li>
  <li><a href=”article-2016-05-26.html”>Owls are looking towards the summer</li>
</ul>
```

In this example looking at the words that come after URL will be far more beneficial than looking at URL itself; therefore, obtaining words that come before and after URL should be considered when deciding if URL is semantically relevant or not.

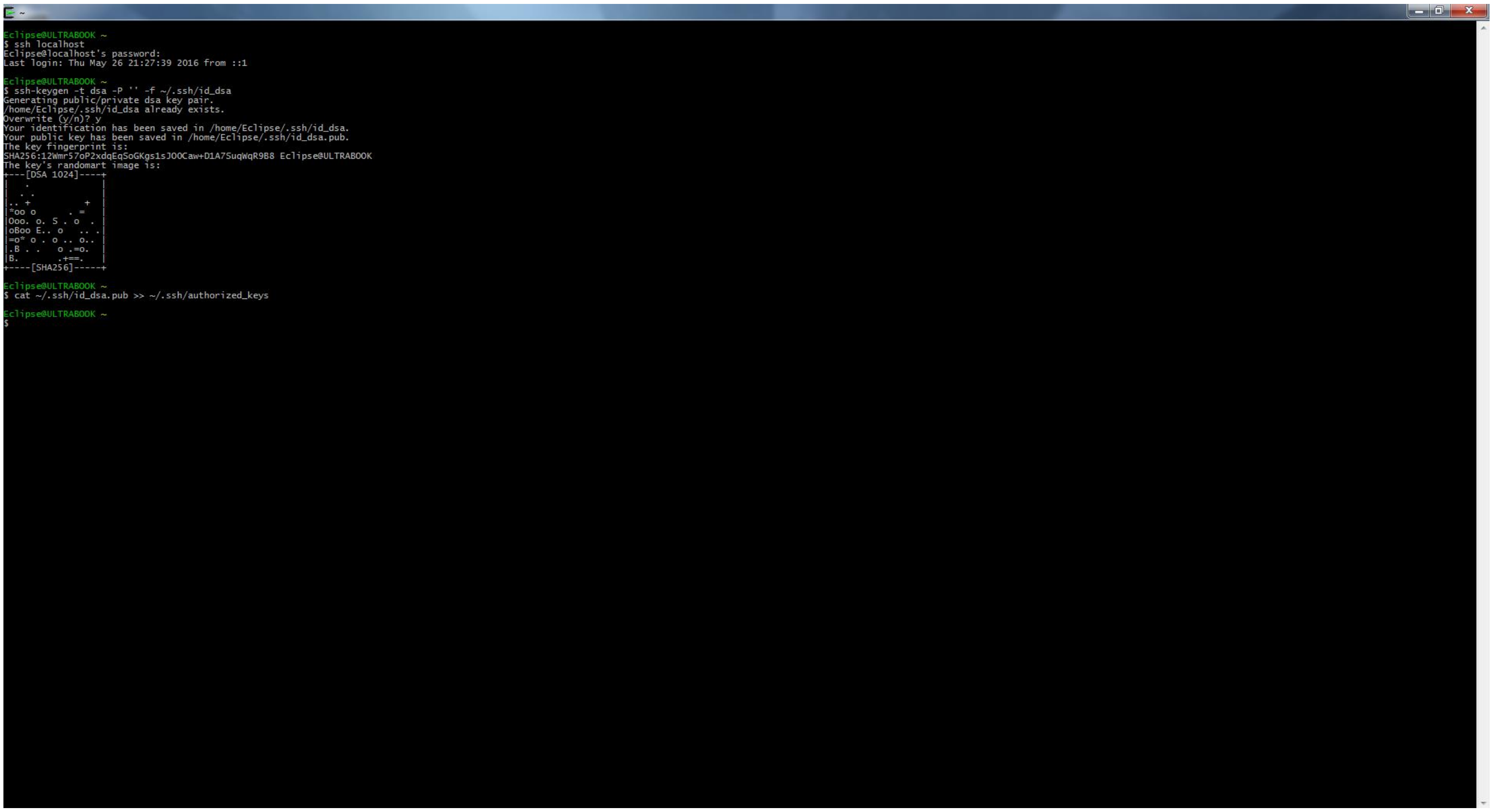
Question 3 – [3.75pt]

Please follow the instructions show in the “Apache.Nutch.installation.docx” file, install and configure a Nutch web crawler. Please capture three screenshots to show that

a. Cygwin has been properly installed and running on your computer [0.50pt]

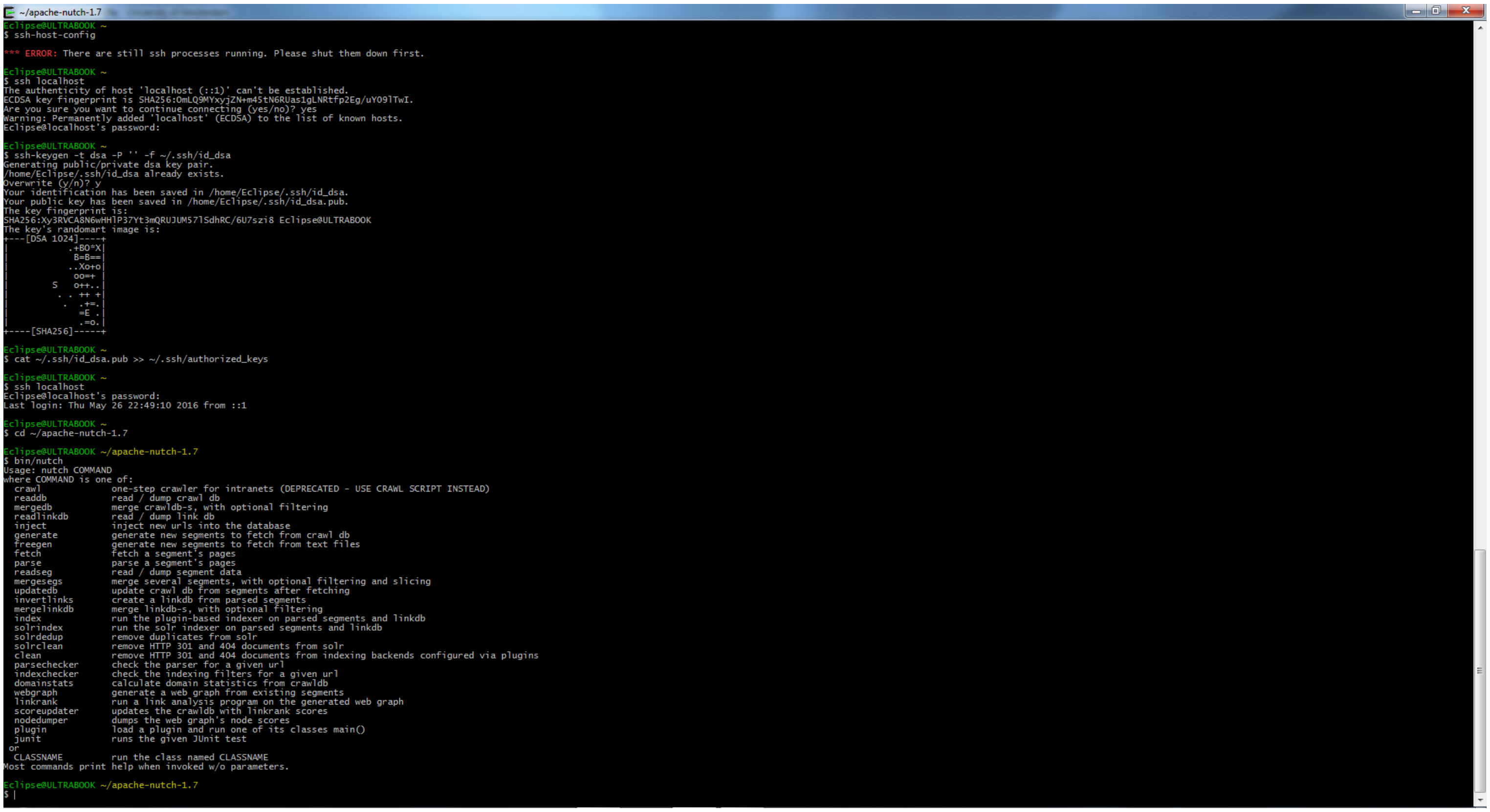


