„FOLLOW THE RIVER TO REACH THE SEA”

The Analysis Department of the Ministry of Finance of the Republic of Poland invites you to face a challenge that requires programming and web mining skills, data base knowledge and critical thinking.

The task is to create a working prototype of the identification system of give phrases (offers) on the Internet (sites, auctions, announcements, forums, home pages, etc.), containing contact details e.g. telephone number, e-mail addresses, NIP data, REGON, VIN etc. with power indication (frequency) of these connections. Basing on a list of keywords, you have to find all possible identifiers and the subsequent levels of associations.

The prototype should accept a list of key phrases (e.g. "sell costs", "Bitcoin", "Iphone" etc.) in the form of a text file (or an internal editor that enables multi-line import of text files) as a source, and then search the content of the websites for contact identifiers in the vicinity of the above-mentioned phrases or other identifiers defined in the detailed assumptions of the project. In addition, the application should enable the parameterization of inquiries regarding the scope of the search (global search, limited to specific national domains or monitoring of the indicated list of websites, where, after finding contact data, the transition to global verification is carried out in order to find possible connections). The system should allow for additional parameterization supporting the search efficiency, e.g. regarding the number of subpages (in the case of implementing a cascade scanning module based on link tracking), the possibility of accepting the results of search engines only, the type of identifiers searched, the number of degrees of links / or other functionalities influencing the search process.

The program should present the search result in a clear and ergonomic form, e.g. a list of identified first degree contact details (directly related to the searched phrases) along with their frequency. Additionally, it should be possible to open links to specific webpages where a "hit" occurred to confirm the result, or classify it as a "false positive" and remove it from the set of results.

The next step should be the identification of the second degree relationships, i.e. verification whether the searched data coexist with other types of identifiers on websites including announcements, auctions, internet forums, social media, WHOIS databases, publicly available registers, etc.

The results (pairs) should be presented in a form that allows for an immediate assessment of the likelihood of a link, e.g. on the basis of the frequency of occurrences of the same pair on different webpages (along with links).

Subsequent levels of linkage identification should be extended in the same way until the defined limit of linkage levels is reached or the algorithm is manually stopped by the user. Contact details can be supplemented with additional information flags, e.g.

("Is Allegro", "Is eBay", "Is Forum", "Has its own website" etc.). The application should present the results in a table with the possibility of exporting these data to a flat file, as well as in the form of a diagram enabling the visualization of relationships between individual data types (ultimately with the possibility of making a simple data analysis, e.g. searching for the shortest path, all paths, filtering and deleting objects, etc., and saving the results to the database). As an additional functional element of the tool, it is expected to be able to identify connections for a specific value of a defined identifier, e.g. a telephone number, without referring to the previously submitted phrase list.

Sample solution

The analyst introduces several phrases with a possibility of choosing whether they are to be treated at the first level of search as AND (all occurrences) or OR (the page is indexed if any of the phrases appears). Then, after setting additional parameters, a first-level scan takes place, as a result of which a database of identifiers that were detected on webpages containing keywords is created, along with their frequency.

Keywords	**Type**	**Identifier**	**Frequency (number of appearances)**	**0-1 flag #1 (e.g. CZY_Allegro)**	**0-1 flag #n (e.g. CZY_forum)**
Sell (250)	e-mail	mailjakis@xyz.pl	120	5	0
Product X (300)	Phone number	500 500 500	90	7	0
Product X (300)	Phone number	600 600 600	45	3	0
Product X (300)	e-mail	poczta@abc.com.pl	40	2	2
Low price (3000)	Landline phone	(32) 123 45 67	30	0	0
No intermediaries (800)	e-mail	jakisinny@zyx.pl	12	3	2
Other phrase (30)	…	…	…		

The table above shows an example illustrating the structure of a single-level search result. The algorithm displays the most frequent contact details found on the webpages containing the indicated keyword. Additionally, on the basis of the information flags described in the previous section, it is possible to quickly verify what type of websites a given identifier is in (additional information - identified frequencies). The user has the option to define a "cut-off point" beyond which a given contact will be treated as irrelevant and not included in further web data mining (or can make a selection manually).

