

Tutorial JSP

Isabel Segura Bedmar,
Grupo LABDA, Departamento de Informática,
Universidad Carlos III de Madrid,
17/10/2016

Web scraping

- Conjunto de técnicas automáticas para recopilar grandes cantidades de datos de diferentes páginas web.
- Primera fase de recolección de datos en Big Data

Qué vamos a hacer hoy?

- Conocer la **librería jsoup** de Java, que nos va a permitir fácilmente recuperar datos de páginas html.
- Es necesario tener conocimiento de java y también de html y css

Cómo importar la librería jsoup

1) **(opción recomendada)** En tu workspace de eclipse, crear un **proyecto maven** y añadir la **dependencia a jsoup** en el fichero **pom.xml** del proyecto:

```
<!--  
https://mvnrepository.com/artifact/org.jsoup/jsoup -->  
<dependency>  
  <groupId>org.jsoup</groupId>  
  <artifactId>jsoup</artifactId>  
  <version>1.9.2</version>  
</dependency>
```

Cómo importar la librería jsoup

2) Descargar la librería de jsoup

<https://jsoup.org/packages/jsoup-1.9.2.jar>

y añadela a tu proyecto. Pasos:

a) Crea una carpeta 'lib' en tu proyecto y guarda la librería en esa carpeta.

b) Víncula la librería a tu proyecto: Build path -> Add Library...

- All Classes
- Packages
- org.jsoup
org.jsoup.examples
org.jsoup.helper
org.jsoup.nodes
org.jsoup.parser
org.jsoup.safety
org.jsoup.select
- All Classes
- Attribute
Attributes
BooleanAttribute
Cleaner
Collector
Comment
Connection
Connection.Base
Connection.KeyVal
Connection.Method
Connection.Request
Connection.Response
DataNode
DataUtil
DescendableLinkedList
Document
Document.OutputSettings
Document.OutputSettings.S
Document.QuirksMode
DocumentType
Element
Elements
Entities

jsoup 1.9.2 API

jsoup: Java HTML parser that makes sense of real-world HTML soup.

See: [Description](#)

Packages

Package	Description
org.jsoup	Contains the main Jsoup class, which provides convenient static access to the jsoup functionality.
org.jsoup.examples	Contains example programs and use of jsoup.
org.jsoup.helper	
org.jsoup.nodes	HTML document structure nodes.
org.jsoup.parser	Contains the HTML parser, tag specifications, and HTML tokeniser.
org.jsoup.safety	Contains the jsoup HTML cleaner, and whitelist definitions.
org.jsoup.select	Packages to support the CSS-style element selector.

jsoup: Java HTML parser that makes sense of real-world HTML soup.

jsoup is a Java library for working with real-world HTML. It provides a very convenient API for extracting and manipulating data, using the best of DOM, CSS, and jquery-like methods.

1º ejercicio: mostrar el título de una página

- Crea una clase **Scraper** y añade un método que reciba como parámetro la url de una página html y muestre por pantalla el título de dicha página.
- Pista: lo primero que debe hacer es conectarte a la ur:

`Document doc= Jsoup.connect(url).get()` (ver javadoc)

Nota: Eclipse nos va a sugerir directamente los paquetes qué debemos importar.

- *¿La clase Document tiene algún método que nos permita obtener el título?*

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

public class Sraper {
    public static final String URL_WIKIE="https://es.wikipedia.org/";

    public static void main(String args[]) throws IOException {
        getTitle(URL_WIKIE);
    }

    /**
     * Este método debería recuperar el título de cualquier página html
     * @param url
     * @throws IOException
     */
    public static void getTitle(String url) throws IOException{
        Document doc = Jsoup.connect(url).get();
        System.out.println(url+":\t"+doc.title());
        System.out.println();
    }
}
```


2º ejercicio: recuperar todos los links

- Añade un método que reciba por parámetro una url y muestre por pantalla todos sus links.

Pista: En el javadoc (<https://jsoup.org/apidocs/>) estudia el **método select()** de la clase *Document*.

