

ПОБУДОВА ІНСТРУМЕНТУ ЦИФРОВОГО АНАЛІЗУ ТЕКСТІВ



Методи визначення подібності текстів знаходяться на передньому краї у таких областях досліджень, як комп'ютерна лінгвістика, літературознавство, комунікаційні науки, філософія, а також науки про здоров'я. Визначення подібності користується попитом у закладах вищої освіти та академічних спільнотах. Але на відміну від комерціалізованих сервісів «антиплагіатної» перевірки, інтеграція інструменту визначення подібності у великі національні текстові архіви надає значно ширші можливості для його використання з позицій доступу до контенту та урізноманітнення дослідницьких цілей.

Під час виконання дослідження у рамках спільного українсько-латвійського проєкту* «Методи текстового аналізу та інструменти визначення подібності у великих національних текстових архівах: на прикладі Латвійської Національної цифрової бібліотеки та Національного репозитарію академічних текстів України» нами розроблений власний інструмент співставлення й аналізу текстів. Він побудований на розподіленому пошуковому та аналітичному двигуні з відкритим вихідним кодом, написаним на Java, який підтримує велику кількість типів даних, включаючи текстові, числові, геопросторові, структуровані та неструктуровані ElasticSearch.



Новий інструмент, який отримав робочу назву «Antic», нами протестовано на повній базі академічних текстів, що містяться у Національному репозитарії академічних текстів України. У результаті порівняння завантаженого для перевірки тексту та бази даних НРАТ, користувач отримує звіт з переліком усіх знайдених співпадінь текстових фрагментів, вказуванням кількості таких співпадінь й активними гіперпосиланнями на джерела інформації, де такі збіги було віднайдено - звіти про НДДКР, автореферати та дисертації на здобуття наукового ступеня.

**Дослідження виконується за фінансової підтримки
Міністерства освіти і науки України**

Володимир Камишин, директор УкрІНТЕІ

<https://orcid.org/0000-0002-8832-9470>

Олексій Сухий, заступник директора УкрІНТЕІ

<https://orcid.org/0000-0002-3479-4123>

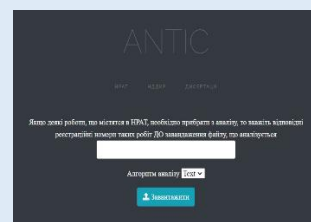
ua_lv@ukrintei.ua, uintei@uintei.kiev.ua

CONSTRUCTION OF A TOOL FOR THE DIGITAL ANALYSIS OF TEXTS



Methods for determining the similarity of texts are at the forefront of such fields of research as computational linguistics, literary studies, communication sciences, philosophy, and health sciences. Similarity determination is in demand in higher education institutions and academic communities. But unlike commercialized "anti-plagiarism" checking services, the integration of the similarity detection tool into large national text archives provides much wider opportunities for its use in terms of access to content and diversification of research purposes.

During the research in the framework of the joint Ukrainian-Latvian project* "Methods of text analysis and tools for determining similarities in large national text archives: on the example of the Latvian National Digital Library and the National Repository of Academic Texts of Ukraine", we developed our own text matching and analysis tool. It is built on an open-source distributed search and analytics engine written in Java that supports a large number of data types including text, numeric, geospatial, structured, and unstructured ElasticSearch.



The new tool, which received the working name "Antic", was tested by us on the full database of academic texts contained in the National Repository of Academic Texts of Ukraine. As a result of comparing the text downloaded for verification and the NRAT database, the user receives a report with a list of all found matches of text fragments, indicating the number of such matches and active hyperlinks to sources of information where such matches were found - R&D reports, abstracts and dissertations for obtaining a scientific degree.

The research is carried out with the financial support of the Ministry of Education and Science of Ukraine

Volodymyr Kamyshyn, Director UkrISTEI

<https://orcid.org/0000-0002-8832-9470>

Olexsii Suhyi, Deputy Director UkrISTEI

<https://orcid.org/0000-0002-3479-4123>

ua_lv@ukrintei.ua, uintei@uintei.kiev.ua