An Intelligent Information System for Generating a Scientist's Scientometrics Using Content Analysis Methods

Mykola Dyvak [1,†], Andriy Yushko[1,*,†], Andriy Melnyk[1,†] and Tiande Pan[1,†]

[1] *West Ukrainian National University, 11 Lvivska Street, Ternopil, 46001, Ukraine*

Abstract

The paper proposes methods and software tools for developing a scientometric profile of a researcher using content analysis techniques. A scientometric profile is a system of indicators that assesses a researcher's scientific productivity and influence. The growing volume of scientific information in various databases, such as Scopus and Web of Science, has made it challenging to manually track and analyze individual publishing activities. For scientific and higher education institutions, monitoring both the quantity and quality of publications is crucial. Additionally, understanding researchers' main areas of interest helps support their professional development and foster interdisciplinary collaboration. Existing tools for monitoring scientific metrics typically offer limited functionality, lack the ability to process large volumes of data efficiently, and struggle to filter irrelevant information automatically. This paper presents an approach to building a researcher's scientometric profile using content analysis, supported by large language models, specifically Ollama. A mathematical model was developed to filter out irrelevant publications based on the researcher's scientometric profile. The system for collecting and analyzing scientometric indicators was implemented, and experimental studies were conducted using profiles of researchers from West Ukrainian National University.

Keywords

intelligent information system, scientometrics, researcher, content analysis methods, large language model, irrelevant publications.

1. Introduction

In the modern scientific environment, the problem of managing the publishing activity of professors and teaching staff is gaining more and more importance. The increase in the volume of scientific information placed in various scientometric databases, such as Scopus, Web of Science, and others, greatly complicates the manual process of tracking and analyzing the publication activity of each individual researcher. For scientific institutions and institutions of higher education, it is important not only to monitor the number and quality of publications, but also to understand the main scientific interests of employees, which allows to support their professional activities and develop interdisciplinary cooperation [1-4].

Existing tools for monitoring scientific publications, as a rule, have limited functionality, are not able to efficiently process large volumes of information and automatically filter irrelevant publications. They often require significant human resources to analyze and select relevant data, which is not always possible for large scientific teams [1,5].

The purpose of this article is to present the developed software system that automates the collection, analysis and processing of the publishing activity of the teaching staff. The system not

[*] 8th International Scientific and Practical Conference Applied Information Systems and Technologies in the Digital Society AISTDS'2024, 2024, October 1, Kyiv, Ukraine

[*] Corresponding author.

[†] These authors contributed equally.

🖂 mdy@wunu.edu.ua (M. Dyvak); a.yushko@wunu.edu.ua (A. Yushko); ame@wunu.edu.ua (A. Melnyk); pan54453@163.com (T. Pan)

0000-0002-9049-4993 (M. Dyvak); 0009-0003-6431-3479 (A. Yushko); 0000-0001-7799-9877 (A. Melnyk); 0009-0005-4371-8084 (T. Pan);

only allows you to collect information from scientific databases about publications, projects, grants and participation in scientific events, but also forms a profile of a scientist, determining his scientific interests. Using this profile, the system is able to filter irrelevant publications, automatically assessing their relevance to the scientist's interests. This decision contributes to increasing the efficiency of scientific activity, allowing to focus attention on really important and relevant scientific achievements.

2. Algorithms and approaches for selecting keywords and determining the researcher's scientific interests

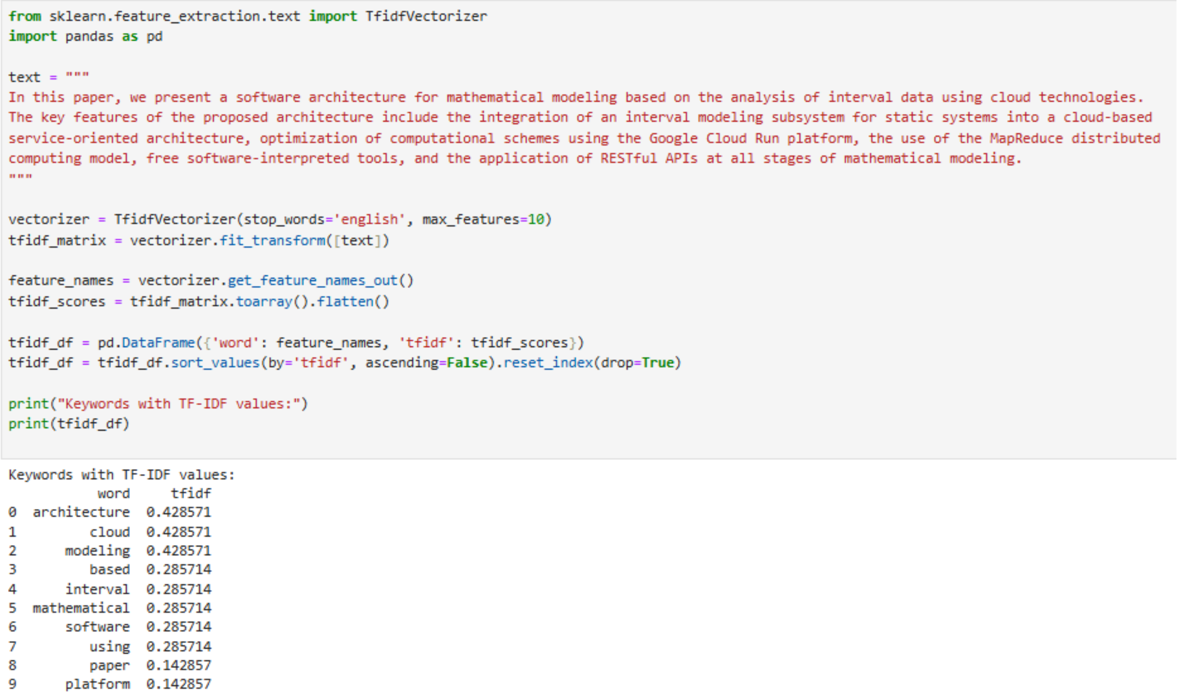
Modern research actively uses algorithms for automatic analysis of text data to select keywords that reflect the main scientific interests of the researcher. The development of such approaches is aimed at simplifying the process of collecting, analyzing and systematizing scientific materials, which allows not only to identify the main areas of work, but also to identify interdisciplinary connections.

The main methods used to analyze texts for the purpose of extracting keywords can be divided into several categories:

3. Statistical methods

One of the basic approaches is to calculate the frequency of use of terms in texts. The TF-IDF metric (Term Frequency-Inverse Document Frequency) is the most popular among statistical methods and allows taking into account both the frequency of a term in a document and its significance in the context of the entire corpus of texts [6]. This increases the accuracy of extracting significant terms, as frequent but insignificant words are given less weight.

Figure 1 shows an example of the implementation of the TF-IDF metric in the Python programming language using the scikit-learn library.



```python
from sklearn.feature_extraction.text import TfidfVectorizer
import pandas as pd

text = """
In this paper, we present a software architecture for mathematical modeling based on the analysis of interval data using cloud technologies.
The key features of the proposed architecture include the integration of an interval modeling subsystem for static systems into a cloud-based
service-oriented architecture, optimization of computational schemes using the Google Cloud Run platform, the use of the MapReduce distributed
computing model, free software-interpreted tools, and the application of RESTful APIs at all stages of mathematical modeling.
"""

vectorizer = TfidfVectorizer(stop_words='english', max_features=10)
tfidf_matrix = vectorizer.fit_transform([text])

feature_names = vectorizer.get_feature_names_out()
tfidf_scores = tfidf_matrix.toarray().flatten()

tfidf_df = pd.DataFrame({'word': feature_names, 'tfidf': tfidf_scores})
tfidf_df = tfidf_df.sort_values(by='tfidf', ascending=False).reset_index(drop=True)

print("Keywords with TF-IDF values:")
print(tfidf_df)
```

```
Keywords with TF-IDF values:
           word     tfidf
0  architecture  0.428571
1         cloud  0.428571
2      modeling  0.428571
3         based  0.285714
4      interval  0.285714
5  mathematical  0.285714
6      software  0.285714
7         using  0.285714
8         paper  0.142857
9      platform  0.142857
```

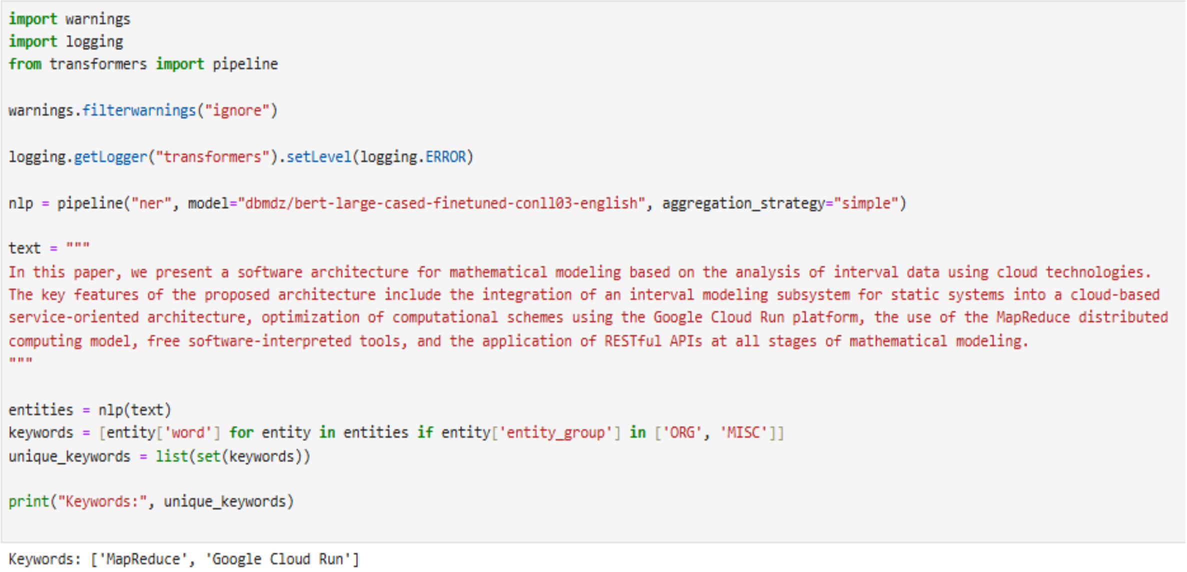
Figure 1: Implementation of the statistical method of selecting keywords from the test using the TF-IDF metric.

