DEMONSTRATION OF WP2 RELATED SOFTWARE PL (POLAND)

AGENDA

- Prerequisites
- Tools and Framework
- Future work and conclusions

PREREQUISITES

FACEBOOK/TWITTER

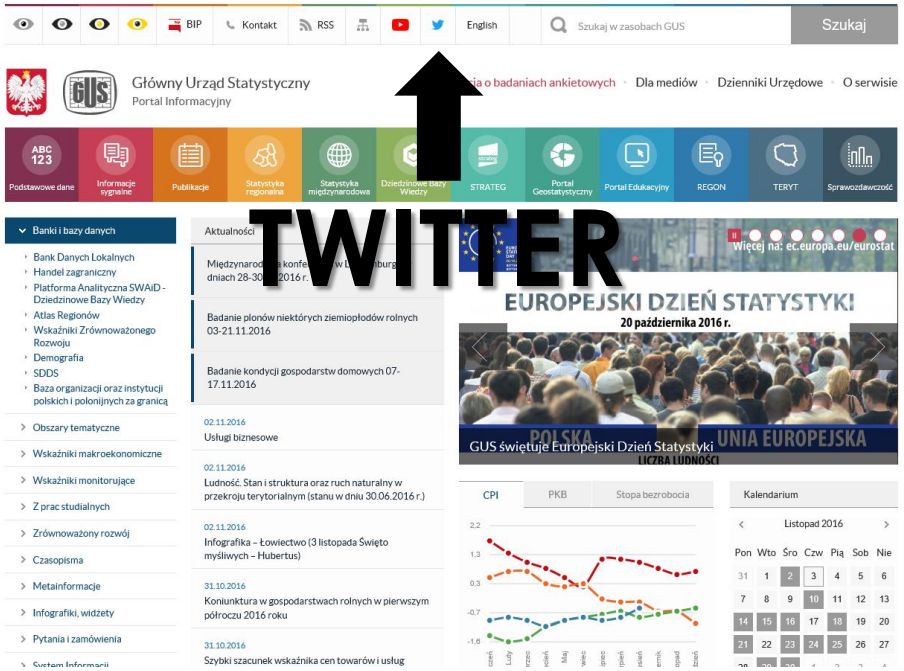
HOW TO COLLECT?

CRAWLERS VS. WEB SCRAPING

Social Media Presence

Central Statistical Office of Poland

- http://www.stat.gov.pl



University of Gdańsk

- http://www.ug.edu.pl

Facebook Code

1. Embedded Comments
2. Embedded Posts
3. Embedded Videos
4. Follow Button
5. Like Button
6. *Comments*
7. *Page Plugin*
8. *Quote Plugin*
9. *Save Button*
10. *Send Button*
11. *Share Button*
12. *oEmbed Endpoints*
13. *Child-Directed Sites*

