Four-day workshop
Automated Content Analysis with Python
Day 4

Damian Trilling

d.c.trilling@uva.nl
@damian0604
`www.damiantrilling.net`

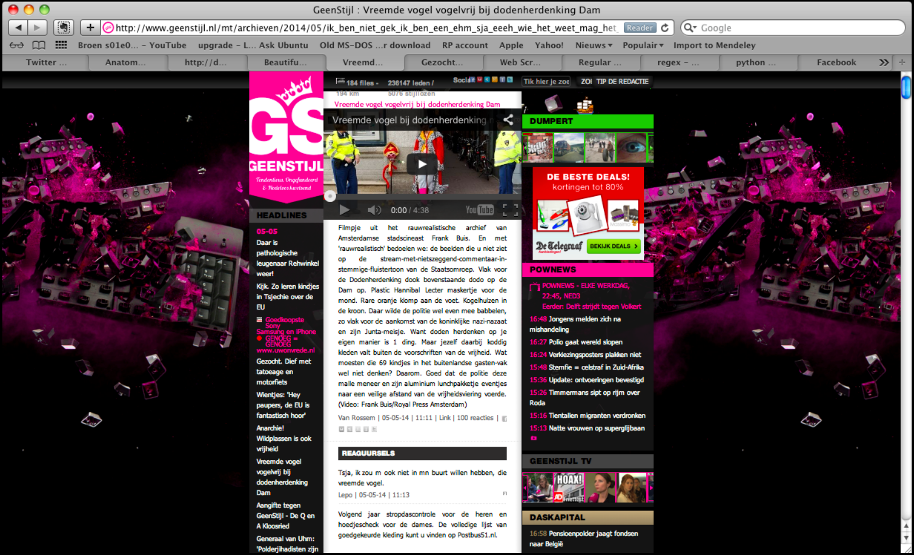
Afdeling Communicatiewetenschap
Universiteit van Amsterdam

26–9–2017

Today

1. We put on our magic cap, pretend we are Firefox, scrape all comments from GeenStijl, clean up the mess, and put the comments in a neat CSV table
2. OK, but this surely can be doe more elegantly? Yes!
3. Next steps

We put on our magic cap, pretend we are Firefox, scrape all comments from GeenStijl, clean up the mess, and put the comments in a neat CSV table



GeenStijl : Vreemde vogel vogelvrij bij dodenherdenking Dam

http://www.geenstijl.nl/mt/archieven/2014/05/ik_ben_niet_gek_ik_ben_een_ehm_sja_eeeh_wie_het_weet_mag_het

Vreemde vogel vogelvrij bij dodenherdenking Dam

Filmpje uit het rauwrealistische archief van Amsterdamse stadscineast Frank Buis. En met 'rauwrealistisch' bedoelen we: de beelden die u niet ziet op de stream-met-nietszeggend-commentaar-in-stemmige-fluistertoon van de Staatsomroep. Vlak voor de Dodenherdenking dook bovenstaande dodo op de Dam op. Plastic Hannibal Lecter maskertje voor de mond. Rare oranje klomp aan de voet. Kogelhulzen in de kroon. Daar wilde de politie wel even mee babbelen, zo vlak voor de aankomst van de koninklijke nazi-nazaat en zijn Junta-meisje. Want doden herdenken op je eigen manier is 1 ding. Maar jezelf daarbij kodig kleden valt buiten de voorschriften van de vrijheid. Wat moesten die 69 kindjes in het buitenlandse gasten-vak wel niet denken? Daarom. Goed dat de politie deze malle meneer en zijn aluminium lunchpakketje eventjes naar een veilige afstand van de vrijheidsviering voerde. (Video: Frank Buis/Royal Press Amsterdam)

Van Rossem | 05-05-14 | 11:11 | Link | 100 reacties |

REAGUURSELS

Tsja, ik zou m ook niet in mn buurt willen hebben, die vreemde vogel.

Lepo | 05-05-14 | 11:13

Volgend jaar stropdascontrole voor de heren en hoedjescheck voor de dames. De volledige lijst van goedgekeurde kleding kunt u vinden op Postbus51.nl.

HEADLINES

05-05

Daar is pathologische leugenaar Rehwinkel weer!

Kijk. Zo leren kindjes in Tsjechie over de EU

Goedkoopste Sony Samsung en iPhone GENOEG = GENOEG www.uwomvrede.nl

Gezocht. Dief met tatoeage en motorfiets

Wientjes: 'Hey paupers, de EU is fantastisch hoor'

