Scraping enterprises characteristics: state of play NL

Statistics Netherlands

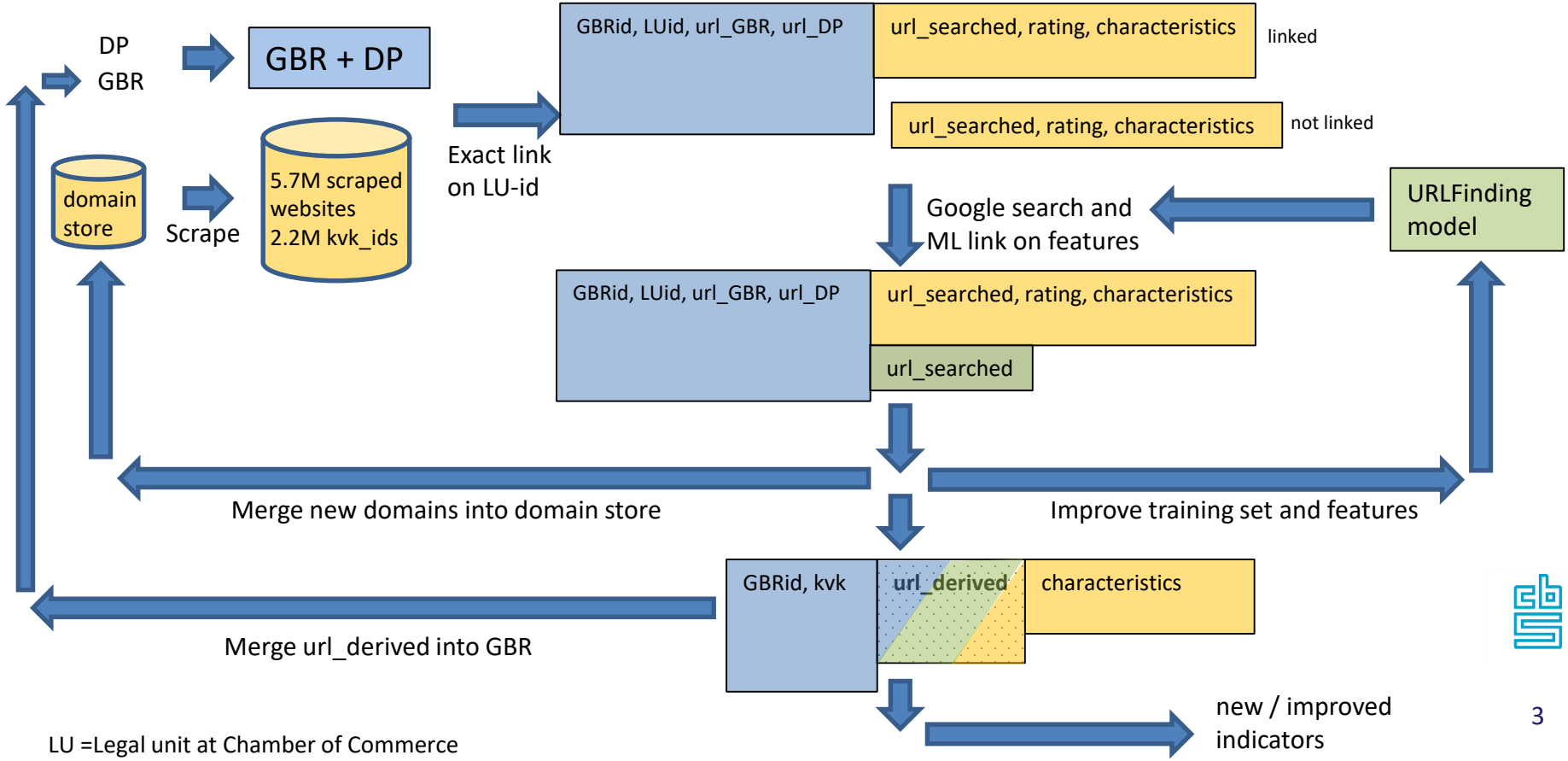
Olav ten Bosch, Dick Windmeijer, Arnout van Delden, Eelco van Vliet, Shirley Ortega

ESSnet BD II WP C meeting, Sofia, 17-06-2019

Contents

- Main process
- URL finding
- Scraping
- Enterprise characteristics from social media
- Generic software
- Wrap up

