EUROPYTHON 2011
FLORENCE, JUNE 20-26

Scraping Techniques to Extract Advertisements From Web Pages

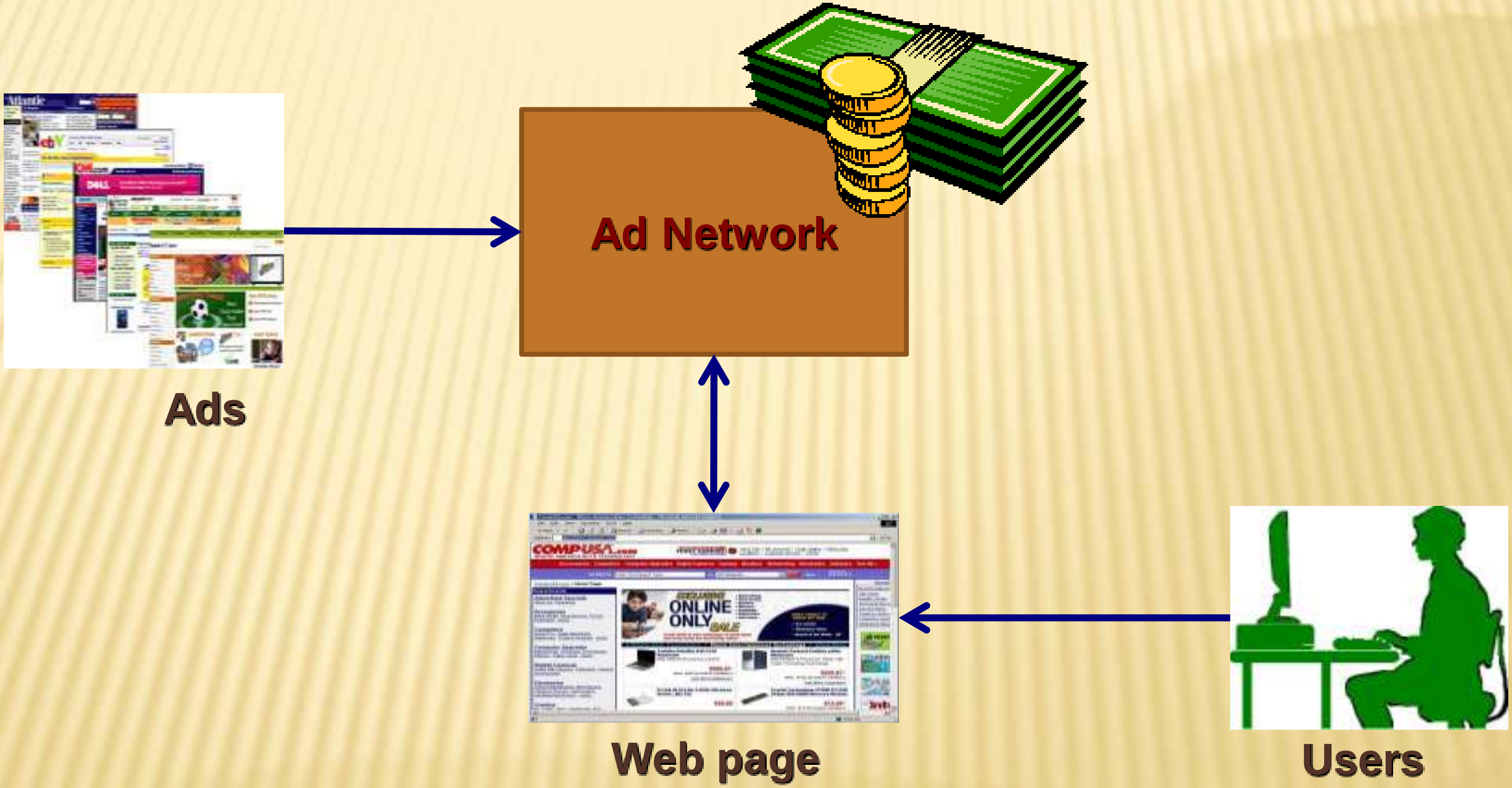
Mirko Urru
mirko.urru@hotmail.it

Stefano Cotta Ramusino
whitone@gmail.com

OutLine

- Scraping techniques applied to contextual advertising
- Background
- What is scraping?
- The form of scraping
- The most famous scraping techniques
- Application of scraping techniques
  to contextual advertising
- Conclusions

Online advertising is a major source of income for most of the websites currently online



Ads

Ad Network

Web page

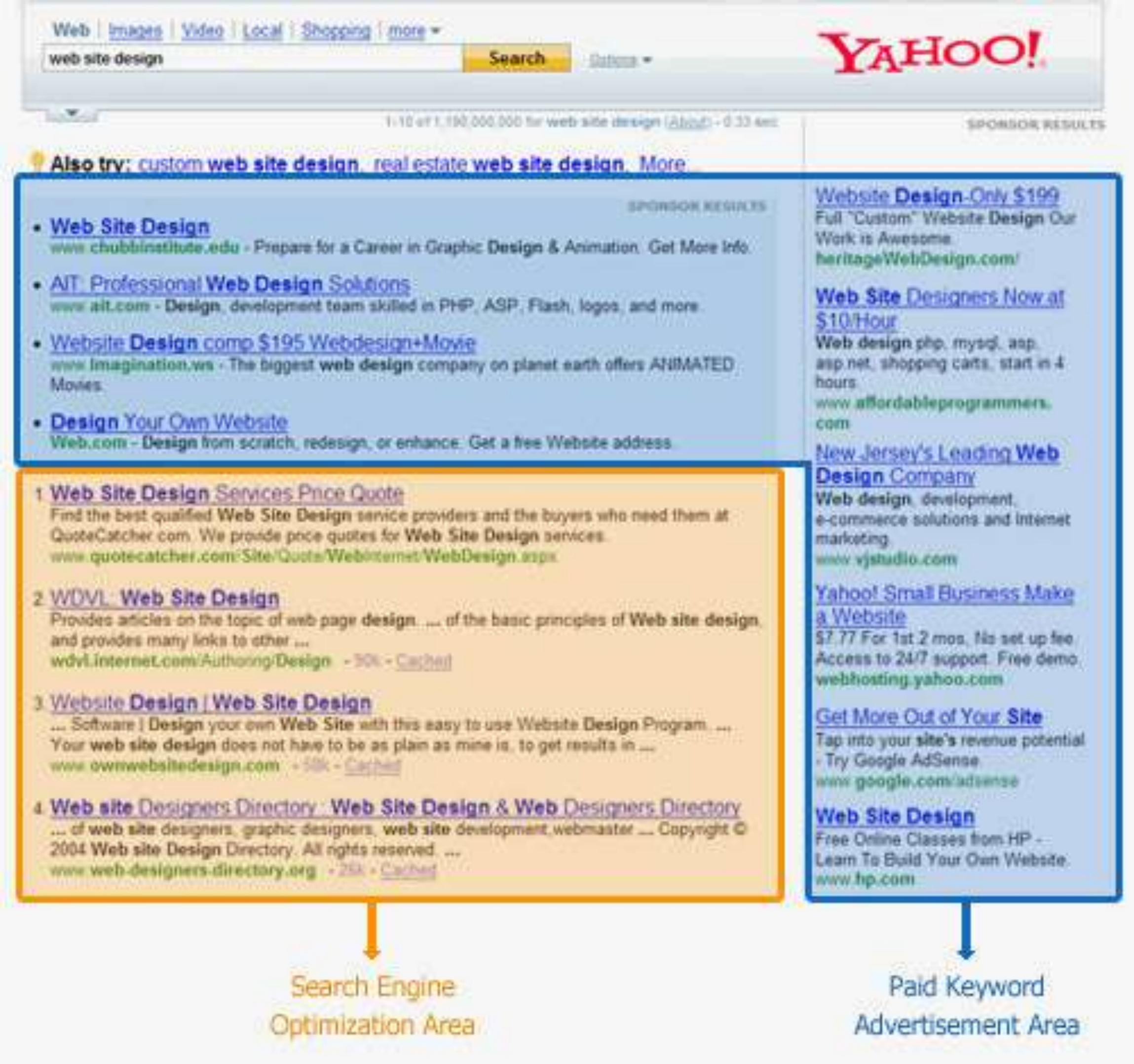
Users

The Adopted Techniques

Sponsored Search

Contextual Advertising

Sponsored Search



web site design

1-10 of 1,190,000,000 for web site design (About) - 0.33 sec

Also try: custom **web site design**, real estate **web site design**, More

- **Web Site Design**
  www.chubbinstitute.edu - Prepare for a Career in Graphic **Design** & Animation. Get More Info.
- AIT: Professional **Web Design** Solutions
  www.ait.com - **Design**, development team skilled in PHP, ASP, Flash, logos, and more.
- Website **Design** comp $195 Webdesign+Movie
  www.Imagination.ws - The biggest **web design** company on planet earth offers ANIMATED Movies.
- **Design** Your Own Website
  Web.com - **Design** from scratch, redesign, or enhance. Get a free Website address.