Anarchie! Wildplassen is ook vrijheid

Vreemde vogel vogelvrij bij dodenherdenking Dam

Aangifte tegen GeenStijl - De Q en A Kloosried

Generaal van Uhm: 'Polderjihadisten zijn

DUMPERT

DE BESTE DEALS! kortingen tot 80%

De Telegraaf Aanbiedingen! BEKIJK DEALS

POWNEWS

POWNEWS - ELKE WERKDAG, 22:45, NED3

Eerder: Delft strijdt tegen Volkert

16:48 Jongens melden zich na mishandeling

16:27 Polio gaat wereld slopen

16:24 Verkiezingsposters plakken niet

15:48 Stemfie = celstraf in Zuid-Afrika

15:36 Update: ontvoeringen bevestigd

15:26 Timmermans sipt op rijm over Roda

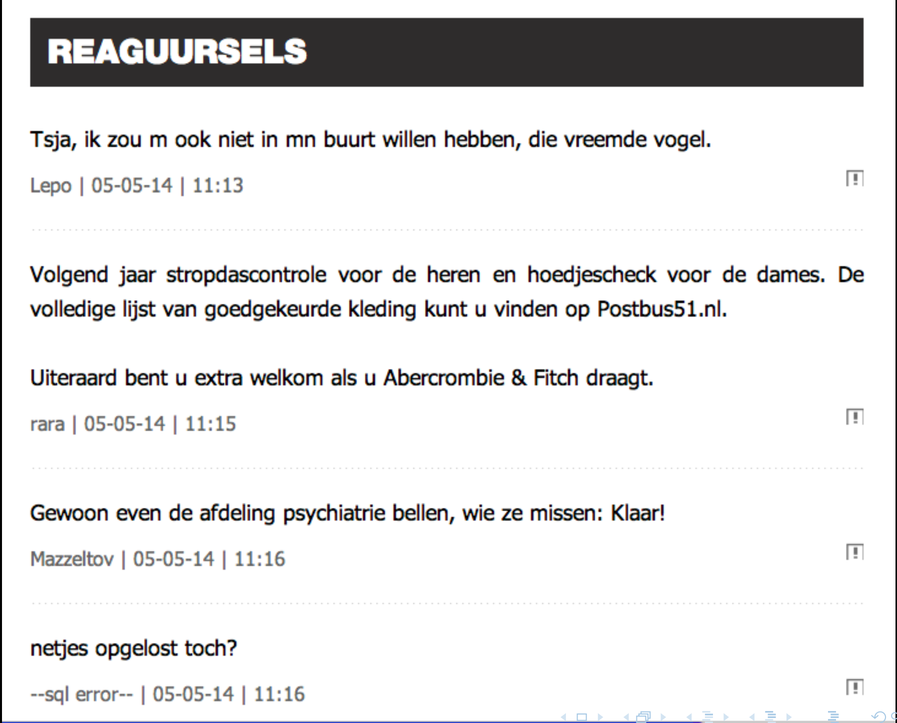
15:16 Tientallen migranten verdronken

15:13 Natte vrouwen op superglijbaan

GEENSTIJL TV

DASKAPITAL

16:58 Pensioenpolder jaagt fondsen naar België



REAGUURSELS

Tsja, ik zou m ook niet in mn buurt willen hebben, die vreemde vogel.

Lepo | 05-05-14 | 11:13

Volgend jaar stropdascontrole voor de heren en hoedjescheck voor de dames. De volledige lijst van goedgekeurde kleding kunt u vinden op Postbus51.nl.

Uiteraard bent u extra welkom als u Abercrombie & Fitch draagt.

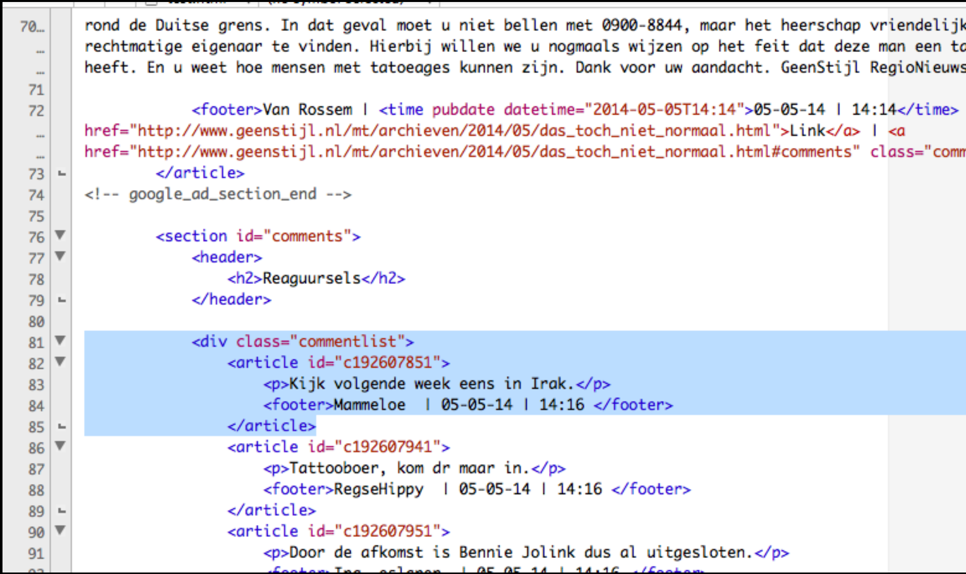
rara | 05-05-14 | 11:15

Gewoon even de afdeling psychiatrie bellen, wie ze missen: Klaar!

Mazzeltov | 05-05-14 | 11:16

netjes opgelost toch?

--sql error-- | 05-05-14 | 11:16



```
rond de Duitse grens. In dat geval moet u niet bellen met 0900-8844, maar het heerschap vriendelijk
rechtmatige eigenaar te vinden. Hierbij willen we u nogmaals wijzen op het feit dat deze man een ta
heeft. En u weet hoe mensen met tatoeages kunnen zijn. Dank voor uw aandacht. GeenStijl RegioNieuws

            <footer>Van Rossem | <time pubdate datetime="2014-05-05T14:14">05-05-14 | 14:14</time>
href="http://www.geenstijl.nl/mt/archieven/2014/05/das_toch_niet_normaal.html">Link</a> | <a
href="http://www.geenstijl.nl/mt/archieven/2014/05/das_toch_niet_normaal.html#comments" class="comm
        </article>
<!-- google_ad_section_end -->

        <section id="comments">
            <header>
                <h2>Reaguursels</h2>
            </header>

            <div class="commentlist">
                <article id="c192607851">
                    <p>Kijk volgende week eens in Irak.</p>
                    <footer>Mammeloe  | 05-05-14 | 14:16 </footer>
                </article>
                <article id="c192607941">
                    <p>Tattooboer, kom dr maar in.</p>
                    <footer>RegseHippy  | 05-05-14 | 14:16 </footer>
                </article>
                <article id="c192607951">
                    <p>Door de afkomst is Bennie Jolink dus al uitgesloten.</p>
```

Let's make a plan!

Which elements from the page do we need?

- What do they mean?
- How are they represented in the source code?

How should our output look like?

- What *lists* do we want?
- ...