Main process



LU =Legal unit at Chamber of Commerce

URL finding: global setup

- Continuation from research Istat/Essnet BD I wp2 / CBS
- GBR: 1/3 with url, 2/3 without
- ML model (SVM) based on training/test set of 3300 enterprises from GBR with url + data from scraping company: dataprovider
- Multiple search queries on Google (custom search API):
  1. Name
  2. Name + ‘contact’
  3. Name + Street + ‘contact’
  4. Name + PostalCode + ‘contact’
  5. Full address
  6. Name + ‘inanchor:contact’
- Analyse first 10 results for each search query
- About 30 features defined on all Google search results

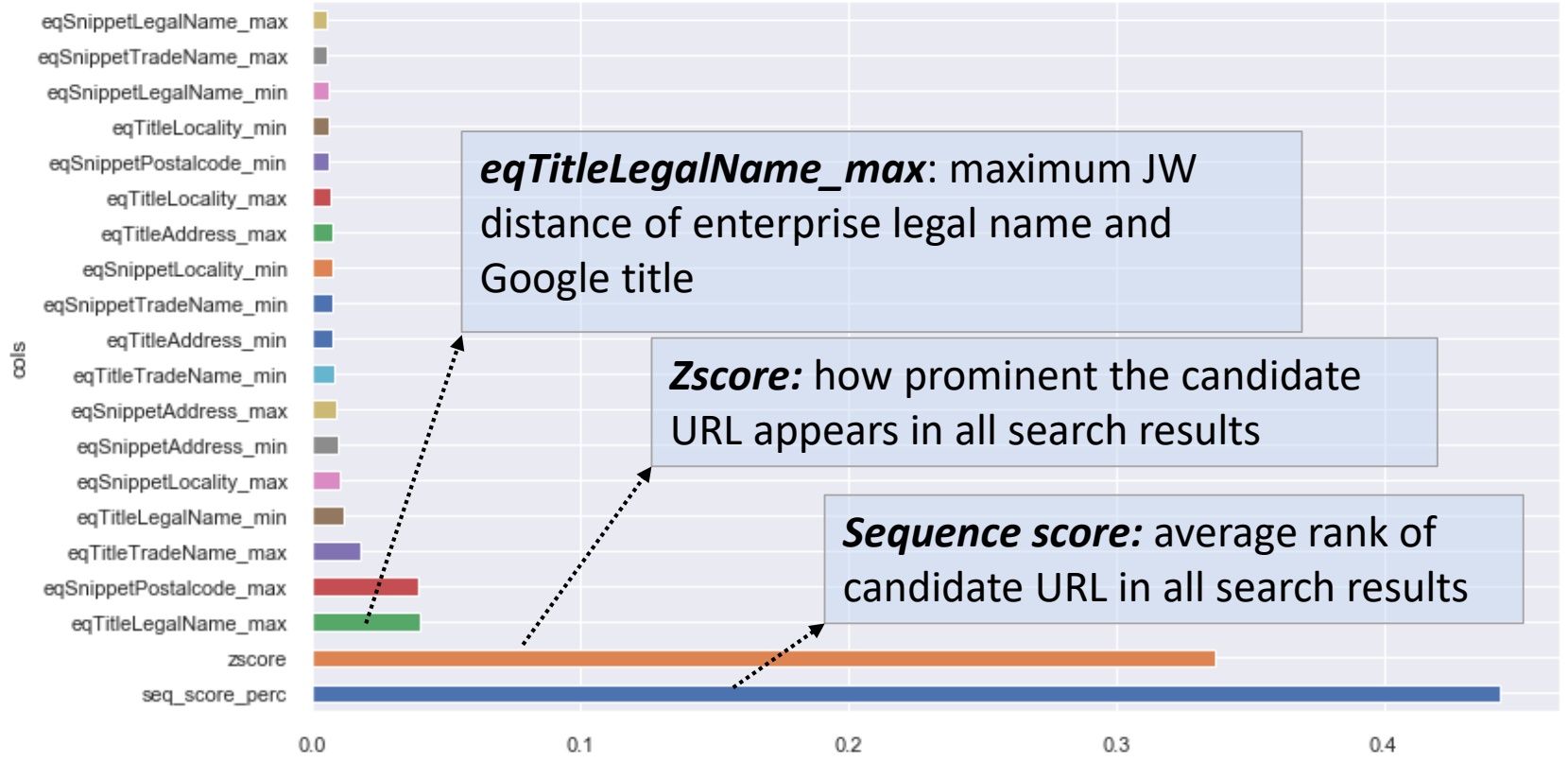
URL finding: search results

PageMap: structured json description of result with ***microformat*** if available