1 **Web Site Design** Services Price Quote
Find the best qualified **Web Site Design** service providers and the buyers who need them at QuoteCatcher.com. We provide price quotes for **Web Site Design** services.
www.quotecatcher.com/Site/Quote/WebInternet/WebDesign.aspx

2 WDVL: **Web Site Design**
Provides articles on the topic of web page **design**. ... of the basic principles of **Web site design** and provides many links to other ...
wdvl.internet.com/Authoring/Design - 50k - Cached

3 Website **Design** | **Web Site Design**
... Software | **Design** your own **Web Site** with this easy to use Website **Design** Program. ... Your **web site design** does not have to be as plain as mine is, to get results in ...
www.ownwebsitedesign.com - 58k - Cached

4 **Web site** Designers Directory : **Web Site Design** & **Web** Designers Directory
... of **web site** designers, graphic designers, **web site** development,webmaster ... Copyright © 2004 **Web site Design** Directory. All rights reserved. ...
www.web-designers-directory.org - 25k - Cached

Website **Design**-Only $199
Full "Custom" Website **Design** Our Work is Awesome.
heritageWebDesign.com/

Web Site Designers Now at $10/Hour
Web design php, mysql, asp, asp.net, shopping carts, start in 4 hours.
www.affordableprogrammers.com

New Jersey's Leading **Web Design** Company
Web design, development, e-commerce solutions and Internet marketing.
www.vjstudio.com

Yahoo! Small Business Make a Website
$7.77 For 1st 2 mos, No set up fee Access to 24/7 support. Free demo.
webhosting.yahoo.com

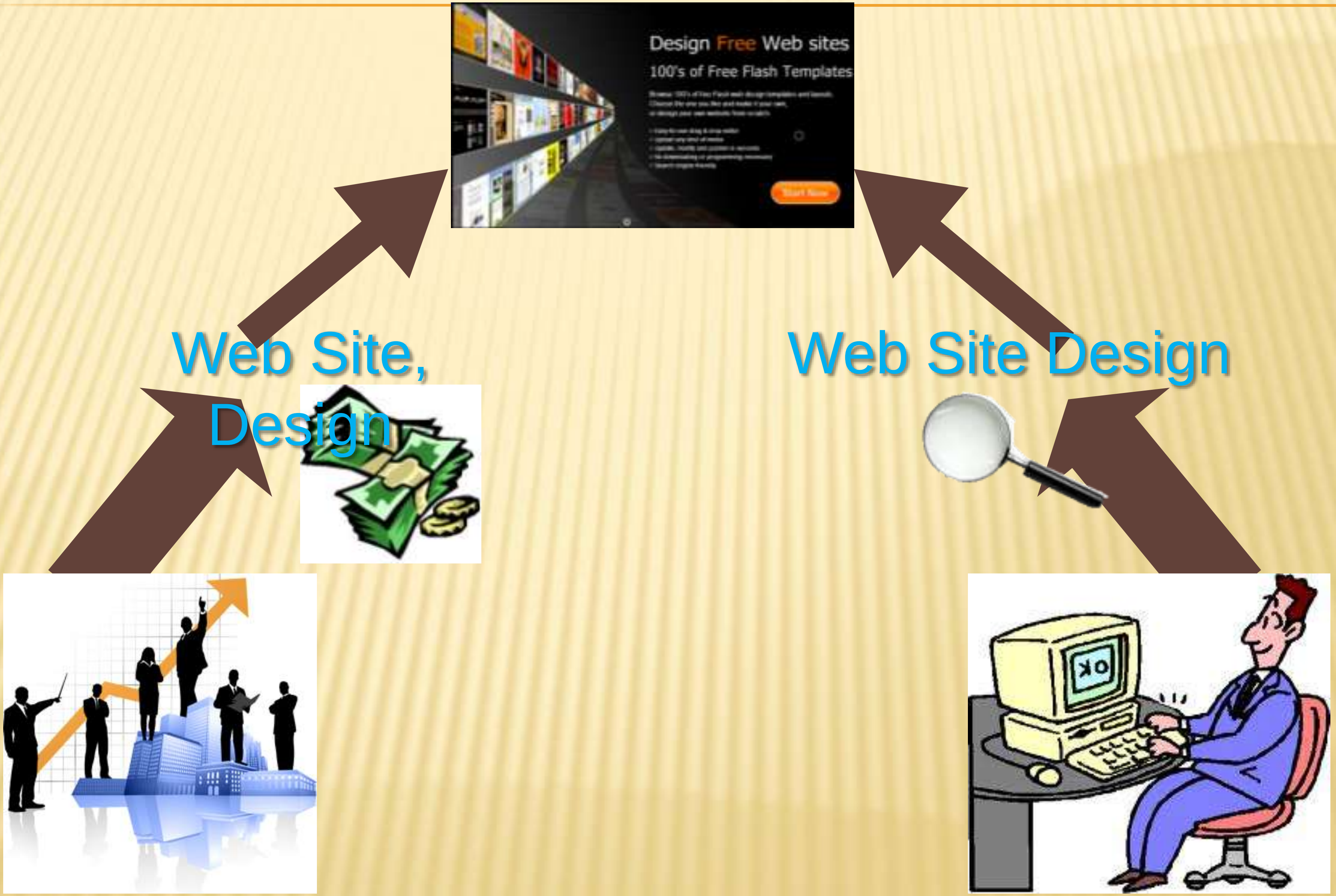
Get More Out of Your **Site**
Tap into your **site's** revenue potential - Try Google AdSense.
www.google.com/adsense

Web Site Design
Free Online Classes from HP - Learn To Build Your Own Website.
www.hp.com

Search Engine Optimization Area

Paid Keyword Advertisement Area

Sponsored Search



Design Free Web sites

100's of Free Flash Templates

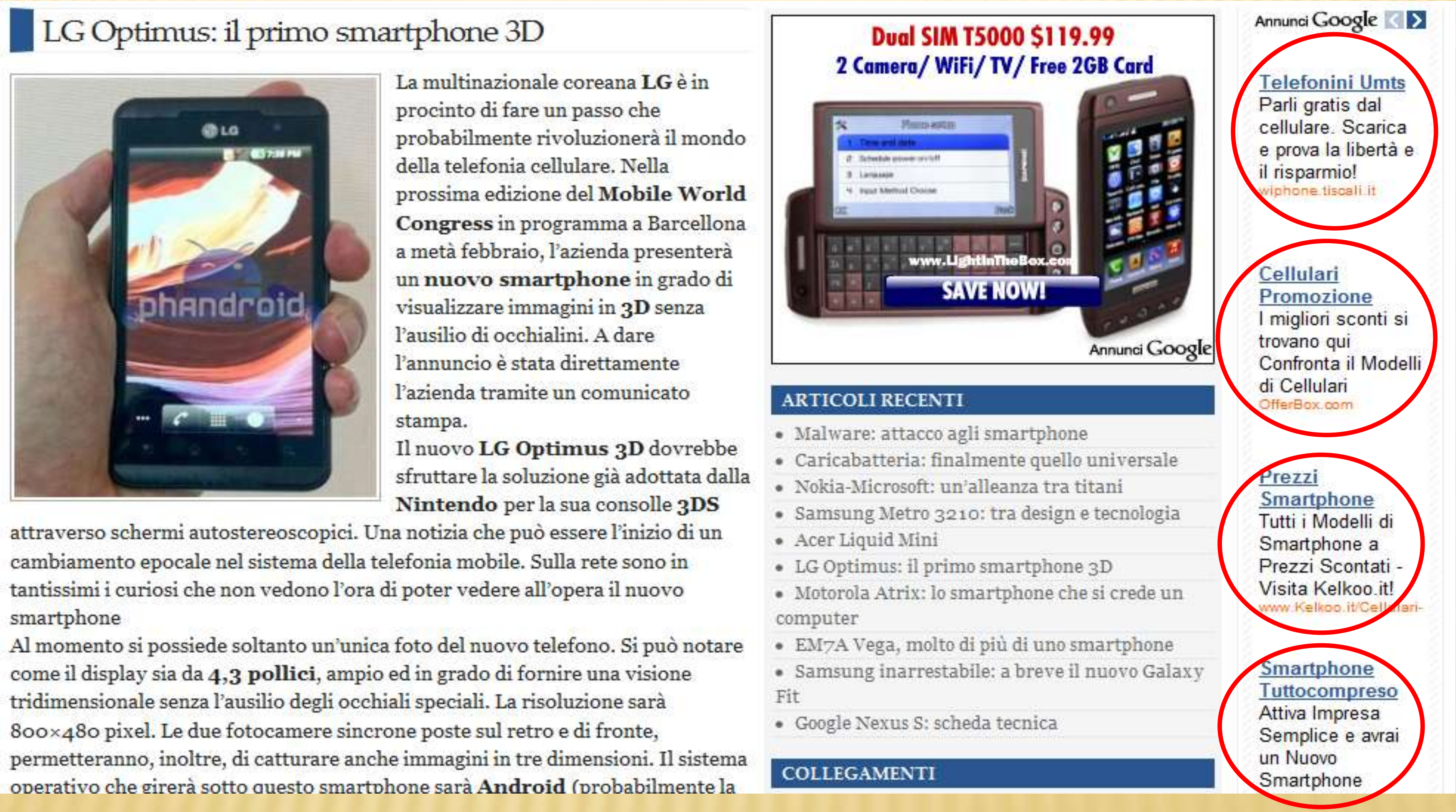
Browse 100's of Free Flash web design templates and layouts.
Choose the one you like and make it your own,
or design your own website from scratch.

