Statistics Poland

WP-C
Webscraping Enterprise Characteristics

AGENDA

1. Web scraping
2. Changes since WP2
3. Summary and future work

Statistics Poland

CASE STUDY – TECHNICAL ASPECTS

- **Tools**

Apache Spark

Language

Python with *pyspark*

Results

CSV files

Statistics Poland

WEB SCRAPING – URL

Results | Messages

		WWW	poprawny_url	nowy_url	e_mail	WWW_z_maila	scraping_url	req...
1	erzyce	www.wpkspzoo.eu	1	https://www.wpkspzoo.eu	wpkspzoo@interia.pl	interia.pl	https://www.wpkspzoo.eu	0
2		www.visusoptyk.pl	1	https://www.visusoptyk.pl	drutplast@drutplast.com.pl	drutplast.com.pl	https://www.visusoptyk.pl	0
3		NULL	0	NULL	empol_ksawerow@interia.pl	interia.pl	NULL	N...
4		WWW.RUNOTEX.PL	1	https://www.runotex.pl	poczta@runotex.pl	runotex.pl	https://www.runotex.pl	0
5		NULL	0	NULL	woda129@wp.pl	wp.pl	NULL	N...

Number of employees: 10 and more

Enterpr.	www	nowy_url	www_z_maila	scraping_url
19068	4429	4393	19063	12704

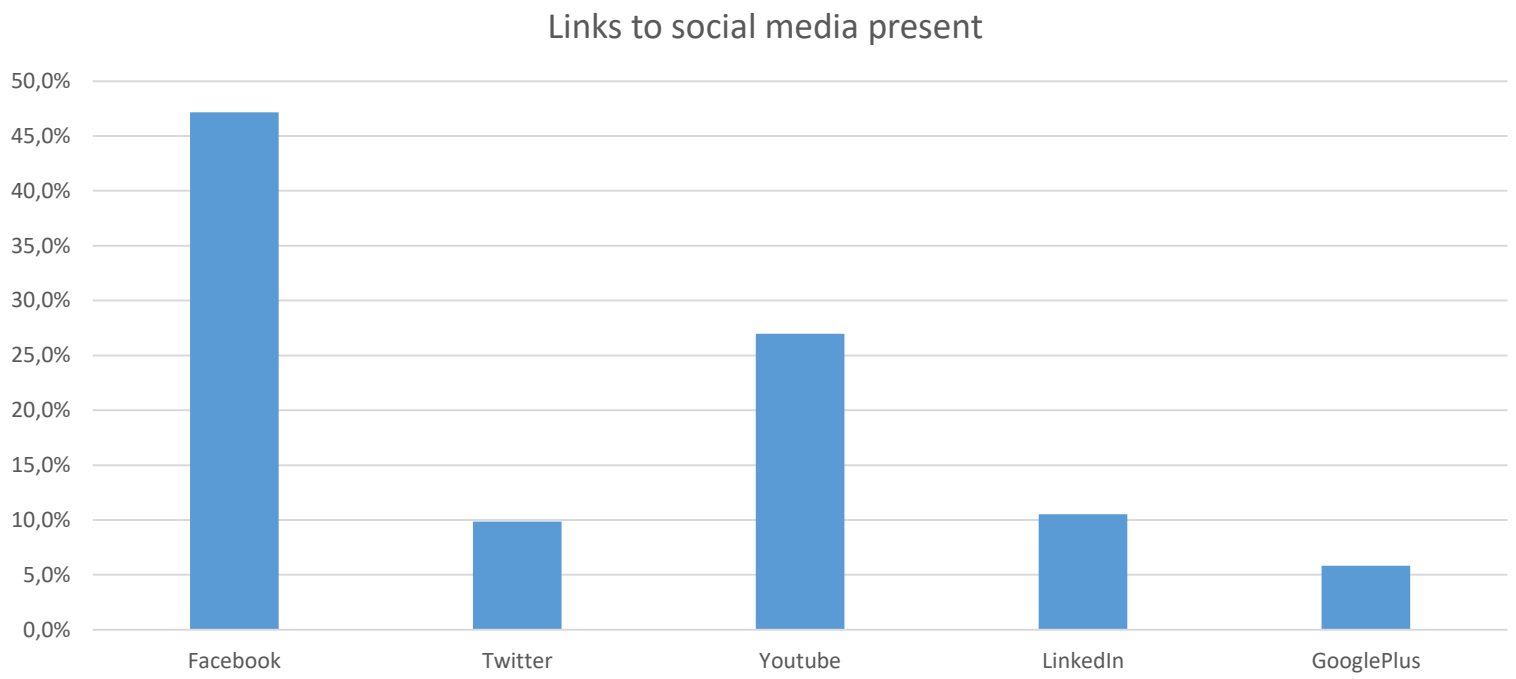
Statistics Poland

WEB SCRAPING – SCHEDULE

March/April 2019 - 6 times every 1-2 weeks

Some websites with no response

RESULTS



Statistics Poland

CHANGES SINCE WP2

Vulnerabilities (PDF, no-html files)

Response within 5 seconds, 60 seconds to download the webpage

Improvements will be here:

http://github.com/jmaslankowski

Statistics Poland

SUMMARY AND FUTURE WORK

- Other attributes of enterprise
- The use of the machine learning
- Extending the population
- Supplement of non-response websites

Statistics Poland

Statistics Poland

Thank you for your attention