Big Data and Automated Content Analysis

Week 6 – Monday

»Web scraping«

Damian Trilling

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

Afdeling Communicatiewetenschap
Universiteit van Amsterdam

12 March 2018

Today

1. A first intro to parsing webpages
2. OK, but this surely can be doe more elegantly? Yes!
3. Next steps

Let's have a look of a webpage with comments and try to understand the underlying structure.


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
Reader
Google
Broen s01e0... - YouTube
upgrade - L... Ask Ubuntu
Old MS-DOS ...r download
RP account
Apple
Yahoo!
Nieuws
Populair
Import to Mendeley
Twitter ...
Anatom...
http://d...
Beautifu...
Vreemd...
Gezocht...
Web Scr...
Regular ...
regex - ...
python ...
Facebook
GEENSTIJL
Tendentieus. Ongefundeerd & Nodeloos kwetsend
184 files - 236147 leden /
194 km
5076 stijlfiezen
Tik hier je zoe
ZOI TIP DE REDACTIE
Vreemde vogel vogelvrij bij dodenherdenking Dam
Vreemde vogel bij dodenherdenking
0:00 / 4:38
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
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 koddig 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.
DUMPERT
DE BESTE DEALS!
kortingen tot 80%
De Telegraaf
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
HOAX!
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


view-source:https://www.geenstijl.nl/3945571/das_toch_niet_normaal/
Aan de slag
Press This
Atlassian Cloud
an class="divider">|
an class="datetime">05-05-14 | 23:34
class="reportcomment" title="Deze reactie is in overtreding met de huisregels.">
er>
e>
article class="comment col-xs-12 no-y-padding"
commentid="192631041"
92631041">
ss="anchor-pos" id="cid_192631041">
class="cmt-content"><p>Een VZ800 Marauder, wat een giga kutmotor is dat. Na 10 km. heb je al
r>
pan class="username">tiswat

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?

A first plan

A first plan

1. Download the page

A first plan

1. Download the page
   - Possibly taking measures to deal with cookie walls, being blocked, etc.

A first plan

1. Download the page
   - Possibly taking measures to deal with cookie walls, being blocked, etc.
2. Remove all line breaks (\n, but maybe also \n\r or \r) and TABs (\t): We want one long string

A first plan

1. Download the page
   - Possibly taking measures to deal with cookie walls, being blocked, etc.
2. Remove all line breaks (\n, but maybe also \n\r or \r) and TABs (\t): We want one long string
3. Try to isolate the comments
   - Do we see any pattern in the source code? $\Rightarrow$ *last week: if we can see a pattern, we can describe it with a regular expression*

```
1  import requests
2  import re
3
4  response = requests.get('http://kudtkoekiewet.nl/setcookie.php?t=http://
       www.geenstijl.nl/mt/archieven/2014/05/das_toch_niet_normaal.html')
5
6  tekst=response.text.replace("\n"," ").replace("\t"," ")
7
8  comments=re.findall(r'<div class="cmt-content">(.*?)</div>',tekst)
9  print("There are",len(comments),"comments")
10 print("These are the first two:")
11 print(comments[:2])
```

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 `<div>` and `</div>`)

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 `<div>` and `</div>`)

Optimization

- Parse usernames, date, time, ...
- Replace `<p>` tags

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)

Samsung Galaxy A5 (2017) zwart / 16 GB - Reviews

Reviewscore | 8,7 Lees 1049 reviews | Schrijf een review

Favorieten | Vergelijk | Prijsalert

6 Afbeeldingen

Testpanel score 8,2 - april 2017

Laagste prijzen -1% prijsdaling

Midimedia	€ 267,00
Belsimpel.nl	€ 275,00
Mobiel.nl	€ 279,00

Specificaties

- 4.7 inch scherm
- 16 GB opslag
- Android

Bekijk 3 andere uitvoeringen

Samsung Galaxy A5 (2017) zwart / 16 GB

Prijzen (9) | Overzicht | Abonnementen | Specificaties | Reviews (1049)

Review type: **Consumenten reviews (1049)**

Gemiddelde beoordeling

8,7

De gemiddelde beoordeling is gebaseerd op de 1049 reviews die zijn achtergelaten op Kieskeurig.nl.