And how can we achieve this?

Operation Magic Cap

Operation Magic Cap

1. Download the page

Operation Magic Cap

1. Download the page
   - They might block us, so let's do as if we were a web browser!

Operation Magic Cap

1. Download the page
   - They might block us, so let's do as if we were a web browser!
2. Remove all line breaks (\n, but maybe also \n\r or \r) and TABs (\t): We want one long string

Operation Magic Cap

1. Download the page
   - They might block us, so let's do as if we were a web browser!
2. Remove all line breaks (`\n`, but maybe also `\n\r` or `\r`) and TABs (`\t`): We want one long string
3. Isolate the comment section (it started with `<div class="commentlist">` and ended with `</div>`)

Operation Magic Cap

1. Download the page
   - They might block us, so let's do as if we were a web browser!
2. Remove all line breaks (`\n`, but maybe also `\n\r` or `\r`) and TABs (`\t`): We want one long string
3. Isolate the comment section (it started with `<div class="commentlist">` and ended with `</div>`)
4. Within the comment section, identify each comment (`<article>`)

Operation Magic Cap

1. Download the page
    - They might block us, so let's do as if we were a web browser!
2. Remove all line breaks (\n, but maybe also \n\r or \r) and TABs (\t): We want one long string
3. Isolate the comment section (it started with `<div class="commentlist">` and ended with `</div>`)
4. Within the comment section, identify each comment (`<article>`)
5. Within each comment, seperate the text (`<p>`) from the metadata `<footer>`)

Operation Magic Cap

1. Download the page
   - They might block us, so let's do as if we were a web browser!
2. Remove all line breaks (\n, but maybe also \n\r or \r) and TABs (\t): We want one long string
3. Isolate the comment section (it started with `<div class="commentlist">` and ended with `</div>`)
4. Within the comment section, identify each comment (`<article>`)
5. Within each comment, seperate the text (`<p>`) from the metadata `<footer>`)
6. Put text and metadata in lists and save them to a csv file

Operation Magic Cap

1. Download the page
    - They might block us, so let's do as if we were a web browser!
2. Remove all line breaks (\n, but maybe also \n\r or \r) and TABs (\t): We want one long string
3. Isolate the comment section (it started with `<div class="commentlist">` and ended with `</div>`)
4. Within the comment section, identify each comment (`<article>`)
5. Within each comment, seperate the text (`<p>`) from the metadata `<footer>`)
6. Put text and metadata in lists and save them to a csv file

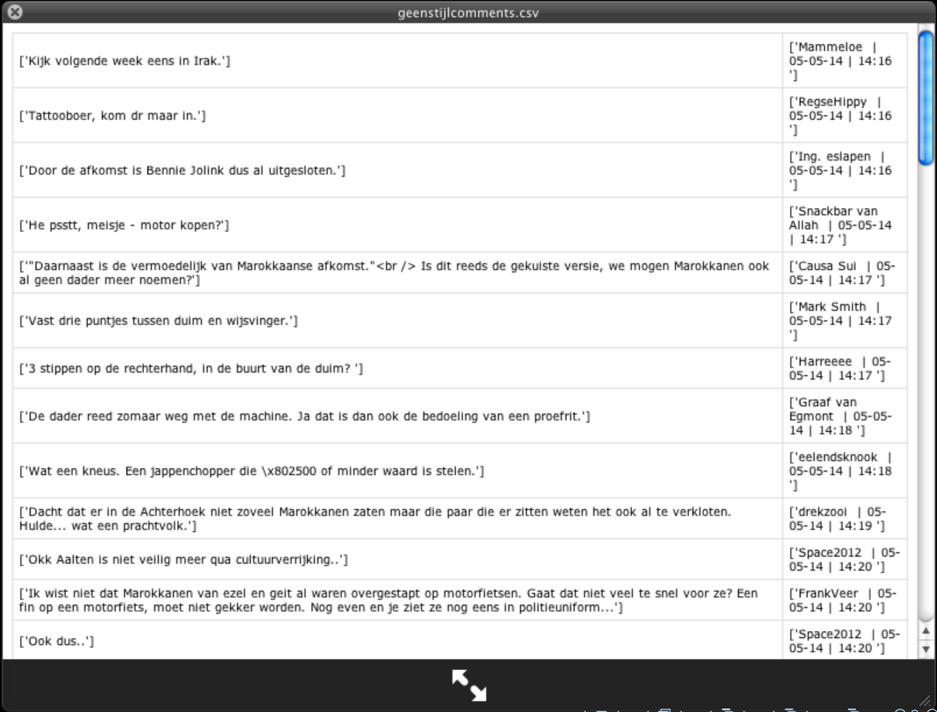
And how can we achieve this?

```
from urllib import request
import re
import csv

onlycommentslist=[]
metalist=[]

req = request.Request("http://www.geenstijl.nl/mt/archieven/2014/05/
    das_toch_niet_normaal.html", headers={"User-Agent" : "Mozilla
    /5.0"})
tekst=request.urlopen(req).read()
tekst=tekst.decode(encoding="utf-8",errors="ignore").replace("\n"," ").
    replace("\t"," ")

commentsection=re.findall(r'<div class="commentlist">.*?</div>',tekst)
print (commentsection)
comments=re.findall(r'<article.*?>(.*?)</article>',commentsection[0])
print (comments)
print ("There are",len(comments),"comments")
for co in comments:
   metalist.append(re.findall(r'<footer>(.*?)</footer>',co))
   onlycommentslist.append(re.findall(r'<p>(.*?)</p>',co))
writer=csv.writer(open("geenstijlcomments.csv",mode="w",encoding="utf
    -8"))
output=zip(onlycommentslist,metalist)
writer.writerows(output)
```