As a result of executing the code, we will get a table with the top 10 keywords and their TF-IDF values.

The TF-IDF value of each keyword reflects its weight in the context of the article's annotation – the higher the TF-IDF value, the more important the term is for this text

4. Rule-based methods

Rule-based approaches, such as Named Entity Recognition (NER), allow the extraction of certain categories of words, such as names of organizations, names of people, geographic locations, and other important entities [7]. In the Python programming language, you can use the transformers library from Hugging Face, which allows you to load a pre-trained model for recognizing named entities (Fig. 2).



```
import warnings
import logging
from transformers import pipeline

warnings.filterwarnings("ignore")

logging.getLogger("transformers").setLevel(logging.ERROR)

nlp = pipeline("ner", model="dbmdz/bert-large-cased-finetuned-conll03-english", aggregation_strategy="simple")

text = """
In this paper, we present a software architecture for mathematical modeling based on the analysis of interval data using cloud technologies.
The key features of the proposed architecture include the integration of an interval modeling subsystem for static systems into a cloud-based
service-oriented architecture, optimization of computational schemes using the Google Cloud Run platform, the use of the MapReduce distributed
computing model, free software-interpreted tools, and the application of RESTful APIs at all stages of mathematical modeling.
"""

entities = nlp(text)
keywords = [entity['word'] for entity in entities if entity['entity_group'] in ['ORG', 'MISC']]
unique_keywords = list(set(keywords))

print("Keywords:", unique_keywords)
```

```
Keywords: ['MapReduce', 'Google Cloud Run']
```

Figure 2: Implementation of the NER method for searching keywords.

As you can see from the code above, we use the pipeline method with a pre-trained dbmdz/bert-large-cased-finetuned-conll03-english model that is specially tuned for Named Entity Recognition (NER). The aggregation_strategy="simple" parameter allows you to aggregate the results for greater convenience.

The next step is to run NER on the text. This pulls up a list of colds with the specified types (eg organizations, technology names, scientific concepts).

After that, keyword filtering is performed by selecting entities that may be relevant. For example ORG (organizations) and MISC (various terms such as technology or scientific concepts).

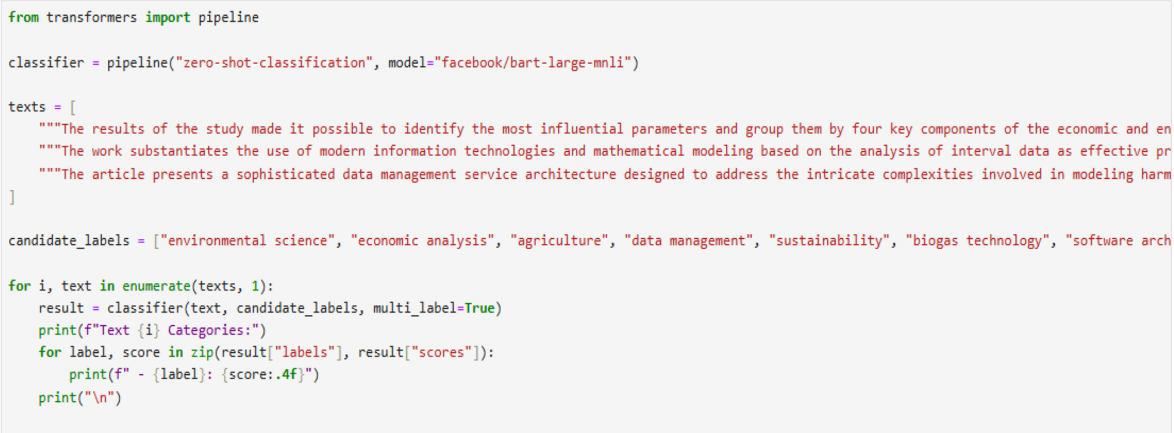
After passing all the stages, we get a list of keywords selected from the text, in our case it is: Google Cloud Reduce and MapReduce.

5. Natural language processing (NLP) models

Thanks to the development of natural language processing methods and the emergence of deep models such as BERT, GPT and others, it became possible to significantly improve the accuracy of text analysis [8]. These models take into account the context of words, which allows not only to highlight keywords, but also to understand their relationship and semantic meaning.

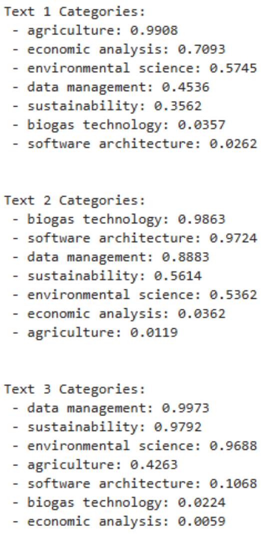
Figure 3 shows a code fragment for selecting article categories by their annotations using the ready-made facebook/bart-large-mnli model from the Transformers library.

In Figure 4, the output shows which categories most closely match each text, as well as the confidence level of the model corresponding to each category.



```python
from transformers import pipeline

classifier = pipeline("zero-shot-classification", model="facebook/bart-large-mnli")

texts = [
    """The results of the study made it possible to identify the most influential parameters and group them by four key components of the economic and en
    """The work substantiates the use of modern information technologies and mathematical modeling based on the analysis of interval data as effective pr
    """The article presents a sophisticated data management service architecture designed to address the intricate complexities involved in modeling harm
]

candidate_labels = ["environmental science", "economic analysis", "agriculture", "data management", "sustainability", "biogas technology", "software arch

for i, text in enumerate(texts, 1):
    result = classifier(text, candidate_labels, multi_label=True)
    print(f"Text {i} Categories:")
    for label, score in zip(result["labels"], result["scores"]):
        print(f" - {label}: {score:.4f}")
    print("\n")
```

Figure 3: Implementation of the BART method.

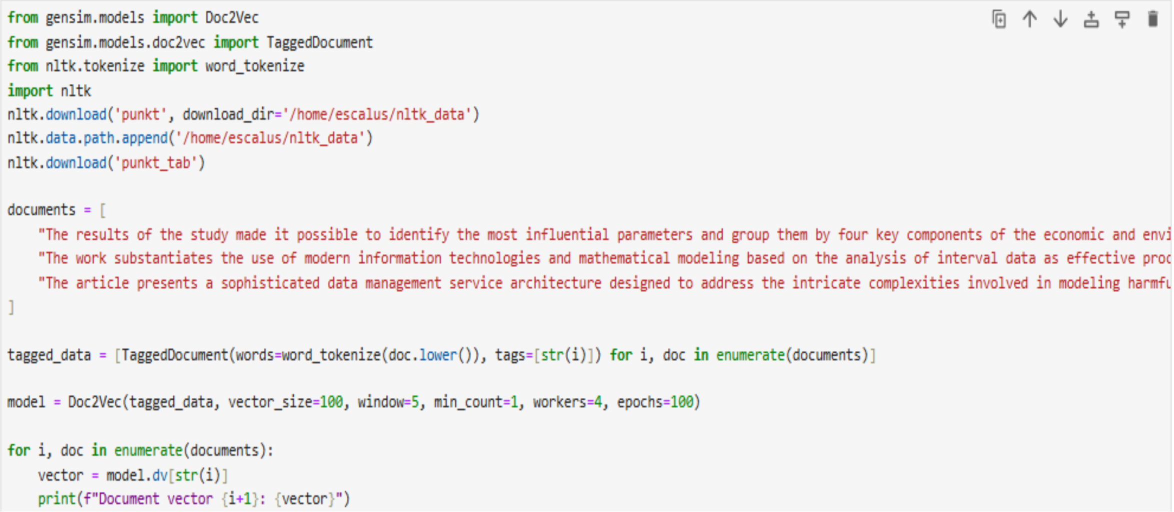


```
Text 1 Categories:
 - agriculture: 0.9908
 - economic analysis: 0.7093
 - environmental science: 0.5745
 - data management: 0.4536
 - sustainability: 0.3562
 - biogas technology: 0.0357
 - software architecture: 0.0262


Text 2 Categories:
 - biogas technology: 0.9863
 - software architecture: 0.9724
 - data management: 0.8883
 - sustainability: 0.5614
 - environmental science: 0.5362
 - economic analysis: 0.0362
 - agriculture: 0.0119


Text 3 Categories:
 - data management: 0.9973
 - sustainability: 0.9792
 - environmental science: 0.9688
 - agriculture: 0.4263
 - software architecture: 0.1068
 - biogas technology: 0.0224
 - economic analysis: 0.0059
```

Figure 4: The result of text categorization using the BART method.

