ABSTRACT

Explanatory note size – 107 pages, contains 20 illustrations, 28 tables, 3 applications, 21 references.

Topicality. Every year, the amount of data is increasing, it can be useful in any area of our life, provided it is properly processed. The topic of the work is relevant, because today there is no universal tool for collecting extremely large arrays of text data from various sources.

The goal of the work is to unify the structure and format of super-large arrays of text data through the use of architectural solutions that allow users to expand it for their own purposes with minimal effort. To achieve this goal, it is necessary to solve the following problems:

- perform the comparative analysis of available solutions for collecting super-large arrays of text data;
- formulation of the technical features of the collection of extremely large arrays of text data;
- development of a unified structure of super-large text data collected from various sources;
- development of software for collecting extremely large arrays of text data;
- implementation of modular architecture in a software solution;
- evaluation of the effectiveness of the proposed solution.

The object of research of the work is mathematical, informational and software for collecting super-large arrays of text data.

The subject of research is methods of collecting extremely large arrays of textual data.

The scientific novelty of the work is the creation of a unified data structure for the sources of large text data of various nature, which includes the storage of the time stamp and data source, as well as the declaration of a strict structure.

The practical significance of the obtained results lies in the possibility of using the proposed unified structure for integration between different systems for collecting extremely large arrays of text data.

Relationship with working with scientific programs, plans, topics. The work was performed at the Department of Informatics and Software Engineering of the National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute" in the framework of the topic "Methods and technologies of high-performance computing and big data processing". State registration number 0117U000924.

Approbation. The main provisions of the work were reported and discussed at the III All-Ukrainian scientific and practical conference of young scientists and students "Software engineering and advanced information technologies (Soft-Tech-2022)".

Publications. The scientific provisions of the dissertation were published in:

1) Kuvichka M.Y. Unification of the structure of super-large arrays of text data collected from various sources / M.Y. Kuvichka, Yu.O. Oliinyk // Materials of the III All-Ukrainian scientific and practical conference of young scientists and students "Software engineering and advanced information technologies" (SoftTech-2022 autumn) - Kyiv: NTUU "KPI them. Igor Sikorsky", November 23-25, 2022.

Keywords: BIG DATA, DATA COLLECTION, DATA STRUCTURING, WEBSCRAPING.