['Kijk volgende week eens in Irak.']	['Mammeloe | 05-05-14 | 14:16 ']
['Tattooboer, kom dr maar in.']	['RegseHippy | 05-05-14 | 14:16 ']
['Door de afkomst is Bennie Jolink dus al uitgesloten.']	['Ing. eslapen | 05-05-14 | 14:16 ']
['He psstt, meisje - motor kopen?']	['Snackbar van Allah | 05-05-14 | 14:17 ']
['"Daarnaast is de vermoedelijk van Marokkaanse afkomst."<br /> Is dit reeds de gekuiste versie, we mogen Marokkanen ook al geen dader meer noemen?']	['Causa Sui | 05-05-14 | 14:17 ']
['Vast drie puntjes tussen duim en wijsvinger.']	['Mark Smith | 05-05-14 | 14:17 ']
['3 stippen op de rechterhand, in de buurt van de duim? ']	['Harreeee | 05-05-14 | 14:17 ']
['De dader reed zomaar weg met de machine. Ja dat is dan ook de bedoeling van een proefrit.']	['Graaf van Egmont | 05-05-14 | 14:18 ']
['Wat een kneus. Een jappenchopper die \x802500 of minder waard is stelen.']	['eelendsknook | 05-05-14 | 14:18 ']
['Dacht dat er in de Achterhoek niet zoveel Marokkanen zaten maar die paar die er zitten weten het ook al te verkloten. Hulde... wat een prachtvolk.']	['drekzooi | 05-05-14 | 14:19 ']
['Okk Aalten is niet veilig meer qua cultuurverrijking..']	['Space2012 | 05-05-14 | 14:20 ']
['Ik wist niet dat Marokkanen van ezel en geit al waren overgestapt op motorfietsen. Gaat dat niet veel te snel voor ze? Een fin op een motorfiets, moet niet gekker worden. Nog even en je ziet ze nog eens in politieuniform...']	['FrankVeer | 05-05-14 | 14:20 ']
['Ook dus..']	['Space2012 | 05-05-14 | 14:20 ']

Some remarks

The regexp

- `.*?` instead of `.*` means *lazy* matching. As `.*` matches everything, the part where the regexp should stop would not be analyzed (*greedy* matching) – we would get the whole rest of the document (or the line, but we removed all line breaks).

Some remarks

The regexp

- `.*?` instead of `.*` means *lazy* matching. As `.*` matches everything, the part where the regexp should stop would not be analyzed (*greedy* matching) – we would get the whole rest of the document (or the line, but we removed all line breaks).
- The parentheses in `(.*?)` make sure that the function only returns what's between them and not the surrounding stuff (like `<footer>` and `</footer>`)

Some remarks

The regexp

- `.*?` instead of `.*` means *lazy* matching. As `.*` matches everything, the part where the regexp should stop would not be analyzed (*greedy* matching) – we would get the whole rest of the document (or the line, but we removed all line breaks).
- The parentheses in `(.*?)` make sure that the function only returns what's between them and not the surrounding stuff (like `<footer>` and `</footer>`)

Optimization

- Only save the 0th (and only) element of the list
- Seperate the username and interpret date and time

Further reading

Doing this with other sites?

- It's basically puzzling with regular expressions.
- Look at the source code of the website to see how well-structured it is.

OK, but this surely can be doe more elegantly? Yes!

Scraping

Geenstijl-example

- Worked well (and we could do it with the knowledge we already had)

Scraping

Geenstijl-example

- Worked well (and we could do it with the knowledge we already had)
- But we can also use existing parsers (that can interpret the structure of the html page)

Scraping

Geenstijl-example

- Worked well (and we could do it with the knowledge we already had)
- But we can also use existing parsers (that can interpret the structure of the html page)
- especially when the structure of the site is more complex

Scraping

Geenstijl-example

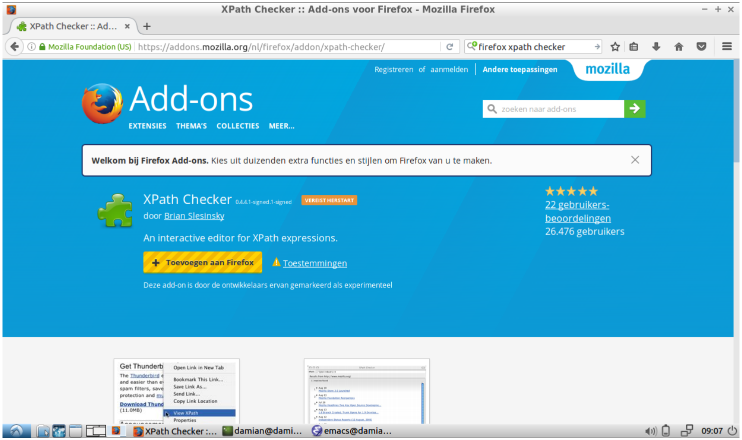
- Worked well (and we could do it with the knowledge we already had)
- But we can also use existing parsers (that can interpret the structure of the html page)
- especially when the structure of the site is more complex

The following example is based on `http://www.chicagoreader.com/chicago/best-of-chicago-2011-food-drink/BestOf?oid=4106228`. It uses the module `lxml`

What do we need?

- the URL (of course)
- the XPATH of the element we want to scrape (you'll see in a minute what this is)

Let’s install a useful add-on: the Firefox XPath Checker


XPath Checker :: Add-ons voor Firefox - Mozilla Firefox

https://addons.mozilla.org/nl/firefox/addon/xpath-checker/

Welkom bij Firefox Add-ons. Kies uit duizenden extra functies en stijlen om Firefox van u te maken.

