ESSnet BD SGA2

WP2: Web Scraping Enterprises, NL plans

Gdansk meeting

Olav ten Bosch, Dick Windmeijer, Oct 4th 2017

Statistics Netherlands

Contents

- SGA1 results
- SGA2 plans

Legal (3)

LinkedIn can't block scrapers from monitoring user activity

August 14th 2017:

- The world is still struggling with legal aspects on scraping
- Who owns public community data? The community (LinkIn users) or the community provider (MS)?
- Keep an eye on what is happening for an official statistics interpretation

Legal (4)

James Densmore Follow
Data Science, Analytics and all things tech
Jul 23 · 3 min read

Ethics in Web Scraping

The Ethical Scraper

I, the web scraper will live by the following principles:

- If you have a public API that provides the data I'm looking for, I'll use it and avoid scraping all together.
- I will always provide a User Agent string that makes my intentions clear and provides a way for you to contact me with questions or concerns.
- I will request data at a reasonable rate. I will strive to never be confused for a DDoS attack.
- I will only save the data I absolutely need from your page. If all I need it OpenGraph meta-data, that's all I'll keep.
- I will respect any content I do keep. I'll never pass it off as my own.
- I will look for ways to return value to you. Maybe I can drive some (real) traffic to your site or credit you in an article or post.
- I will respond in a timely fashion to your outreach and work with you towards a resolution.
- I will scrape for the **purpose of creating new value from the data**, not to duplicate it.

The Ethical Site Owner

I, the site owner will live by the following principles:

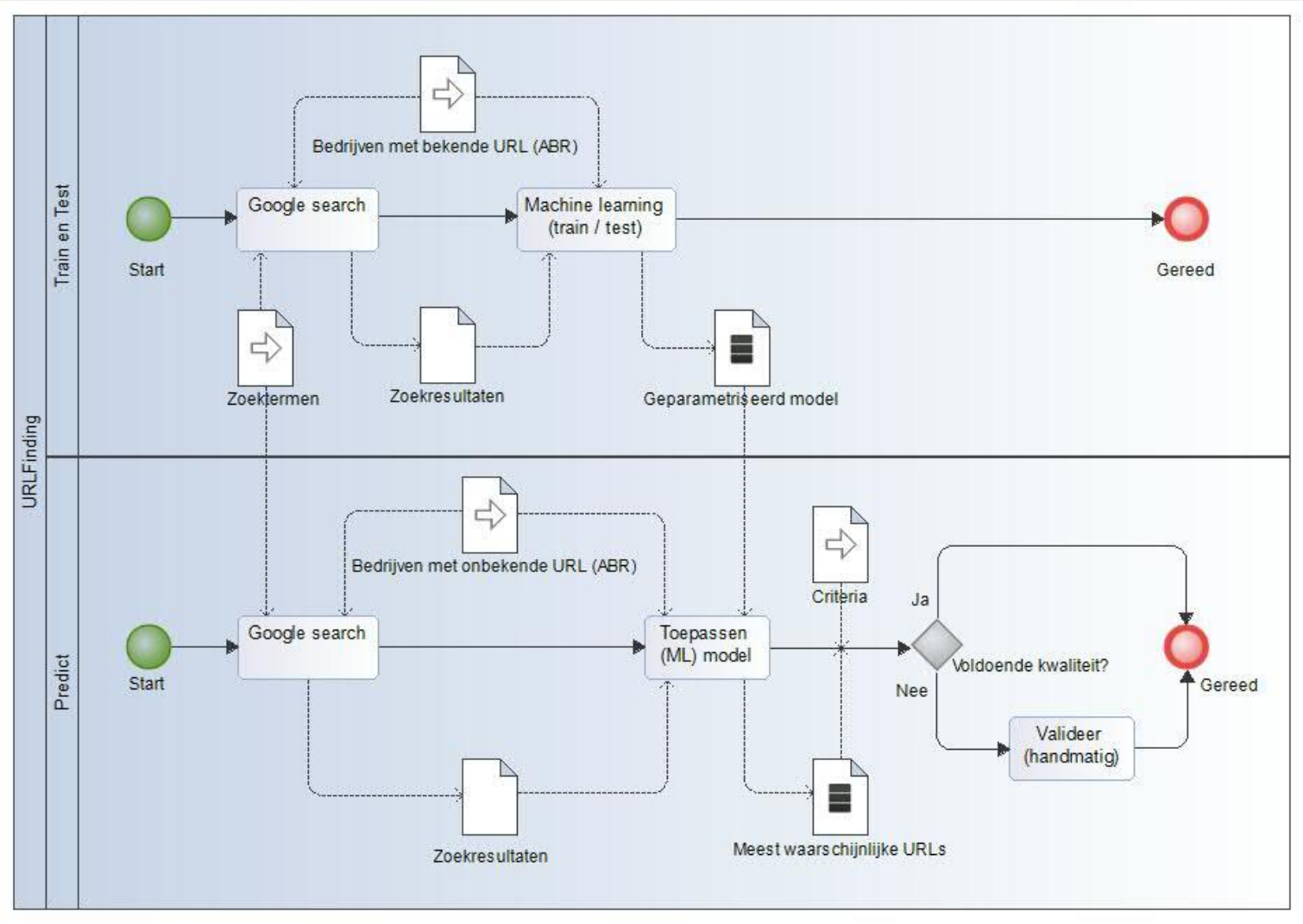
- I will allow ethical scrapers to access my site as long as they are not a burden on my site's performance.
- I will respect transparent User Agent strings rather than blocking them and encouraging use of scrapers masked as human visitors.
- I will reach out to the owner of the scraper (thanks to their ethical User Agent string) before blocking permanently. A temporary block is acceptable in the case of site performance or ethical concerns.
- I understand that scrapers are a reality of the open web.
- I will consider public APIs to provide data as an alternative to scrapers.