The dynamic fields with a list of websites are displayed after clicking on a given identifier (with the possibility of opening them for verification or excluding from the set of results) and a list of keywords that have been associated with the identifier.

The user should be able to define the parameters for level 1 search – maximum search time, number of hits or manual termination. Then the user can select which identifiers to search in the next level scan (possible options "select all", searching for associations only for a specific type of identifier, manual selection, etc.).

The search of subsequent levels should take place without taking into account the keywords (i.e. global search, possibly narrowed down in accordance with the initial settings). The sample result of a second level scan should logically reflect the structure:

Type Level 1	**Identifier Level 1**	**Type Level 2**	**Identifier Level 2**	**Number of appearances (pairs)**
e-mail	mailjakis@xyz.pl	Mobile phone	700 700 700	20
e-mail	mailjakis@xyz.pl	Landline phone	(22) 333 44 55	15
Landline phone	(32) 123 45 67	e-mail	Jakisinny2@yyy.pl	10
e-mail	jakisinny@zyx.pl	Mobile phone	500 500 500	10
…	…	…	…	…
Mobile phone	500 500 500	Mobile phone	-	0

The result of the second level scan (and subsequent levels) should enable the identification of connections between identifiers along with the possibility of dynamic verification (opening a specific website in order to possibly exclude the link). Additional data such as information flags are displayed in a way that enables effective management of results.

After the connections identification process is finished, the program should be able to display the results in the form of:

• table - with the option of exporting to an external file, possible to open and further analyze in commonly available programs (e.g. txt, csv, xls, etc.),

• graph - in the form of a diagram of connections (e.g. based on a graph base), enabling a clear visualization of connections (it may contain a visualization of clusters consisting of individual identifiers). The user should be able to edit the diagram, e.g. moving, deleting objects and saving the result to a file (e.g. bmp, jpg, png).

Functional requirements:

1. Keywords import from an external file (or the ability to copy them to a dedicated dialog box),

2. Scan of connections based on a single identifier (or list of identifiers),

3. Classification of types of connections,

4. Parameterization of queries (at least the search scope: global or limited to the list of domains - separately for level 1, 2 and the subsequent ones, defining the types of identifiers searched, criteria for the search stop),

5. Graphical user interface,

6. Export the result to an external file,

7. Anti-robot mechanisms on the websites,

8. Visualization in the form of a graph (additionally scored).

Evaluation criteria:

On November 29, 2020 at 8:00 am, on a dedicated Discord channel (# -follow-the-river), the input data (keywords for the search algorithm) will be provided, giving the possibility to verify the created solution at the second stage of the evaluation.

The result of the algorithm will be one of the elements of the overall evaluation, which consists of:

Stage 1

- Idea - 30%,
- Technical aspects - 30%,
- design - 20%,
- relation to task category - 10%,
- wow! Factor - 10%.

Stage 2

- algorithm efficiency (0-40 points),
- interface and ergonomics (20 points),

- configurability and parameterization of the algorithm (0-15 points),
- presentation of results (0-10 points),
- innovative solution, including additional functions not indicated in the requirements, which significantly contribute to the functionality (0-15 points)

At each stage of the evaluation, the Jury will award more points to solutions developed from scratch or using open-source tools. Any use of commercial solutions must allow for the quick evaluation of the solution, without the need to incur additional costs.

We ask you to submit your final solutions to our challenge in the form the following files:

- "1.INFO.zip" - an archive containing all kinds of information materials, descriptions, screenshots, presentations, links to films and other visualizations related to the task,
- "2.CODE.zip" - an archive containing source codes or addresses of file locations with source codes,
- "3.FOLLOW_THE_RIVER.zip" - an archive containing a compiled version of the tool or a link with its location, as well as a description of the steps and possible configuration of the environment to run the solution on Windows or Linux in a short time,
- "4.TEST.zip" - archive containing the test scenario result,
- "5.OTHERS.zip" - an archive containing all other information.