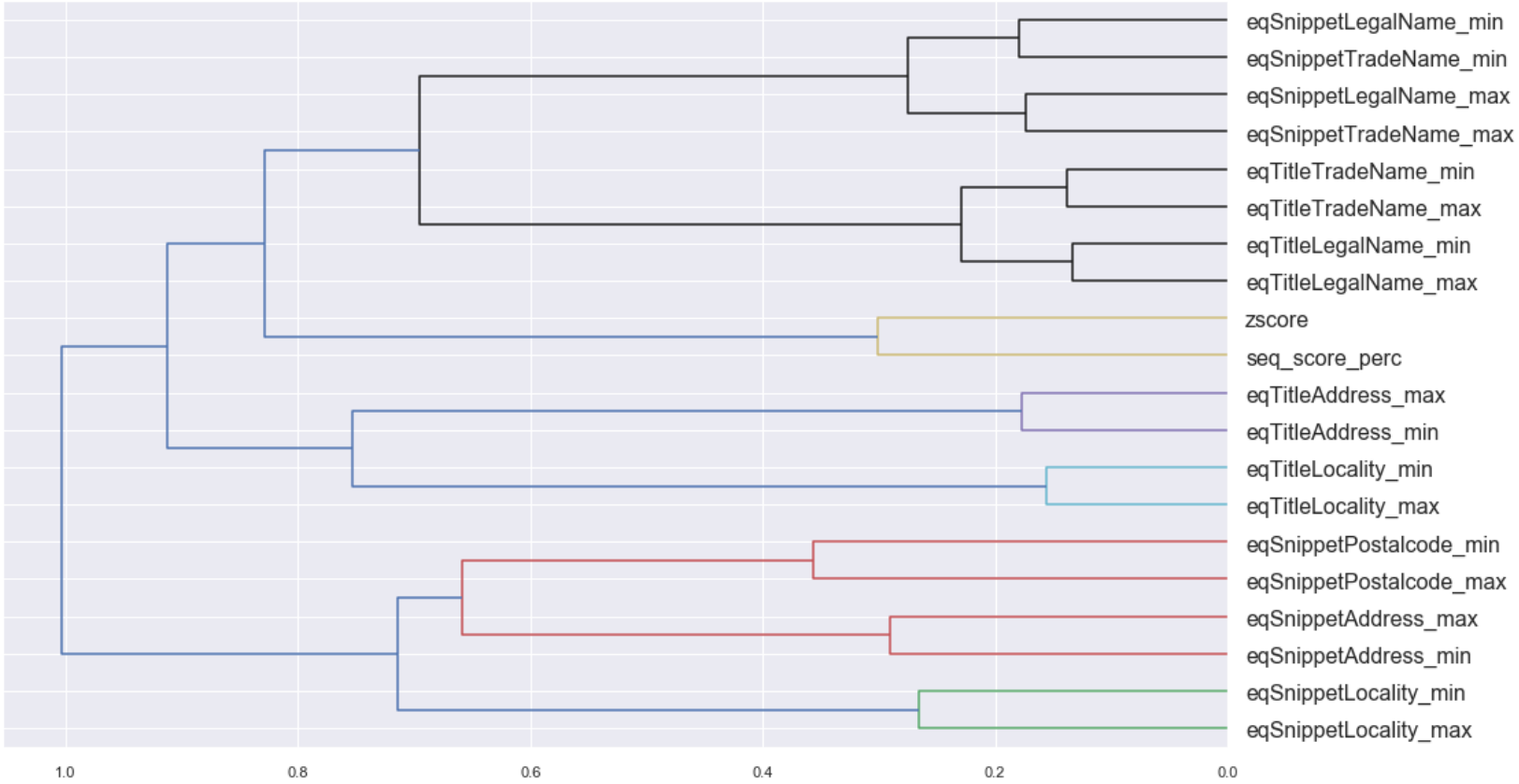
Search results of one query:

	date	seqno	query	title	snippet	urlGoogle	pagemap	Id	queryType
0	20190314	1	HIFA contact	Hifa - Trollenwol	Hifa is een oud Noors familiebedrijf, waar inm...	http://www.trollenwol.nl/garens/hifa	{"cse_thumbnail": [{"width": "110", "height": ...	40234711	0
1	20190314	2	HIFA contact	Restaurant Haifa	Restaurant Haifa. ... Restaurant Haifa. 010 - ...	http://www.restauranthaifa.nl/	{"cse_thumbnail": [{"width": "397", "height": ...	40234711	0
2	20190314	3	HIFA contact	Find Coworking Office Space In Haifa | WeWork	Browse 1 coworking office space in Haifa.	https://www.wework.com/nl-NL/l/haifa	{"cse_thumbnail": [{"width": "275", "height": ...	40234711	0
3	20190314	4	HIFA contact	Haifa Cars | Occasions, tweedehans auto kopen	Welkom bij Autobedrijf- Haifa Cars te Groninge...	http://haifacars.nl/	{"cse_thumbnail": [{"width": "305", "height": ...	40234711	0
4	20190314	5	HIFA contact	Crowne Plaza Haifa - Haifa, Israël	Officiële site van Crowne Plaza Haifa - lees b...	https://www.ihg.com/crowneplaza/hotels/nl/nl/h...	{"cse_thumbnail": [{"width": "318", "height": ...	40234711	0

URL finding: feature importance



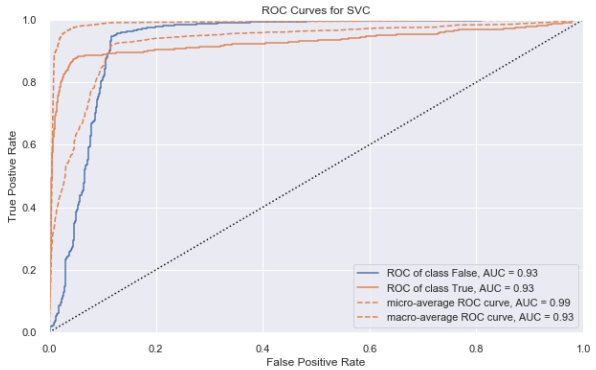
URL finding: feature correlation



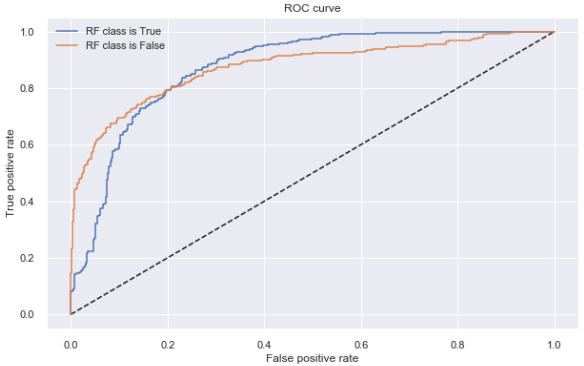
URL finding: tests

ROC curve:

Search result



Legal units



Scores per label:

	f1-score	precision	recall	support
False	0,98	0,98	0,99	7740
True	0,77	0,81	0,72	641
micro avg	0,97	0,97	0,97	8381
macro avg	0,87	0,89	0,86	8381
weighted avg	0,97	0,96	0,97	8381

	f1-score	precision	recall	support
False	0,77	0,77	0,77	392
True	0,84	0,84	0,84	547
micro avg	0,81	0,81	0,81	939
macro avg	0,80	0,80	0,80	939
weighted avg	0,81	0,81	0,81	939