```
*** Info: account.  This script will help you do so.

*** Info: It's not possible to use the LocalSystem account for services
*** Info: that can change the user id without an explicit password
*** Info: (such as passwordless logins [e.g. public key authentication]
*** Info: via sshd) when having to create the user token from scratch.
*** Info: For more information on this requirement, see
*** Info: https://cygwin.com/cygwin-ug-net/ntsec.html#ntsec-nopasswd1

*** Info: If you want to enable that functionality, it's required to create
*** Info: a new account with special privileges (unless such an account
*** Info: already exists). This account is then used to run these special
*** Info: servers.

*** Info: Note that creating a new user requires that the current account
*** Info: have Administrator privileges itself.

*** Info: No privileged account could be found.

*** Info: This script plans to use 'cyg_server'.
*** Info: 'cyg_server' will only be used by registered services.
*** Query: Do you want to use a different name? (yes/no) no
*** Query: Create new privileged user account 'ULTRABOOK\cyg_server' (Cygwin name: 'cyg_server')? (yes/no) yes
*** Info: Please enter a password for new user cyg_server.  Please be sure
*** Info: that this password matches the password rules given on your system.
*** Info: Entering no password will exit the configuration.
*** Query: Please enter the password:
*** Query: Please enter the password:
*** Query: Reenter:

*** Info: User 'cyg_server' has been created with password 'viper550'.
*** Info: If you change the password, please remember also to change the
*** Info: password for the installed services which use (or will soon use)
*** Info: the 'cyg_server' account.


*** Info: The sshd service has been installed under the 'cyg_server'
*** Info: account.  To start the service now, call `net start sshd' or
*** Info: `cygrunsrv -S sshd'.  Otherwise, it will start automatically
*** Info: after the next reboot.

*** Info: Host configuration finished. Have fun!

Eclipse@ULTRABOOK ~
$ ssh-host-config

*** Info: Generating missing SSH host keys
*** Query: Overwrite existing /etc/ssh_config file? (yes/no) yes
*** Info: Creating default /etc/ssh_config file
*** Query: Overwrite existing /etc/sshd_config file? (yes/no) yes
*** Info: Creating default /etc/sshd_config file

*** Info: StrictModes is set to 'yes' by default.
*** Info: This is the recommended setting, but it requires that the POSIX
*** Info: permissions of the user's home directory, the user's .ssh
*** Info: directory, and the user's ssh key files are tight so that
*** Info: only the user has write permissions.
*** Info: On the other hand, StrictModes don't work well with default
*** Info: Windows permissions of a home directory mounted with the
*** Info: 'noacl' option, and they don't work at all if the home
*** Info: directory is on a FAT or FAT32 partition.
*** Query: Should StrictModes be used? (yes/no) no

*** Info: Privilege separation is set to 'sandbox' by default since
*** Info: OpenSSH 6.1.  This is unsupported by Cygwin and has to be set
*** Info: to 'yes' or 'no'.
*** Info: However, using privilege separation requires a non-privileged account
*** Info: called 'sshd'.
*** Info: For more info on privilege separation read /usr/share/doc/openssh/README.privsep.
*** Query: Should privilege separation be used? (yes/no) no
*** Info: Updating /etc/sshd_config file

*** Info: Sshd service is already installed.

*** Info: Host configuration finished. Have fun!

Eclipse@ULTRABOOK ~
$ net start sshd
The CYGWIN sshd service is starting.
The CYGWIN sshd service was started successfully.


