

Analiza zawartości stron internetowych

Kamil Chełminiak

Maj 2020

- org.jsoup.JSoup
- javax.mail

HTML jest podstawowym formatem stosowanym w dokumentach sieci WWW. Jest on przypadkiem SGMLa (ang. Standard Generalized Markup Language). Podstawowym założeniem SGML, a w konsekwencji HTML jest przedstawienie znaczenia, a nie wyglądu informacji.

- Znacznik- specjalny tekst umieszczony w ostrych nawiasach, będący częścią składni języka HTML, pozwalający na sterowanie jej wyglądem
- Atrybut- wartość, powiązana z elementem, składająca się z nazwy i wartości tekstowej

HTML w Javie

Wykorzystanie biblioteki JSoup w praktyce

```
public String information(String phrase){  
    addresses.add("https://wp.pl");  
    addresses.add("https://tvn24.pl");  
    addresses.add("https://gov.pl");  
    try{  
        for (String url: addresses) {  
            //Connect to URL address  
            Connection connect = Jsoup.connect(url);  
            //Start each block with URL address of the currently browsed www site  
            site.append("<h1>").append(url).append("</h1>");  
            //Get www site as Document object  
            Document document = connect.get();  
        }  
    }  
}
```

HTML w Javie

Wykorzystanie biblioteki JSoup w praktyce

```
<div class="sc-Inch14j-0 ymfzd sc-7ruplj-0 LiPgo">
  
<div class="sc-1430vqz-0 TvQvX sc-1ed14w-0-1 eitCIE">
  <div></div>
  <div class="zviox-0 ka0aUk">
    Był w lesie bez maseczki. Jest reakcja prezydenta Dudy
  </div>
</div></a>
</div>
</div>
<div data-st-area="Events-mozaika" class="sc-1lvjlv-2 sc-1lvjlv-3 c0sv0e">
  <div class="sc-1ed14w-0 bthvOV">
    <a data-st-clk="{&quot;teaserId&quot;:&quot;6514002001152129&quot;,&quot;sgcat&quot;:200,&quot;sgcatid&quot;:3}" id="6514002001152129" label="" href="https://sportowefakty.wp.pl/pilka"
    <div class="sc-Inch14j-0 ymfzd sc-7ruplj-0 LiPgo">
      
    <div class="sc-1430vqz-0 TvQvX sc-1ed14w-0-1 eitCIE">
      <div></div>
      <div class="zviox-0 ka0aUk">
        Wszyscy zazdroszczą Czechom. A on mówi, gdzie leży ich przewaga
      </div>
    </div></a>
  </div>
```

HTML w Javie

Wykorzystanie biblioteki JSoup w praktyce

```
"C:\Program Files\jdk-8.0_221\bin\java.exe" ...  
<a id="6513877163837569" href="https://wiadomosci.wp.pl/the-new-york-times-w-holdzie-dla-ofiar-koronawirusa-pierwsza-strona-epelajona-nazwiskami-651386598852825?r" data-st-clk="{&quot;teaserId&quot;  
<div class="zb&quot;1 QwkPL"> <!-- -->"NYT" w holdzie ofiarom koronawirusa  
</div></a>  
<a id="6514887611065985" data-st-clk="{&quot;teaserId&quot;:&quot;6514887611065985&quot;;&quot;sgcat&quot;:10,&quot;sgcatid&quot;:5}" data-st-area="Sport-mozaika" style="margin-bottom:0" href="ht  
<div class="sc-1nch14j-0 ymFzd sc-7rupij-0 LiPgo">  
  
<div class="sc-1430vqz-0 TvQvX">  
<div class="zfmlaB-3 jt1bNt">  
<div></div>  
</div>  
<div class="zviqxx-0 kaOaUk">  
Coraz większe problemy Barcelony przez koronawirusa  
</div>  
</div></a>  
<a id="6513899834607233" data-st-clk="{&quot;teaserId&quot;:&quot;6513899834607233&quot;;&quot;sgcat&quot;:10,&quot;sgcatid&quot;:4}" data-st-area="Sport-mozaika" style="margin-bottom:20px" href=  
<div class="sc-1nch14j-0 ymFzd sc-7rupij-0 LiPgo"></div>  
<div class="sc-1430vqz-0 TvQvX">  
<div class="zfmlaB-3 jt1bNt">  
<div></div>  
</div>  
<div class="zviqxx-0 kaOaUk">  
Afera w Anglii. "Wstrzyknąłem im koronawirusa"  
</div>  
</div></a>  
<a id="oce-4592919" title="przejdź do Strzygli klientów, mając koronawirusa. Narazili na zakażenie nawet 140 osób " class="default-teaser__link" href="https://tvn24.pl/swiat/koronawirus-w-usa-Nisz  
<figure class="default-teaser__container">  
<figcaption class="default-teaser__description">  
-----<div class="main__caption">
```