XPath Checker 0.4.4.1-signed.1-signed VEREIST HERSTART

door Brian Slesinsky

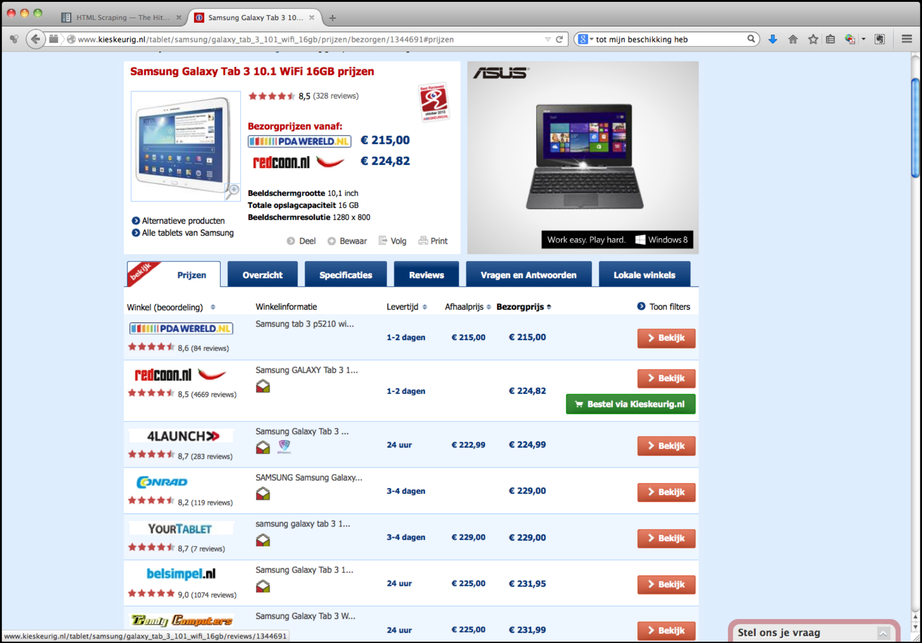
An interactive editor for XPath expressions.

22 gebruikers-beoordelingen

26.476 gebruikers

Deze add-on is door de ontwikkelaars ervan gemarkeerd als experimenteel

. . . and after that, inspect the website we’re interested in:



HTML Scraping — The Hit... | Samsung Galaxy Tab 3 10...

www.kieskeurig.nl/tablet/samsung/galaxy_tab_3_101_wifi_16gb/prijzen/bezorgen/1344691#prijzen

tot mijn beschikking heb

Samsung Galaxy Tab 3 10.1 WiFi 16GB prijzen

★★★★½ 8,5 (328 reviews)

Bezorgprijzen vanaf:

PDA WERELD.NL	**€ 215,00**
redcoon.nl	**€ 224,82**

Beeldschermgrootte 10,1 inch
Totale opslagcapaciteit 16 GB
Beeldschermresolutie 1280 x 800

Alternatieve producten
Alle tablets van Samsung

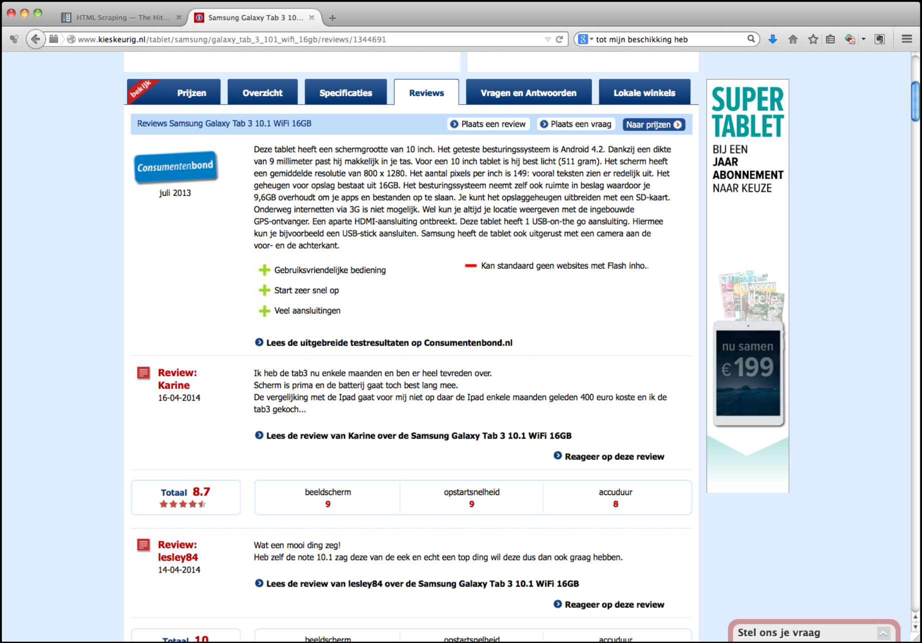
Deel Bewaar Volg Print

Prijzen | Overzicht | Specificaties | Reviews | Vragen en Antwoorden | Lokale winkels

Winkel (beoordeling)	Winkelinformatie	Levertijd	Afhaalprijs	**Bezorgprijs**	Toon filters
PDA WERELD.NL ★★★★½ 8,6 (84 reviews)	Samsung tab 3 p5210 wi...	**1-2 dagen**	**€ 215,00**	**€ 215,00**	Bekijk
redcoon.nl ★★★★½ 8,5 (4669 reviews)	Samsung GALAXY Tab 3 1...	**1-2 dagen**		**€ 224,82**	Bekijk / Bestel via Kieskeurig.nl
4LAUNCH ★★★★½ 8,7 (283 reviews)	Samsung Galaxy Tab 3 ...	**24 uur**	**€ 222,99**	**€ 224,99**	Bekijk
CONRAD ★★★★½ 8,2 (119 reviews)	SAMSUNG Samsung Galaxy...	**3-4 dagen**		**€ 229,00**	Bekijk
YOURTABLET ★★★★½ 8,7 (7 reviews)	samsung galaxy tab 3 1...	**3-4 dagen**	**€ 229,00**	**€ 229,00**	Bekijk
belsimpel.nl ★★★★★ 9,0 (1074 reviews)	Samsung Galaxy Tab 3 1...	**24 uur**	**€ 225,00**	**€ 231,95**	Bekijk
Dandy Computers	Samsung Galaxy Tab 3 W...	**24 uur**	**€ 225,00**	**€ 231,99**	Bekijk