Eclipse@ULTRABOOK ~
$
```



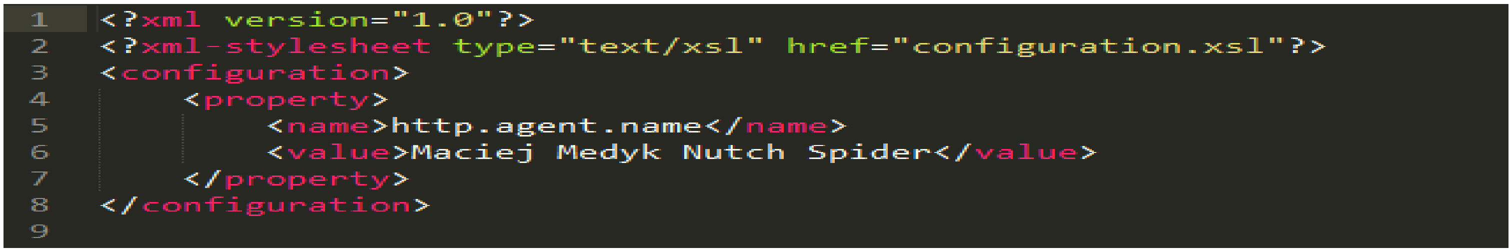
```
Eclipse@ULTRABOOK ~
$ ssh localhost
Eclipse@localhost's password:
Last login: Thu May 26 21:27:39 2016 from ::1

Eclipse@ULTRABOOK ~
$ ssh-keygen -t dsa -P '' -f ~/.ssh/id_dsa
Generating public/private dsa key pair.
/home/Eclipse/.ssh/id_dsa already exists.
Overwrite (y/n)? y
Your identification has been saved in /home/Eclipse/.ssh/id_dsa.
Your public key has been saved in /home/Eclipse/.ssh/id_dsa.pub.
The key fingerprint is:
SHA256:12Wmr57oP2xdqEqSoGKgs1sJOOCaw+D1A7SuqWqR9B8 Eclipse@ULTRABOOK
The key's randomart image is:
+---[DSA 1024]----+
|    .            |
|   . .           |
|.. +         +   |
|*oo o      . =   |
|Ooo. o. S . o  . |
|oBoo E.. o   .. .|
|=o* o . o .. o.. |
|.B . .   o .=o.  |
|B.       .+==.   |
+----[SHA256]-----+

Eclipse@ULTRABOOK ~
$ cat ~/.ssh/id_dsa.pub >> ~/.ssh/authorized_keys

Eclipse@ULTRABOOK ~
$
```

b. Notch has been downloaded and configured as showing in the instructions [0.50pt]

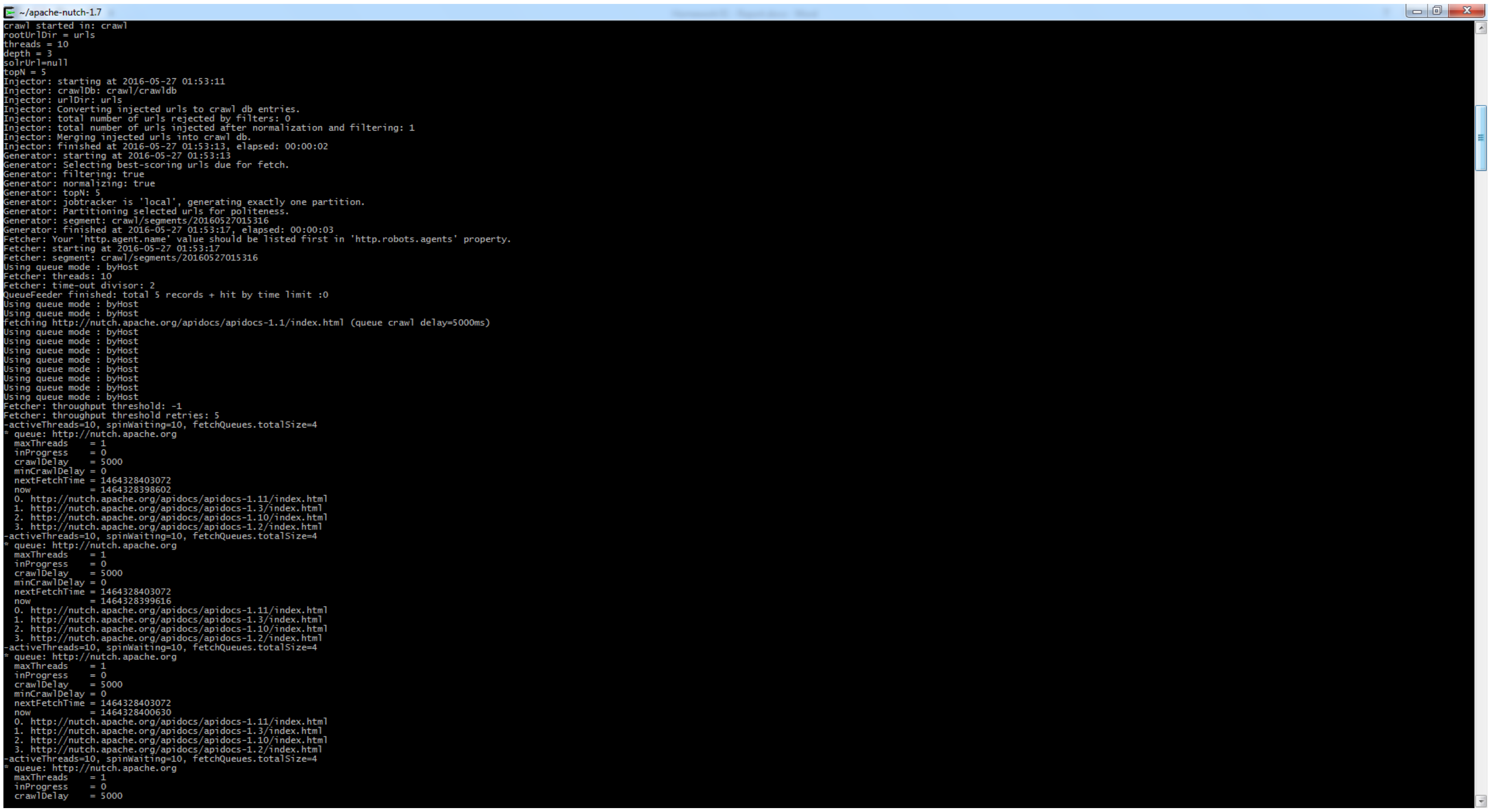


~/apache-nutch-1.7