- Easy to use drag & drop editor
- Upload any kind of media
- Update, modify and publish in seconds
- No downloading or programming necessary
- Search engine friendly

Web Site, Design

Web Site Design

The contextual advertising is the integration of content (text) and advertising information(text)



LG Optimus: il primo smartphone 3D

La multinazionale coreana **LG** è in procinto di fare un passo che probabilmente rivoluzionerà il mondo della telefonia cellulare. Nella prossima edizione del **Mobile World Congress** in programma a Barcellona a metà febbraio, l'azienda presenterà un **nuovo smartphone** in grado di visualizzare immagini in **3D** senza l'ausilio di occhialini. A dare l'annuncio è stata direttamente l'azienda tramite un comunicato stampa.

Il nuovo **LG Optimus 3D** dovrebbe sfruttare la soluzione già adottata dalla **Nintendo** per la sua consolle **3DS** attraverso schermi autostereoscopici. Una notizia che può essere l'inizio di un cambiamento epocale nel sistema della telefonia mobile. Sulla rete sono in tantissimi i curiosi che non vedono l'ora di poter vedere all'opera il nuovo smartphone

Al momento si possiede soltanto un'unica foto del nuovo telefono. Si può notare come il display sia da **4,3 pollici**, ampio ed in grado di fornire una visione tridimensionale senza l'ausilio degli occhiali speciali. La risoluzione sarà 800×480 pixel. Le due fotocamere sincrone poste sul retro e di fronte, permetteranno, inoltre, di catturare anche immagini in tre dimensioni. Il sistema operativo che girerà sotto questo smartphone sarà **Android** (probabilmente la

Dual SIM T5000 $119.99
2 Camera/ WiFi/ TV/ Free 2GB Card
www.LightInTheBox.com
SAVE NOW!
Annunci Google

ARTICOLI RECENTI

- Malware: attacco agli smartphone
- Caricabatteria: finalmente quello universale
- Nokia-Microsoft: un'alleanza tra titani
- Samsung Metro 3210: tra design e tecnologia
- Acer Liquid Mini
- LG Optimus: il primo smartphone 3D
- Motorola Atrix: lo smartphone che si crede un computer
- EM7A Vega, molto di più di uno smartphone
- Samsung inarrestabile: a breve il nuovo Galaxy Fit
- Google Nexus S: scheda tecnica

COLLEGAMENTI

Annunci Google

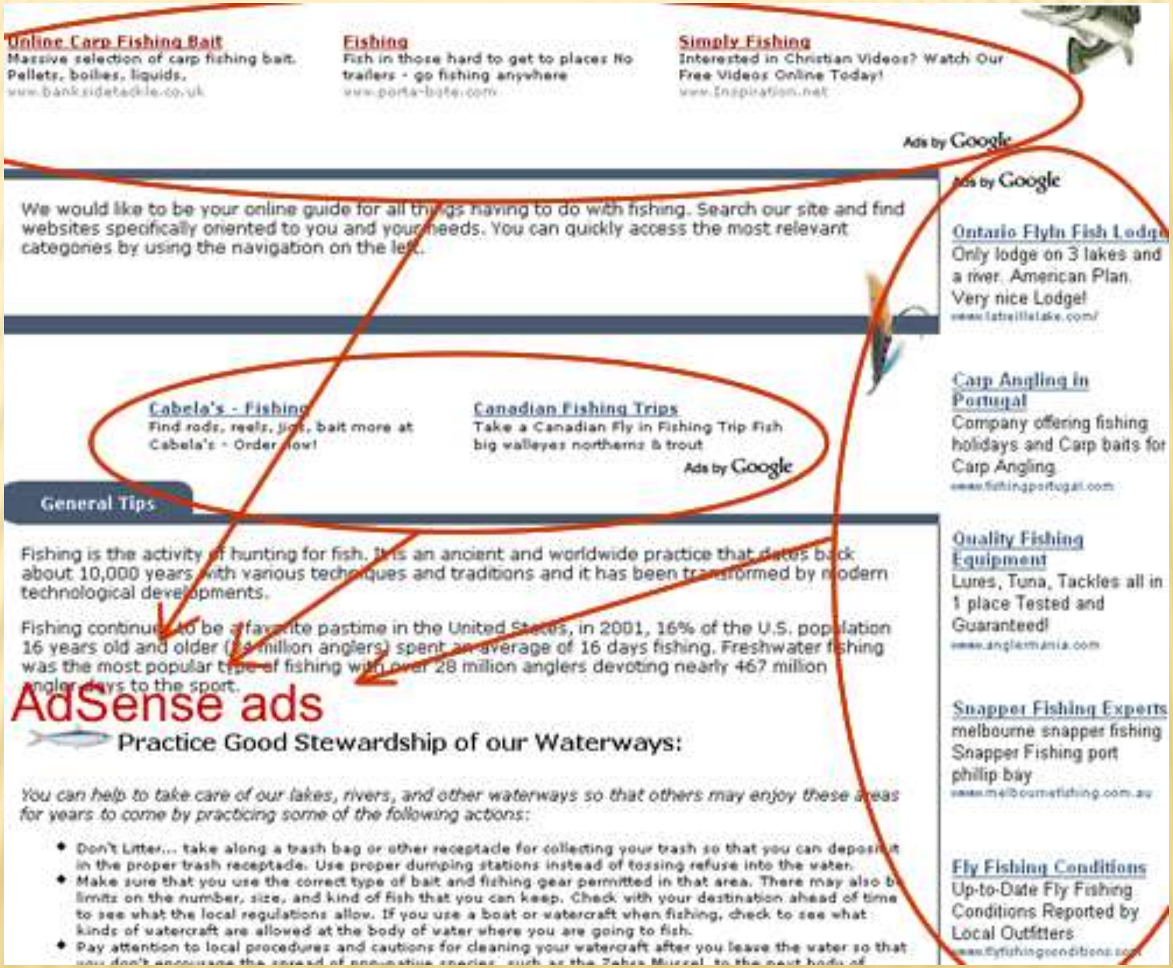
Telefonini Umts
Parli gratis dal cellulare. Scarica e prova la libertà e il risparmio!
wiphone.tiscali.it

Cellulari Promozione
I migliori sconti si trovano qui Confronta il Modelli di Cellulari
OfferBox.com

Prezzi Smartphone
Tutti i Modelli di Smartphone a Prezzi Scontati - Visita Kelkoo.it!
www.Kelkoo.it/Cellulari-

Smartphone Tuttocompreso
Attiva Impresa Semplice e avrai un Nuovo Smartphone

Contextual Advertising



Online Carp Fishing Bait
Massive selection of carp fishing bait. Pellets, boilies, liquids.
www.banksidetackle.co.uk

Fishing
Fish in those hard to get to places No trailers - go fishing anywhere
www.porta-bote.com

Simply Fishing
Interested in Christian Videos? Watch Our Free Videos Online Today!
www.Inspiration.net

Ads by Google

We would like to be your online guide for all things having to do with fishing. Search our site and find websites specifically oriented to you and your needs. You can quickly access the most relevant categories by using the navigation on the left.

Cabela's - Fishing
Find rods, reels, jigs, bait more at Cabela's - Order Now!

Canadian Fishing Trips
Take a Canadian Fly in Fishing Trip Fish big walleyes northerns & trout

Ads by Google

General Tips

Fishing is the activity of hunting for fish. It is an ancient and worldwide practice that dates back about 10,000 years with various techniques and traditions and it has been transformed by modern technological developments.

Fishing continues to be a favorite pastime in the United States, in 2001, 16% of the U.S. population 16 years old and older (34 million anglers) spent an average of 16 days fishing. Freshwater fishing was the most popular type of fishing with over 28 million anglers devoting nearly 467 million angler days to the sport.