Help anderen en geef jouw mening over dit product. We verloten iedere maand onder alle reviewers 2 x € 250,-.

Schrijf een review

Kieskeurig.nl review criteria:

Geluidskwaliteit	Goed
Mogelijkheden	Goed
Gebruiksgemak	Uitstekend
Vormgeving	Uitstekend
Degelijkheid	Goed

Echte reviews van echte gebruikers, door ons gecontroleerd op betrouwbaarheid.

Sorteer op: Meest nuttig eerst +

BartTilburg
21-03-2017

· Product paar weken in bezit

Verliefd op de Samsung Galaxy A5 2017

✓ Ja, ik beveel dit product aan

Wat een fenomenale smartphone is de nieuwe Samsung Galaxy A5 2017 geworden, lees hier waarom.

Bij het openen van de verpakking valt direct op dat dit toestel niet onder doet voor een smartphone uit het duurste segment. Ik heb gekozen voor de kleur sand/gold, een mooie champagneachtige gouden kleur die erg opvalt. Het toestel heeft een aluminium behuizing en zowel de voorkant als de achterkant van de Samsung Galaxy A5 2017 zijn geheel van glas en er is zelfs een

Algemene score

Reviewscore | 10,0

Review criteria:

Geluidskwaliteit	Uitstekend
Mogelijkheden	Uitstekend
Gebruiksgemak	Uitstekend

Verliefd op de Samsung Galaxy A5 2017

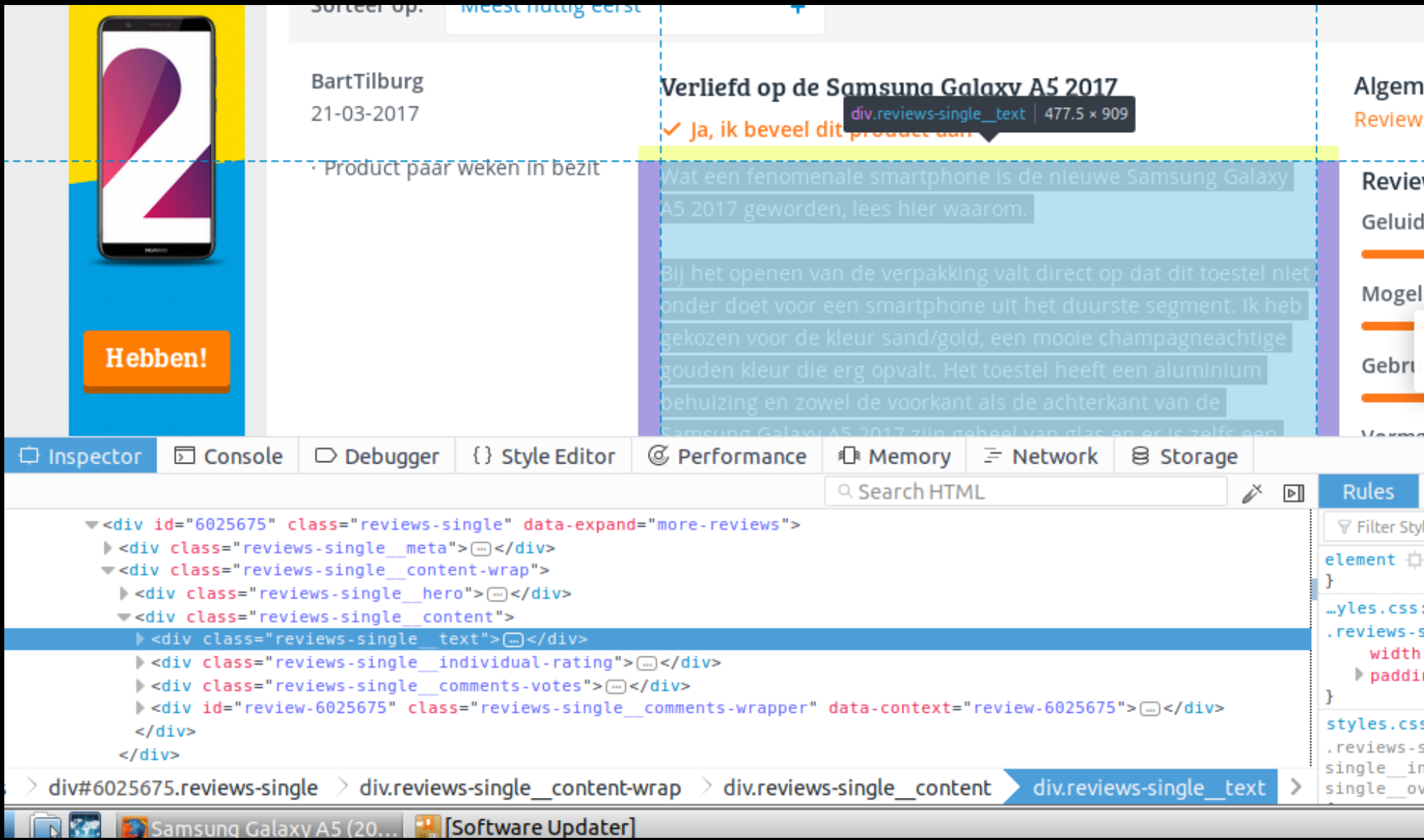
✓ Ja, ik beveel dit product aan

Wat een fenomenale smartphone is de nieuwe Samsung Galaxy A5 2017 geworden, lees hier waarom.

Bij het openen van de verpakking valt direct op dat dit toestel niet onder doet voor een smartphone uit het duurste segment. Ik heb gekozen voor de kleur sand/gold, een mooie champagneachtige gouden kleur die erg opvalt. Het toestel heeft een aluminium behuizing en zowel de voorkant als de achterkant van de Samsung Galaxy A5 2017 zijn gehee
vingerafdruksensor aanwezig. Dit a
premium look and feel, prachtig! In
accessoires die je nodig hebt zoals
hoofdtelefoon én de bijzondere US
USB-C kabel kan je dit toestel versn
toestel heeft echt alles aan boord wat je mag verwachten en zelfs

