НАЦИОНАЛЕН СТАТИСТИЧЕСКИ ИНСТИТУТ
www.nsi.bg

State of the work (SGA-I) Bulgarian National Statistical Institute

WP2, November 8th 2016, Rome

Legal issues

- The legal framework for the use of data for Official Statistics in Bulgaria:

- Statistics Act;
- Personal Data Protection Act;
- Copyright and Neighbouring Rights Act;
- Electronic Commerce Act;

Statistics Act, art.7: "The National Statistical Institute shall collect process and store individual and personal data and statistical information"

Statistics Act

- Copyright protection

- Individual data received and collected through statistical surveys shall constitute a statistical secret and may be used only for statistical purposes;
- Individual data received for the purposes of statistical surveys may not be used as evidence before the bodies of the executive and the judiciary.

- Privacy protection

- The National Statistical Institute and statistical authorities shall have the obligation to ensure protection of individual data and prevention of any misuse thereof by undertaking appropriate organizational and technical measures and allowing such data to be handled only by persons who have signed an affidavit for protection of statistical secrecy.

Legal issues: conclusions

- We don’t expect to arise legal constraints according to different use-cases
- The BNSI’s legal department considers that there are no legal constraints on the web-scrapping activities at present.

What we have done?

- URLs Retrieval

- A list of enterprises with available URLs and e-mails from BR (26836, 2006 URLs, 20649 e-mails);
- Result: verified 7038 URLs (2006 URLs obtained from BR and generated from the domain name of e-mail);
- All verified URLs are stored within DB;
- The Jabse Search API was used to suggest a possible URLs of the enterprises on the base of the enteprise's name with some text analysis (in Bulgarian and transliterated in Latin);
- The Google Custom Search API was used to suggest a possible URLs of the enterprises on the base of the enteprise's name;
- The URL Inventory with validated 9809 URLs.

Next steps…..

E-commerce

1. Developing of web-spider script for web-scraping (in real time) has started recently.
2. Web-scraping activity will be started to check whether an enterprise performs e-commerce or not for validated 9809 URLs.
3. The statistical hypothesis testing is planning to tests how precise and comprehensive the web-scraping algorithm would be (***precise*** means companies identified as e-traders are really e-traders; ***comprehensiveness*** is the ability of algorithm to capture all e-traders in the population of companies).
4. Assuming 90 % precision and 80 % comprehensiveness are OK for the algorithm
5. Creating a 10% sample from both populations: e-traders and remaining non e-traders and verified statistically whether precision is lower than 90 percent and comprehensiveness is lower than 80 percent. Employing normal distribution for both hypothesis we will get to the conclusion that our filter is both precise and comprehensive.

Social media……- the same steps as E-commerce use-case

Thank you
for your attention!