AdSense ads

Practice Good Stewardship of our Waterways:

You can help to take care of our lakes, rivers, and other waterways so that others may enjoy these areas for years to come by practicing some of the following actions:

- Don't Litter... take along a trash bag or other receptacle for collecting your trash so that you can deposit it in the proper trash receptacle. Use proper dumping stations instead of tossing refuse into the water.
- Make sure that you use the correct type of bait and fishing gear permitted in that area. There may also be limits on the number, size, and kind of fish that you can keep. Check with your destination ahead of time to see what the local regulations allow. If you use a boat or watercraft when fishing, check to see what kinds of watercraft are allowed at the body of water where you are going to fish.
- Pay attention to local procedures and cautions for cleaning your watercraft after you leave the water so that you don't encourage the spread of non-native species, such as the Zebra Mussel, to the next body of

Ads by Google

Ontario Flyin Fish Lodge
Only lodge on 3 lakes and a river. American Plan. Very nice Lodge!
www.latreillelake.com/

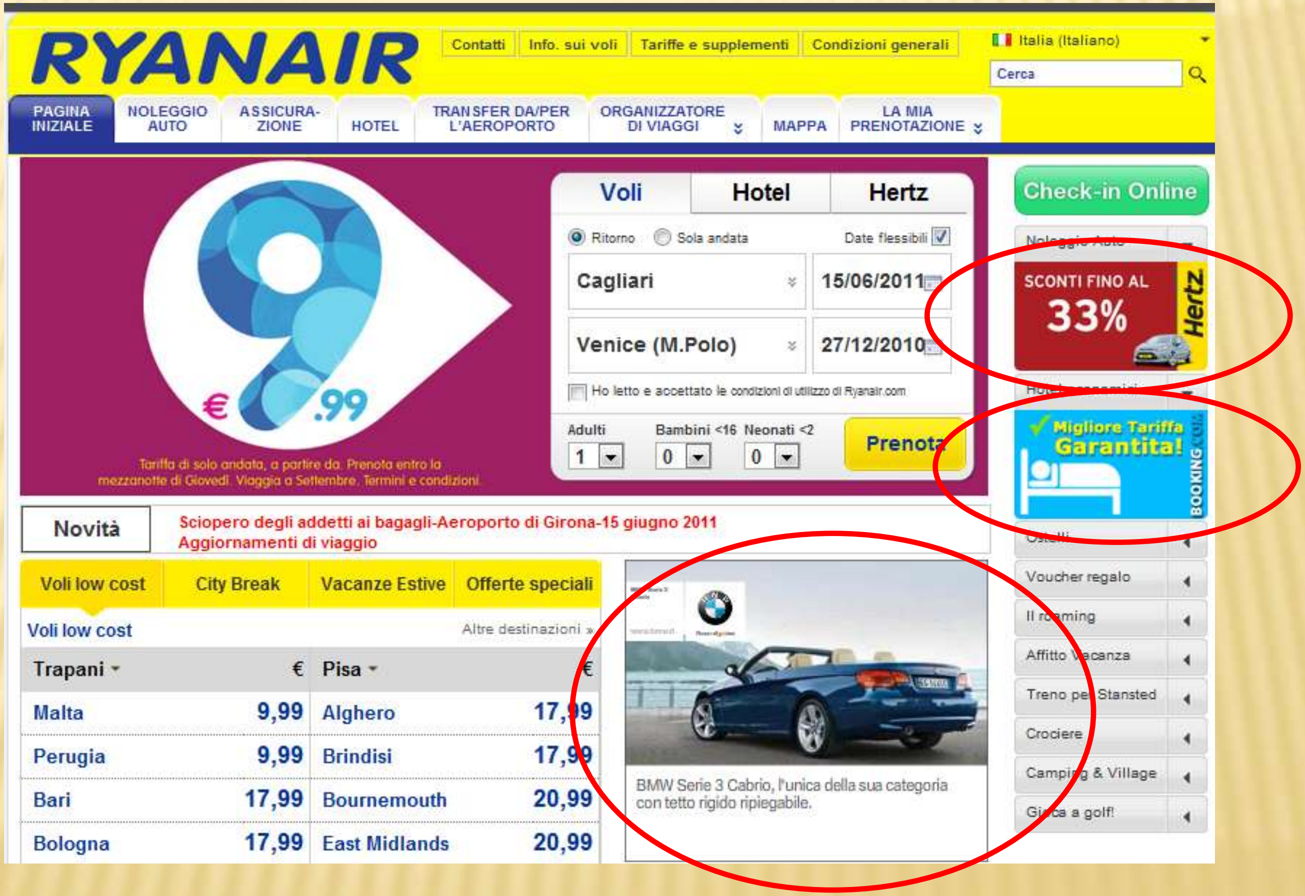
Carp Angling in Portugal
Company offering fishing holidays and Carp baits for Carp Angling.
www.fishingportugal.com

Quality Fishing Equipment
Lures, Tuna, Tackles all in 1 place Tested and Guaranteed!
www.anglermania.com

Snapper Fishing Experts
melbourne snapper fishing Snapper Fishing port phillip bay
www.melbournefishing.com.au

Fly Fishing Conditions
Up-to-Date Fly Fishing Conditions Reported by Local Outfitters
www.flyfishingconditions.com

Scraping advertising is a technique to obtain ads from a Web page



Tariffa di solo andata, a partire da. Prenota entro la mezzanotte di Giovedì. Viaggio a Settembre. Termini e condizioni

● Ritorno ○ Sola andata Date flessibili ☑

Cagliari 15/06/2011

Venice (M.Polo) 27/12/2010

□ Ho letto e accettato le condizioni di utilizzo di Ryanair.com

Adulti 1 Bambini <16 0 Neonati <2 0

Sciopero degli addetti ai bagagli-Aeroporto di Girona-15 giugno 2011
Aggiornamenti di viaggio

Voli low cost Altre destinazioni

Trapani	€	Pisa	€
Malta	9,99	Alghero	17,99
Perugia	9,99	Brindisi	17,99
Bari	17,99	Bournemouth	20,99
Bologna	17,99	East Midlands	20,99

BMW Serie 3 Cabrio, l'unica della sua categoria con tetto rigido ripiegabile.

What’s the scraping?

Origin:

scraping data from mainframes
from green texts on black screens
to new data structures or API

Nowadays:

forcing data from old websites in
something new (web ≥ 2.0)