6. Text vectorization

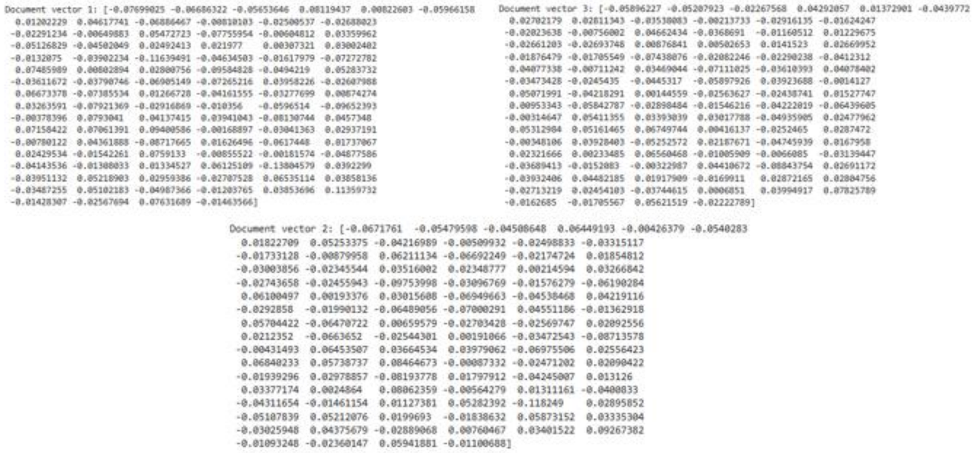
To vectorize the text and create its numerical representation, you can use the Word2Vec or Doc2Vec methods from the gensim library in Python. Word2Vec creates vectors for individual words, while Doc2Vec allows you to get a vector representation for an entire document[9] (Fig. 5).



```python
from gensim.models import Doc2Vec
from gensim.models.doc2vec import TaggedDocument
from nltk.tokenize import word_tokenize
import nltk
nltk.download('punkt', download_dir='/home/escalus/nltk_data')
nltk.data.path.append('/home/escalus/nltk_data')
nltk.download('punkt_tab')

documents = [
    "The results of the study made it possible to identify the most influential parameters and group them by four key components of the economic and envi
    "The work substantiates the use of modern information technologies and mathematical modeling based on the analysis of interval data as effective proc
    "The article presents a sophisticated data management service architecture designed to address the intricate complexities involved in modeling harmfu
]

tagged_data = [TaggedDocument(words=word_tokenize(doc.lower()), tags=[str(i)]) for i, doc in enumerate(documents)]

model = Doc2Vec(tagged_data, vector_size=100, window=5, min_count=1, workers=4, epochs=100)

for i, doc in enumerate(documents):
    vector = model.dv[str(i)]
    print(f"Document vector {i+1}: {vector}")
```

Figure 5: Implementation of the Doc2Vec vectorization method.

As a result of executing this code, we will get a vector representation for three annotation texts, which is shown in Figure 6.



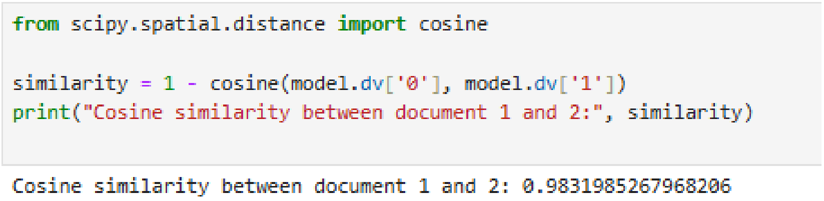
```
Document vector 1: [-0.07699025 -0.06686322 -0.05653646  0.08119437  0.00822603 -0.05966158
  0.01202229  0.04617741 -0.06886467 -0.00810103 -0.02500537 -0.02688023
 -0.02291234 -0.00649883  0.05472723 -0.07755954 -0.00604812  0.03359962
 -0.05126829 -0.04502049  0.02492413  0.021977    0.00307321  0.03002402
 -0.0132075  -0.03902234 -0.11639491 -0.04634503 -0.01617979 -0.07272782
  0.07485989  0.00802894  0.02800756 -0.09584828 -0.0494219   0.05283732
 -0.03611672 -0.03790746 -0.06905149 -0.07265216  0.03958226 -0.02607988
  0.06673378 -0.07385534  0.01266728 -0.04161555 -0.03277699  0.00874274
  0.03263591 -0.07921369 -0.02916869 -0.010356   -0.0596514  -0.09652393
 -0.00378396  0.0793041   0.04137415  0.03941043 -0.08130744  0.0457348
  0.07158422  0.07061391  0.09400586 -0.00168897 -0.03041363  0.02937191
 -0.00780122  0.04361888 -0.08717665  0.01626496 -0.0617448   0.01737067
  0.02429534 -0.01542261  0.0759133  -0.00855522 -0.00181574 -0.04877586
 -0.04143536 -0.01308033  0.01334527  0.06125109 -0.13804579  0.0392299
 -0.03951132  0.05218903  0.02959386 -0.02707528  0.06535114  0.03858136
 -0.03487255  0.05102183 -0.04987366 -0.01203765  0.03853696  0.11359732
 -0.01428307 -0.02567694  0.07631689 -0.01463566]
```

```
Document vector 3: [-0.05896227 -0.05207923 -0.02267568  0.04292057  0.01372901 -0.0439772
  0.02702179  0.02811343 -0.03530083 -0.00213733 -0.02916135 -0.01624247
 -0.02023638 -0.00756002  0.04662434 -0.0368691  -0.01160512  0.01229675
 -0.02661203 -0.02693748  0.00876841  0.00502653  0.0141523   0.02669952
 -0.01876479 -0.01705549 -0.07438076 -0.02082246 -0.02290238 -0.0412312
  0.04077338 -0.00711242  0.03469044 -0.07111025 -0.03610393  0.04078402
 -0.03473428 -0.0245435  -0.0445317  -0.05897926  0.03923688 -0.0014127
  0.05071991 -0.04218291  0.00144559 -0.02563627 -0.02438741  0.01527747
  0.00953343 -0.05842787 -0.02898484 -0.01546216 -0.04222019 -0.06439605
 -0.00314647  0.05411355  0.03393039  0.03017788 -0.04935905  0.02477962
  0.05312984  0.05161465  0.06749744  0.00416137 -0.0252465   0.0287472
 -0.00348106  0.03928403 -0.05252572  0.02187671 -0.04745939  0.0167958
  0.02321666  0.00233485  0.06560468 -0.01005909 -0.0066085  -0.03139447
 -0.03689413 -0.0152083  -0.00322987  0.04410672 -0.08843754  0.02691172
 -0.03932406  0.04482185  0.01917909 -0.0169911   0.02872165  0.02804756
 -0.02713219  0.02454103 -0.03744615  0.0006851   0.03994917  0.07825789
 -0.0162685  -0.01705567  0.05621519 -0.02222789]
```

```
Document vector 2: [-0.0671761  -0.05479598 -0.04508648  0.06449193 -0.00426379 -0.0540283
  0.01822709  0.05253375 -0.04216989 -0.00509932 -0.02498833 -0.03315117
 -0.01733128 -0.00879958  0.06211134 -0.06692249 -0.02174724  0.01854812
 -0.03003856 -0.02345544  0.03516002  0.02348777  0.00214594  0.03266842
 -0.02743658 -0.02455943 -0.09753998 -0.03096769 -0.01576279 -0.06190284
  0.06100497  0.00193376  0.03015608 -0.06949663 -0.04538468  0.04219116
 -0.0292858  -0.01990132 -0.06489056 -0.07000291  0.04551186 -0.01362918
  0.05704422 -0.06470722  0.00659579 -0.02703428 -0.02569747  0.02092556
  0.0212352  -0.0663652  -0.02544301  0.00191066 -0.03472543 -0.08713578
 -0.00431493  0.06453507  0.03664534  0.03979062 -0.06975506  0.02556423
  0.06840233  0.05738737  0.08464673 -0.00087332 -0.02471202  0.02090422
 -0.01939296  0.02978857 -0.08193778  0.01797912 -0.04245007  0.013126
  0.03377174  0.0024864   0.08062359 -0.00564279  0.01311161 -0.0400833
 -0.04311654 -0.01461154  0.01127381  0.05282392 -0.118249    0.02895852
 -0.05107839  0.05212076  0.0199693  -0.01838632  0.05873152  0.03335304
 -0.03025948  0.04375679 -0.02889068  0.00760467  0.03401522  0.09267382
 -0.01093248 -0.02360147  0.05941881 -0.01100688]
```

Figure 6: Vector representation of annotations

In the future, the obtained vectors can be used to compare the similarities between documents or to cluster documents based on topics. For example, we can calculate cosine similarity between vectors to find out how similar documents are to each other (Fig. 7).



```
from scipy.spatial.distance import cosine

similarity = 1 - cosine(model.dv['0'], model.dv['1'])
print("Cosine similarity between document 1 and 2:", similarity)
```

```
Cosine similarity between document 1 and 2: 0.9831985267968206
```

Figure 7: Calculation of cosine similarity of two documents.

