**Федеральное агентство связи**

**Ордена Трудового Красного Знамени**

**Федеральное государственное бюджетное образовательное учреждение высшего образования**

**«Московский технический университет связи и информатики»**

Кафедра «Информатики»

**Отчет по лабораторной работе №7**

Выполнил: студент группы БВТ1901

Кускова А. Е.

Руководитель:

Мосева М. С.

Москва 2020

**Цель работы:**

Реализовать элементарный веб-сканер, который будет искать новые URL-адреса (местоположения веб-страниц) на каждой странице, собирать их и выводить в конце работы программы.

**Задание:**

1. Программа должна принимать в командной строке два параметра:

* Строку, которая представляет собой URL-адрес, с которого можно начать просмотр страницы.
* Положительное целое число, которое является максимальной глубиной поиска.

1. Программа должна хранить URL-адрес в виде строки вместе с его

глубиной. Необходимо создать класс для представления пар [URL, depth].

1. Программа должна подключиться к указанному сайту в URL-адресе на порт 80 с использованием сокета и запросить указанную веб-страницу.
2. Программа должна проанализировать возвращаемый текст, построчно для любых подстрок, имеющих формат: <a href="[любой\_URL-адрес\_начинающийся\_с\_http://]">. Найденные URL-адреса должны быть сохранены в паре с новым значением глубины в LinkedList (URL, depth).
3. Программа должна повторять шаги с 3 по 6 для каждого нового URL-адреса, если глубина, соответствующая URL-адресу, меньше максимальной.
4. Программа должна вывести все посещенные URL-страницы вместе с их глубиной поиска.

**Решение:**

1. Код для класса Crawler:

import java.io.\*;  
import java.net.Socket;  
import java.net.URL;  
import java.util.Arrays;  
import java.util.LinkedList;  
import java.util.regex.Matcher;  
import java.util.regex.Pattern;  
  
public class Crawler {  
 public LinkedList<UrlDepthPair> proccessed;  
  
 public Crawler() {  
 proccessed = new LinkedList<>();  
 }  
  
 public void startParsing(URL baseUrl, int maxDepth, int currentDepth) {  
 if (currentDepth > maxDepth) return;  
 LinkedList<UrlDepthPair> links = *getAllLinks*(baseUrl, currentDepth);  
 for (UrlDepthPair link: links) {  
 startParsing(link.realUrl, maxDepth, currentDepth + 1);  
 }  
  
 proccessed.addAll(links);  
 }  
  
 private static LinkedList<UrlDepthPair> getAllLinks(URL url, int depth) {  
 try {  
 LinkedList<UrlDepthPair> links = new LinkedList<>();  
  
 int port = 80;  
 String hostname = url.getHost();  
 Socket socket = new Socket(hostname, port);  
 socket.setSoTimeout(3000);  
 OutputStream outStream = socket.getOutputStream();  
 PrintWriter writer = new PrintWriter(outStream, true);  
  
 if (url.getPath().length() == 0) {  
 writer.println("GET / HTTP/1.1");  
 writer.println("Host: " + hostname);  
 writer.println("Accept: text/html");  
 writer.println("Accept-Language: en,en-US;q=0.9,ru;q=0.8");  
 writer.println("Connection: close");  
 writer.println();  
 }  
 else {  
 writer.println("GET " + url.getPath() + " HTTP/1.1");  
 writer.println("Host: " + hostname);  
 writer.println("Accept: text/html");  
 writer.println("Accept-Language: en,en-US;q=0.9,ru;q=0.8");  
 writer.println("Connection: close");  
 writer.println();  
 }  
  
 InputStream input = socket.getInputStream();  
 BufferedReader reader = new BufferedReader(new

InputStreamReader(input));  
  
 String htmlLine;

Pattern patternURL = Pattern.*compile*("(href=\"http|href=\"" +

"https)://([\\w\_-]+(?:(?:\\.[\\w\_-]+)+))([\\w.,@?^=%&:/~+#- " +

"]\*[\\w@?^=%&/~+#-])?"  
 );  
  
 while ((htmlLine = reader.readLine()) != null) {  
 Matcher matcherURL = patternURL.matcher(htmlLine);  
 while (matcherURL.find()) {  
 String link = htmlLine.substring(matcherURL.start() + 6,  
 matcherURL.end());  
  
 links.add(new UrlDepthPair(link, depth));  
}  
 }  
 socket.close();  
  
 return links;  
 } catch (Exception e) {  
 System.*out*.print(e.getMessage());  
 System.*out*.print(Arrays.*toString*(e.getStackTrace()));  
  
 return new LinkedList<>();  
 }  
 }  
}

1. Код для класса UrlDepthPair:

import java.net.MalformedURLException;  
import java.net.URL;  
  
public class UrlDepthPair {  
 public String url;  
 public int depth;  
 public URL realUrl;  
  
 public UrlDepthPair (String url, int depth) throws MalformedURLException {  
 realUrl = new URL(url);  
 this.url = url;  
 this.depth = depth;  
 }  
}

1. Код для main:

import java.net.MalformedURLException;  
import java.net.URL;  
import java.util.LinkedList;  
  
public class main {  
 public static void main(String[] args) throws MalformedURLException {  
 URL url = new URL("http://mtuci.ru/");  
 var parser = new Crawler();  
 parser.startParsing(url, 2, 0);  
 LinkedList<UrlDepthPair> result = parser.proccessed;  
 result.forEach(urlDepthPair -> {  
 System.*out*.println(  
 "URL is " + urlDepthPair.url +  
 " | depth: " + urlDepthPair.depth);  
 });  
 }  
}

**Результат работы программы**

На рисунке 1 изображен результат работы сканера:

**Изображение выглядит как текст

Автоматически созданное описание**

**Рисунок 1** – Работа сканера.

**Вывод:**

Реализовали элементарный веб-сканер, который автоматически загружает веб-страницы из Интернета, ищет новые ссылки на этих страницах и повторяет их.