The Form of Scraping

Web scraping is the process of automatically collecting Web information

The most famous scraping techniques

Beautiful Soup
mechanize
lxml
html5lib
scrapemark
pyquery
scrapy

The most famous scraping techniques

QtWebKit

HtmlUnit

Selenium

Crowbar

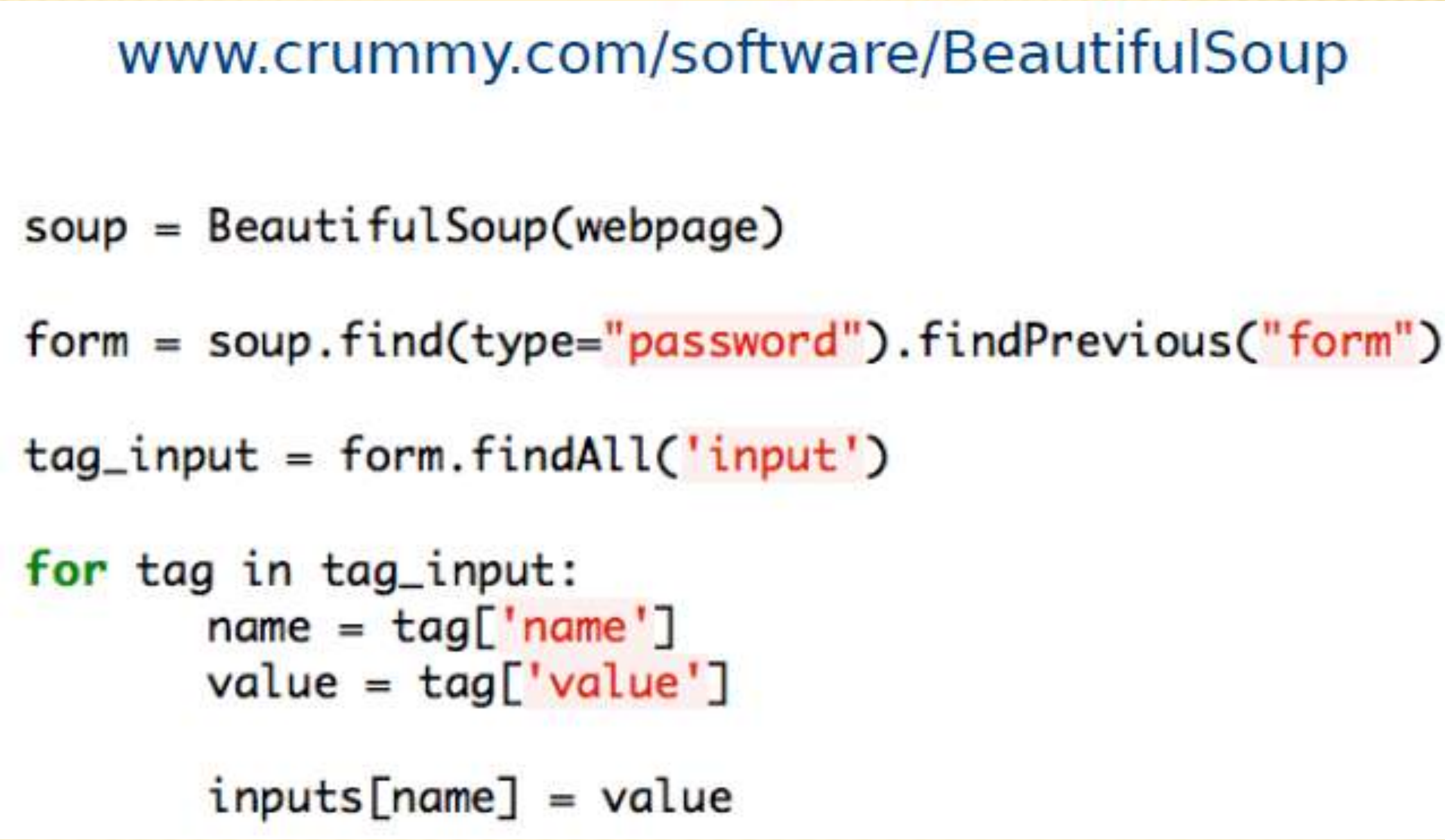
Chrome Remote Shell

PyAuto

FireWatir

Spidermonkey

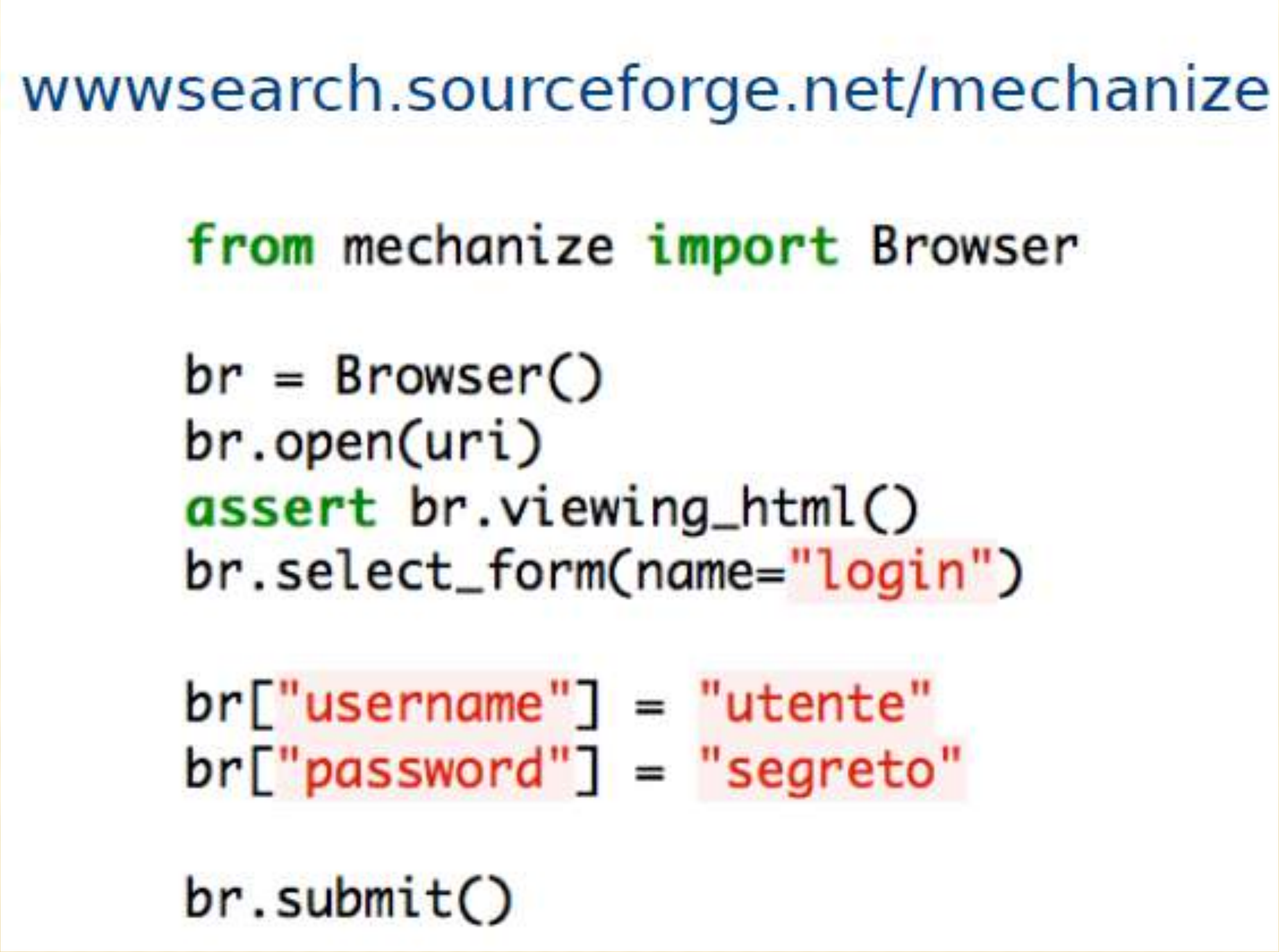
Beautiful Soup



www.crummy.com/software/BeautifulSoup

```
soup = BeautifulSoup(webpage)

form = soup.find(type="password").findPrevious("form")

tag_input = form.findAll('input')

for tag in tag_input:
        name = tag['name']
        value = tag['value']

        inputs[name] = value
```

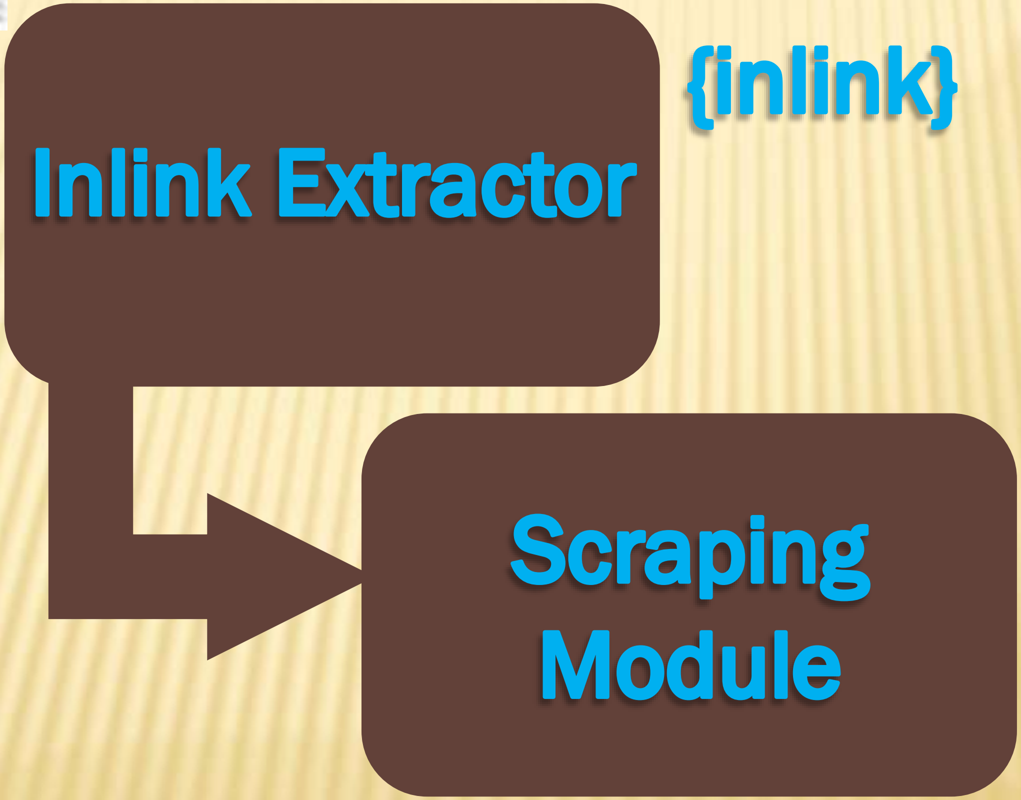
Mechanize



wwwsearch.sourceforge.net/mechanize

```
from mechanize import Browser

br = Browser()
br.open(uri)
assert br.viewing_html()
br.select_form(name="login")

br["username"] = "utente"
br["password"] = "segreto"

br.submit()
```

