

Web scraping in the Norwegian CPI

DATA FOR NOW WORKSHOP

7 APRIL 2022

GUNNAR LARSSON



Statistisk sentralbyrå
Statistics Norway

Web scraping

- The extraction and transformation of unstructured data from the web into structured data
- Scraping tools – external software tools online, import.io, in-house solutions based on software programs like R, Python
- Increase of efficiency and productivity
 - Collecting more data at the same web page/online store
- Shift of tasks: less resources needed for the collection of data, more resources needed for the maintenance of programs and analysis of retrieved data



Web scraping II

- Many countries develop their own technical solution
 - Increased control compared to external software
 - More and more NSOs make their own solutions in R or Python
- Ethics of scraping
 - Minimization of harms
 - Check website's Terms of Use



Generalized web scraper

- Able to web scrape a range of online stores covering multiple areas for the consumer price index
 - Gone from specific tool for each web site to generalized approach
- Built using Python, output stored in Excel for further use
 - Data must often be adjusted and cleaned before use
 - Mainly used to compile price indices for the CPI



Pros and cons

- Potential for improved price indices
 - Larger sample
 - Extended time dimension
- Challenging to find common traits among different web sites
 - Requires testing the web scraping tool on different types of web sites
- Web sites change over time
 - Can lead to maintenance

Demonstration

ASKINER OVERLOCK MASKINER COVER PRO/COVERSTITCH MASKINER SCANN CUT MASKINER ALLE KATEGORIER VAREMERKE

Viser 17 produkter

Løgsteste pris



SMARTER by Pfaff Symaskin 130s

KR 2 895



SMARTER by Pfaff Symaskin 140s

KR 3 295



SMARTER by Pfaff Symaskin 160s

KR 3 395



Pfaff Symaskin Passport 2.0

KR 7 295



Pfaff Symaskin Select 3.2, Nyheit: LED lys

KR 7 695



Pfaff Symaskin Passport 3.0

KR 7 895



Pfaff Symaskin Select 4.2, Nyheit: LED lys

KR 8 695



APRIL KAMPANJE: Pfaff Ambition 620 Symaskin

KR 10 495
KR 8 990

```
<product-promo-label _ngcontent-hyoga-v2-c27 _ngghost-hyoga-v2-c42>...</product-promo-label>  
<div _ngcontent-hyoga-v2-c27 class="product-card-image product-card-item_item">...</div>  
</div>  
<div _ngcontent-hyoga-v2-c27 class="product-card-content-holder" tabindex="0"> <flex>  
  <span _ngcontent-hyoga-v2-c27 class="product-card-label product-card-label_top-selling"> mest solgte </span>  
  <h4 _ngcontent-hyoga-v2-c27 class="product-card-title product-card-item_item"> SMARTER by Pfaff Symaskin 130s </h4>  
  <!-->  
  <product-star-rating _ngcontent-hyoga-v2-c27 _ngghost-hyoga-v2-c44>...</product-star-rating> <flex>  
  <div _ngcontent-hyoga-v2-c27 class="product-card-price product-card-item_item product-card-item_price">  
    <product-price _ngcontent-hyoga-v2-c27 _ngghost-hyoga-v2-c45>  
      <!-->  
      <!-->  
      <div _ngcontent-hyoga-v2-c45 class="product-card-price ng-star-inserted"> <flex>  
        <!-->  
        <div _ngcontent-hyoga-v2-c45 class="product-card-price_item product-card-price_item_final-price ng-star-inserted"> <kr&nbsp;2&nbsp;895 </div>  
        <!-->  
        <!-->  
      </div>  
    </div>  
  <!-->  
  <!-->  
</product-price>  
</div>  
<!-->  
<!-->  
<div _ngcontent-hyoga-v2-c27 class="product-item-holder product-item-holder_hoverable ng-star-inserted">...</div>  
<!-->  
</div>  
</div>  
<!-->
```

Thank you!