```
Eclipse@ULTRABOOK ~
$ ssh-host-config

*** ERROR: There are still ssh processes running. Please shut them down first.

Eclipse@ULTRABOOK ~
$ ssh localhost
The authenticity of host 'localhost (::1)' can't be established.
ECDSA key fingerprint is SHA256:OmLQ9MYxyjZN+m45tN6RUas1gLNRtfp2Eg/uYO9lTwI.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added 'localhost' (ECDSA) to the list of known hosts.
Eclipse@localhost's password:

Eclipse@ULTRABOOK ~
$ ssh-keygen -t dsa -P '' -f ~/.ssh/id_dsa
Generating public/private dsa key pair.
/home/Eclipse/.ssh/id_dsa already exists.
Overwrite (y/n)? y
Your identification has been saved in /home/Eclipse/.ssh/id_dsa.
Your public key has been saved in /home/Eclipse/.ssh/id_dsa.pub.
The key fingerprint is:
SHA256:Xy3RVCA8N6wHHlP37Yt3mQRUJUM57lSdhRC/6U7szi8 Eclipse@ULTRABOOK
The key's randomart image is:
+---[DSA 1024]----+
|           .+BO*X|
|            B=B==|
|           ..Xo+o|
|            oo=+ |
|        S   o++..|
|         . . ++ +|
|          .  .+=.|
|             =E .|
|             .=o.|
+----[SHA256]-----+

Eclipse@ULTRABOOK ~
$ cat ~/.ssh/id_dsa.pub >> ~/.ssh/authorized_keys

Eclipse@ULTRABOOK ~
$ ssh localhost
Eclipse@localhost's password:
Last login: Thu May 26 22:49:10 2016 from ::1

Eclipse@ULTRABOOK ~
$ cd ~/apache-nutch-1.7

Eclipse@ULTRABOOK ~/apache-nutch-1.7
$ bin/nutch
Usage: nutch COMMAND
where COMMAND is one of:
  crawl             one-step crawler for intranets (DEPRECATED - USE CRAWL SCRIPT INSTEAD)
  readdb            read / dump crawl db
  mergedb           merge crawldb-s, with optional filtering
  readlinkdb        read / dump link db
  inject            inject new urls into the database
  generate          generate new segments to fetch from crawl db
  freegen           generate new segments to fetch from text files
  fetch             fetch a segment's pages
  parse             parse a segment's pages
  readseg           read / dump segment data
  mergesegs         merge several segments, with optional filtering and slicing
  updatedb          update crawl db from segments after fetching
  invertlinks       create a linkdb from parsed segments
  mergelinkdb       merge linkdb-s, with optional filtering
  index             run the plugin-based indexer on parsed segments and linkdb
  solrindex         run the solr indexer on parsed segments and linkdb
  solrdedup         remove duplicates from solr
  solrclean         remove HTTP 301 and 404 documents from solr
  clean             remove HTTP 301 and 404 documents from indexing backends configured via plugins
  parsechecker      check the parser for a given url
  indexchecker      check the indexing filters for a given url
  domainstats       calculate domain statistics from crawldb
  webgraph          generate a web graph from existing segments
  linkrank          run a link analysis program on the generated web graph
  scoreupdater      updates the crawldb with linkrank scores
  nodedumper        dumps the web graph's node scores
  plugin            load a plugin and run one of its classes main()
  junit             runs the given JUnit test
 or
  CLASSNAME         run the class named CLASSNAME
Most commands print help when invoked w/o parameters.

