JSoup

Cargar una página web

```java
import java.io.IOException;
import org.jsoup.Jsoup;
import org.jsoup.nodes.Document;;

public class ExamplesJsoup {

    public static void loadDoc(String url) {
        try {
            Document doc = Jsoup.connect("http://eldiario.es/").get();
        } catch (IOException e) {
            // TODO Auto-generated catch block
            e.printStackTrace();
        }

    }

}
```

Como recupera el título de una página web

```
try {
    Document doc = Jsoup.connect("http://eldiario.es/").get();
    System.out.println(doc.title());
} catch (IOException e) {
    e.printStackTrace();
}
```

Puedes usar métodos DOM para navegar por el documento

```
try {
    Document doc = Jsoup.connect("http://eldiario.es/").get();
    System.out.println(doc.title());
    Element content = doc.getElementById("main");
    Elements links = content.getElementsByTag("a");
    for (Element link : links) {
      String linkText = link.text();
      String linkHref = link.attr("href");
      System.out.println(linkText+": "+linkHref);
    }

} catch (IOException e) {
    e.printStackTrace();
}
```

También puedes usar select...

```
try {
    Document doc = Jsoup.connect("http://eldiario.es/").get();
    System.out.println(doc.title());

    Elements links = doc.select("a[href]");
    for (Element link : links) {
      String linkText = link.text();
      String linkHref = link.attr("href");
      System.out.println(linkText+": "+linkHref);
    }

} catch (IOException e) {
    e.printStackTrace();
}
```

También puedes usar select...

```
try {
    Document doc = Jsoup.connect("http://eldiario.es/").get();
    System.out.println(doc.title());

    Elements links = doc.select("a[href*=Puigdemont]");
    for (Element link : links) {
      String linkText = link.text();
      String linkHref = link.attr("href");
      System.out.println(linkText+": "+linkHref);
    }

} catch (IOException e) {
    e.printStackTrace();
}
```

Método select – opciones

- **tagname**: buscamos elementos por su etiqueta, (por ejemplo *a, img, p* o *table)*.

```
try {
    Document doc = Jsoup.connect("http://eldiario.es/").get();
    System.out.println(doc.title());

    Elements links = doc.select("img");
    for (Element link : links) {
      String src= link.attr("src");
      System.out.println("src: "+src);
    }

} catch (IOException e) {
    e.printStackTrace();
}
```

- *El código anterior va a devolver todas las imágenes.*

Método select – opciones

- **tagname**: buscamos elementos por su etiqueta, (por ejemplo *a*, *img, p o table).*
- ***ns|tag:*** *Si tenemos varios namespace (ns), podemos especificar en que namespace queremos buscar.*

Método select – opciones

- **#id**: podemos buscar por el id de un elemento

```
try {
    Document doc = Jsoup.connect("http://eldiario.es/").get();
    System.out.println(doc.title());
    //buscar por el id de un elemento
    Elements links = doc.select("#cintillo-internacional");
    for (Element link : links) {
      String html= link.outerHtml();
      System.out.println(html);
    }

} catch (IOException e) {
    e.printStackTrace();
}
```

Método select – opciones

- **tag.class**: podemos seleccionar los elementos indicando su clase

```
try {
    Document doc = Jsoup.connect("http://eldiario.es/").get();
    System.out.println(doc.title());
    //buscar elementos por su nombre de clase
    Elements links = doc.select("div.mg");
    for (Element link : links) {
      String html= link.outerHtml();
      System.out.println(html);
    }

} catch (IOException e) {
    e.printStackTrace();
}
```

Método select – opciones

- **[attribute]**: podemos seleccionar los elementos que contengan un determinado atributo (por ejemplo, src, height, href)

```
//buscar elementos que tengan el atributo href
Elements links = doc.select("[href]");
for (Element link : links) {
  String text= link.text();
  String src= link.attr("href");
  System.out.println(text+ ": " + src);
}
```

- *El código anterior devuelve todos los elementos que tengan el atributo href*

Método select – opciones

- **[attribute=value]**: podemos seleccionar los elementos que contengan un determinado atributo (por ejemplo, src, height, href)

```
//buscar elementos que tengan el atributo href
Elements links = doc.select("[width=290]");
for (Element link : links) {
  String html= link.outerHtml();
  System.out.println(html);
}
```

- *El código anterior va a devolver todas las imágenes con ancho 290.*

Método select – opciones

- **[attribute^=value]**: los elementos cuyo valor empieza con
- **[attribute$=value]**: los elementos cuyo valor termina con
- **[attribute*=value]**: los elementos cuyo valor contiene
- **[attr~=regex]:** los elementos que cumplan la expresión regular reges (por ejemplo img[src~=\.(png|jpe?g)])

Intentalo tu mismo

- Muestra los nombres de las imágenes con extensión png.
- Muestra los nombres de las imágenes con extensión jpg.
- Muestra los nombres de todas las imágenes que contienen “Puigdemont” en la página http://www.eldiario.es. \\w
- Muestra los nombres de todas las imágenes que en su propiedad *alt* contengan algún número real (0,25).