```

import java.io.IOException;

import org.jsoup.Jsoup;
import org.jsoup.nodes.Document;
import org.jsoup.nodes.Element;
import org.jsoup.select.Elements;

public class Sraper {
    public static final String URL_WIKIE="https://es.wikipedia.org/";

    public static void main(String args[]) throws IOException {
        getTitle(URL_WIKIE);
        getHref(URL_WIKIE);
    }

    /**
     * Este método debe recuperar todos los a href y mostrar su texto asociado.
     * @param url
     * @throws IOException
     */
    public static void getHref(String url) throws IOException{
        Document doc = Jsoup.connect(url).get();
        System.out.println("a href from:\t"+doc.title());
        Elements lst = doc.select("a[href]"); // a with href
        for (Element elem: lst)
            System.out.println("\t"+elem.text());
    }
}

```

3º ejercicio: recuperar todos las imágenes

- Añade un método que reciba por parámetro una url y muestre por pantalla todos el atributo src (ruta) de todas las imágenes que cumplan:
 - 3.1.- Su extensión es .jpg.*
 - 3.2.- Contiene la cadena 'logo' (uso de expresiones regulares Java)*

```

public class Sraper {
    public static final String URL_WIKIE="https://es.wikipedia.org/";

    public static void main(String args[]) throws IOException {
        getTitle(URL_WIKIE);
        getImgs(URL_WIKIE);
    }
    /**
     * Este método debe recuperar todos los a href y mostrar su texto asociado.
     * @param url
     * @throws IOException
     */
    public static void getImgs(String url) throws IOException{
        Document doc = Jsoup.connect(url).get();
        System.out.println("Imgs from:\t"+doc.title());
        //Elements lst = doc.select("img[src~=[\\w\\d\\W].jpg]");
        Elements lst = doc.select("img[src~=[\\w\\d\\W]logo[\\w\\d\\W]]");
        for (Element elem: lst)
