

```

1 package edu.upvehu.gbg.scrapers;
2
3 import java.io.*;
4 import java.net.MalformedURLException;
5 import java.net.URL;
6 import java.util.ArrayList;
7 import java.util.List;
8 import java.util.regex.Matcher;
9 import java.util.regex.Pattern;
10
11
12 //Ejemplo de extracción de información de una página web.
13 //Se recogen los datos de una tabla en una lista de arrays de Strings
14
15 public class Scraper {
16
17     private static final String URL="http://www.bolsamadrid.es/esp/asp/Mercados/Precios.aspx?indice=ESI100000000";
18     private static final int NUM_COLUMNAS=9;
19
20     //Patrón para extraer la tabla que nos interesa y no otras.
21     private static final String TABLE_CONTENT="ct100_Contenido_tblAcciones.*?>(.*?)</table>";
22     //Extraeremos las filas de la tabla
23     private static final String ROW_PATTERN="<tr.*?>(.*?)</tr>";
24     //Y de ellas el contenido de cada celda, descartando los elementos <a> (links) aplicados a los nombres de empresas
25     private static final String DATA_PATTERN="<td.*?>(?:.*?<a.*?>)?(.*?)?(?:</a>.*?)?</td>";
26
27
28     public static void main(String[] args) throws MalformedURLException, IOException {
29         List<String[]> empresas=new ArrayList<>();
30
31         //La página web a un buffer
32         BufferedReader br=new BufferedReader(new InputStreamReader(new URL(URL).openStream()));
33         StringBuffer sb=new StringBuffer();
34         for(String linea;(linea=br.readLine())!=null;) sb.append(linea);
35         br.close();
36
37         //Cogemos el contenido de la tabla
38         Matcher tableContentMatcher=Pattern.compile(TABLE_CONTENT).matcher(sb);
39         tableContentMatcher.find();
40         //Para todas las filas (que salen de procesar el contenido de la tabla)
41         for (Matcher rowMatcher=Pattern.compile(ROW_PATTERN).matcher(tableContentMatcher.group(1));rowMatcher.find();){
42             //Cogemos el contenido de cada celda en un array de Strings
43             try{
44                 Matcher dataMatcher=Pattern.compile(DATA_PATTERN).matcher(rowMatcher.group(1));
45                 String[] empresa=new String[NUM_COLUMNAS];
46                 for (int i=0;i<empresa.length;i++) {dataMatcher.find();empresa[i]=dataMatcher.group(1);}
47                 empresas.add(empresa);
48             } catch(java.lang.IllegalStateException ex) {System.out.println("Esta fila parece no ser una empresa: "+rowMatcher.group(1));}
49         }
50
51         //Mostramos el resultado
52         for (String[] s:empresas){
53             for (String s2:s) System.out.print(s2+" ");
54             System.out.println("");
55         }
56     }
57
58 }

```