Eclipse@ULTRABOOK ~/apache-nutch-1.7
$ |
```



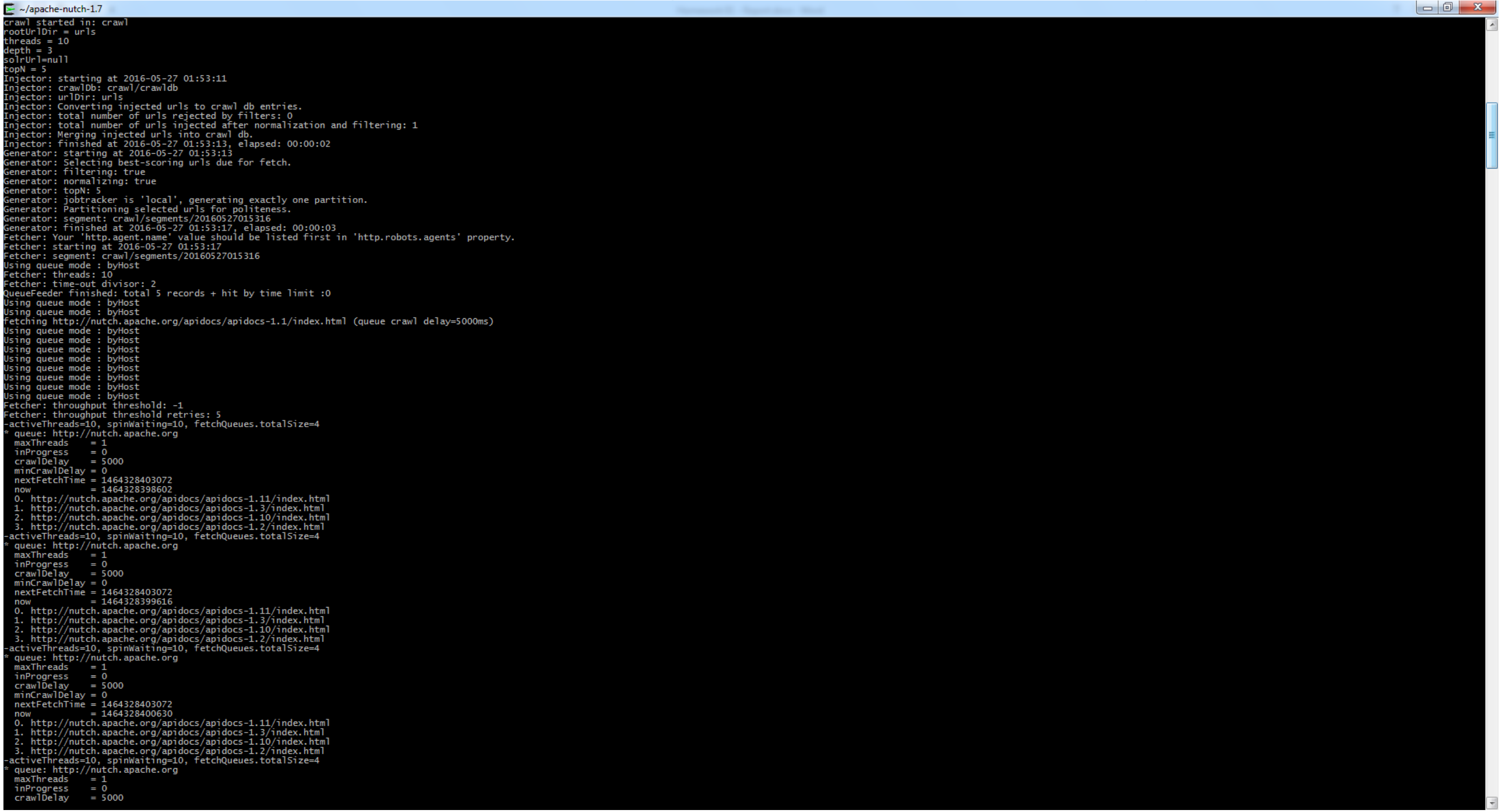
```xml
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<configuration>
    <property>
        <name>http.agent.name</name>
        <value>Maciej Medyk Nutch Spider</value>
    </property>
</configuration>

```



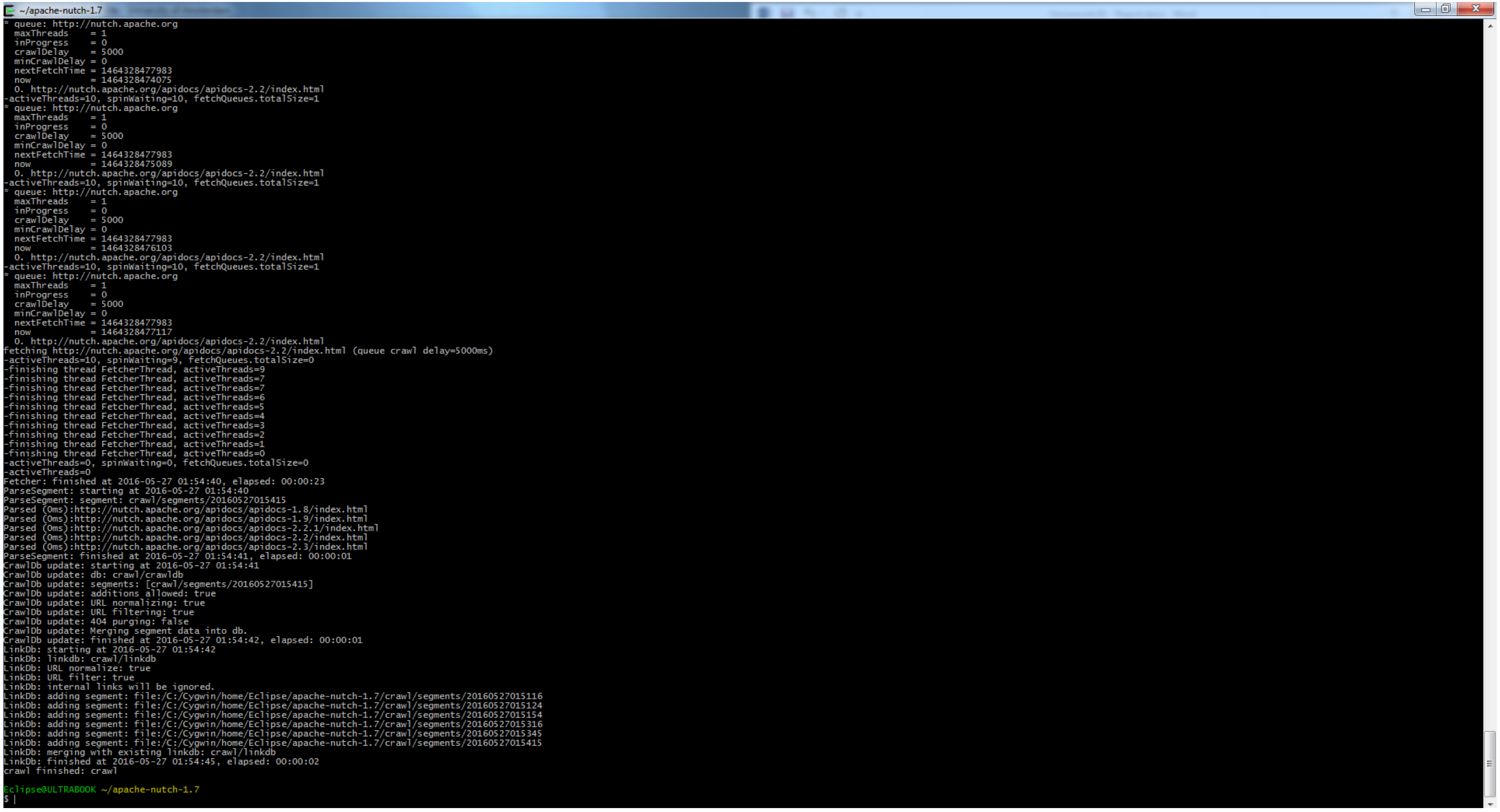
~/apache-nutch-1.7