            System.out.println("\t"+elem.attr("src"));
    }
}

```

http://www.laenfermeria.es/docuwiki/doku.php?id=siglas_medicas

siglas_medicas [laen x]

www.laenfermeria.es/docuwiki/doku.php?id=siglas_medicas

[[siglas_medicas]]

LAENFERMERIA WIKI

Ver fuente

Revisiones anteriores

Cambios recientes

Buscar

Traza: » siglas_medicas

Introducción

Este apartado de siglas y abreviaturas médicas, es en si mismo un importante recurso dentro del archivo documental, de interés tanto para las personas vinculadas al mundo de la salud (técnicos, diplomados, licenciados), como, particularmente, a los Documentalistas Sanitarios, ya que en su día a día codificando informes de alta, se encontrarán frecuentemente con términos abreviados, por lo que existen referencias múltiples a la CIE-9-MC.

A

A: Abdomen / Aborto / Aguas (meconiales) / Analítica / Anestesia / Anexo / Antecedentes / Años / Aurícula.

A-: Prefijo negativo.

A00: Marcapasos con estimulación auricular asincrónica.

a. Ce.: Antes de la cena.

a. Co.: Antes de la comida.

a. De.: Antes del desayuno.

A. Gral.: Analítica general / Anestesia general.

a. m.: Ante meridiem (por la mañana).

AA: Abdomen agudo / Alcohólicos anónimos / Amenaza de aborto / Aminoácido / Anemia aplásica / Aorta abdominal / Aorta ascendente / Apendicitis aguda.

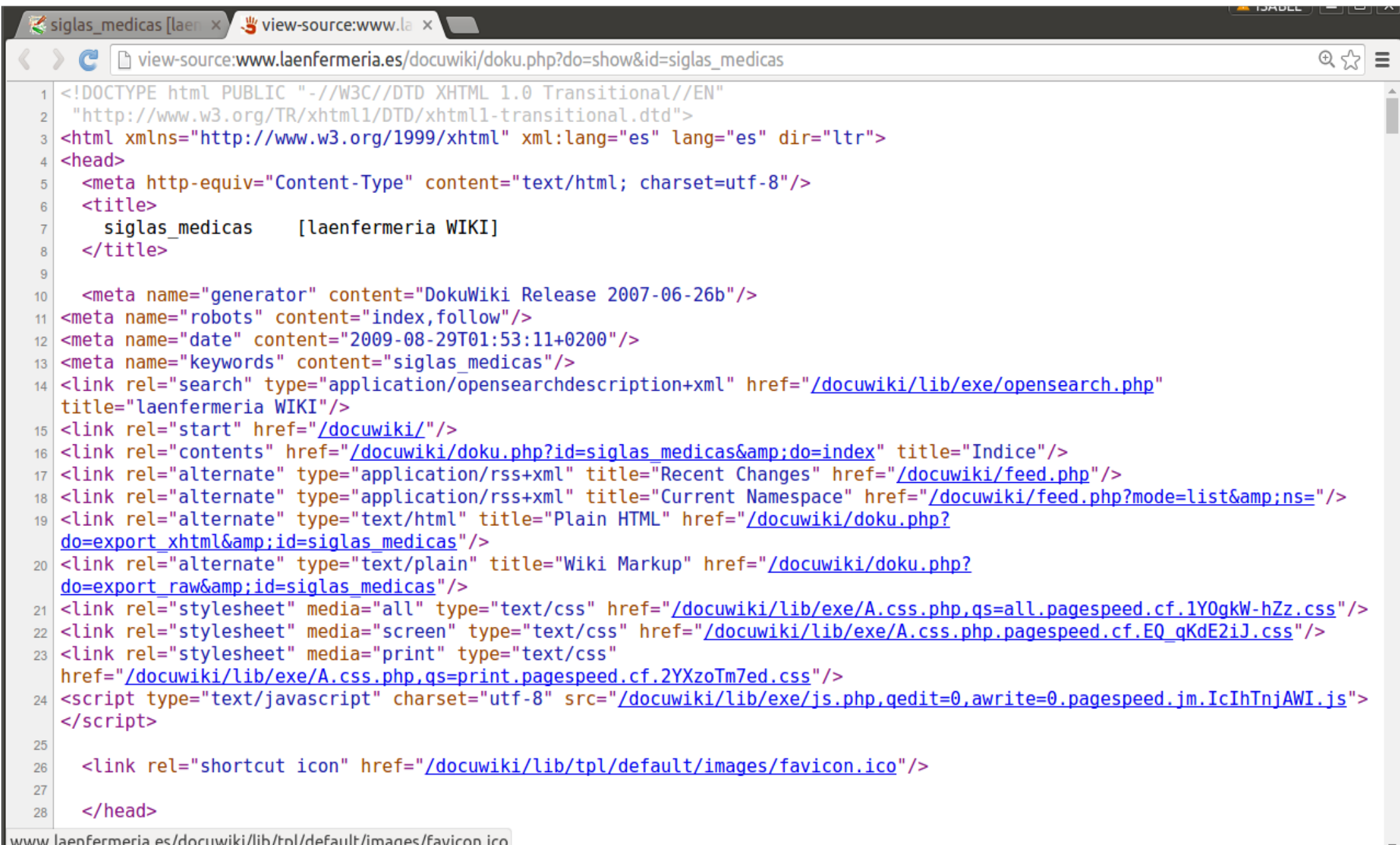
Tabla de Contenidos

- Introducción
- A
- B
- C
- D
- E
- F
- G
- H
- I
- J
- K
- L
- M
- N
- O
- P
- Q
- R
- S
- T
- U
- V
- W
- X
- Y
- Z

3º ejercicio: Recuperar el diccionario de acrónimos y volcarlo en un fichero

- Observa el código fuente.
- En el método java tendrás que usar el método select. ¿qué elementos tenemos que recuperar del documento?
- Escribe una función que vuelque el contenido en un fichero.

http://www.laenfermeria.es/docuwiki/doku.php?id=siglas_medicas (Código fuente)