www.kieskeurig.nl/tablet/samsung/galaxy_tab_3_101_wifi_16gb/reviews/1344691

Stel ons je vraag



Reviews Samsung Galaxy Tab 3 10.1 WiFi 16GB

Consumentenbond

juli 2013

Deze tablet heeft een schermgrootte van 10 inch. Het geteste besturingssysteem is Android 4.2. Dankzij een dikte van 9 millimeter past hij makkelijk in je tas. Voor een 10 inch tablet is hij best licht (511 gram). Het scherm heeft een gemiddelde resolutie van 800 x 1280. Het aantal pixels per inch is 149: vooral teksten zien er redelijk uit. Het geheugen voor opslag bestaat uit 16GB. Het besturingssysteem neemt zelf ook ruimte in beslag waardoor je 9,6GB overhoudt om je apps en bestanden op te slaan. Je kunt het opslaggeheugen uitbreiden met een SD-kaart. Onderweg internetten via 3G is niet mogelijk. Wel kun je altijd je locatie weergeven met de ingebouwde GPS-ontvanger. Een aparte HDMI-aansluiting ontbreekt. Deze tablet heeft 1 USB-on-the go aansluiting. Hiermee kun je bijvoorbeeld een USB-stick aansluiten. Samsung heeft de tablet ook uitgerust met een camera aan de voor- en de achterkant.

- Gebruiksvriendelijke bediening
- Start zeer snel op
- Veel aansluitingen

− Kan standaard geen websites met Flash inho..

Lees de uitgebreide testresultaten op Consumentenbond.nl

Review: Karine
16-04-2014

Ik heb de tab3 nu enkele maanden en ben er heel tevreden over.
Scherm is prima en de batterij gaat toch best lang mee.
De vergelijking met de Ipad gaat voor mij niet op daar de Ipad enkele maanden geleden 400 euro koste en ik de tab3 gekoch...

Lees de review van Karine over de Samsung Galaxy Tab 3 10.1 WiFi 16GB

Reageer op deze review

Totaal 8.7	beeldscherm	opstartsnelheid	accuduur
★★★★½	9	9	8

Review: lesley84
14-04-2014

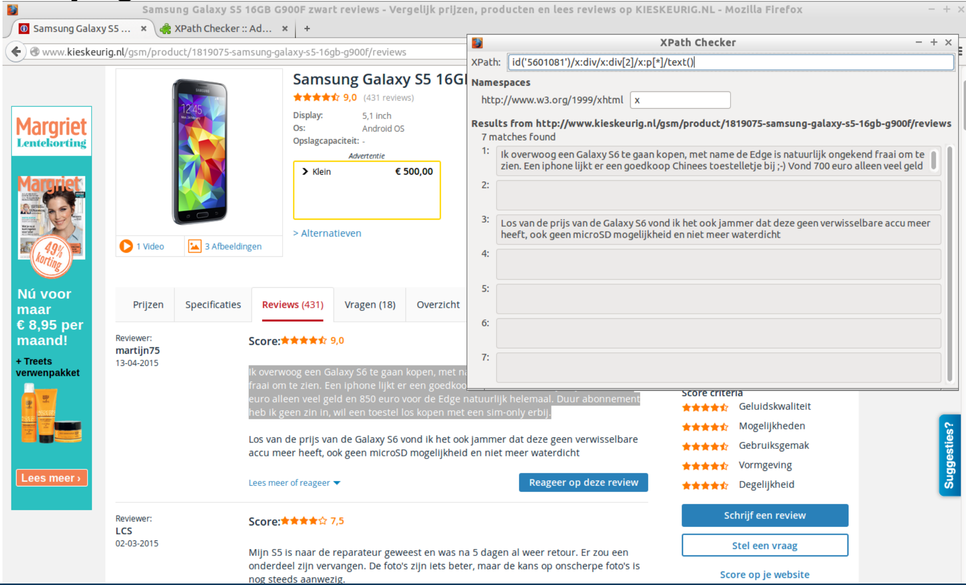
Wat een mooi ding zeg!
Heb zelf de note 10.1 zag deze van de eek en echt een top ding wil deze dus dan ook graag hebben.

Lees de review van lesley84 over de Samsung Galaxy Tab 3 10.1 WiFi 16GB

Reageer op deze review

Totaal 10	beeldscherm	opstartsnelheid	accuduur

Playing around with the Firefox XPath Checker



Samsung Galaxy S5 16GB G900F zwart reviews - Vergelijk prijzen, producten en lees reviews op KIESKEURIG.NL - Mozilla Firefox

www.kieskeurig.nl/gsm/product/1819075-samsung-galaxy-s5-16gb-g900f/reviews

XPath Checker

XPath: id('5601081')/x:div/x:div[2]/x:p[*]/text()

Namespaces

http://www.w3.org/1999/xhtml x

Results from http://www.kieskeurig.nl/gsm/product/1819075-samsung-galaxy-s5-16gb-g900f/reviews

7 matches found

1: Ik overwoog een Galaxy S6 te gaan kopen, met name de Edge is natuurlijk ongekend fraai om te zien. Een iphone lijkt er een goedkoop Chinees toestelletje bij ;-) Vond 700 euro alleen veel geld

2:

3: Los van de prijs van de Galaxy S6 vond ik het ook jammer dat deze geen verwisselbare accu meer heeft, ook geen microSD mogelijkheid en niet meer waterdicht

4:

5:

6:

7:

Samsung Galaxy S5 16G

9,0 (431 reviews)

Display: 5,1 inch
Os: Android OS
Opslagcapaciteit: -

Reviewer: martijn75 13-04-2015

Score: 9,0

Ik overwoog een Galaxy S6 te gaan kopen, met na… fraai om te zien. Een iphone lijkt er een goedkoo… euro alleen veel geld en 850 euro voor de Edge natuurlijk helemaal. Duur abonnement heb ik geen zin in, wil een toestel los kopen met een sim-only erbij.