```
crawl started in: crawl
rootUrlDir = urls
threads = 10
depth = 3
solrUrl=null
topN = 5
Injector: starting at 2016-05-27 01:53:11
Injector: crawlDb: crawl/crawldb
Injector: urlDir: urls
Injector: Converting injected urls to crawl db entries.
Injector: total number of urls rejected by filters: 0
Injector: total number of urls injected after normalization and filtering: 1
Injector: Merging injected urls into crawl db.
Injector: finished at 2016-05-27 01:53:13, elapsed: 00:00:02
Generator: starting at 2016-05-27 01:53:13
Generator: Selecting best-scoring urls due for fetch.
Generator: filtering: true
Generator: normalizing: true
Generator: topN: 5
Generator: jobtracker is 'local', generating exactly one partition.
Generator: Partitioning selected urls for politeness.
Generator: segment: crawl/segments/20160527015316
Generator: finished at 2016-05-27 01:53:17, elapsed: 00:00:03
Fetcher: Your 'http.agent.name' value should be listed first in 'http.robots.agents' property.
Fetcher: starting at 2016-05-27 01:53:17
Fetcher: segment: crawl/segments/20160527015316
Using queue mode : byHost
Fetcher: threads: 10
Fetcher: time-out divisor: 2
QueueFeeder finished: total 5 records + hit by time limit :0
Using queue mode : byHost
Using queue mode : byHost
fetching http://nutch.apache.org/apidocs/apidocs-1.1/index.html (queue crawl delay=5000ms)
Using queue mode : byHost
Using queue mode : byHost
Using queue mode : byHost
Using queue mode : byHost
Using queue mode : byHost
Using queue mode : byHost
Using queue mode : byHost
Using queue mode : byHost
Fetcher: throughput threshold: -1
Fetcher: throughput threshold retries: 5
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=4
* queue: http://nutch.apache.org
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1464328403072
  now           = 1464328398602
  0. http://nutch.apache.org/apidocs/apidocs-1.11/index.html
  1. http://nutch.apache.org/apidocs/apidocs-1.3/index.html
  2. http://nutch.apache.org/apidocs/apidocs-1.10/index.html
  3. http://nutch.apache.org/apidocs/apidocs-1.2/index.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=4
* queue: http://nutch.apache.org
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1464328403072
  now           = 1464328399616
  0. http://nutch.apache.org/apidocs/apidocs-1.11/index.html
  1. http://nutch.apache.org/apidocs/apidocs-1.3/index.html
  2. http://nutch.apache.org/apidocs/apidocs-1.10/index.html
  3. http://nutch.apache.org/apidocs/apidocs-1.2/index.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=4
* queue: http://nutch.apache.org
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1464328403072
  now           = 1464328400630
  0. http://nutch.apache.org/apidocs/apidocs-1.11/index.html
  1. http://nutch.apache.org/apidocs/apidocs-1.3/index.html
  2. http://nutch.apache.org/apidocs/apidocs-1.10/index.html
  3. http://nutch.apache.org/apidocs/apidocs-1.2/index.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=4
* queue: http://nutch.apache.org
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
```

c. Notch can successfully lunch a web crawler task. [1.00pt] (see Appendix D and E for output results via Dropbox Link)
Appendix D Link : https://www.dropbox.com/s/jdue95bx7gsoxrc/Appendix%20D%20-%20nutch.apache.org-report-23.txt?dl=0
Appendix E Link : https://www.dropbox.com/s/d59sfblv8dv1gyv/Appendix%20E%20-%20nutch.apache.org-dump-23.txt?dl=0