```
1 <!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"
2 "http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
3 <html xmlns="http://www.w3.org/1999/xhtml" xml:lang="es" lang="es" dir="ltr">
4 <head>
5   <meta http-equiv="Content-Type" content="text/html; charset=utf-8"/>
6   <title>
7     siglas_medicas    [laenfermeria WIKI]
8   </title>
9
10  <meta name="generator" content="DokuWiki Release 2007-06-26b"/>
11  <meta name="robots" content="index, follow"/>
12  <meta name="date" content="2009-08-29T01:53:11+0200"/>
13  <meta name="keywords" content="siglas_medicas"/>
14  <link rel="search" type="application/opensearchdescription+xml" href="/docuwiki/lib/exe/opensearch.php"
15    title="laenfermeria WIKI"/>
16  <link rel="start" href="/docuwiki/" />
17  <link rel="contents" href="/docuwiki/doku.php?id=siglas_medicas&do=index" title="Indice"/>
18  <link rel="alternate" type="application/rss+xml" title="Recent Changes" href="/docuwiki/feed.php"/>
19  <link rel="alternate" type="application/rss+xml" title="Current Namespace" href="/docuwiki/feed.php?mode=list&ns="/>
20  <link rel="alternate" type="text/html" title="Plain HTML" href="/docuwiki/doku.php?
21    do=export_xhtml&id=siglas_medicas"/>
22  <link rel="alternate" type="text/plain" title="Wiki Markup" href="/docuwiki/doku.php?
23    do=export_raw&id=siglas_medicas"/>
24  <link rel="stylesheet" media="all" type="text/css" href="/docuwiki/lib/exe/A.css.php,qs=all.pagespeed.cf.1Y0gkW-hZz.css"/>
25  <link rel="stylesheet" media="screen" type="text/css" href="/docuwiki/lib/exe/A.css.php.pagespeed.cf.EQ_qKdE2iJ.css"/>
26  <link rel="stylesheet" media="print" type="text/css"
27    href="/docuwiki/lib/exe/A.css.php,qs=print.pagespeed.cf.2YXzoTm7ed.css"/>
28  <script type="text/javascript" charset="utf-8" src="/docuwiki/lib/exe/js.php,qedit=0,awrite=0.pagespeed.jm.IcIhTnjAWI.js">
29  </script>
30
31  <link rel="shortcut icon" href="/docuwiki/lib/tpl/default/images/favicon.ico"/>
32
33 </head>
```



```

/**
 * buscar todos las siglas y su definición
 * @param url
 * @throws IOException
 */
public static void getSiglas(String url) throws IOException{
    Document doc = Jsoup.connect(url).get();
    Elements lst = doc.select("p");
    StringBuffer out=new StringBuffer();
    for (Element elem: lst) {
        System.out.println("\t"+elem.text());
        out.append(elem.text()+"\n");
    }