The cosine similarity value of two documents of 0.98 indicates that these documents have a very high level of similarity in terms of their vector representations. Cosine similarity measures the angle between the vectors of two texts: a value close to 1 means that the vectors are nearly parallel, indicating a high degree of similarity between the texts.

These methods can be used both separately and in combination to achieve more accurate results in determining the researcher's key scientific interests. The use of these approaches allows automating the processes of analysis of scientific activity, which, in turn, contributes to the formation of a comprehensive profile of a researcher capable of reflecting the dynamics of his scientific work and interdisciplinary connections.

Each of the described methods has its own unique application and can complement other methods in complex tasks of text analysis. In the next section, we will look at how you can use the Ollama model with its powerful language models to identify keywords in text. This approach will make it possible to apply the latest capabilities of deep learning to improve the accuracy of extracting relevant terms and analyzing complex textual data.

7. Methodology for creating a scientometric portrait of scientist using large Ollama language models

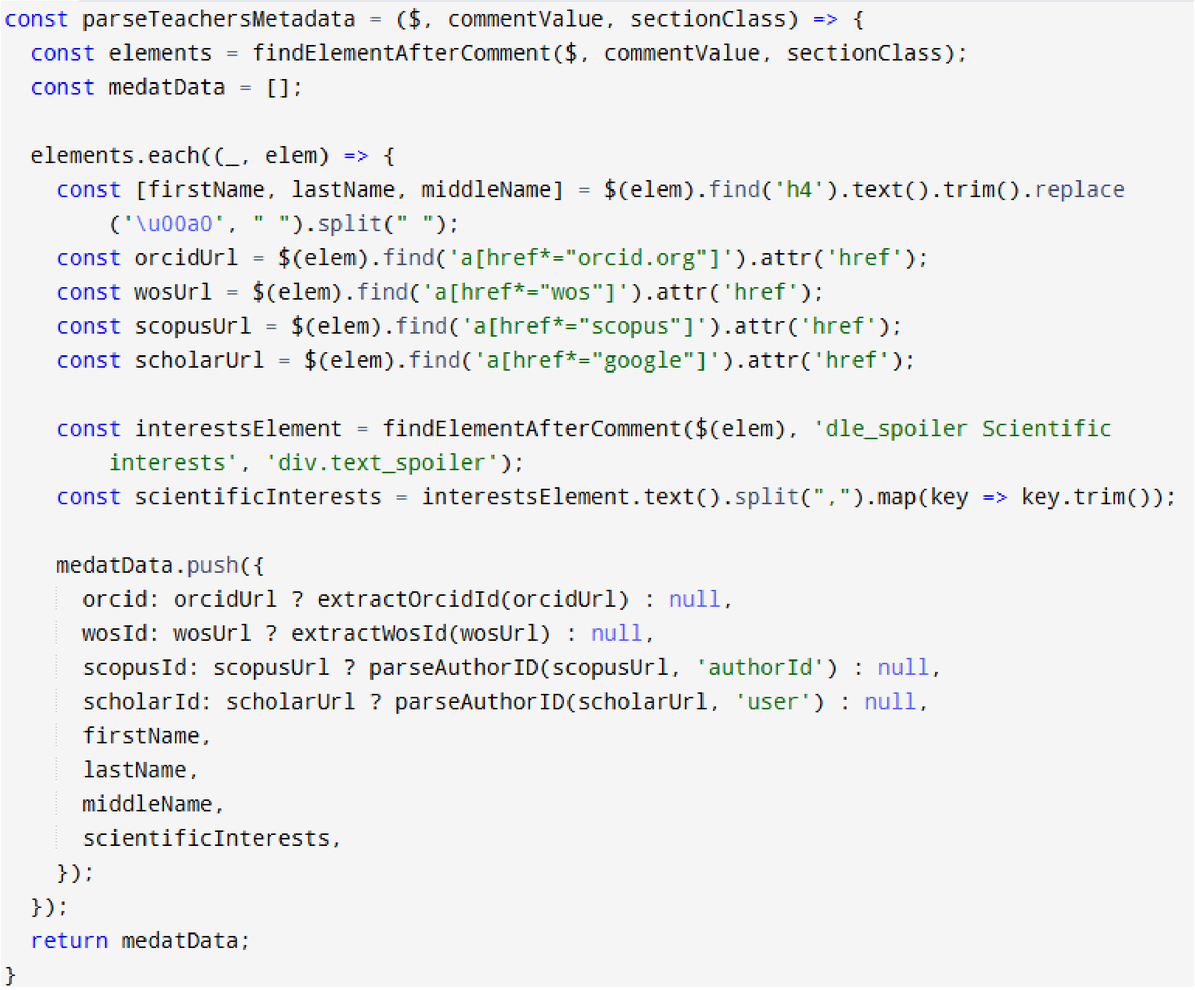
A researcher's profile is a comprehensive description of the researcher's professional activities, scientific achievements, and interests. It includes such key elements as name, surname, position, academic title, scientific interests, number of published works, participation in scientific grants and projects. The formation of a scientist's profile is an important task, since it can be used to solve such tasks as automated filtering of publications that match the researcher's scientific interests, and optimized selection of a scientific supervisor for young scientists or graduate students whose scientific activity coincides with the topic.

To form a profile of a scientist, first of all, it is necessary to collect basic metadata, which will become the foundation for further processing. The web scraping method can help us in this, which will allow us to collect basic information from the official website of the organization where the scientist works. This method provides automated extraction of such data as name, surname, position, academic title, circle of scientific interests, links to scientometric profiles of the author (Scopus, Web of Science, ORCID, Google Scholar, DSpace).

The use of web scraping at the initial stage provides automatic filling of the profile with publicly available information, which significantly reduces the time spent on manual data collection and creates an accurate starting point for further analysis.

To implement the web scraping process, you can use specialized libraries that allow you to automatically read and extract information from web pages [10,11]. For example, using the Cheerio library in JavaScript, it is possible to retrieve and process the HTML content of a page, extracting the required metadata such as name, title, academic interests, etc. The following example demonstrates the basic code for obtaining information about a scientist from the official website of the Western Ukrainian National University, focusing on the necessary profile elements.

Figure 8 shows a fragment of the code for parsing the metadata of scientists from the official website of the organization.



```
const parseTeachersMetadata = ($, commentValue, sectionClass) => {
  const elements = findElementAfterComment($, commentValue, sectionClass);
  const medatData = [];

  elements.each((_, elem) => {
    const [firstName, lastName, middleName] = $(elem).find('h4').text().trim().replace
        ('\u00a0', " ").split(" ");
    const orcidUrl = $(elem).find('a[href*="orcid.org"]').attr('href');
    const wosUrl = $(elem).find('a[href*="wos"]').attr('href');
    const scopusUrl = $(elem).find('a[href*="scopus"]').attr('href');
    const scholarUrl = $(elem).find('a[href*="google"]').attr('href');

    const interestsElement = findElementAfterComment($(elem), 'dle_spoiler Scientific
        interests', 'div.text_spoiler');
    const scientificInterests = interestsElement.text().split(",").map(key => key.trim());

    medatData.push({
      orcid: orcidUrl ? extractOrcidId(orcidUrl) : null,
      wosId: wosUrl ? extractWosId(wosUrl) : null,
      scopusId: scopusUrl ? parseAuthorID(scopusUrl, 'authorId') : null,
      scholarId: scholarUrl ? parseAuthorID(scholarUrl, 'user') : null,
      firstName,
      lastName,
      middleName,
      scientificInterests,
    });
  });
  return medatData;
}
```

Figure 8: Code listing for parsing metadata.

Now that we have a basic set of web-scraping metadata, we can move on to the next step — fleshing out a scientist's profile using Ollama's large language models.

Large Language Models (LLM) are a powerful tool for analyzing and processing textual data. Thanks to the ability to understand the context and extract meaningful units.

The main advantage of Ollama is the ability to run and manage large language models (LLM) locally on a computer, without the need for cloud services. This ensures increased confidentiality of data, reduces costs and allows users to fully control information processing processes [10].

The models presented in the Ollama platform are specialized in the processing of scientific texts and have a wide range of applications, such as automatic text classification, extraction of keywords and phrases, identification of scientific interests and creation of a generalized profile.

Table 1 provides a comparative analysis of the major language models supported by the Ollama platform [12,13].

The above table shows the main features of the models, their advantages and disadvantages and allows you to choose the appropriate model for solving this or that problem.

Table 1
Comparison table of major language models in Ollama

Model	Number of parameters	Size (GB)	Advantages	Disadvantages
Llama 3.2	1,2B	~2	High productivity in general NLP tasks; support for many languages.	Requires significant computing resources for efficient operation.
Mistral	1,3B	~2,5	Effective in code generation and analysis of large volumes of data.	Narrow specialization; may be less effective in other tasks.
Code Llama	1,5B	~3	Optimized for tasks related to programming; support for various programming languages.	Narrow specialization; not suitable for general NLP tasks.
Gemma 2	1,1B	~2	Multimodal model; processes text and images simultaneously.	Requires more resources to process multimodal data.
Phi-3	1,4B	~2,8	Specialized in scientific and research tasks; high accuracy.	May be less effective in general tasks; needs setting.
Phi-3	1,4B	~2,8	Specialized in scientific and research tasks; high accuracy.	May be less effective in general tasks; needs setting.