HTML w Javie

Wykorzystanie biblioteki JSoup w praktyce

```
if (elem.text().toLowerCase().contains(phrase.toLowerCase())) {  
    //Handle relative addresses  
    site.append("<div class = 'container'><a href='").append(elem.attr( attributeKey: "href"))  
        .append("'>").append("<h2>").append(elem.text()).append("</h2></a></div>");  
}
```


HTML w Javie

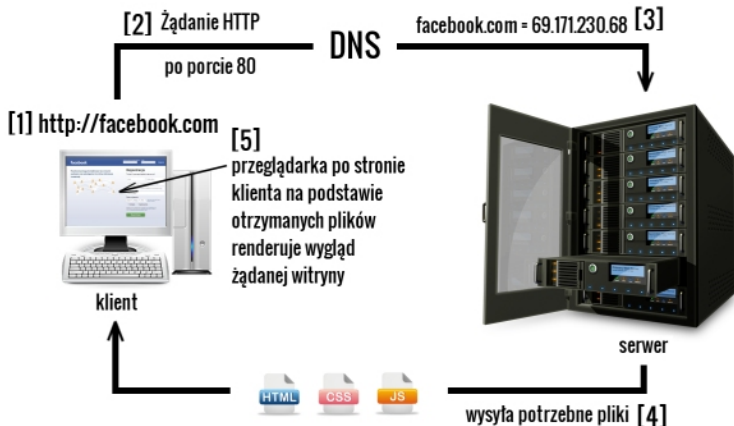
Wykorzystanie biblioteki JSoup w praktyce

```
for (Element img : allImg.subList(0, Math.min(amount, allImg.size()))) {  
    //Handle relative addresses  
    if (img.attr(attributeKey: "src").startsWith("/")){  
        site.append("<div class = 'container'><img src='").append(url);  
        site.append(img.attr(attributeKey: "src")).append("></div>");  
    } else {  
        site.append("<div class = 'container'><img src='").append(img.attr(attributeKey: "src"));  
        site.append("></div>");  
    }  
}
```

String vs StringBuilder

- String jest niemodyfikowalny (immutable)
- Oznacza to, że po wykonaniu kodu dopisującego tekst do utworzonego Stringa tworzony jest nowy obiekt String
- StringBuilder jest modyfikowalny (mutable)
- StringBuilder pozwala na dynamiczną modyfikację ciągów znaków

Komunikacja Klient- Server



Source: <https://pasja-informatyki.pl/programowanie-webowe/architektura-klient-serwer/>

Sockety w Javie

Czyli przedsmak tego, co czeka Was na 6 semestrze

```
timer.start();  
try {  
    //Create server socket  
    serverSocket = new ServerSocket(port);  
} catch (IOException e) {  
    System.err.println("Could not listen on a port: " + port);  
}
```

Sockety w Javie

Co dalej?

- Zaakceptuj przychodzące od klientów połączenia
- Utwórz strumień danych: wejściowych i wyjściowych
- Przy tworzeniu strumieni jako argumenty podaj gettery wybranego typu strumienia, wywołane na sockecie klienta
- Uruchom kod, obsługujący żądania klienta

Sockety w Javie

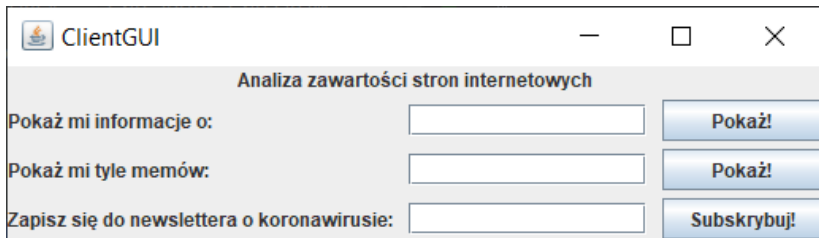
ClientHandler

ClientHandler musi odczytać wysyłane przez klientów żądania.
Warto od razu zdefiniować kodowanie tych żądań.

```
case "subscription":
    String newSubscriber = in.readUTF();
    //Variable for checking existence of client on subscriber list
    Boolean newAddress = true;
    //Check is client appears on subscriber list
    for (String subscriber: server.getCoronavirusSubscribers()){
        if (newSubscriber.equals(subscriber)){
            newAddress = false;
        }
    }
    //If it is truly a new client add him to subscriber list
    if (newAddress) {
        server.addSubscriber(newSubscriber.toString());
    }
    break;
default:
    System.err.println("Something went wrong!");
}
socket.close();
in.close();
out.close();
```

Sockety w Javie

Klient też chce mieć coś do powiedzenia



ClientGUI

Analiza zawartości stron internetowych

Pokaż mi informacje o: **Pokaż!**

Pokaż mi tyle memów: **Pokaż!**

Zapisz się do newslettera o koronawirusie: **Subskrybuj!**

Sockety w Javie

Klient też chce mieć coś do powiedzenia

```
public void connection(String type, String title, String args) {  
    try {  
        //Getting localhost ip  
        host = InetAddress.getByName("localhost");  
  
        //Establish connection with Server  
        socket = new Socket(host, port);  
    }  
}
```