URL finding: results

	Id	TradeName	LegalName	Address	Postalcode	Locality	Location	Url	Url_redirect	eqPred	pTrue	pFalse	eqUrl2	CbsPersoonIdentificatie	Bron	host
0	23168633	WINANS DAKWERKEN B.V.	winans dakwerken	scheepsdiep 12	8356VJ	BLOKZYL	scheepsdiep 12 blokzyl	winans.nl	winans.nl	True	0.924913	0.075087	True	23168633	DP	winans.nl
1	24794554	VAN DER KOLK EN VOS STAALBOUW B.V.	van der kolk en vos staalbouw	bagynhof 30	4264AZ	VEEN	bagynhof 30 veen	kolk-vos.nl	kolk-vos.nl	True	0.909778	0.090222	True	24794554	DP	kolk-vos.nl
2	10080066	DE KONINGH CODING & LABELING B.V.	DE KONINGH CODING & LABELING B.V.	postbus 137	6920AC	DUIVEN	postbus 137 duiven	dekoningh.nl	dekoningh.nl	True	0.880613	0.119387	True	10080066	CBS	dekoningh.nl
3	36136948	WESTREENEN	H. VAN WESTREENEN ELECTROTECHNISCH INSTALLATIE...	tolsestraat 12	4043KB	OPHEUSDEN	tolsestraat 12 opheusden	westreenen.nl	westreenen.nl	True	0.698753	0.301247	True	36136948	CBS	westreenen.nl
4	76270777	TSHR INTERNATIONAL B.V.	tshr international	strickledeweg 44	3044EK	ROTTERDAM	strickledeweg 44 rotterdam	tshrinternational.com	tshrinternational.com	False	0.127624	0.872376	False	76270777	DP	cavallaronapoli.nl