Figure 9 demonstrates the process of forming a profile of a scientist, which includes the main stages: data collection, analysis of text documents using Ollama models, parsing of scientific interests, classification of information and final creation of a profile.

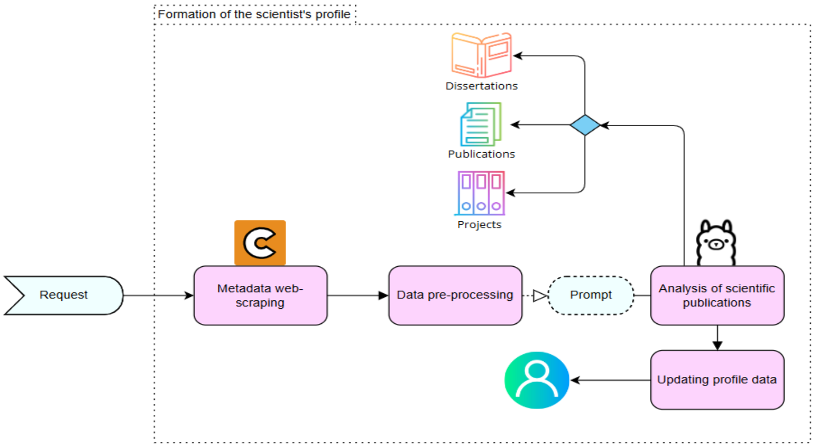


Figure 9: Scheme of formation of a scientist's profile.

As can be seen from the diagram above, the profile formation process can be divided into six main stages:

1. Request – the process begins with formulating a request to create a profile of a scientist;
2. Metadata web scraping – the first stage of information collection. Web scraping is used to obtain basic metadata, such as name, title, research interests, etc., from organizations' websites or other available sources;
3. Data pre-processing – collected data undergo a process of cleaning and structuring to prepare them for further analysis;
4. Prompt – a generated request or command for data analysis using a language model;
5. Analysis of scientific publications – the language model analyzes scientific publications, projects and dissertations to highlight key topics, scientific interests and specializations;
6. Update of profile data – based on the analysis of publications and additional data, the profile of a scientist is formed or supplemented.

8. A mathematical model of filtering irrelevant publications based on the profile of a scientist

In scientometric databases, a problem often arises when, due to the coincidence of the author's last name, first name, and patronymic, publications that do not belong to the scientist are added to the scientist's profile. This distorts the indicators of scientific activity and complicates the objective assessment of the researcher's contribution. The development of a mathematical model for filtering irrelevant publications based on the profile of a scientist allows to effectively solve this problem. To build mathematical models in conditions of limited data sampling, it is advisable to use methods based on interval analysis [14-18]. Using detailed scholarly profile data, such as author research interests, affiliations, and other unique characteristics, it is possible to accurately identify publications that actually belong to a particular scholar. This ensures an increase in the accuracy of scientometric indicators and will contribute to a more objective analysis of scientific activity.

The process of building a mathematical model for filtering irrelevant publications based on the profile of a scientist can be divided into several steps:

Step 1. Formulation of the author's scientific interests.

The author's scientific interests can be represented as a vector of keywords that provides an opportunity to describe the main areas of research. Let $I = \{k_1, k_2, \ldots, k_n\}$, where k_i is a keyword or phrase describing the author's interests.

Step 2. Vector representation of the publication.

Each post can also be represented as a vector of keywords. Let $P_j = \{p_1, p_2, \ldots, p_m\}$, where p_i is a keyword or phrase associated with post P_j.

Step 3. Calculating relevance using cosine similarity.

To measure the similarity between the scientific interests of the author I and the publication vector P_j, you can use the cosine similarity:

$$elevance\left(I, P_j\right) = \frac{\sum_{i=1}^{n} k_i \times p_i}{\sqrt{\sum_{i=1}^{n} k_i^2} \times \sqrt{\sum_{i=1}^{m} p_i^2}}, \qquad (1)$$

where × is the scalar product operation. The value of $relevance\left(I, P_j\right)$ ranges from 0 to 1, where a value close to 1 means high relevance.

Step 4. Filtering of irrelevant publications.

If the value $relevance\left(I, P_j\right)$ is less than some threshold T, then the publication is considered irrelevant and is filtered out.

P_j is relevant if $relevance\left(I, P_j\right) \geq T$

So, as we can see, the model we received allows us to automatically filter out irrelevant publications based on the scientific profile of the author.

9. Software implementation of the system for collecting and analyzing scientific and scientific-pedagogical activities of the academic team

In the modern conditions of the information society, it is important to have effective tools for collecting and analyzing scientific activity [19-21]. The developed system is aimed at automating the processes of data collection about the scientific and scientific and pedagogical achievements of the academic staff, filtering this data based on relevance to their interests, and forming reports by university, faculty, department, which allows to improve the quality of management and planning of scientific work.

Conventionally, our system can be divided into several interacting modules, namely:

1. The authorization and authentication module, which ensures secure user access to the system, access management and protection of personal data;
2. Data collection module: responsible for obtaining information from scientometric databases (for example, Scopus, Crossref, NRAT) and the profile of a scientist;
3. Data processing and analysis module: cleans, normalizes and pre-processes collected data for preparation for further analysis;
4. Filtering module: implements a mathematical filtering model using machine learning algorithms and criteria defined on the basis of the scientist's profile;
5. Reporting module: provides an opportunity to generate a general report on the scientific activity of the university, faculty or department;
6. User interface: provides user interaction with the system, providing the ability to perform CRUD operations with the main entities (for example, publications, dissertations, grants, projects, scientific activities).

The system architecture was implemented using advanced technologies that ensure reliability, performance and flexibility. The core technology stack is based on JavaScript as both client-side and server-side programming languages, which helps ensure codebase consistency and eases application development. The server part was developed using Node.js, which allows you to create high-performance and scalable server applications with high request processing capabilities even in real time.

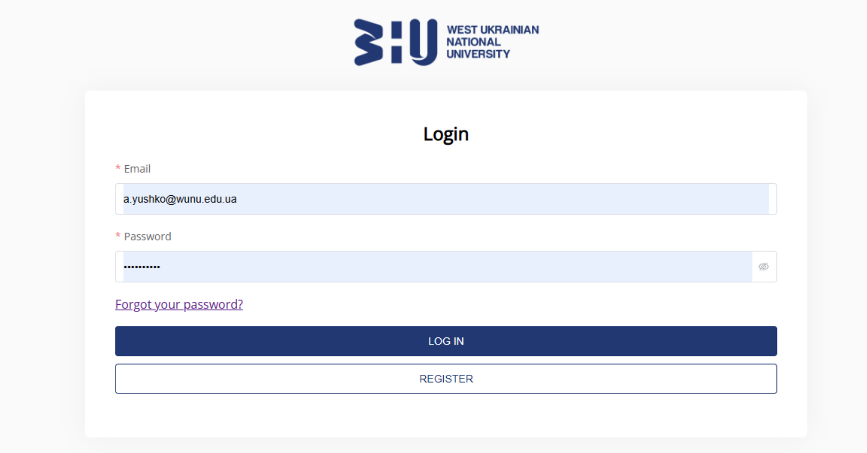
To optimize the interaction between the client and the server, GraphQL is used, which gives the client the opportunity to get only the data that is needed, which reduces the load on the network and server resources. Which, in turn, will improve system performance from the point of view of building complex queries.

The MongoDB database acts as a storage, which provides speed and flexibility when working with large volumes of unstructured data. It also provides efficient work with various data types used in describing the structure for data from various scientometric information systems, and provides easy scalability of the database in accordance with the load and needs.

In addition, the Ollama platform is integrated into the system, which provides a mechanism for working with various models of machine learning and artificial intelligence. Thanks to such capabilities, the system can more accurately determine the relevance of publications based on the profile of a scientist, calculating complex relationships between data.

The system interface is developed based on the principles of building intuitiveness and ease of use, which provides convenient user access to the main functionality without the need for additional training of personnel.

Figure 10 shows the initial screen of the page with the authorization and authentication forms.