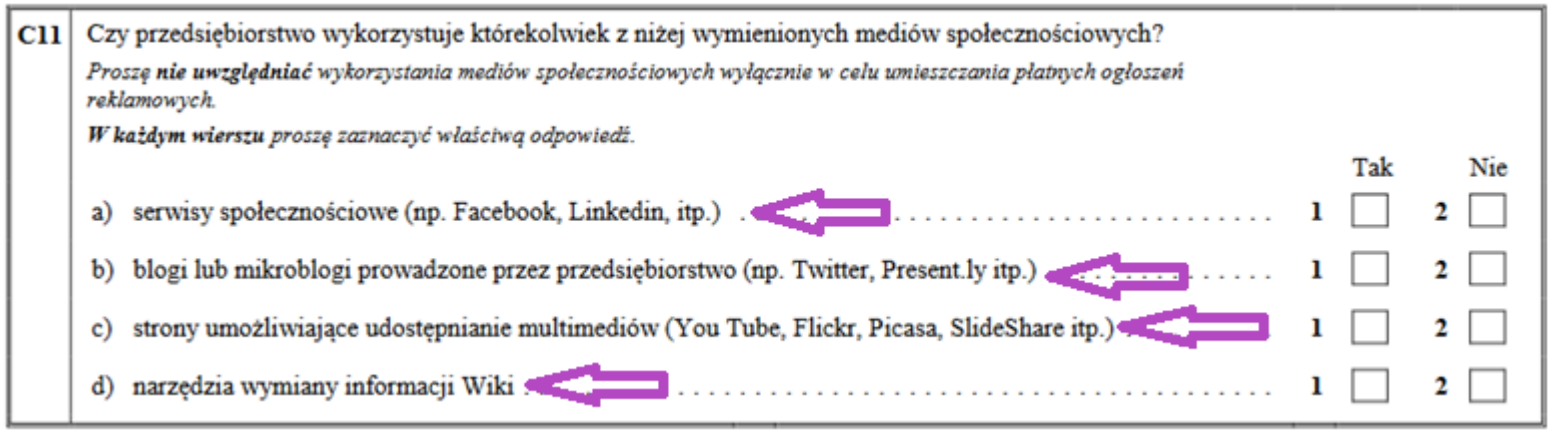
14. MORE? Facebook Login?

```
<!-- Load Facebook SDK for JavaScript -->
  <div id="fb-root"></div>
  <script>(function(d, s, id) {
    var js, fjs = d.getElementsByTagName(s)[0];
    if (d.getElementById(id)) return;
    js = d.createElement(s); js.id = id;
    js.src = "//connect.facebook.net/en_US/sdk.js#xfbml=1";
    fjs.parentNode.insertBefore(js, fjs);
  }(document, 'script', 'facebook-jssdk'));</script>
  <!-- Your like button code -->
  <div class="fb-like"
          data-href="http://www.your-domain.com/your-page.html"
          data-layout="standard"
          data-action="like"
          data-show-faces="true">
  </div>
```

Issues to consider

- What social media should be included in the data analysis? Twitter/Facebook? Instagram? Youtube? LinkedIn? Xing, Viadeo, Yammer, Present.ly, Flickr, Picasa, SlideShare, Wiki?
- What groups of social media activity should we monitor?

- External .js libraries
- Different functions



C11 Czy przedsiębiorstwo wykorzystuje którekolwiek z niżej wymienionych mediów społecznościowych?

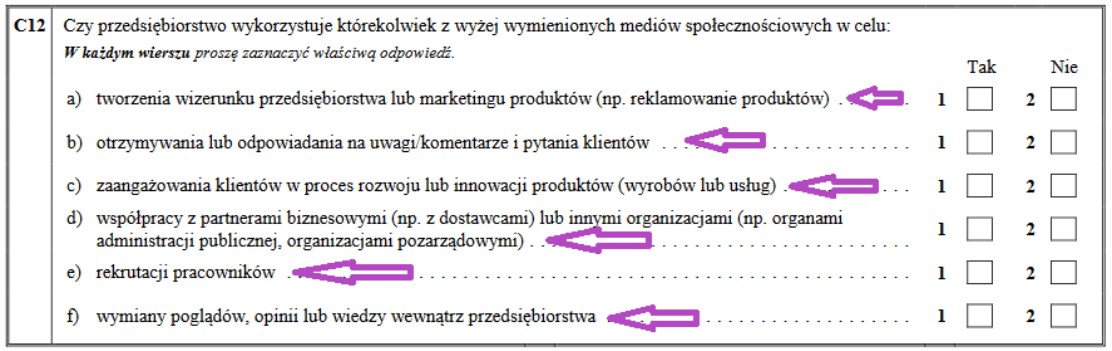
*Proszę **nie uwzględniać** wykorzystania mediów społecznościowych wyłącznie w celu umieszczania płatnych ogłoszeń reklamowych.*

***W każdym wierszu** proszę zaznaczyć właściwą odpowiedź.*

	Tak	Nie
a) serwisy społecznościowe (np. Facebook, Linkedin, itp.)	1 ☐	2 ☐
b) blogi lub mikroblogi prowadzone przez przedsiębiorstwo (np. Twitter, Present.ly itp.)	1 ☐	2 ☐
c) strony umożliwiające udostępnianie multimediów (You Tube, Flickr, Picasa, SlideShare itp.)	1 ☐	2 ☐
d) narzędzia wymiany informacji Wiki	1 ☐	2 ☐

Facebook comments

- Some elements can be changed (e.g., names of functions, attributes names).
- Destination address of Facebook will not be modified.
- Therefore in the results of analysis, the data can be provided by particular Facebook objects presented on the webpage.