SGA1 results: URLFinding

URLFinding:

- Goal: find missing URLs of enterprises using web scraping and machine learning
- 1/3 of 1.5 Million enterprises in Dutch BR have a URL
- Method:
  - Automated Google Search, 5 different search queries per enterprise using combinations of enterprise name and address
  - Calculate scores (features)
  - Train and test model
  - Apply model
- A test on a sample of 1000 enterprises (>10 persons) indicated:
  - The search found in 70 % of the cases a valid URL
  - The model predicted URL validity correctly in 73 % of the cases

SGA1 results: URLFinding workflow

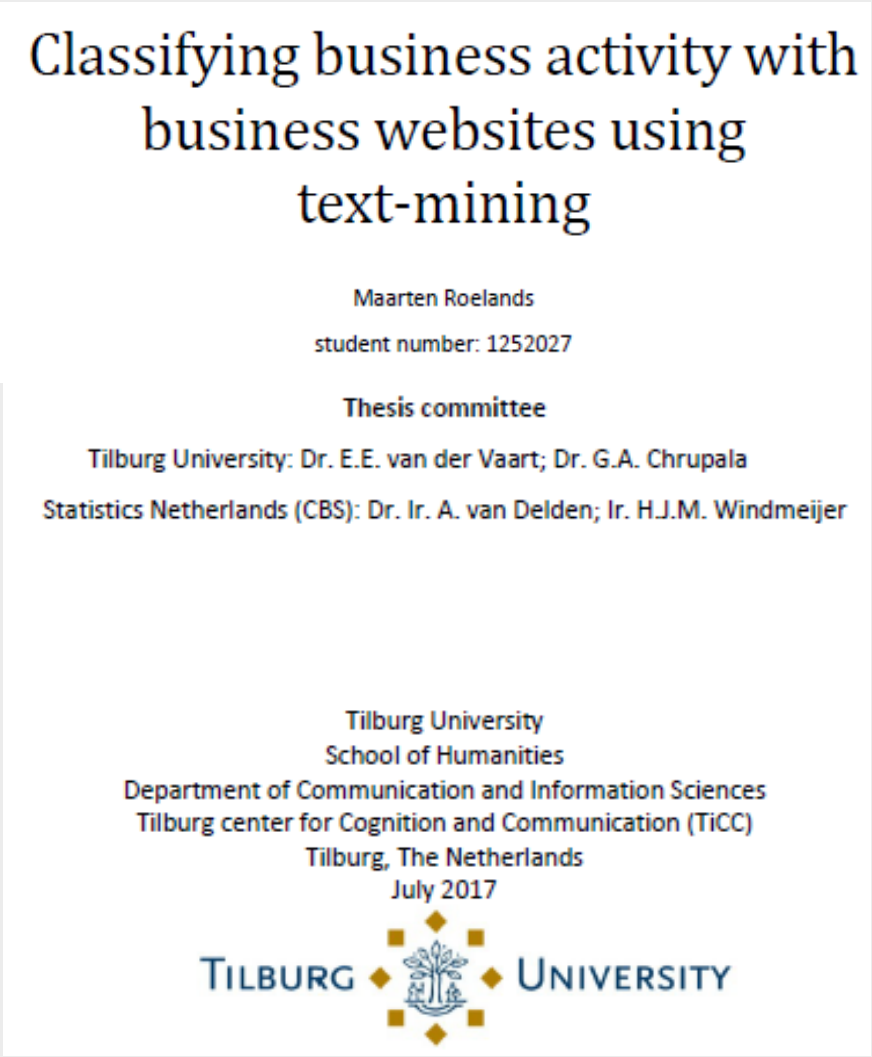


SGA1 results: Ecommerce

Ecommerce detection:

- Goal: automatically detect web shops from a list of 1200 foreign companies paying Dutch VAT
- URL finding step found in 97 % of the cases a domain with indication of correctness (Good, Fair, Mediocre, Poor)
- About 1000 sites crawled to determine E-commerce activities automatically via text analysis
- Manual check on sample indicated:
  - URL finding in class “Good”: accuracy of about 90 %
  - Webshop detection algorithm: accuracy of about 85 %
- Executed spring 2017 on 22 000 enterprises
- Plans to repeat twice a year

SGA1 results: (related) student thesis



Classifying business activity with business websites using text-mining

Maarten Roelands

student number: 1252027

Thesis committee

Tilburg University: Dr. E.E. van der Vaart; Dr. G.A. Chrupala

Statistics Netherlands (CBS): Dr. Ir. A. van Delden; Ir. H.J.M. Windmeijer

Tilburg University
School of Humanities
Department of Communication and Information Sciences
Tilburg center for Cognition and Communication (TiCC)
Tilburg, The Netherlands

July 2017

TILBURG UNIVERSITY

- On sample of 1918 instances, predict "top sector" classification
- Model could predict 10 top sector with sufficient precision
- For predicting more detailed classes (30 subsectors) performance was poor
- CBS paper to appear

SGA1 results

Conclusions and further work:

- Finding URLs using a search engine and machine learning proofed to be valuable
- Can be used as a starting point for retrieving additional characteristics of a population / developing new indicators
- URL finding and text mining successfully applied to detect E-commerce of foreign companies

Further work:

- Improve URL finding accuracy by additional scraping
- Experiment with other search queries and search engines
- Scale up experiments
- Keep methodology aligned with other NSI's as much as possible: ***circumstances may differ but concepts are the same***

SGA2: Scale up URLFinding

- On the Dutch Business Register as of 20170101 enterprises > 10 persons
- Known URL: 20 000 (= 33%), unknown URL: 40 000
- Perform URLFinding on enterprises with unknown URL
- Using the ML model (DT, RF, NB or SVM) that was created in SGA1 from the sample of 1000 enterprises > 10 persons
- Using the 5 search queries designed in SGA1, extended with one new query:
  - 'inanchor:contact' and a blacklist of sites we are not interested in
- Spread over multiple days (because of day limits):
  - 6 * 40 000 queries = 240 000
  - Our key is limited to 10 000 per day
  - Hence 24 days
  - 240 * $5 = $1200
- Executed from a server with clearly defined inputs and outputs:
  - Input: csv of enterprise name (trade name and fiscal name), address, phone nr., email, kvk, VAT nr, NACE
  - Output: csv's of enterprise URL's with score, multiple URL's possible
- Create baseline:
  - Select all 'found URLs' above a threshold score
  - Combine 'known URLs' with the 'found URLs' into a baseline set
  - Expectations SGA1: 20 000 + 70% of 40 000 = 48 000 URLs in total (= 80%)

Check Baseline

- Manually / confrontation with other sources?

SGA2: Perform Ecommerce detection

- On baseline:
  - Execute webshop detection
  - Using the E-commerce detection software for foreign Enterprises developed in SGA1 (or alternative ESSnet sw)
  - Home page only
  - Using some simple text analysis (keywords)
  - Designed for multiple languages
  - Compare with other statistics on Ecommerce
- Output:
  - Experimental indicator (to be interpreted)

SGA2: Social media presence analysis

- On baseline:
  - Perform social media presence analysis
  - Using the Social media presence software from Poland
  - Compare with statistics on social media use in NL from an earlier BD project
- Output:
  - Experimental indicator (to be interpreted)

SGA2: Sustainability reporting

On Baseline:

- If available perform sustainability reporting analysis on the baseline, using the generic software to be developed in this project?
- Problem: do we have a training set?