* Email

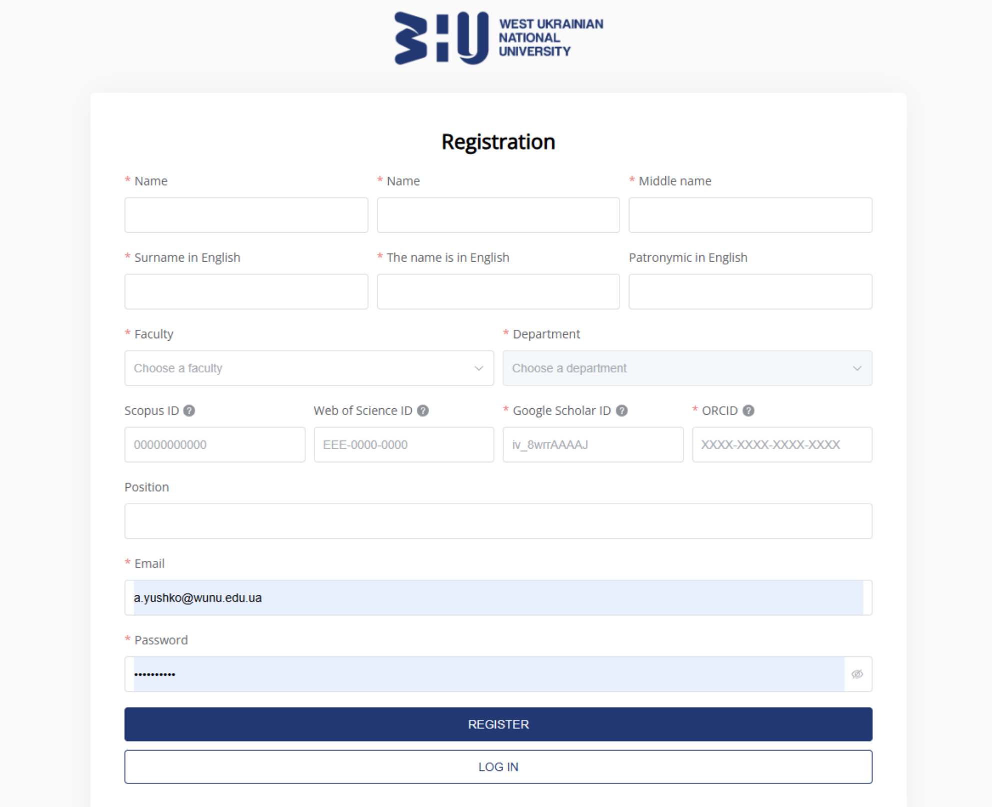
a.yushko@wunu.edu.ua

* Password

••••••••••

Figure 10: Authorization form.

As you can see from the screenshot above, the authorization form is quite simple, as it requires the user to enter only the e-mail address and the password that was created during registration in the system (Fig. 11).



* Name

* Name

* Middle name

* Surname in English

* The name is in English

Patronymic in English

* Faculty: Choose a faculty

* Department: Choose a department

Scopus ID: 00000000000

Web of Science ID: EEE-0000-0000

* Google Scholar ID: iv_8wrrAAAAJ

* ORCID: XXXX-XXXX-XXXX-XXXX

Position

* Email

a.yushko@wunu.edu.ua

* Password

••••••••••

Figure 11: Registration form in the system.

The registration form requires the user to fill out a database of information about himself, such as surname, first name, patronymic, position, faculty, department, etc. Also, when registering, the employee must specify his identifiers in other scientometric databases to ensure the process of automated information collection. Another of the main fields of this form is the scientist's last name and first name in Latin, as these data will be needed to search for publications in the Crossref database.

After successful authorization in the system, the user gets to the "Overview" page (Fig. 12) where he can see quantitative indicators of publication activity.

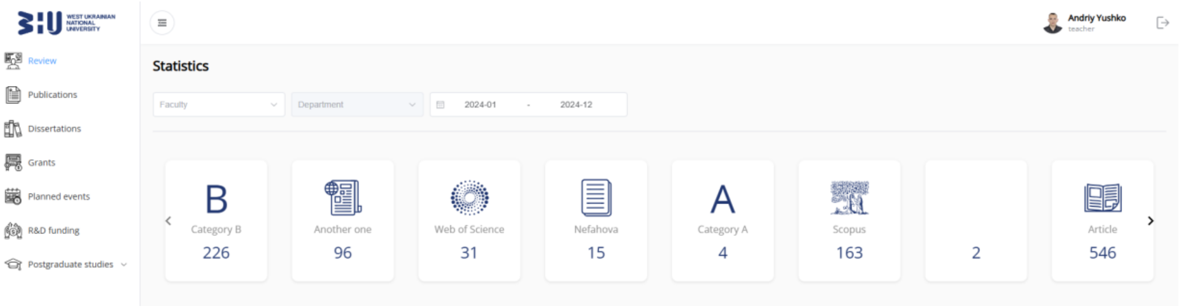
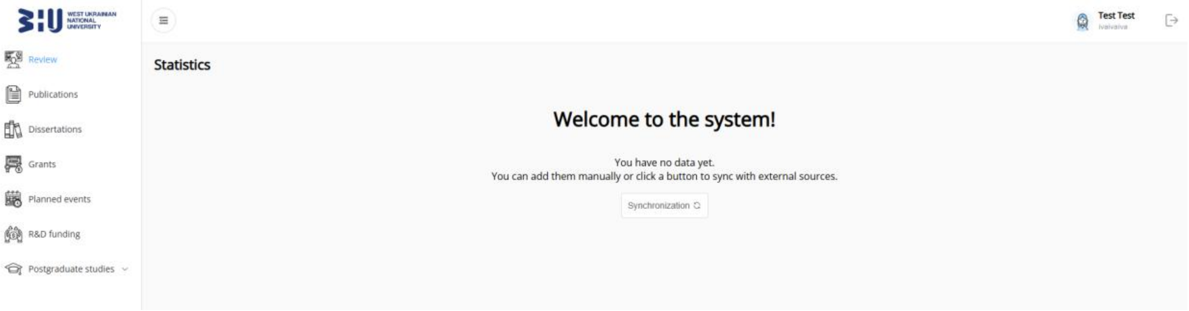


Figure 12: Viewing the number of publications in each of the author's categories.

As can be seen from the screenshot above, the user has the opportunity to filter all indicators by faculty, department and publication period.

There is also an opportunity to create a world based on a specific division by clicking the "Download report" button. This opportunity is available only to employees with appropriate access rights (for example, the head of the department, the dean of the faculty, the vice-rector for scientific work).

If the user has entered the system and there is no added data yet, he will see a welcome window and a button that will allow synchronization of all publication activity from other scientometric databases (Fig. 13).

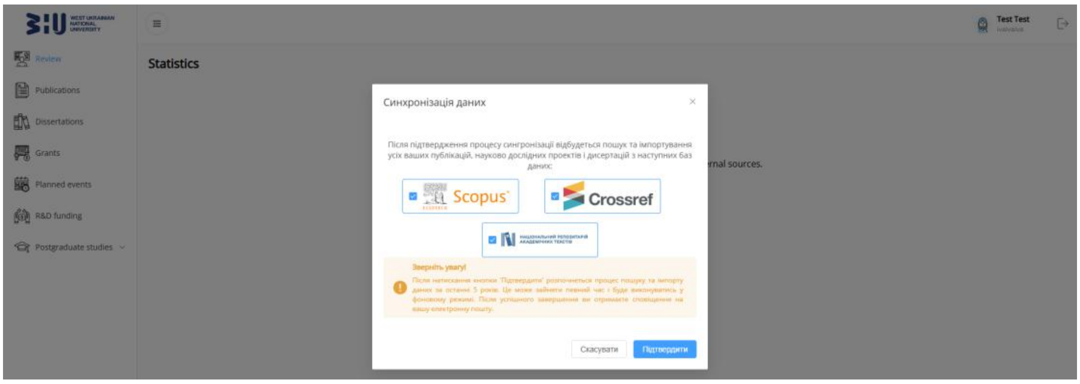


Welcome to the system!

You have no data yet.
You can add them manually or click a button to sync with external sources.

Figure 3: The user's welcome page in the system.

After pressing the "Synchronization" button, a window will open with a description of the databases in which information will be searched (Fig. 14).



Після підтвердження процесу сингронізації відбудеться пошук та імпортування усіх ваших публікацій, науково дослідних проектів і дисертацій з наступних баз даних:

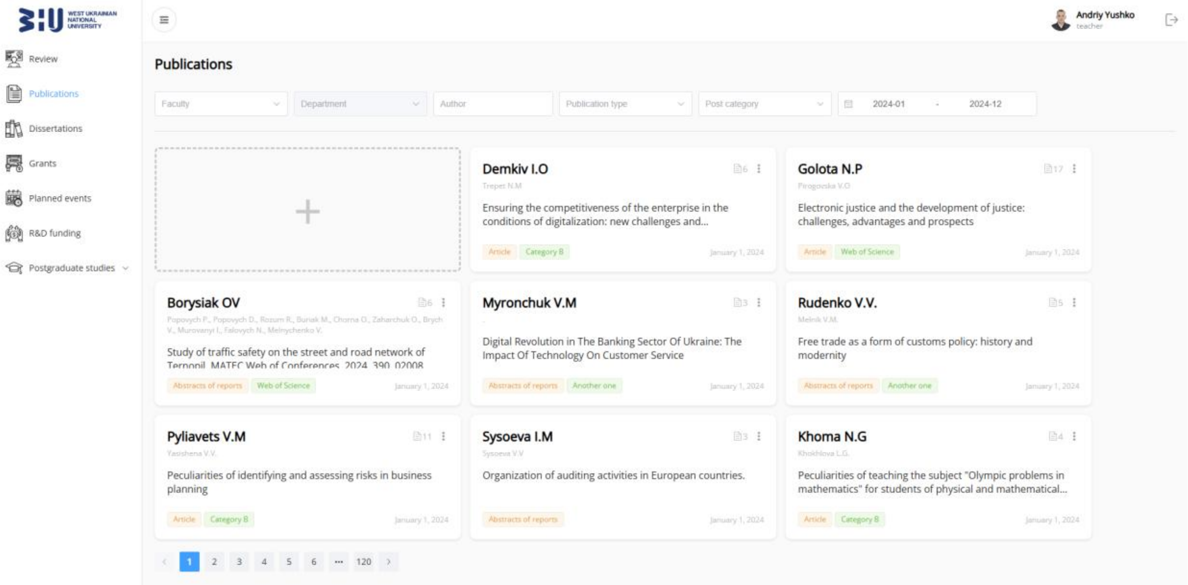
Зверніть увагу!
Після натискання кнопки "Підтвердити" розпочнеться процес пошуку та імпорту даних за останні 5 років. Це може зайняти певний час і буде виконуватись у фоновому режимі. Після успішного завершення ви отримаєте сповіщення на вашу електронну пошту.