Application of scraping techniques to contextual advertising


Inlink Extractor

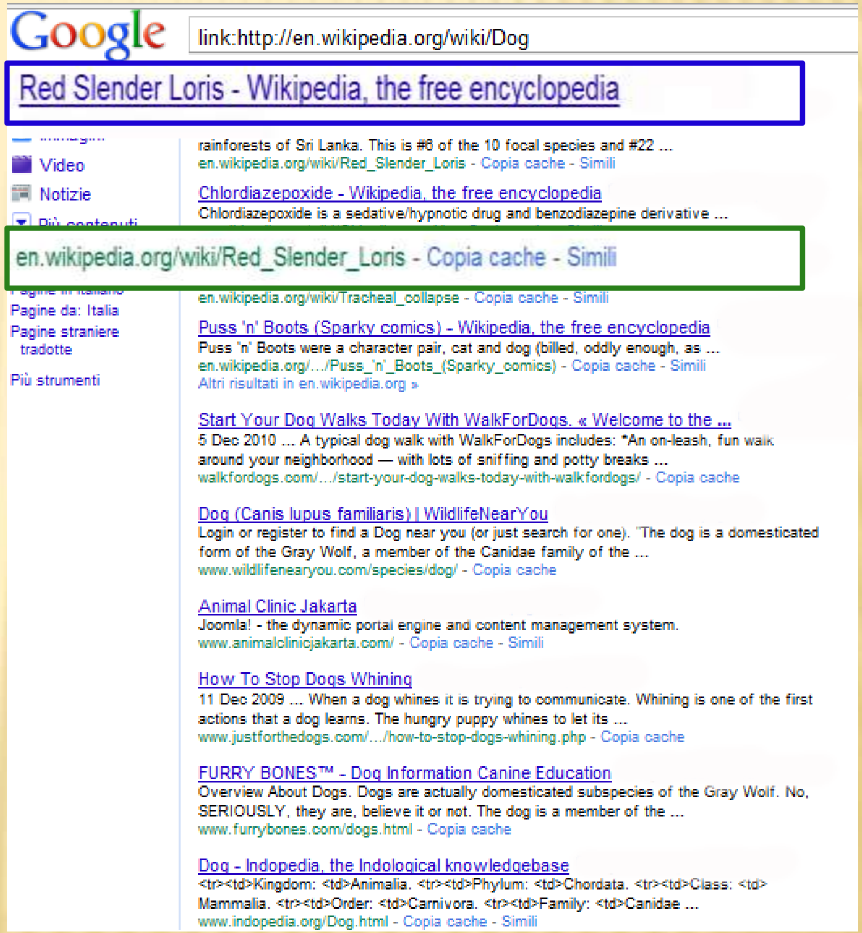
Scraping Module

{inlink}

{ads}

Application of scraping techniques to contextual advertising

Given
a generic page
the
module extract
s the p inlink
Each inlink is
displayed
with the title
and url



titolo

url

Google link:http://en.wikipedia.org/wiki/Dog

Red Slender Loris - Wikipedia, the free encyclopedia

Video
Notizie
Più contenuti
Pagine da: Italia
Pagine straniere tradotte
Più strumenti

rainforests of Sri Lanka. This is #6 of the 10 focal species and #22 ...
en.wikipedia.org/wiki/Red_Slender_Loris - Copia cache - Simili

Chlordiazepoxide - Wikipedia, the free encyclopedia
Chlordiazepoxide is a sedative/hypnotic drug and benzodiazepine derivative ...

en.wikipedia.org/wiki/Red_Slender_Loris - Copia cache - Simili

en.wikipedia.org/wiki/Tracheal_collapse - Copia cache - Simili

Puss 'n' Boots (Sparky comics) - Wikipedia, the free encyclopedia
Puss 'n' Boots were a character pair, cat and dog (billed, oddly enough, as ...
en.wikipedia.org/.../Puss_'n'_Boots_(Sparky_comics) - Copia cache - Simili
Altri risultati in en.wikipedia.org »

Start Your Dog Walks Today With WalkForDogs. « Welcome to the ...
5 Dec 2010 ... A typical dog walk with WalkForDogs includes: *An on-leash, fun walk around your neighborhood — with lots of sniffing and potty breaks ...
walkfordogs.com/.../start-your-dog-walks-today-with-walkfordogs/ - Copia cache

Dog (Canis lupus familiaris) | WildlifeNearYou
Login or register to find a Dog near you (or just search for one). "The dog is a domesticated form of the Gray Wolf, a member of the Canidae family of the ...
www.wildlifenearyou.com/species/dog/ - Copia cache

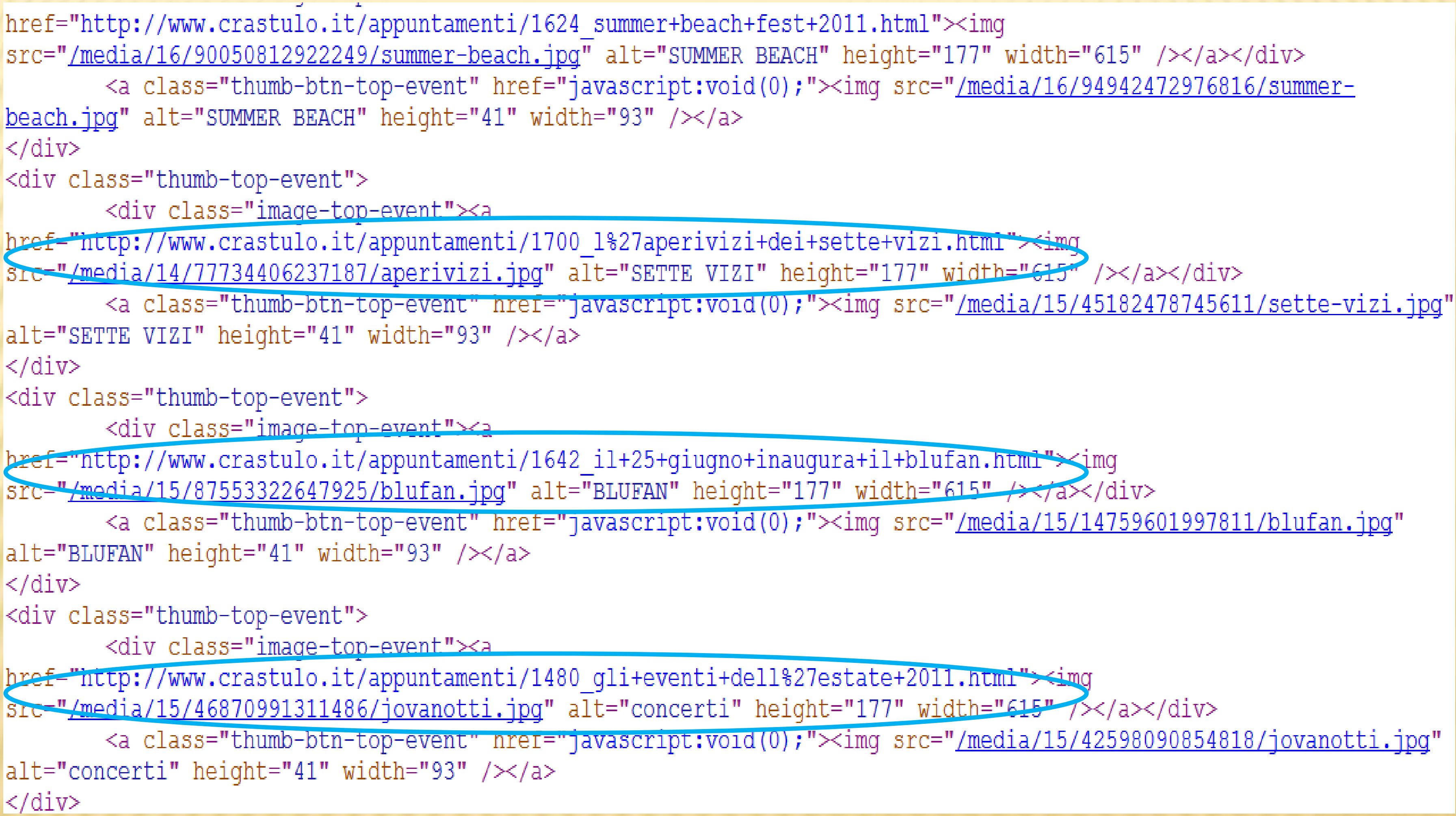
Animal Clinic Jakarta
Joomla! - the dynamic portal engine and content management system.
www.animalclinicjakarta.com/ - Copia cache - Simili

How To Stop Dogs Whining
11 Dec 2009 ... When a dog whines it is trying to communicate. Whining is one of the first actions that a dog learns. The hungry puppy whines to let its ...
www.justforthedogs.com/.../how-to-stop-dogs-whining.php - Copia cache

FURRY BONES™ - Dog Information Canine Education
Overview About Dogs. Dogs are actually domesticated subspecies of the Gray Wolf. No, SERIOUSLY, they are, believe it or not. The dog is a member of the ...
www.furrybones.com/dogs.html - Copia cache

