Thilina Buddhika

March 2, 2015
Agenda

- Course logistics
- Quiz 5 review
- ASG 02 - FAQ
Course Logistics

1. HW1-PC grades are released. To check the feedback, 
   \texttt{\sim cs455/bin/peek HW1-PC feedback.txt}

2. HW1-WC grades will be released during this week

3. Collect previous graded quizzes.
1. One way of coping with thread-safety issues for a data structure that needs to be returned from a method is to perform a deep-copy on that structure and return a reference to the copy. [True/False]

2. Suppose the TCPReceiver class from your first programming assignment implemented the Runnable interface. Is the following constructor safe? [True/False]

```java
public TCPReceiver(Socket socket){
    this.socket = socket;
    new Thread(this).start();
}
```
The snippet below, vector is an instance of java.util.Vector. The snippet is thread-safe even when vector is accessed by multiple threads concurrently since all methods of the class java.util.Vector are synchronized. [True/False]

```java
for (int i=0; i < vector.size(); i++) {
    doSomething(vector.get(i));
}
```
4 Synchronized collections (instantiated by Collections.synchronizedX() factory method) provides thread-safe compound operations. [True/False]

5 Can this snippet result in a ConcurrentModificationException? [True/False]

```
List<Widget> widgetList = Collections.synchronizedList(
    new ArrayList<Widget>());
...

for (Widget w: widgetList)
    doSomething(w);
```
6. Lock striping a data structure increases the number of threads that can be active in that data structure concurrently. [True/False]

7. For a lock-striped data structure, locking the data structure for exclusive access involves the acquisition of just 1 lock. [True/False]

8. ConcurrentHashMap is an example for a data structure, which provides higher concurrency for critical operations at the expense of strong consistency for less critical operations. [True/False]
9. Synchronizers are objects that coordinate the flow of other threads based on the internal state of that synchronizer. [True/False]

10. Once a Latch reaches its terminal state, it can be reset. [True/False]
Handling Request Redirects

```java
public String resolveRedirects(String url) throws IOException {
    HttpURLConnection con = (HttpURLConnection)(new URL(url).openConnection());
    con.setInstanceFollowRedirects(false);
    con.connect();
    int responseCode = con.getResponseCode();
    if(responseCode == 301){
        return con.getHeaderField( "Location" );
    } else {
        return url;
    }
}
```
Resolving relative URLs

```java
if(!new URI(outGoingUrl).isAbsolute()){
    URI resolvedUrl = new URI(currentPageUrl).resolve(outGoingUrl);
    System.out.println("Resolved URL: " + resolvedUrl.toString());
}
```
Output directory with the username is conflicting with the directory created for Hadoop.

- Use cs455-{username} instead of just {username}. For instance \\tmp\cs455-jdoe\\www.bmb.colostate.edu

When a task is handed off to another crawler, it is possible that the corresponding URL is already crawled possibly at a different recursion level. Is it required to crawl it again?
No, consider it as a duplicate task.

Can we use ConcurrentHashMap in our implementations?
You are not allowed to use any classes from java.util.concurrent package.
Contents of the `node` directory follows a flat structure.

```
|\tmp\cs455-jdoo\www.bmb.colostate.edu
 |--nodes
  |  |--www.bmb.colostate.edu
  |     |--in
  |     |--out
  |
  |--news
  |  |--in
  |  |--out
  |
  |--news-previous
  |  |--in
  |  |--out
  |
  |--news-latest
  |  |--in
  |  |--out
```
How do I make sure a particular page URL belongs to a particular domain?

```java
public static boolean checkDomain(String pageUrl, String rootUrl) throws MalformedURLException {
    return new URL(pageUrl).getHost().equals(new URL(rootUrl).getHost());
}
```

This does not work with the Psychology department web site (http://www.colostate.edu/Depts/Psychology/).

It is fine to handle this particular web site using a custom logic.
Am I allowed to use threads outside of the thread pool to handle communicating with other machines? Use separate threads for TCPServerSocket and the TCPReceivers. Task handoff should be performed using the worker threads.

What is the recursion depth for root url? Consider it as 1.
Wrap Up

Questions?