    writeToFile(out.toString(),"siglas.txt");
}

private static void writeToFile(String str,String nameFile) throws IOException {
    BufferedWriter bwr = new BufferedWriter(new FileWriter(new File(nameFile)));
    bwr.write(str);
    bwr.flush();
    bwr.close();
    System.out.println(nameFile + " saved!!!");
}

```


4º ejercicio: Obtener el glosario de la EMA (Agencia Europea del Medicamento)

- Buscar en google (“European Medicines Agency Glossary”)
- http://www.ema.europa.eu/ema/index.jsp?curl=pages/document_library/landing/glossary.jsp

Primer paso: recuperar el link de cada letra



EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

An agency of the European Union



Text size: [A](#) [A](#) [A](#)

Site-wide search

GO ▶

Search document library

Follow us:

[Home](#) [Find medicine](#) [Human regulatory](#) [Veterinary regulatory](#) [Committees](#) [News & events](#) [Partners & networks](#) [About us](#)

What we do

Who we are

How we work

History of EMA

Careers

Procurement

Support to research

Contact

Legal

▼ About this website

▶ Glossary

FAQs

▶ [Home](#) ▶ [About Us](#) ▶ [About this website](#) ▶ [Glossary](#)

Glossary

Email Print Help Share

This glossary gives definitions for the main regulatory terms used on this website and in European Medicines Agency documents.

Click on a letter below to see terms beginning with that letter.

Disclaimer: The definitions in this glossary have been developed to help this website's users understand regulatory terminology. Definitions may differ from those given in European Union legislation.

Glossary

[A](#) [B](#) [C](#) [D](#) [E](#) [F](#) [G](#) [H](#) [I](#) [J](#) [K](#) [L](#) [M](#) [N](#) [O](#) [P](#) [Q](#) [R](#) [S](#) [T](#) [U](#) [V](#) [W](#) [X](#) [Y](#) [Z](#)



What we do

Who we are

How we work

History of EMA

Careers

Procurement

Support to research

Contact

Legal

▼ About this website

► Glossary

FAQs

[Home](#) [About Us](#) [About this website](#) [Glossary](#)

Glossary

[Email](#) [Print](#) [Help](#) [Share](#)

This glossary gives definitions for the main regulatory terms used on this website and in European Medicines Agency documents.

Click on a letter below to see terms beginning with that letter.

Disclaimer: The definitions in this glossary have been developed to help this website's users understand regulatory terminology. Definitions may differ from those given in European Union legislation.

Glossary

[A](#) [B](#) [C](#) [D](#) [E](#) [F](#) [G](#) [H](#) [I](#) [J](#) [K](#) [L](#) [M](#) [N](#) [O](#) [P](#) [Q](#) [R](#) [S](#) [T](#) [U](#) [V](#) [W](#) [X](#) [Y](#) [Z](#)

accelerated assessment	Rapid assessment of medicines in the centralised procedure that are of major interest for public health, especially ones that are therapeutic innovations. Accelerated assessment usually takes 150 evaluation days, rather than 210. For more information, see: accelerated assessment .
access to documents	Activities relating to public access to European Medicines Agency documents. For more information, see access to documents .
active substance	The substance responsible for the activity of a medicine.
active substance master file	Documentation providing detailed information on the manufacturing of the active substance of a medicine.

4º ejercicio: Recuperar el listado de medicinas de la web MedLinePlus y de cada fármaco la sección “¿Cómo se debe usar este medicamento?”

- <https://medlineplus.gov/spanish/druginformation.html>

Primer paso: recuperar el link de cada letra

 Biblioteca Nacional de Medicina de los EE. UU.



MedlinePlus
Información de salud para usted

Busque en MedlinePlus

BUSCAR

[Sobre MedlinePlus](#) [Índice](#) [FAQs](#) [Contáctenos](#)

[Temas de salud](#) [Medicinas y suplementos](#) [Videos y multimedia](#) [English](#)

[Página Principal](#) → [Medicinas, hierbas y suplementos](#)

Medicinas, hierbas y suplementos

Medicinas

Aprenda sobre sus medicamentos de receta y de venta libre incluyendo efectos secundarios, dosis, precauciones especiales y mucho más.

Busque por marca o nombre genérico

A B C D E F G H I J K L M N O P Q R S T U

V W X Y Z 0-9



Hierbas y suplementos

Hojee información sobre suplementos dietarios y hierbas para aprender sobre su efectividad, dosis e interacciones con otras medicinas.

Todas las hierbas y suplementos

AHFS® Consumer Medication Information provee información sobre centenares de medicinas de receta y venta libre y es propiedad de la [American Society of Health-System Pharmacists, Inc.](#), Bethesda, Maryland. Está protegida por la ley de derechos de autor. Copyright© 2016. Todos los derechos reservados.