Dog - Indopedia, the Indological knowledgebase
<tr><td>Kingdom: <td>Animalia. <tr><td>Phylum: <td>Chordata. <tr><td>Class: <td> Mammalia. <tr><td>Order: <td>Carnivora. <tr><td>Family: <td>Canidae ...
www.indopedia.org/Dog.html - Copia cache - Simili

Application of scraping techniques to contextual advertising



```
href="http://www.crastulo.it/appuntamenti/1624_summer+beach+fest+2011.html"><img
src="/media/16/90050812922249/summer-beach.jpg" alt="SUMMER BEACH" height="177" width="615" /></a></div>
        <a class="thumb-btn-top-event" href="javascript:void(0);"><img src="/media/16/94942472976816/summer-
beach.jpg" alt="SUMMER BEACH" height="41" width="93" /></a>
</div>
<div class="thumb-top-event">
        <div class="image-top-event"><a
href="http://www.crastulo.it/appuntamenti/1700_1%27aperivizi+dei+sette+vizi.html"><img
src="/media/14/77734406237187/aperivizi.jpg" alt="SETTE VIZI" height="177" width="615" /></a></div>
        <a class="thumb-btn-top-event" href="javascript:void(0);"><img src="/media/15/45182478745611/sette-vizi.jpg"
alt="SETTE VIZI" height="41" width="93" /></a>
</div>
<div class="thumb-top-event">
        <div class="image-top-event"><a
href="http://www.crastulo.it/appuntamenti/1642_il+25+giugno+inaugura+il+blufan.html"><img
src="/media/15/87553322647925/blufan.jpg" alt="BLUFAN" height="177" width="615" /></a></div>
        <a class="thumb-btn-top-event" href="javascript:void(0);"><img src="/media/15/14759601997811/blufan.jpg"
alt="BLUFAN" height="41" width="93" /></a>
</div>
<div class="thumb-top-event">
        <div class="image-top-event"><a
href="http://www.crastulo.it/appuntamenti/1480_gli+eventi+dell%27estate+2011.html"><img
src="/media/15/46870991311486/jovanotti.jpg" alt="concerti" height="177" width="615" /></a></div>
        <a class="thumb-btn-top-event" href="javascript:void(0);"><img src="/media/15/42598090854818/jovanotti.jpg"
alt="concerti" height="41" width="93" /></a>
</div>
```

I Scraping the Source Code

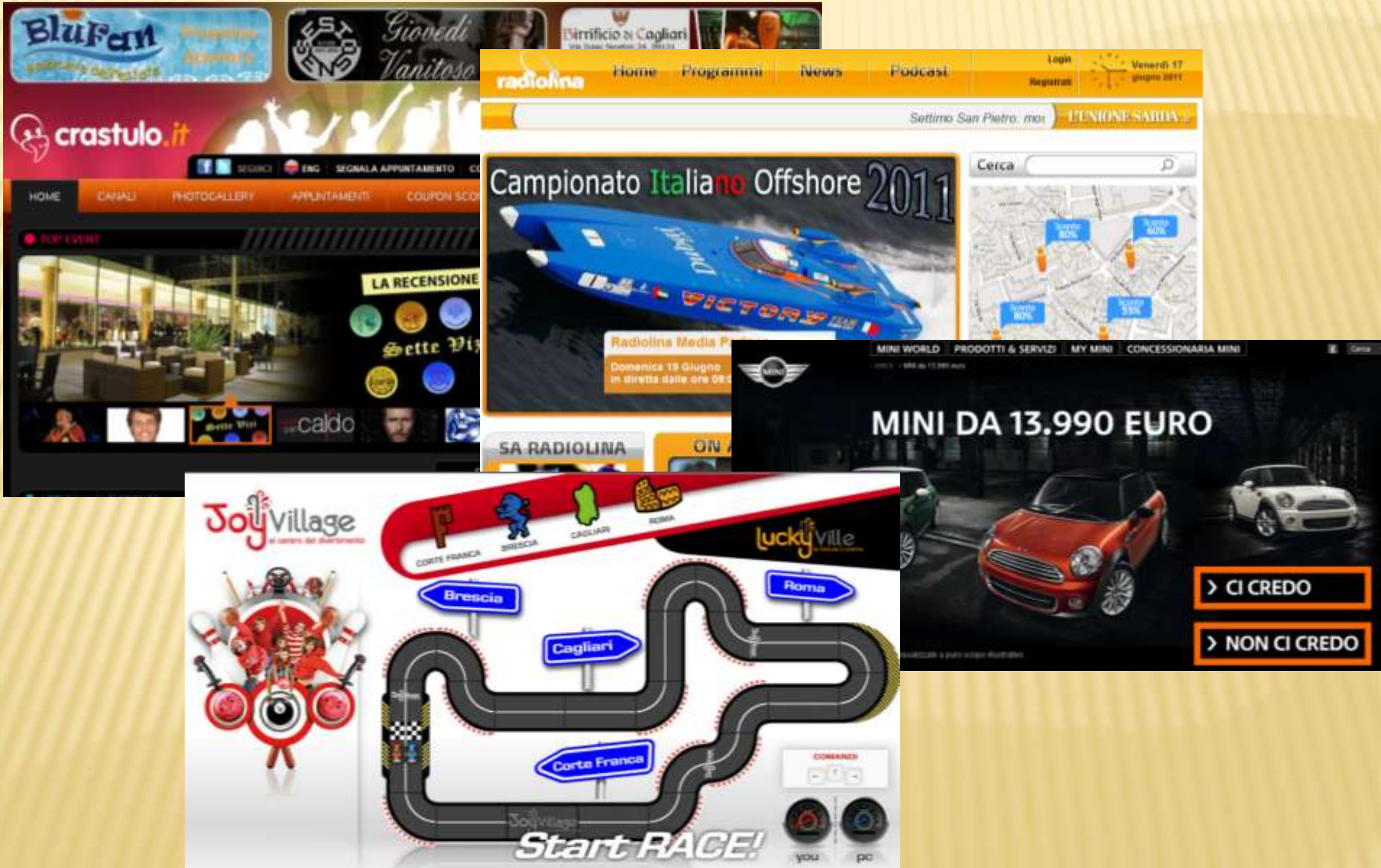
Application of scraping techniques to contextual advertising

```
tobeparsed = mechanize.urlopen(url)
body = BeautifulSoup.BeautifulSoup(tobeparsed)
body = body.prettify()
body = body[:body.find('</div>')]
link = body[body.find('<a')+len('</a>'):body.find('<br />')]
control = body[body.rfind('-->')+len('-->'):body.find('</strong>')]
link = link.replace('&#039;',"'")
link = link.split()
control = control.split()
```

Application of scraping techniques to contextual advertising

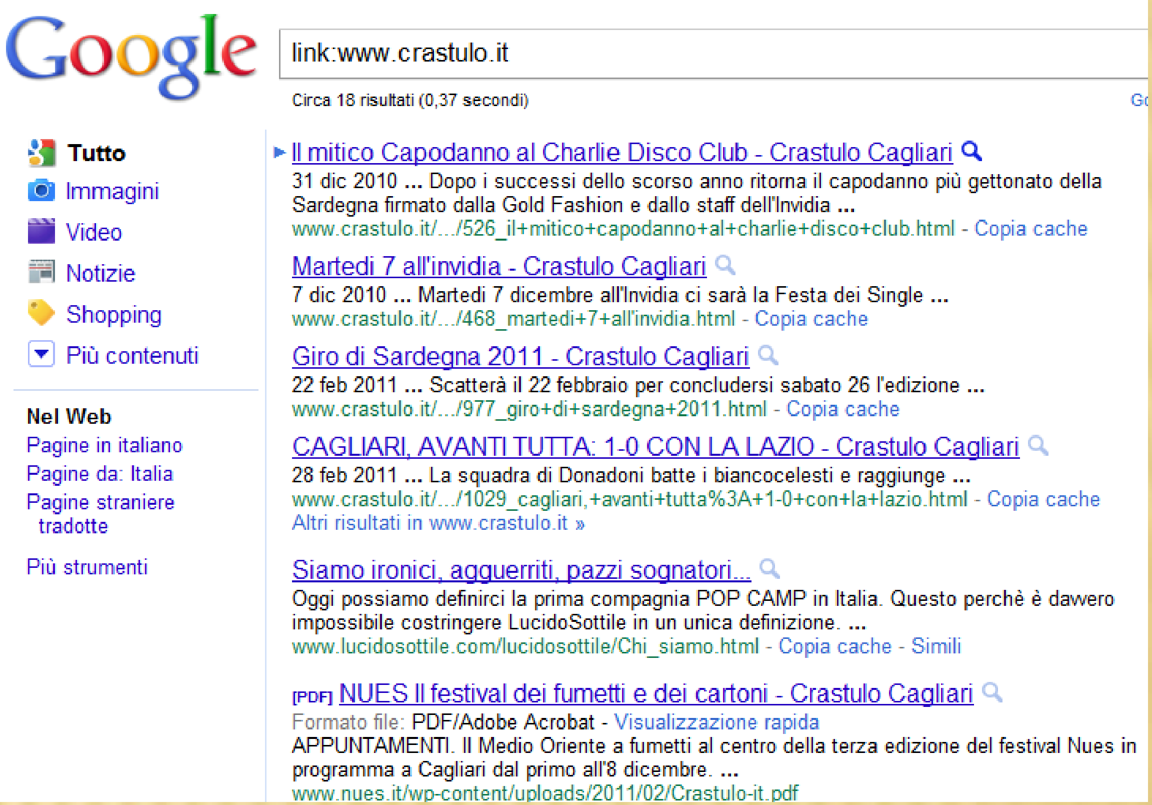
The extracted ads are then randomly selected and displayed in the target Web Page