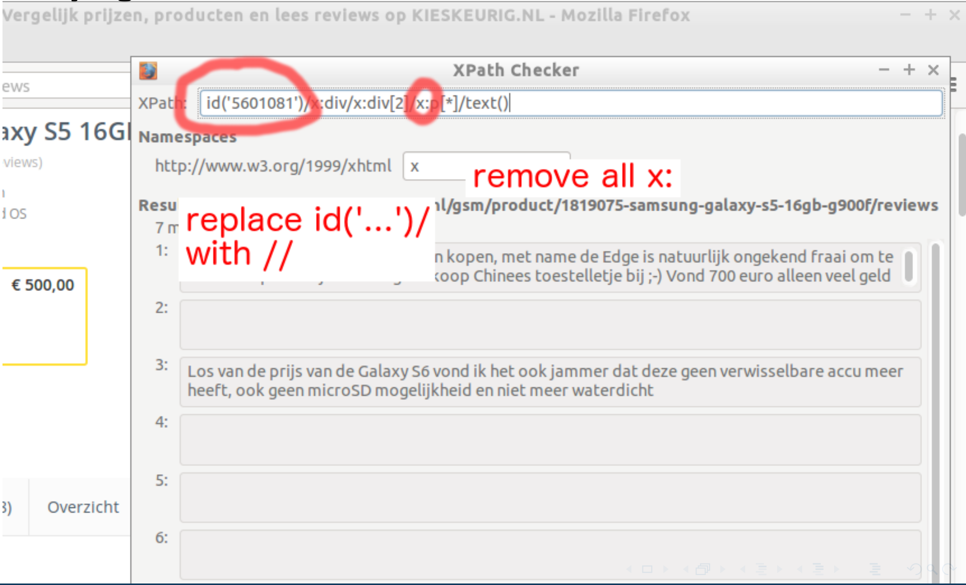
Los van de prijs van de Galaxy S6 vond ik het ook jammer dat deze geen verwisselbare accu meer heeft, ook geen microSD mogelijkheid en niet meer waterdicht

Reviewer: LCS 02-03-2015

Score: 7,5

Mijn S5 is naar de reparateur geweest en was na 5 dagen al weer retour. Er zou een onderdeel zijn vervangen. De foto's zijn iets beter, maar de kans op onscherpe foto's is nog steeds aanwezig.

Playing around with the Firefox XPath Checker



Vergelijk prijzen, producten en lees reviews op KIESKEURIG.NL - Mozilla Firefox

XPath Checker

XPath: id('5601081')/x:div/x:div[2]/x:p[*]/text()

Namespaces

http://www.w3.org/1999/xhtml x

remove all x:

replace id('...')/ with //

Resu l/gsm/product/1819075-samsung-galaxy-s5-16gb-g900f/reviews

7 m

1: n kopen, met name de Edge is natuurlijk ongekend fraai om te koop Chinees toestelletje bij ;-) Vond 700 euro alleen veel geld

2:

3: Los van de prijs van de Galaxy S6 vond ik het ook jammer dat deze geen verwisselbare accu meer heeft, ook geen microSD mogelijkheid en niet meer waterdicht

4:

5:

6:

Playing around with the Firefox XPath Checker

Some things to play around with:

- // means ‘arbitrary depth’ (=may be nested in many higher levels)

Playing around with the Firefox XPath Checker

Some things to play around with:

- // means ‘arbitrary depth’ (=may be nested in many higher levels)
- * means ‘anything’. (`p[2]` is the second paragraph, `p[*]` are all

Playing around with the Firefox XPath Checker

Some things to play around with:

- `//` means ‘arbitrary depth’ (=may be nested in many higher levels)
- `*` means ‘anything’. (`p[2]` is the second paragraph, `p[*]` are all
- If you want to refer to a specific attribute of a HTML tag, you can use `@`. For example, every `*[@id="reviews-container"]` would grap a tag like `<div id=reviews-container" class="’user-content’`

Playing around with the Firefox XPath Checker

Some things to play around with:

- // means 'arbitrary depth' (=may be nested in many higher levels)
- * means 'anything'. (`p[2]` is the second paragraph, `p[*]` are all
- If you want to refer to a specific attribute of a HTML tag, you can use `@`. For example, every `*[@id="reviews-container"]` would grap a tag like `<div id=reviews-container" class="'user-content'`
- Let the XPATH end with `/text()` to get all text

Playing around with the Firefox XPath Checker

Some things to play around with:

- `//` means 'arbitrary depth' (=may be nested in many higher levels)
- `*` means 'anything'. (`p[2]` is the second paragraph, `p[*]` are all
- If you want to refer to a specific attribute of a HTML tag, you can use `@`. For example, every `*[@id="reviews-container"]` would grap a tag like `<div id=reviews-container" class="'user-content'`
- Let the XPATH end with `/text()` to get all text
- Have a look at the source code of the web page to think of other possible XPATHs!

The XPATH

You get something like

```
//*[@id="tabbedReviewsDiv"]/dl[1]/dd
//*[@id="tabbedReviewsDiv"]/dl[2]/dd
```

The XPATH

You get something like

```
//*[@id="tabbedReviewsDiv"]/dl[1]/dd
//*[@id="tabbedReviewsDiv"]/dl[2]/dd
```

The `*` means “every”.
Also, to get the text of the element, the XPATH should end on `/text()`.

The XPATH

You get something like

```
//*[@id="tabbedReviewsDiv"]/dl[1]/dd
//*[@id="tabbedReviewsDiv"]/dl[2]/dd
```

The * means “every”.
Also, to get the text of the element, the XPATH should end on `/text()`.