Soluciones

Imágenes con extensión png.

```
Elements links = doc.select("img[src$=png]");
for (Element elem: links) {
  String src= elem.attr("src");
  String alt= elem.attr("alt");
  System.out.println(src+": "+alt);
}
```

Imágenes que contengan “Puigdemont” con extensión jpg

```
Elements links = doc.select("img[src~=[\\w\\W]*Puigdemont[\\w\\W]*.jpg]");
for (Element elem: links) {
  String src= elem.attr("src");
  String alt= elem.attr("alt");
  System.out.println(src+": "+alt);
}
```

Soluciones

Imágenes que en su atributo alt contenga algun número decimal

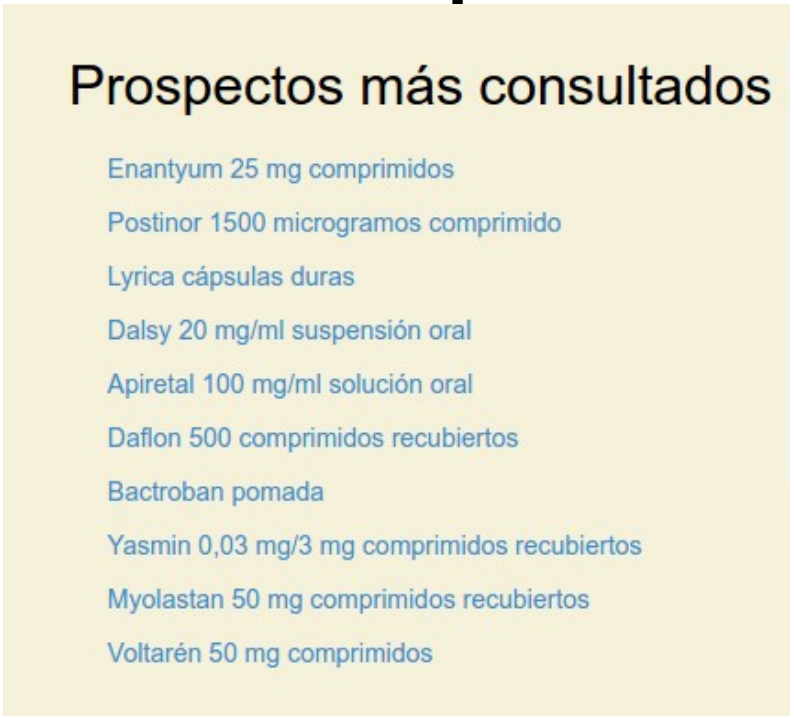
```
Elements links = doc.select("img[alt~=\\d,\\d]");
for (Element elem: links) {
  String src= elem.attr("src");
  String alt= elem.attr("alt");
  System.out.println(src+": "+alt);
}
```

Método select – más opciones

- **parent child**: devuelve todos los elementos child que cuelgan de parent. Por ejemplo, *div p*

- **parent > child**: devuelve todos los elementos child que cuelgan directamente de parent. Por ejemplo, *div > p*

http://www.prospectos.net




```
<div class="col-xs-10 col-xs-push-2 col-sm-6 col-sm-push-0 col-md-6 col-md-push-1 col-lg-6 col-lg-push-2">
    <h2>Prospectos más consultados</h2>
    <ul class="list-unstyled top-prospectos">
        <li><a href="/enantyum_25_mg_comprimidos">Enantyum 25 mg comprimidos</a></li>
        <li><a href="/postinor_1500_microgramos_comprimido">Postinor 1500 microgramos comprimido</a></li>
        <li><a href="/lyrica_capsulas_duras">Lyrica cápsulas duras</a></li>
        <li><a href="/dalsy_suspension_oral">Dalsy 20 mg/ml suspensión oral</a></li>
        <li><a href="/apiretal_solucion_gotas">Apiretal 100 mg/ml solución oral</a></li>
        <li><a href="/daflon_500_comprimidos_recubiertos">Daflon 500 comprimidos recubiertos</a></li>
        <li><a href="/bactroban_pomada">Bactroban pomada</a></li>
        <li><a href="/yasmin_comprimidos_recubiertos">Yasmin 0,03 mg/3 mg comprimidos recubiertos</a></li>
        <li><a href="/myolastan_50_mg_comprimidos_recubiertos">Myolastan 50 mg comprimidos recubiertos</a></li>
        <li><a href="/voltaren_50_mg_comprimidos">Voltarén 50 mg comprimidos</a></li>
    </ul>
</div>
```

http://www.prospectos.net

```
try {
    Document doc = Jsoup.connect("https://www.prospectos.net/").get();
    System.out.println(doc.title());
    Elements links = doc.select("ul>li");
    for (Element elem: links) {
        System.out.println(elem.text());
    }

} catch (IOException e) {
    e.printStackTrace();
}
```

```
    <ul class="list-unstyled top-prospectos">
        <li><a href="/enantyum_25_mg_comprimidos">Enantyum 25 mg comprimidos</a></li>
        <li><a href="/postinor_1500_microgramos_comprimido">Postinor 1500 microgramos comprimido</a></li>
        <li><a href="/lyrica_capsulas_duras">Lyrica cápsulas duras</a></li>
        <li><a href="/dalsy_suspension_oral">Dalsy 20 mg/ml suspensión oral</a></li>
        <li><a href="/apiretal_solucion_gotas">Apiretal 100 mg/ml solución oral</a></li>
        <li><a href="/daflon_500_comprimidos_recubiertos">Daflon 500 comprimidos recubiertos</a></li>
        <li><a href="/bactroban_pomada">Bactroban pomada</a></li>
        <li><a href="/yasmin_comprimidos_recubiertos">Yasmin 0,03 mg/3 mg comprimidos recubiertos</a></li>
        <li><a href="/myolastan_50_mg_comprimidos_recubiertos">Myolastan 50 mg comprimidos recubiertos</a></li>
        <li><a href="/voltaren_50_mg_comprimidos">Voltarén 50 mg comprimidos</a></li>
    </ul>
</div>
```

Método select – más opciones

- **A + B**: devuelve los elementos B que están inmediatamente precedidos por A.
- **A ~ B**: devuelve los elementos B precedidos por A.

  (A y B son hermanos, es decir, están en el mismo nivel de anidamiento)