Figure 14: Synchronization confirmation window.

Also, as can be seen from the image above, the user is provided with brief information about the period for which his data will be collected (by default 5 years) and how exactly this process will take place.

After successful synchronization, the user of the system receives a message in his e-mail box, which informs that his publishing activity has been successfully collected.

By going to the "Publications" section, the user will be able to view all the publications that the system managed to find (Fig. 15).



Publications

Demkiv I.O 6
Trepet N.M
Ensuring the competitiveness of the enterprise in the conditions of digitalization: new challenges and...
Article | Category B | January 1, 2024

Golota N.P 17
Pirogovska V.O
Electronic justice and the development of justice: challenges, advantages and prospects
Article | Web of Science | January 1, 2024

Borysiak OV 6
Popovych P., Popovych D., Rozum R., Buriak M., Chorna O., Zaharchuk O., Brych V., Murovanyi I., Falovych N., Melnychenko V.
Study of traffic safety on the street and road network of Ternopil. MATEC Web of Conferences, 2024, 390, 02008
Abstracts of reports | Web of Science | January 1, 2024

Myronchuk V.M 3
-
Digital Revolution in The Banking Sector Of Ukraine: The Impact Of Technology On Customer Service
Abstracts of reports | Another one | January 1, 2024

Rudenko V.V. 5
Melnik V.M.
Free trade as a form of customs policy: history and modernity
Abstracts of reports | Another one | January 1, 2024

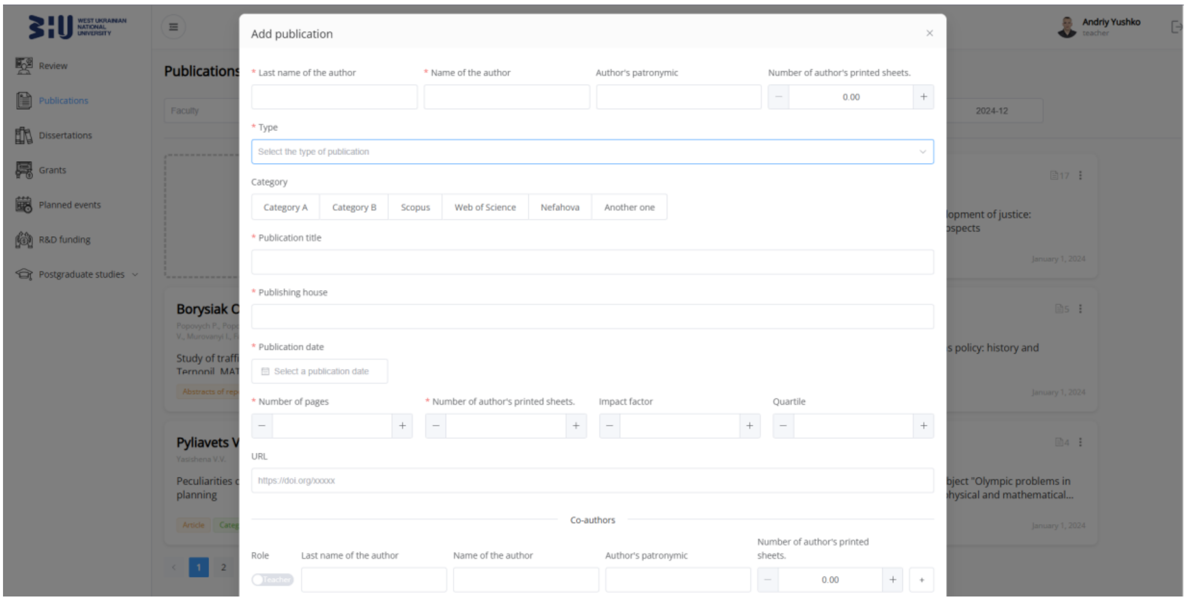
Pyliavets V.M 11
Yasishena V.V.
Peculiarities of identifying and assessing risks in business planning
Article | Category B | January 1, 2024

Sysoeva I.M 3
Sysoeva V.V
Organization of auditing activities in European countries.
Abstracts of reports | January 1, 2024

Khoma N.G 4
Khokhlova L.G.
Peculiarities of teaching the subject "Olympic problems in mathematics" for students of physical and mathematical...
Article | Category B | January 1, 2024

Figure 15: Publications display page.

If the system could not find any publication of the author, he can add it manually by pressing the corresponding button. After that, the user will open a form where he will need to fill in all the necessary fields (Fig. 16).



Add publication

* Last name of the author
* Name of the author
Author's patronymic
Number of author's printed sheets. 0.00

* Type: Select the type of publication

Category

* Publication title

* Publishing house

* Publication date: Select a publication date

* Number of pages
* Number of author's printed sheets.
Impact factor
Quartile

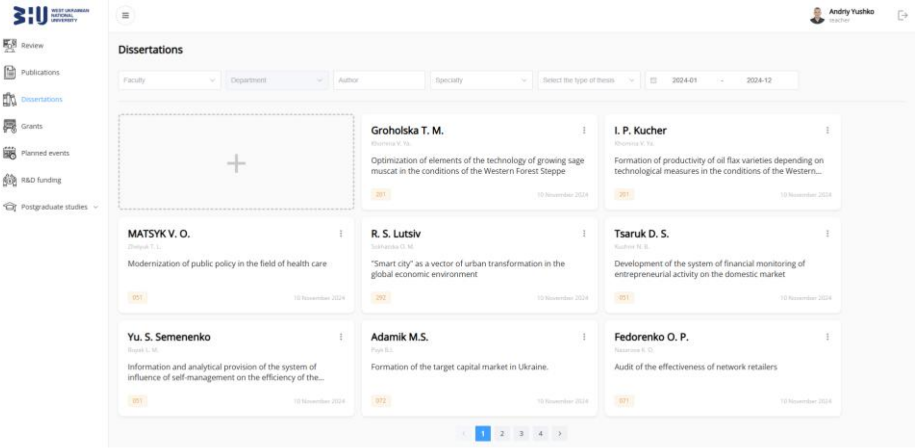
URL: https://doi.org/xxxxx

Co-authors

Role | Last name of the author | Name of the author | Author's patronymic | Number of author's printed sheets. 0.00

Figure 16: Form for adding a new publication.

The page for viewing dissertations protected by the user, where there is an addition form, has a similar appearance (Fig. 17).



Dissertations

Groholska T. M.
Optimization of elements of the technology of growing sage muscat in the conditions of the Western Forest Steppe
201 — 10 November 2024

I. P. Kucher
Formation of productivity of oil flax varieties depending on technological measures in the conditions of the Western...
201 — 10 November 2024

MATSYK V. O.
Modernization of public policy in the field of health care
051 — 10 November 2024

R. S. Lutsiv
"Smart city" as a vector of urban transformation in the global economic environment
292 — 10 November 2024

Tsaruk D. S.
Development of the system of financial monitoring of entrepreneurial activity on the domestic market
051 — 10 November 2024

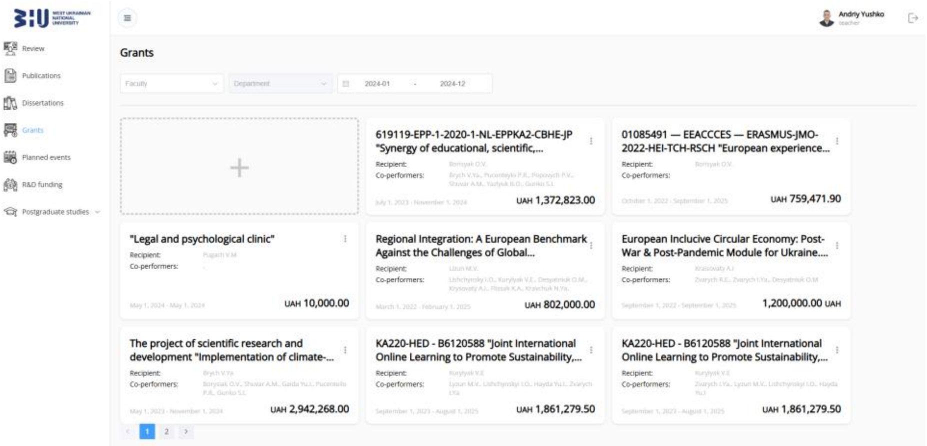
Yu. S. Semenenko
Information and analytical provision of the system of influence of self-management on the efficiency of the...
051 — 10 November 2024

Adamik M.S.
Formation of the target capital market in Ukraine.
072 — 10 November 2024

Fedorenko O. P.
Audit of the effectiveness of network retailers
071 — 10 November 2024

Figure 17: Dissertation viewing page.