We can infer that we (probably) get all comments with

```
//*[@id="tabbedReviewsDiv"]/dl[*]/dd/text()
```

Let's scrape them!

```
from lxml import html
from urllib import request

req=request.Request("http://www.kieskeurig.nl/smartphone/product
     /2334455-apple-iphone-6/reviews")
tree =

html.fromstring(request.urlopen(req).read().decode(encoding="utf-8",
     errors="ignore"))
reviews = tree.xpath('//*[@class="reviews-single__text"]/text()')

# remove empty reviews and remove leading/trailing whitespace
reviews = [r.strip() for r in reviews if r.strip()!=""]

print (len(reviews),"reviews scraped. Showing the first 60 characters of
      each:")
i=0
for review in reviews:
    print("Review",i,":",review[:60])
    i+=1
```

The output – perfect!

```
63 reviews scraped. Showing the first 60 characters of each:
Review 0 : Apple maakt mooie toestellen met hard- en software uitsteken
Review 1 : Vanaf de iPhone 4 ben ik erg te spreken over de intuitieve i
Review 2 : Helaas ontbreekt het Apple toestellen aan een noodzakelijk i
Review 3 : Met een enorme mate van pech hebben wij als (beschaafd!!) ge
```

Recap

General idea

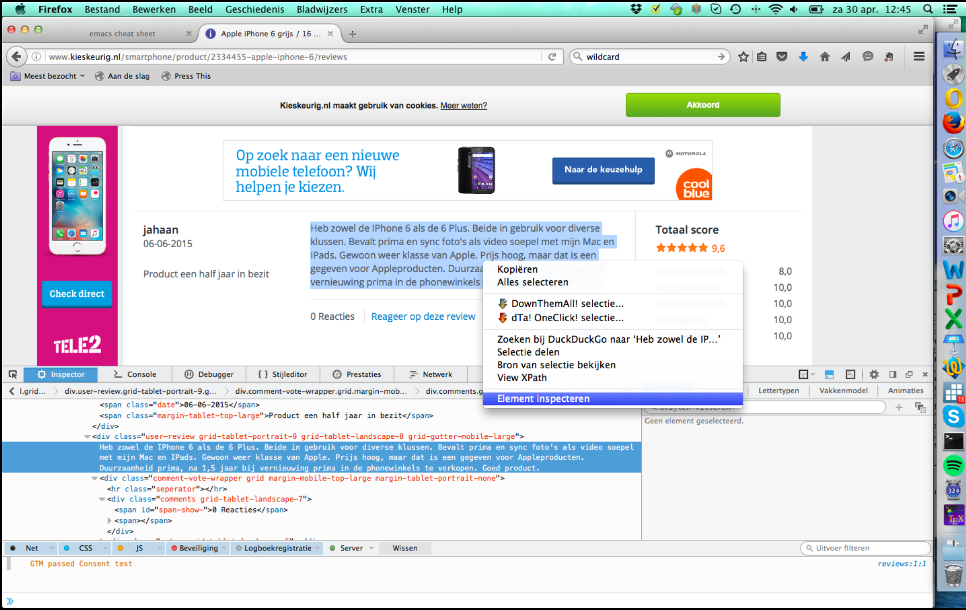
1. Identify each element by its XPATH (look it up in your browser)
2. Read the webpage into a (looooong) string
3. Use the XPATH to extract the relevant text into a list (with a module like lxml)
4. Do something with the list (preprocess, analyze, save)

Alternatives: scrapy, beautifulsoup, regular expressions, ...

Last remarks

There is often more than one way to specify an XPATH

1. You can usually leave away the namespace (the `x:`)
2. Sometimes, you might want to use a different suggestion to be able to generalize better (e.g., using the *attributes* rather than the *tags*)
3. in that case, it makes sense to look deeper into the structure of the HTML code, for example with "Inspect Element" and use that information to play around with in the XPATH Checker



Kieskeurig.nl maakt gebruik van cookies. Meer weten?

Op zoek naar een nieuwe mobiele telefoon? Wij helpen je kiezen.

jahaan
06-06-2015

Product een half jaar in bezit

Heb zowel de IPhone 6 als de 6 Plus. Beide in gebruik voor diverse klussen. Bevalt prima en sync foto's als video soepel met mijn Mac en IPads. Gewoon weer klasse van Apple. Prijs hoog, maar dat is een gegeven voor Appleproducten. Duurzaa

vernieuwing prima in de phonewinkels

0 Reacties Reageer op deze review

Totaal score
★★★★★ 9,6

8,0
10,0
10,0
10,0
10,0

l.grid... > div.user-review.grid-tablet-portrait-9.g... > div.comment-vote-wrapper.grid.margin-mob... > div.comments.g

```
        <span class="date">06-06-2015</span>
        <span class="margin-tablet-top-large">Product een half jaar in bezit</span>
      </div>
    <div class="user-review grid-tablet-portrait-9 grid-tablet-landscape-8 grid-gutter-mobile-large">
      Heb zowel de IPhone 6 als de 6 Plus. Beide in gebruik voor diverse klussen. Bevalt prima en sync foto's als video soepel
      met mijn Mac en IPads. Gewoon weer klasse van Apple. Prijs hoog, maar dat is een gegeven voor Appleproducten.
      Duurzaamheid prima, na 1,5 jaar bij vernieuwing prima in de phonewinkels te verkopen. Goed product.
      <div class="comment-vote-wrapper grid margin-mobile-top-large margin-tablet-portrait-none">
        <hr class="seperator"></hr>
        <div class="comments grid-tablet-landscape-7">
          <span id="span-show-">0 Reacties</span>
          <span></span>
        </div>
```

Geen element geselecteerd.

```
GTM passed Consent test                                                    reviews:1:1
```