```
crawl started in: crawl
rootUrlDir = urls
threads = 10
depth = 3
solrUrl=null
topN = 5
Injector: starting at 2016-05-27 01:53:11
Injector: crawlDb: crawl/crawldb
Injector: urlDir: urls
Injector: Converting injected urls to crawl db entries.
Injector: total number of urls rejected by filters: 0
Injector: total number of urls injected after normalization and filtering: 1
Injector: Merging injected urls into crawl db.
Injector: finished at 2016-05-27 01:53:13, elapsed: 00:00:02
Generator: starting at 2016-05-27 01:53:13
Generator: Selecting best-scoring urls due for fetch.
Generator: filtering: true
Generator: normalizing: true
Generator: topN: 5
Generator: jobtracker is 'local', generating exactly one partition.
Generator: Partitioning selected urls for politeness.
Generator: segment: crawl/segments/20160527015316
Generator: finished at 2016-05-27 01:53:17, elapsed: 00:00:03
Fetcher: Your 'http.agent.name' value should be listed first in 'http.robots.agents' property.
Fetcher: starting at 2016-05-27 01:53:17
Fetcher: segment: crawl/segments/20160527015316
Using queue mode : byHost
Fetcher: threads: 10
Fetcher: time-out divisor: 2
QueueFeeder finished: total 5 records + hit by time limit :0
Using queue mode : byHost
Using queue mode : byHost
fetching http://nutch.apache.org/apidocs/apidocs-1.1/index.html (queue crawl delay=5000ms)
Using queue mode : byHost
Using queue mode : byHost
Using queue mode : byHost
Using queue mode : byHost
Using queue mode : byHost
Using queue mode : byHost
Using queue mode : byHost
Using queue mode : byHost
Fetcher: throughput threshold: -1
Fetcher: throughput threshold retries: 5
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=4
* queue: http://nutch.apache.org
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1464328403072
  now           = 1464328398602
  0. http://nutch.apache.org/apidocs/apidocs-1.11/index.html
  1. http://nutch.apache.org/apidocs/apidocs-1.3/index.html
  2. http://nutch.apache.org/apidocs/apidocs-1.10/index.html
  3. http://nutch.apache.org/apidocs/apidocs-1.2/index.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=4
* queue: http://nutch.apache.org
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1464328403072
  now           = 1464328399616
  0. http://nutch.apache.org/apidocs/apidocs-1.11/index.html
  1. http://nutch.apache.org/apidocs/apidocs-1.3/index.html
  2. http://nutch.apache.org/apidocs/apidocs-1.10/index.html
  3. http://nutch.apache.org/apidocs/apidocs-1.2/index.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=4
* queue: http://nutch.apache.org
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1464328403072
  now           = 1464328400630
  0. http://nutch.apache.org/apidocs/apidocs-1.11/index.html
  1. http://nutch.apache.org/apidocs/apidocs-1.3/index.html
  2. http://nutch.apache.org/apidocs/apidocs-1.10/index.html
  3. http://nutch.apache.org/apidocs/apidocs-1.2/index.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=4
* queue: http://nutch.apache.org
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
```