URL GBR

Redirected URL GBR

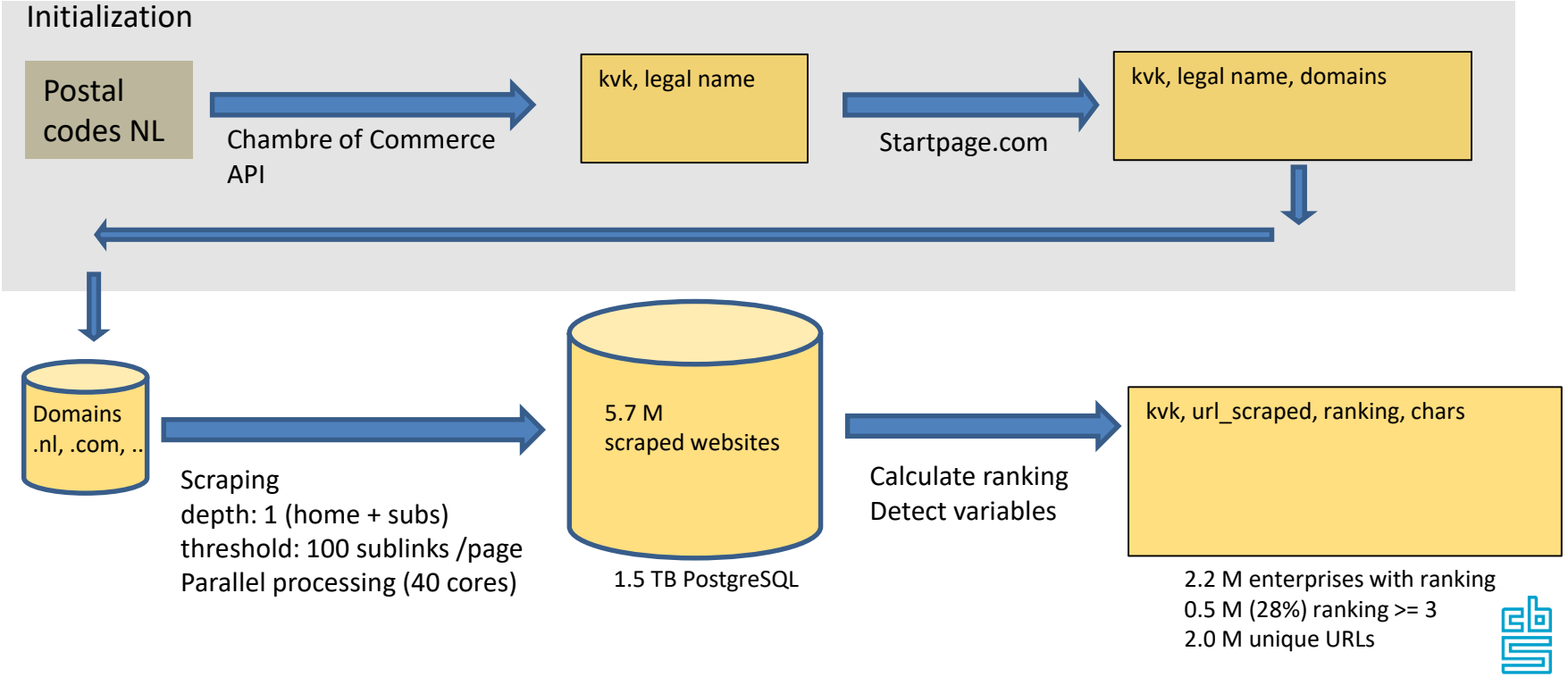
Data source

Predicted URL

True score

False score

Scraping: process



kvk = id of enterprise at Chamber of Commerce

Scraping: variables

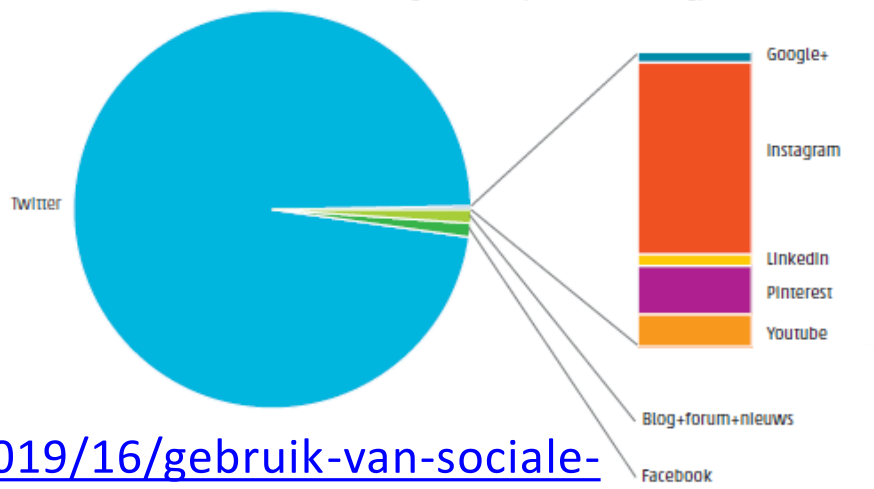
Variables extracted from webpages:

- Identifying variables (kvk, VAT, postal code)
- Ranking 0-10 based on identifying vars + name
- Social media (twitter, facebook, youtube etc.)
- Ecommerce (shopping cart)
- Secure (TLS, certificate)
- Payment options (Paypal, ..)
- “Is your internet uptodate” rating (internet.nl)
- Sitemap, many more …

Enterprise characteristics from social media (1)

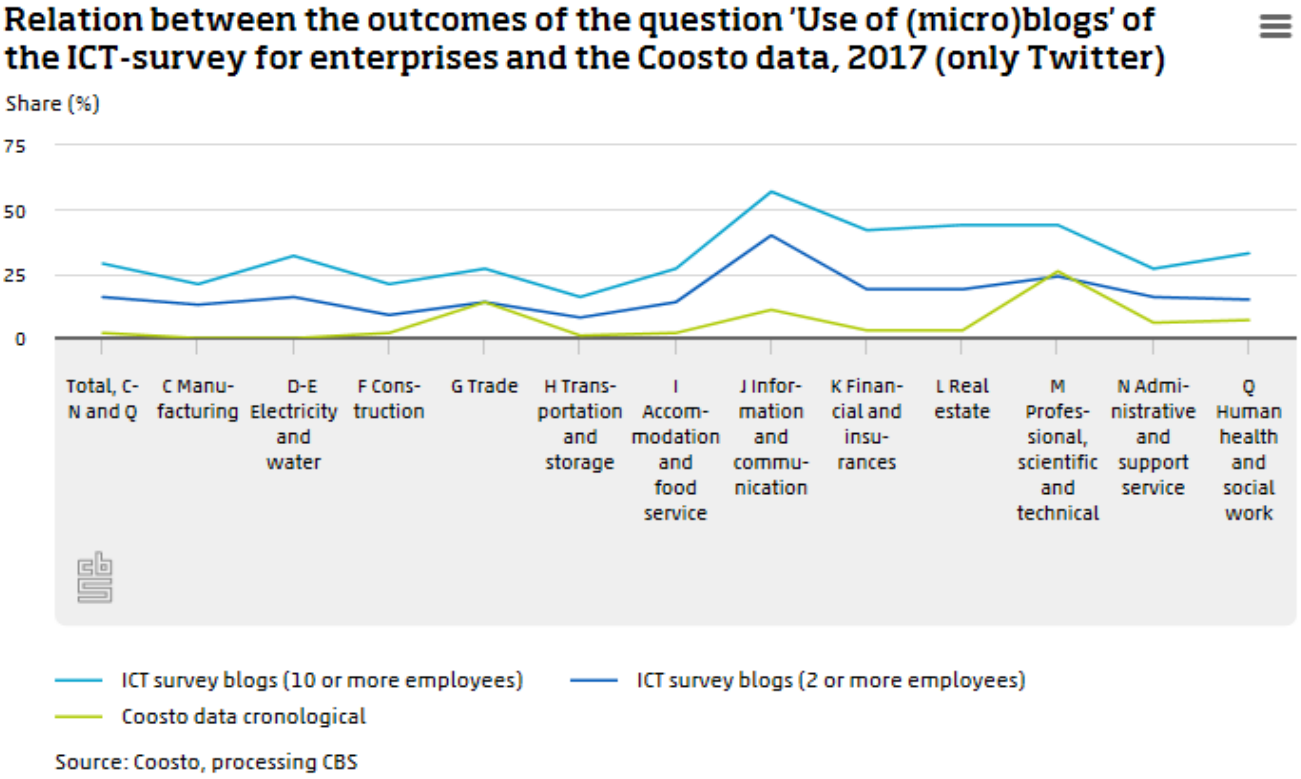
Ortega & Heerschap 2017:

- Use of social media: exploratory research to the possibilities of the use of social media for statistics on enterprises
- Based on messages collected by www.coosto.com
- From 2017-06-26 to 2017-08-04
- 1.6 M workday messages
- 0.2 M weekendday messages



https://www.cbs.nl/nl-nl/achtergrond/2019/16/gebruik-van-sociale-media-2017 (Dutch)

Enterprise characteristics from social media (2)



Enterprise characteristics from social media (3)

Table 3.2.10a Number of social media messages (Twitter) by economic activity and size class, period 26-06-2017 until 04-08-2017

Economic sector (NACE 1-digit)	0	1	2	3	Subtotal	Percentage
A: Agriculture	155	141	12	0	308	0.3
B: Mining	0	0	0	76	76	0.1
C: Manufacturing	1,539	508	374	32	2,453	2.0
D: Electricity	202	45	17	49	313	0.3
E: Water and waste management	36	2	89	56	183	0.2
F: Construction	603	1,015	244	68	1,930	1.6
G: Trade	7,968	1,947	785	181	10,881	9.0
H: Transport and storage	604	112	58	5	779	0.6
I: Accommodation and catering	1,854	2,314	62	6	4,236	3.5
J: Information and communication	13,720	2,196	2,403	858	19,177	15.8
K: Financial and insurance	2,611	360	129	92	3,192	2.6
L : Real estate	3,691	223	96	16	4,026	3.3
M : Professional, scientific and technical	32,551	2,030	618	76	35,275	29.1
N: Rentals and other business services	7,839	1,094	1,116	523	10,572	8.7
O: Public organizations and government	49	53	1,177	395	1,674	1.4
P: Education	4,465	393	257	219	5,334	4.4
Q: Human health and welfare	3,507	733	662	523	5,425	4.5
R: Arts, sport and recreation	5,794	883	652	34	7,363	6.1
S: Other services	6,420	877	708	8	8,013	6.6
Total	**93,608**	**14,926**	**9,459**	**3,217**	**121,210**	100

Source: Coosto and Dataprovider, processing CBS

Some generic software developed

- S4SGoogleSearch Node + Python version
  - https://github.com/SNStatComp/S4SGoogleSearch (Node)
  - Python version in use, not yet published
- Generic scraping software in Python:
  - Python, Postgress, parallel scraping, storing webpages
  - Work in progress!
  - https://github.com/SNStatComp/kvk_url_finder

Wrap up

State of play:

- Webscraping for all enterprises has been set up, many variables derived
- URL finding based on earlier research using Google API and ML is getting mature
- We want to combine both and do maintenance of domains via Google URLfinding
- Social media as a source is also promising

- Other work: NACE detection (Kuhnemann, ms thesis)