Algemene scor
Reviewscore | 10

Review criteria
Geluidskwaliteit
Mogelijkheden
Gebruiksgemak

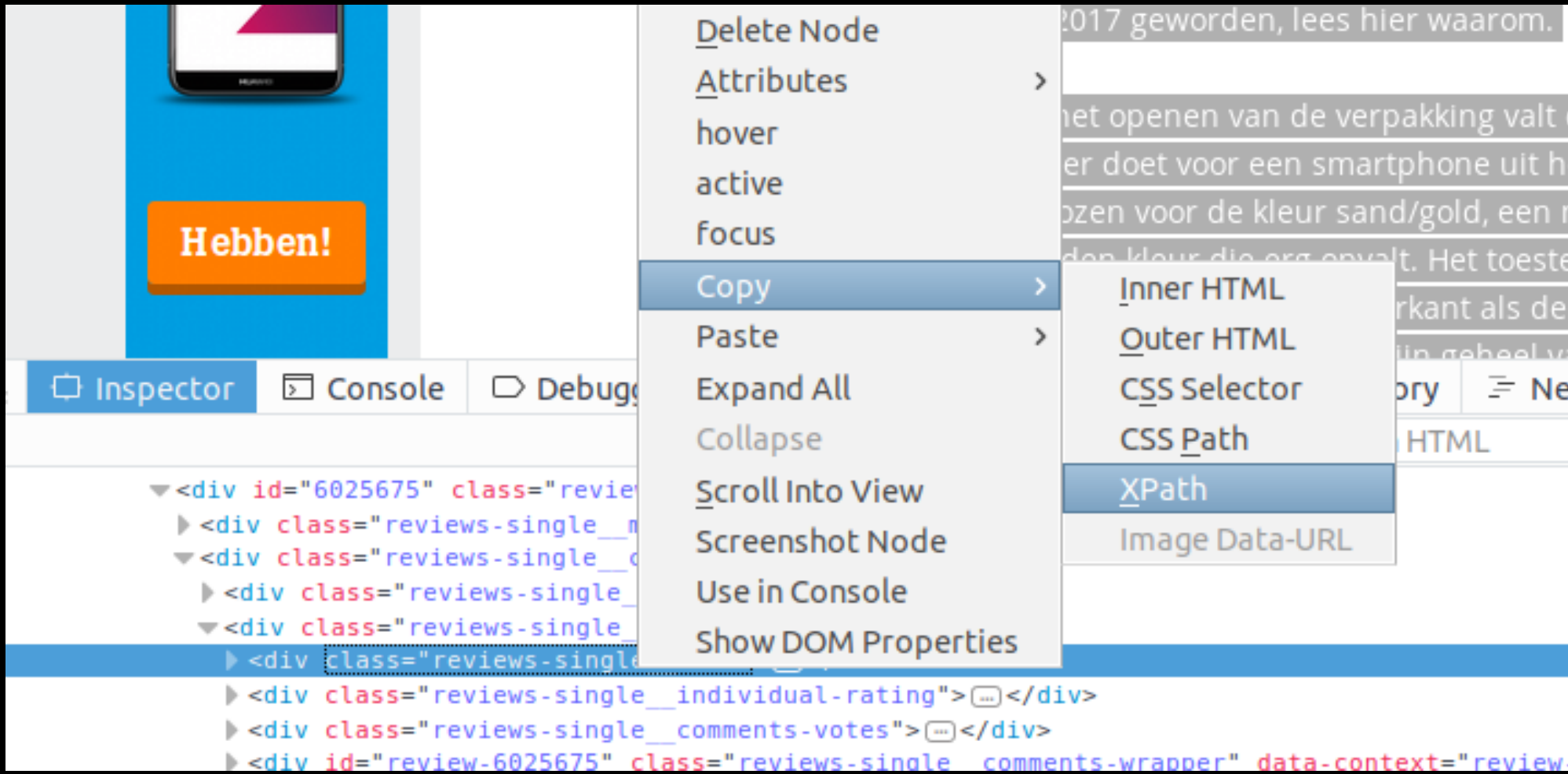

Hebben!
BartTilburg
21-03-2017
Verliefd op de Samsung Galaxy A5 2017
div.reviews-single__text | 477.5 × 909
Ja, ik beveel dit
Product paar weken in bezit
Wat een fenomenale smartphone is de nieuwe Samsung Galaxy A5 2017 geworden, lees hier waarom.
Bij het openen van de verpakking valt direct op dat dit toestel niet onder doet voor een smartphone uit het duurste segment. Ik heb gekozen voor de kleur sand/gold, een mooie champagneachtige gouden kleur die erg opvalt. Het toestel heeft een aluminium behuizing en zowel de voorkant als de achterkant van de
Inspector
Console
Debugger
Style Editor
Performance
Memory
Network
Storage
Search HTML
Rules

…

…

…
…
…
…

element
}
.reviews-s
width:
paddir
}
styles.css
.reviews-s
single__in
single__ov
div#6025675.reviews-single
div.reviews-single__content-wrap
div.reviews-single__content
div.reviews-single__text
Samsung Galaxy A5 (20...
[Software Updater]


Hebben!
Delete Node
Attributes
hover
active
focus
Copy
Paste
Expand All
Collapse
Scroll Into View
Screenshot Node
Use in Console
Show DOM Properties
Inner HTML
Outer HTML
CSS Selector
CSS Path
XPath
Image Data-URL
Inspector
Console
<div id="6025675" class="
…
…

But you can also create your XPATH yourself

There are multiple different XPATHs to address a specific element. Some things to play around with:

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

But you can also create your XPATH yourself

There are multiple different XPATHs to address a specific element. 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

But you can also create your XPATH yourself

There are multiple different XPATHs to address a specific element. 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'`

But you can also create your XPATH yourself

There are multiple different XPATHs to address a specific element. 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

But you can also create your XPATH yourself

There are multiple different XPATHs to address a specific element. 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 (via 'inspect elements') of the web page to think of other possible XPATHs!

Let's test it!