```
* queue: http://nutch.apache.org
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1464328477983
  now           = 1464328474075
  0. http://nutch.apache.org/apidocs/apidocs-2.2/index.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=1
* queue: http://nutch.apache.org
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1464328477983
  now           = 1464328475089
  0. http://nutch.apache.org/apidocs/apidocs-2.2/index.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=1
* queue: http://nutch.apache.org
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1464328477983
  now           = 1464328476103
  0. http://nutch.apache.org/apidocs/apidocs-2.2/index.html
-activeThreads=10, spinWaiting=10, fetchQueues.totalSize=1
* queue: http://nutch.apache.org
  maxThreads    = 1
  inProgress    = 0
  crawlDelay    = 5000
  minCrawlDelay = 0
  nextFetchTime = 1464328477983
  now           = 1464328477117
  0. http://nutch.apache.org/apidocs/apidocs-2.2/index.html
fetching http://nutch.apache.org/apidocs/apidocs-2.2/index.html (queue crawl delay=5000ms)
-activeThreads=10, spinWaiting=9, fetchQueues.totalSize=0
-finishing thread FetcherThread, activeThreads=9
-finishing thread FetcherThread, activeThreads=7
-finishing thread FetcherThread, activeThreads=7
-finishing thread FetcherThread, activeThreads=6
-finishing thread FetcherThread, activeThreads=5
-finishing thread FetcherThread, activeThreads=4
-finishing thread FetcherThread, activeThreads=3
-finishing thread FetcherThread, activeThreads=2
-finishing thread FetcherThread, activeThreads=1
-finishing thread FetcherThread, activeThreads=0
-activeThreads=0, spinWaiting=0, fetchQueues.totalSize=0
-activeThreads=0
Fetcher: finished at 2016-05-27 01:54:40, elapsed: 00:00:23
ParseSegment: starting at 2016-05-27 01:54:40
ParseSegment: segment: crawl/segments/20160527015415
Parsed (0ms):http://nutch.apache.org/apidocs/apidocs-1.8/index.html
Parsed (0ms):http://nutch.apache.org/apidocs/apidocs-1.9/index.html
Parsed (0ms):http://nutch.apache.org/apidocs/apidocs-2.2.1/index.html
Parsed (0ms):http://nutch.apache.org/apidocs/apidocs-2.2/index.html
Parsed (0ms):http://nutch.apache.org/apidocs/apidocs-2.3/index.html
ParseSegment: finished at 2016-05-27 01:54:41, elapsed: 00:00:01
CrawlDb update: starting at 2016-05-27 01:54:41
CrawlDb update: db: crawl/crawldb
CrawlDb update: segments: [crawl/segments/20160527015415]
CrawlDb update: additions allowed: true
CrawlDb update: URL normalizing: true
CrawlDb update: URL filtering: true
CrawlDb update: 404 purging: false
CrawlDb update: Merging segment data into db.
CrawlDb update: finished at 2016-05-27 01:54:42, elapsed: 00:00:01
LinkDb: starting at 2016-05-27 01:54:42
LinkDb: linkdb: crawl/linkdb
LinkDb: URL normalize: true
LinkDb: URL filter: true
LinkDb: internal links will be ignored.
LinkDb: adding segment: file:/C:/Cygwin/home/Eclipse/apache-nutch-1.7/crawl/segments/20160527015116
LinkDb: adding segment: file:/C:/Cygwin/home/Eclipse/apache-nutch-1.7/crawl/segments/20160527015124
LinkDb: adding segment: file:/C:/Cygwin/home/Eclipse/apache-nutch-1.7/crawl/segments/20160527015154
LinkDb: adding segment: file:/C:/Cygwin/home/Eclipse/apache-nutch-1.7/crawl/segments/20160527015316
LinkDb: adding segment: file:/C:/Cygwin/home/Eclipse/apache-nutch-1.7/crawl/segments/20160527015345
LinkDb: adding segment: file:/C:/Cygwin/home/Eclipse/apache-nutch-1.7/crawl/segments/20160527015415
LinkDb: merging with existing linkdb: crawl/linkdb
LinkDb: finished at 2016-05-27 01:54:45, elapsed: 00:00:02
crawl finished: crawl

Eclipse@ULTRABOOK ~/apache-nutch-1.7
$ |
```

Please provide a seed URL (such as www.amazon.com, www.mit.edu, www.yelp.com) to collect at least 1000 web pages online. Please submit the combined file (the original downloaded files from the web servers) in your report [1.00pt].

File too big to fit into report causing word to go into non-responding mode. Report contains 997 pages and 870401 output lines
Search was initiated with command bin/nutch crawl urls –dir crawl –depth 4 –topN 1000
(see Appendix F for output results or follow Dropbox link below)
Link : https://www.dropbox.com/s/dqan9vimenn0s7i/Appendix%20F%20-%20www.encyclopedia.com-dump-0997.txt?dl=0

I also run another report that contained 1788 pages and 1162480 output lines. The file was even bigger than one before.
and used command bin/nutch crawl urls –dir crawl –depth 20 –topN 100
(see Appendix G for output results or follow the Dropbox link below)
Link : https://www.dropbox.com/s/czewebkdv9ypgad/Appendix%20G%20-%20www.encyclopedia.com-dump-1788.txt?dl=0

Please explain the meaning of the Nutch parameters, and change at least three parameters to design three crawling tasks. Please report and explain the crawling results for each parameter setting [0.75pt].

Parameter –dir specifies the directory that segments and crawldb folders will be saved into. If directory doesn't exist then program will create a new directory and if the directory exists then it will use the that directory. It is important however that the directory is empty as if there are segments from prior crawl that will mix later when doing a merge of all the segments.

Parameter –depth specifies how deep to go into webpage outlinks when crawling. For example if user sets the parameter as –depth 1 then Nutch will only index first level. If parameter is set to –depth 2 then Nutch will follow up with number of outlinks. Nutch default value is 5

Parameter –threads enables to choose the user how many threads Nutch should use when crawling.

Parameter –topN defines maximum amount of outlink Nutch will obtain from one page. It determines how many links from each page go into frontier.

Runtime examples:

```
$ bin/nutch crawl urls -dir crawl1 -depth 1 -topN 10
$ bin/nutch mergesegs crawl1/merged crawl1/segments/*
$ bin/nutch readseg -dump crawl1/merged/* Output
```

Produced a combined file into Output folder that contained 1 page. Here we limited the crawl to seed URL as depth was only 1. Depth of 1 created only 1 segment.

```
$ bin/nutch crawl urls -dir crawl2 -depth 10 -topN 1
$ bin/nutch mergesegs crawl2/merged crawl2/segments/*
$ bin/nutch readseg -dump crawl2/merged/* Output
```

Produced a combined file into Output folder that contained 10 pages. Here we limited the each page fetched to only add 1 hyperlink to the frontier. Depth of 10 created 10 segments.

```
$ bin/nutch crawl urls -dir crawl3 -depth 10 -topN 10
$ bin/nutch mergesegs crawl3/merged crawl3/segments/* */
$ bin/nutch readseg -dump crawl3/merged/* */ CombinedDump
```

Produced a combined file into Output folder that contained 91 pages. Here we limited the each page fetched to 10 hyperlink to the frontier and frontier was fully rerun 10 times. By this we should have received maximum 100 pages fetched. Depth of 10 created 10 segments.