Temas relacionados

- [Analgésicos](#)
- [Antibióticos](#)
- [Anticoagulantes y antiplaquetarios](#)
- [Antidepresivos](#)
- [Corticoides](#)
- [Estatinas](#)
- [Medicamentos](#)
- [Medicamentos sin receta médica](#)
- [Medicina alternativa y](#)

Segundo paso: guardar el link de cada medicina

Medicinas: A

A Tan 12x Suspension ® (como combinacion de productos que contiene fenilefrina, mepiramina) ver [Fenilefrina](#)

A-Hydrocort ® ver [Inyección de hidrocortisona](#)

A-Methapred ® ver [Inyección de metilprednisolona](#)

A.R. ® Eye Drops (como combinacion de productos que contiene tetrahidrozolina, sulfato de zinc) ver [Tetrahidrozolina oftálmica](#)
[Abacavir](#)

Abelcet ® ver [Inyección de complejo lipídico de anfotericina B](#)

Abilify ® ver [Aripiprazol](#)

Abilify ® ver [Inyección de aripiprazol](#)

Abilify Maintena ® ver [Inyección de aripiprazol](#)

[Abiraterona](#)

Abitrexate ® ver [Inyección de metotrexato](#)

[AbobotulinumtoxinA inyectable](#)

Abraxane ® ver [Paclitaxel inyectable](#)

Absorbine ver [Tolnaftato](#)

Absorica ® ver [Isotretinoína](#)

Abstral ® ver [Fentanilo](#)

[Acamprosato](#)

Acanya ® (como combinacion de productos que contiene peroxido de benzoilo, clindamicina) ver [Peróxido de benzoilo Tópica](#)

[Acarbose](#)

Accolate ® ver [Zafirlukast](#)

Tercer paso: ojo!!! sólo tienes que recuperar la respuesta a la pregunta “¿Cómo se debe usar este medicamento?”?



Busque en MedlinePlus

BUSCAR

Sobre MedlinePlus Índice FAQs Contáctenos

Temas de salud

Medicinas y suplementos

Videos y multimedia

English

Página Principal → Medicinas, hierbas y suplementos → Tolnaftato

Tolnaftato



¿Para cuáles condiciones o enfermedades se prescribe este medicamento?

¿Cómo se debe usar este medicamento?

¿Cuáles son las precauciones especiales que debo seguir?

¿Qué tengo que hacer si me olvido de tomar una dosis?

¿Cuáles son los efectos secundarios que podría provocar este medicamento?

¿Cómo debo almacenar o desechar este medicamento?

¿Qué otra información de importancia debería saber?

Marcas comerciales

¿Para cuáles condiciones o enfermedades se prescribe este medicamento?

El tolnaftato detiene el crecimiento de los hongos que provocan las infecciones de la piel, incluyendo el pie de atleta, la sarna y la tiña.

Este medicamento también puede ser prescrito para otros usos; pídale más información a su doctor o farmacéutico.

¿Cómo se debe usar este medicamento?

El tolnaftato viene envasado en forma de crema, líquido, polvo, gel, spray en polvo y líquido para aplicar sobre la piel. El tolnaftato por lo general, se aplica 2 veces al día. Siga cuidadosamente las instrucciones en la etiqueta del medicamento y pregúntele a su doctor o farmacéutico cualquier cosa que no entienda. Use el medicamento exactamente como se indica. No use más ni menos que la dosis indicada ni tampoco más seguido que lo prescrito por su doctor.

El ardor y el dolor producido por el pie de atleta o el prurito (picaazón) de la sarna deberían disminuir dentro de 2 a 3 días. Siga el tratamiento durante al menos 2 semanas después de que desaparezcan los síntomas. El tratamiento podría durar un total de 4 a 6 semanas.

Limpie bien la zona afectada, deje que se seque y luego frote suavemente el medicamento hasta que desaparezca la mayoría. Use lo suficiente como para cubrir la zona afectada. Lávese las manos después de aplicar el medicamento.

El spray en polvo y en solución líquida deben aplicarse entre los dedos del pie; también hay que aplicar el medicamento a calcetines y zapatos. El spray debe agitarse mucho antes de cada uso para mezclar el medicamento y luego rociarlo a una distancia de al menos 6 pulgadas.

¿Cuáles son las precauciones especiales que debo seguir?

5º ejercicio: descarga todos los datos de los

Conciertos en Madrid y agenda de ocio 2016



Alejandro Sanz en vivo en Madrid 5 diciembre



Malú concierto en Madrid 17 diciembre



Leiva concierto en Madrid 30 diciembre



Quique González & Los Detectives, concierto 29 diciembre en Madrid



091 concierto en La Riviera 29 octubre



Madrid Live! con The Chemical Brothers en Madrid 28 octubre



Serrat, Ana Belén, Víctor Manuel y Miguel Ríos 27 octubre

Links muy útiles !!!

- <https://jsoup.org/cookbook/>
- <http://jarroba.com/scraping-java-jsoup-ejemplos/>

Web scraping en Python

- **BeautifulSoup**

- <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>
- <http://jarroba.com/scraping-python-beautifulsoup-ejemplos/>
- <http://docs.python-guide.org/en/latest/scenarios/scrape/>