Sockety w Javie

Klient też chce mieć coś do powiedzenia

```
//Sending data to ClientHandler
out.writeUTF(type);
out.writeUTF(args);
if (!type.equals("subscription")) {
    String input = in.readUTF();
```

HyperlinkListener

A niech idzie gdzie chce

```
public void hyperlinkUpdate(HyperlinkEvent evt) {
    //Create hyperlink listener, which opens browser when hyperlink is clicked in JEditorPane
    if (evt.getEventType() == HyperlinkEvent.EventType.ACTIVATED) {
        try {
            openWebpage(evt.getURL());
        } catch (Exception e) {
        }
    }
}

public static boolean openWebpage(URI uri) {
    //Use desktop class to run application from native desktop
    Desktop desktop = Desktop.isDesktopSupported() ? Desktop.getDesktop() : null;
    if (desktop != null && desktop.isSupported(Desktop.Action.BROWSE)) {
        try {
            //Launch default browser and handle and URI to it
            desktop.browse(uri);
            return true;
        } catch (Exception e) {
            e.printStackTrace();
        }
    }
    return false;
}

public static boolean openWebpage(URL url) {
    //Convert URL (Uniform Resource Locator) address to URI (Uniform Resource Identifier)
    try {
        return openWebpage(url.toURI());
    } catch (URISyntaxException e) {
        e.printStackTrace();
    }
}
```

Wysyłanie maili

Czyli nie chcę otwierać przeglądarki, ale muszę go wysłać

```
public class Mailer {  
  
    public void sendEmail(String from, String password, String to, String subject, String text){  
        //Set properties of email message like: protocol and port  
        Properties props = new Properties();  
        props.put("mail.smtp.host", "smtp.gmail.com");  
        props.put("mail.smtp.socketFactory.port", "465");  
        props.put("mail.smtp.socketFactory.class",  
            "javax.net.ssl.SSLSocketFactory");  
        props.put("mail.smtp.auth", "true");  
        props.put("mail.smtp.port", "465");  
        //Create session with smtp server and log in to your account  
        Session session = Session.getDefaultInstance(props,  
            getPasswordAuthentication() → {  
                return new PasswordAuthentication(from, password);  
            });  
        try {  
            //Create email based on session, created earlier  
            MimeMessage message = new MimeMessage(session);
```

Wysyłanie maili

MIME?

MIME

From Wikipedia, the free encyclopedia

For mime as a performing art form, see [Mime artist](#). For the British engineering society, see [Institution of Mechanical Engineers](#). For the type format, see [Media 1](#)

Multipurpose Internet Mail Extensions (MIME) is an [Internet standard](#) that extends the format of [email](#) messages to support text in [character sets](#) other than [ASCII](#), as well as attachments of audio, video, images, and application programs. Message bodies may consist of multiple parts, and header information may be specified in non-ASCII character sets. Email messages with MIME formatting are typically transmitted with standard protocols, such as the [Simple Mail Transfer Protocol](#) (SMTP), the [Post Office Protocol](#) (POP), and the [Internet Message Access Protocol](#) (IMAP).

The MIME standard is specified in a series of [requests for comments](#): [RFC 2045](#), [RFC 2046](#), [RFC 2047](#), [RFC 4288](#), [RFC 4289](#) and [RFC 2049](#). The integration with SMTP email is specified in [RFC 1521](#) and [RFC 1522](#).

Although the MIME formalism was designed mainly for SMTP, its content types are also important in other [communication protocols](#). In the [HyperText Transfer Protocol](#) (HTTP) for the [World Wide Web](#), servers insert a MIME header field at the beginning of any Web transmission. Clients use the [content type](#) or [media type](#) header to select an appropriate viewer application for the type of data indicated. Browsers typically contain GIF and JPEG image viewers.

Source: <https://en.wikipedia.org/wiki/MIME>

Wysyłanie maili

O czym muszę pamiętać?

Niezbędne do wysłania maila są:

- Adres email nadawcy (ustawić)
- Adres email odbiorcy (dodać)
- Temat (ustawić)

Warto ustawić również treść. W końcu klient oczekuje najświeższych wiadomości, a nie pustego spamu.

Wysyłanie maili

Ważna informacja!

W celu prawidłowego działania programu należy udać się pod adres: <https://myaccount.google.com/lesssecureapps> i włączyć na czas testowania kodu dostęp dla mniej bezpiecznych aplikacji.

- "Java. Programowanie sieciowe", Eliotte Rusty Harold
- Oficjalna dokumentacja biblioteki JSoup
- Oficjalna dokumentacja biblioteki javax.mail
- <http://kurshtml.edu.pl>
- <https://www.geeksforgeeks.org/introducing-threads-socket-programming-java/>
- <http://www.w3big.com/pl/java/net-multisoc.html>
- <https://stormit.pl/stringbuilder/>