C12	Czy przedsiębiorstwo wykorzystuje którekolwiek z wyżej wymienionych mediów społecznościowych w celu: *W każdym wierszu proszę zaznaczyć właściwą odpowiedź.*	Tak	Nie
a)	tworzenia wizerunku przedsiębiorstwa lub marketingu produktów (np. reklamowanie produktów)	1 ☐	2 ☐
b)	otrzymywania lub odpowiadania na uwagi/komentarze i pytania klientów	1 ☐	2 ☐
c)	zaangażowania klientów w proces rozwoju lub innowacji produktów (wyrobów lub usług)	1 ☐	2 ☐
d)	współpracy z partnerami biznesowymi (np. z dostawcami) lub innymi organizacjami (np. organami administracji publicznej, organizacjami pozarządowymi)	1 ☐	2 ☐
e)	rekrutacji pracowników	1 ☐	2 ☐
f)	wymiany poglądów, opinii lub wiedzy wewnątrz przedsiębiorstwa	1 ☐	2 ☐

-

-
- <script>(function(d, s, id) {
- var js, fjs = d.getElementsByTagName(s)[0];
- if (d.getElementById(id)) return;
- js = d.createElement(s); js.id = id;
- js.src = "//**connect.facebook.net**/pl_PL/sdk.js#xfbml=1&version=v2.8";
- fjs.parentNode.insertBefore(js, fjs);
- }(document, 'script', 'facebook-jssdk'));</script>

Useful links

- https://developers.facebook.com/docs/plugins
- https://dev.twitter.com/overview/api

- After identifying of the presense on Twitter or Facebook, we can make analysis of the activity of an enterprise on social media using API.

TOOLS AND POSSIBLE FRAMEWORKS

PYTHON 3/
HTML PARSER
OR BEAUTIFUL SOUP
/HADOOP/SPARK

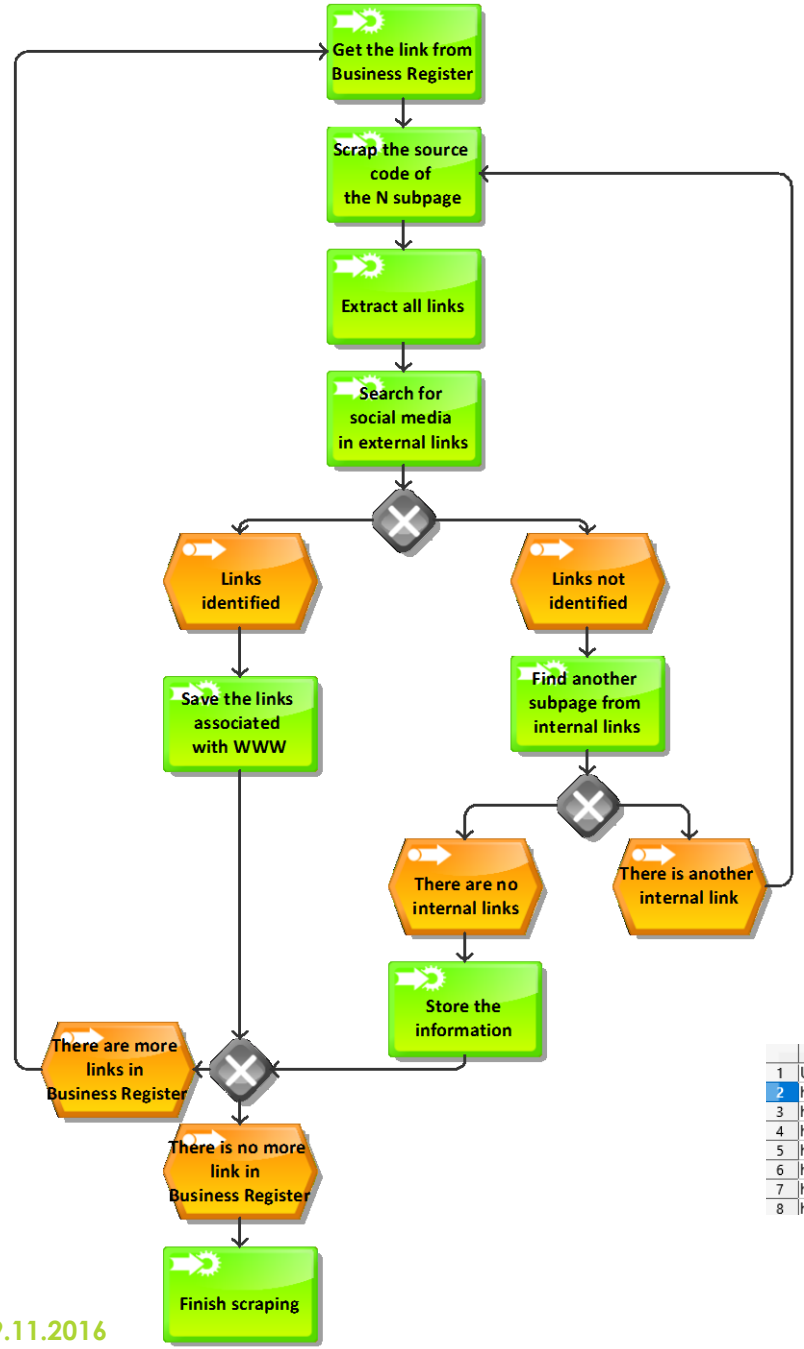
WWW addresses of enterprise webpages in Business Register