`https://www.kieskeurig.nl/smartphone/product/3518003-samsung-galaxy-a5-2017-zwart/reviews`

Let's scrape them!

```
1  from lxml import html
2  from urllib import request
3
4  req=request.Request("https://www.kieskeurig.nl/smartphone/product
       /3518003-samsung-galaxy-a5-2017-zwart/reviews")
5  tree = html.fromstring(request.urlopen(req).read().decode(encoding="utf
       -8",errors="ignore"))
6
7  reviews = tree.xpath('//*[@class="reviews-single__text"]/text()')
8
9  # remove empty reviews
10 reviews = [r.strip() for r in reviews if r.strip()!=""]
11
12 print (len(reviews),"reviews scraped. Showing the first 60 characters of
       each:")
13 i=0
14 for review in reviews:
15     print("Review",i,":",review[:60])
16     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 (looooooong) 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. Sometimes, you might want to use a different suggestion to be able to generalize better (e.g., using the *attributes* rather than the *tags*)
2. 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

Console | Debugger | Style Editor | Performance | Memory | Network | Storage

Search HTML

```
id="6025675" class="reviews-single" data-expand="more-reviews">
 class="reviews-single__meta">…</div>
 class="reviews-single__content-wrap">
iv class="reviews-single__hero">…</div>
iv class="reviews-single__content">
<div class="reviews-single__text">…</div>
<div class="reviews-single__individual-rating">…</div>
<div class="reviews-single__comments-votes">…</div>
<div id="review-6025675" class="reviews-single__comments-wrapper" data-context="review-6025675">…</div>
div>
v>
```

75.reviews-single > div.reviews-single__content-wrap > div.reviews-single__content > div.reviews-single__text

Samsung Galaxy A5 (20... | [Software Updater]

Next steps

From now on...

...focus on individual projects!

Wednesday

- Write a scraper for a website of your choice!
- Prepare a bit, so that you know where to start
- Suggestions: Review texts and grades from `http://iens.nl`, articles from a news site of your choice, ...

Shell Wijzig Weergave Venster Help

Za 11 mrt. 08:51 Damian Tr

Restaurants in Amsterdam ...

https://www.iens.nl/stad/amsterdam/19088

Aan de slag Press This Stan - Interfaces Atlassian Cloud

Proef Italie!

BEKIJK DE RESTAURANTS

damian — ipython3 — 80×26

```
In [35]: ratings = tree.xpath("//*[@class='rating-ratingValue']/text()")

In [36]: restaurants = tree.xpath("//*[@class='restaurantSelection-name']/text()
")

In [37]: len(ratings)
Out[37]: 10

In [38]: len(restaurants)
Out[38]: 10

In [39]: for i in range(10):
    print(restaurants[i].strip(),ratings[i])
   ....:
Risto Enoteca PepeNero 9.3
Cafe Sage 9.2
The Lobby Fizeaustraat Restaurant & Bar (Hotel V) 9.1
Brasserie Ambassade 9.1
Tulsi Indian restaurant 9.1
Zaza's 9
Jacobsz 9
Bistrot des Alpes 9
Restaurant Jaspers 8.9
En Pluche 8.9

In [40]:
```

rdam Voor jou geselecteerd

De Vrienden Live

FRANS

8.5 /10

```
<div class="restaurantSelection-rateContainer">
  <div class="restaurantSelection-avgRate rating">
    <span class="rating-ratingValue">8.6</span>
    /10
  </div>
```

Regels Berekend Animaties

Stijlen filteren

Pseudo-elementen

Dit element

```
element {
}
.restaurantSelection .restaurantSelection-avgRate {
    margin: 8px 0;
    text-align: right;
}
.rating {
    font-size: 11px;
```

li.restaurantSelectionWrapper-item.resta... div.restaurantSelection-rateContainer div.restaurantSelection-avgRate.rating

CSS JS Beveiliging Logboekregistratie Server

Mutation Events wordt niet meer ondersteund. Gebruik in plaats daarvan MutationObserver.

ut: geen hoofdelement gevonden Locatie: https://api.realytics.io/event/track?cb=1489218342118 Regelnummer 1, kolom 1:

naar document.write() van een asynchroon geladen extern script is genegeerd.