Case study: www.crastulo.it



Case study: www.crastulo.it

Extracting inlink from : www.crastulo.i



link:www.crastulo.it

Circa 18 risultati (0,37 secondi)

Il mitico Capodanno al Charlie Disco Club - Crastulo Cagliari
31 dic 2010 ... Dopo i successi dello scorso anno ritorna il capodanno più gettonato della Sardegna firmato dalla Gold Fashion e dallo staff dell'Invidia ...
www.crastulo.it/.../526_il+mitico+capodanno+al+charlie+disco+club.html - Copia cache

Martedì 7 all'invidia - Crastulo Cagliari
7 dic 2010 ... Martedì 7 dicembre all'Invidia ci sarà la Festa dei Single ...
www.crastulo.it/.../468_martedi+7+all'invidia.html - Copia cache

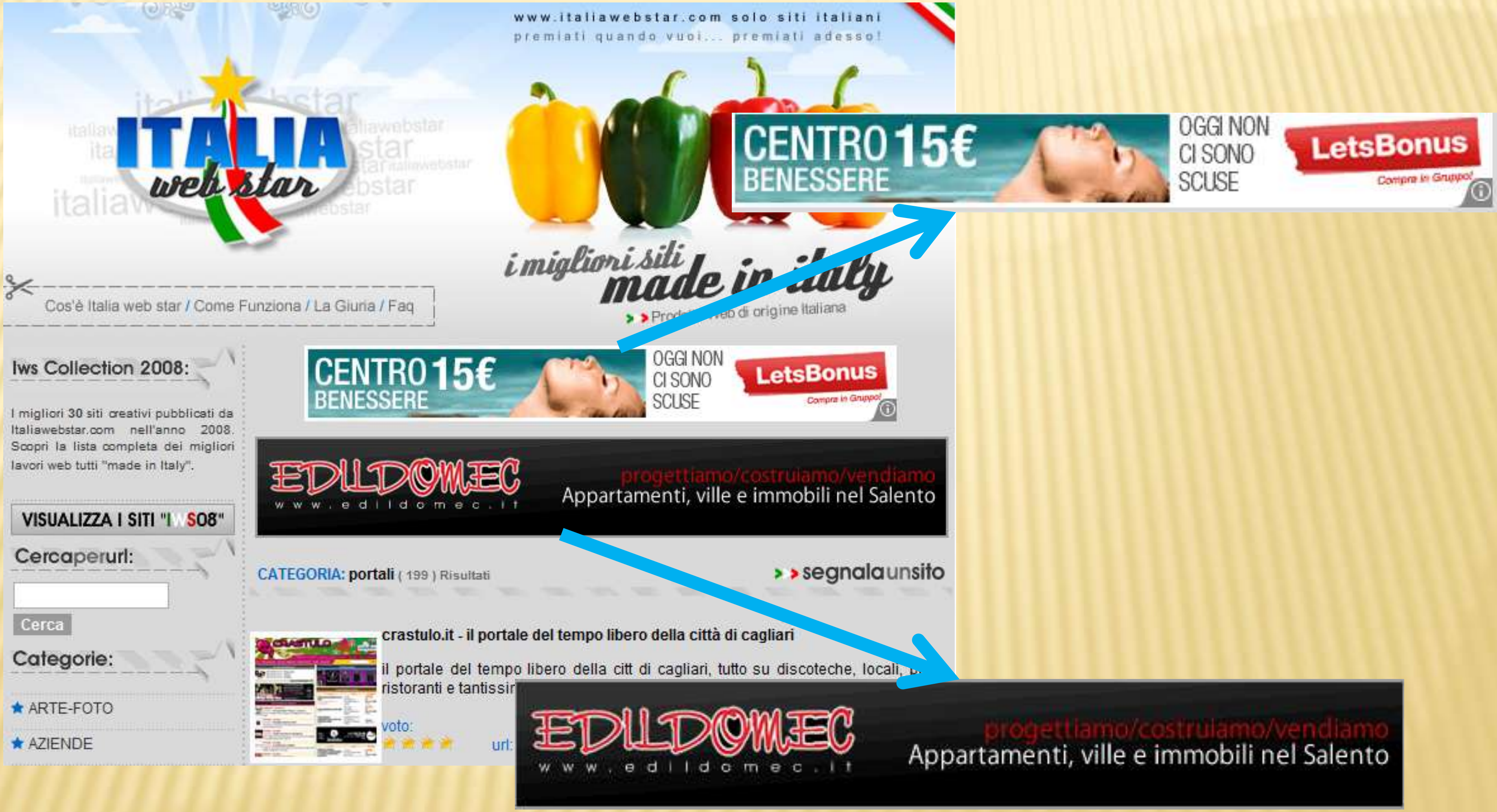
Giro di Sardegna 2011 - Crastulo Cagliari
22 feb 2011 ... Scatterà il 22 febbraio per concludersi sabato 26 l'edizione ...
www.crastulo.it/.../977_giro+di+sardegna+2011.html - Copia cache

CAGLIARI, AVANTI TUTTA: 1-0 CON LA LAZIO - Crastulo Cagliari
28 feb 2011 ... La squadra di Donadoni batte i biancocelesti e raggiunge ...
www.crastulo.it/.../1029_cagliari,+avanti+tutta%3A+1-0+con+la+lazio.html - Copia cache
Altri risultati in www.crastulo.it »

Siamo ironici, agguerriti, pazzi sognatori...
Oggi possiamo definirci la prima compagnia POP CAMP in Italia. Questo perchè è davvero impossibile costringere LucidoSottile in un unica definizione. ...
www.lucidosottile.com/lucidosottile/Chi_siamo.html - Copia cache - Simili

[PDF] NUES Il festival dei fumetti e dei cartoni - Crastulo Cagliari
Formato file: PDF/Adobe Acrobat - Visualizzazione rapida
APPUNTAMENTI. Il Medio Oriente a fumetti al centro della terza edizione del festival Nues in programma a Cagliari dal primo all'8 dicembre. ...
www.nues.it/wp-content/uploads/2011/02/Crastulo-it.pdf

Case study: www.crastulo.it



www.italiawebstar.com solo siti italiani
premiati quando vuoi... premiati adesso!

ITALIA web star

i migliori siti made in italy

Prod... web di origine italiana

Cos'è Italia web star / Come Funziona / La Giuria / Faq

Iws Collection 2008:

I migliori 30 siti creativi pubblicati da Italiawebstar.com nell'anno 2008. Scopri la lista completa dei migliori lavori web tutti "made in Italy".

VISUALIZZA I SITI "IWS08"

Cercaperurl:

Cerca

Categorie:

- ARTE-FOTO
- AZIENDE

CATEGORIA: portali (199) Risultati

segnalaunsito

crastulo.it - il portale del tempo libero della città di cagliari

il portale del tempo libero della citt di cagliari, tutto su discoteche, locali, ... ristoranti e tantissi...

voto:

url:

Case study: www.crastulo.it

Here's the four randomly selected ads

CENTRO 15€
BENESSERE
OGGI NON CI SONO SCUSE
LetsBonus
Compra in Gruppo!

PUOI CREDERCI.
MINI DA 13.990 EURO O DA 198 EURO AL MESE.
PUOI CREDERCI.
Concessionaria MINI
NUOVA SPECIAL CAR

EDILDOMEC
www.edildomec.it
progettiamo/costruiamo/vendiamo
Appartamenti, ville e immobili nel Salento

Gardaland
RESORT
RAPTOR
UN VOLO DA BRIVIDO
LE NUOVE
MONTAGNE RUSSE ALATE
SCOPRI DI PIÙ

Conclusions

- This project was aimed at suggesting suitable ads to a given Web page
- To this end I devised a system written in Python that:
  - extracts a set of inlinks of a given Web page
  - randomly selects four ads previously extracted by scraping

Future work

- To apply scraping techniques also for dynamic advertising
- To suggest ads according to users interests

Thanks to all

Contact

Contact us for details and questions on scraping in Python:

mirko.urru@hotmail.it

whitone@gmail.com

Contact Eloisa Vargiu for details and questions on contextual advertising :

vargiu@diee.unica.it

Thanks to

Università di Cagliari

ERSU Cagliari