- Lots of attributes including the most important:
  - Unique identifier
  - Name
  - WWW
  - Date of last update of the link

The simplest way – Python with HTML Parser

- Parse the source code to find all hyperlinks related to the website
- Remove any duplicates
- Try to search for Twitter and Facebook accounts
- If not identified go to other links
- Identified – store the data

DEMO




```
maslankowski.pl
maslankowski.pl
http://maslankowski.pl
http://wzr.pl
8
brak
http://maslankowski.pl
['http://kie.wzr.pl', 'http://wzr.pl', 'http://www.univ.gda.pl',
', 'index.php?Akcja=consultations&PHPSESSID=3c28a4685f44dde671431
44dde6714317bd8d566310']
Z1:http://kie.wzr.pl
Z1:http://wzr.pl
Z1:http://www.univ.gda.pl
W1:./indexpl.php?Akcja=&PHPSESSID=3c28a4685f44dde6714317bd8d56631
W1:index.php?Akcja=&PHPSESSID=3c28a4685f44dde6714317bd8d566310
W1:index.php?Akcja=consultations&PHPSESSID=3c28a4685f44dde6714317
W1:index.php?Akcja=links&PHPSESSID=3c28a4685f44dde6714317bd8d5663
W1:index.php?Akcja=contact&PHPSESSID=3c28a4685f44dde6714317bd8d56
jacek@bigdata:~/WP2$
```

	A	B	C
1	URL	Facebook	Twitter
2	http://www.fiat.pl	https://www.facebook.com/FiatPL?ref=hl	https://twitter.com/Fiat_Polska
3	http://www.stat.gov.pl		https://twitter.com/GUS_STAT
4	http://www.ug.edu.pl	https://www.facebook.com/UniwersytetGdanski	
5	http://www.wzr.pl	https://www.facebook.com/Wydzia%C5%82-Zarz%C4%85dzania-Uniwersytet-Gda%C5%84ski-207090449326598/	
6	http://maslankowski.pl		
7	http://www.zalando.pl	https://www.facebook.com/zalando.polska	https://twitter.com/ZalandoPL
8	http://answear.com	http://www.facebook.com/ANSWEARcom	

FUTURE WORK AND CONCLUSIONS

FUTURE WORK

- Extend the framework to gather all the issues related to social media activity.
- Validation of the social media sites – if functioning or not.
- Test this environment on a large scale.

CONCLUSIONS AND QUESTIONS

- PL is ready to start webscraping but only experimental, due to legal issues
- Some WWW's need to be identified, according to the provided solution
- Any questions?