Also, our system provides a section where you can view the list of granted grants (Fig. 18).



Grants

619119-EPP-1-2020-1-NL-EPPKA2-CBHE-JP "Synergy of educational, scientific,...
Recipient:
Co-performers:
July 1, 2023 - November 1, 2024 — UAH 1,372,823.00

01085491 — EEACCCES — ERASMUS-JMO-2022-HEI-TCH-RSCH "European experience...
Recipient:
Co-performers:
October 1, 2022 - September 1, 2025 — UAH 759,471.90

"Legal and psychological clinic"
Recipient:
Co-performers:
May 1, 2024 - May 1, 2024 — UAH 10,000.00

Regional Integration: A European Benchmark Against the Challenges of Global...
Recipient:
Co-performers:
March 1, 2022 - February 1, 2025 — UAH 802,000.00

European Inclucive Circular Economy: Post-War & Post-Pandemic Module for Ukraine....
Recipient:
Co-performers:
September 1, 2022 - September 1, 2025 — 1,200,000.00 UAH

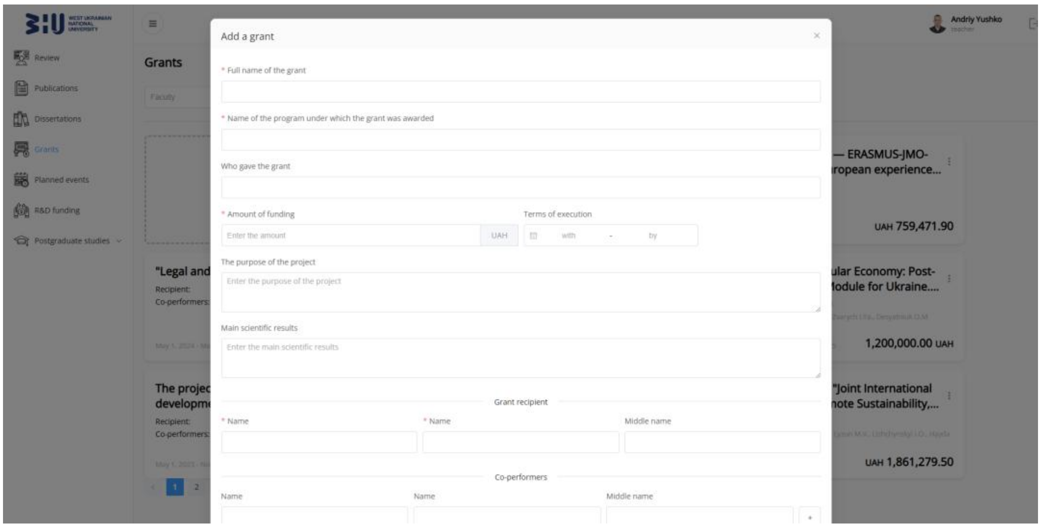
The project of scientific research and development "Implementation of climate-...
Recipient:
Co-performers:
May 1, 2023 - November 1, 2024 — UAH 2,942,268.00

KA220-HED - B6120588 "Joint International Online Learning to Promote Sustainability,...
Recipient:
Co-performers:
September 1, 2023 - August 1, 2025 — UAH 1,861,279.50

KA220-HED - B6120588 "Joint International Online Learning to Promote Sustainability,...
Recipient:
Co-performers:
September 1, 2023 - August 1, 2025 — UAH 1,861,279.50

Figure 18: Grants View Page.

Figure 19 shows the form for adding a new grant.



Add a grant

* Full name of the grant

* Name of the program under which the grant was awarded

Who gave the grant

* Amount of funding: Enter the amount UAH

Terms of execution: with - by

The purpose of the project: Enter the purpose of the project

Main scientific results: Enter the main scientific results

Grant recipient

* Name | * Name | Middle name

Co-performers

Name | Name | Middle name

Figure 19: Grant addition form.

In the planned events section, users can plan new events and view existing ones (Fig. 20).

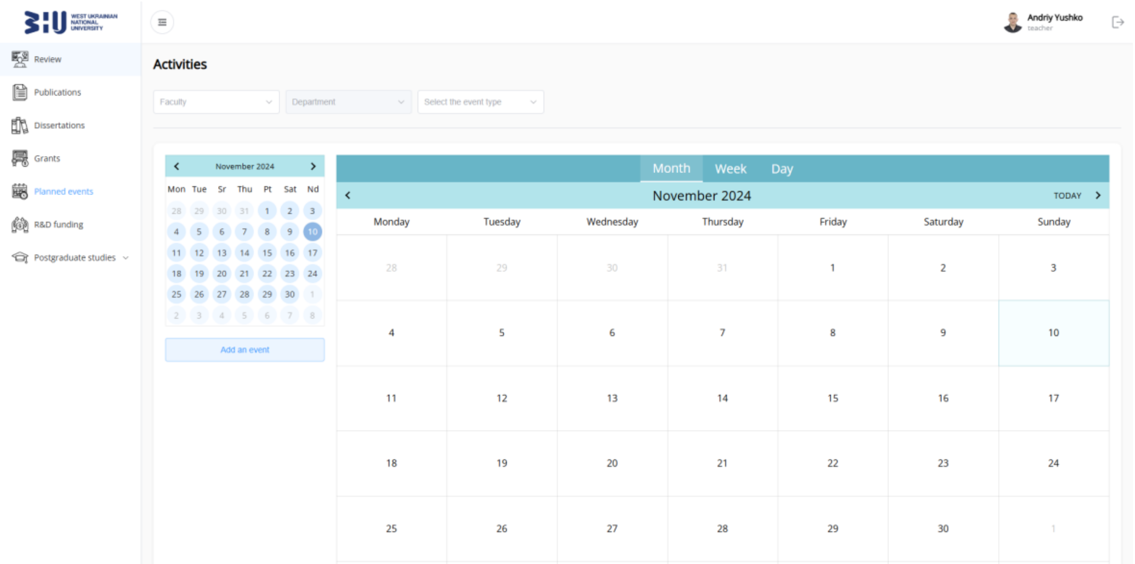
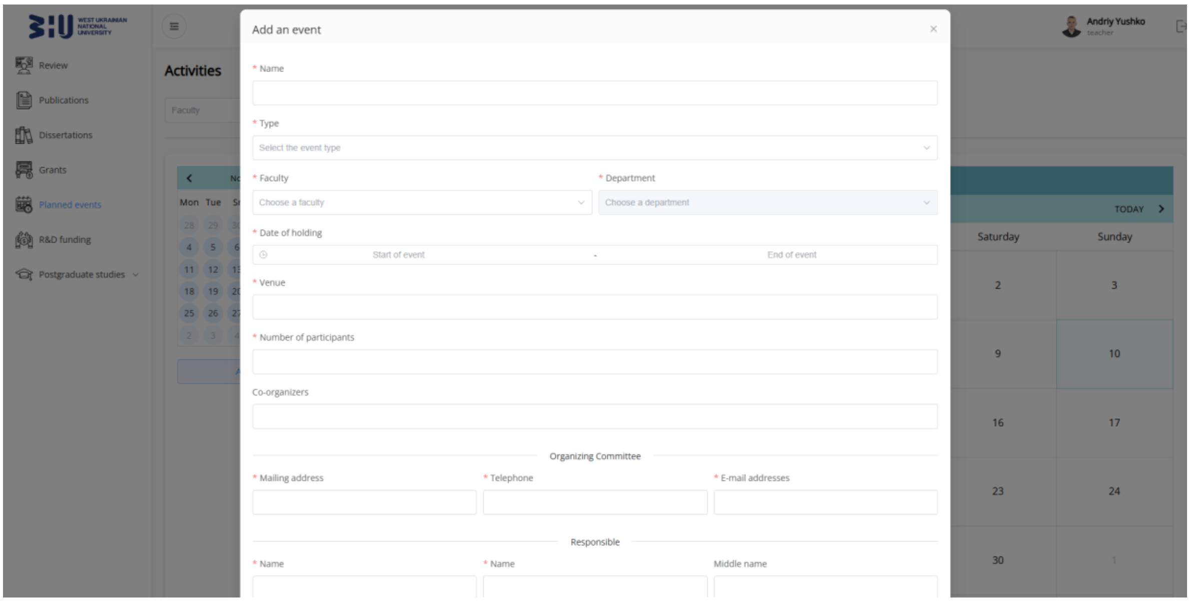


Figure 20: View and add events page.

Figure 21 shows the Add Activities form, which includes a variety of required fields marked with an asterisk to collect information. It includes fields such as "Name", "Type", "Faculty", "Department", "Date", "Place", "Number of participants" and "Co-organizers". There are also sections for the contact details of the organizing committee, where you need to enter a postal address, phone number and email address.


Add an event

* Name

* Type

Select the event type

* Faculty

Choose a faculty

* Department

Choose a department

* Date of holding

Start of event - End of event

* Venue

* Number of participants

Co-organizers

Organizing Committee

* Mailing address

* Telephone

* E-mail addresses

Responsible

* Name

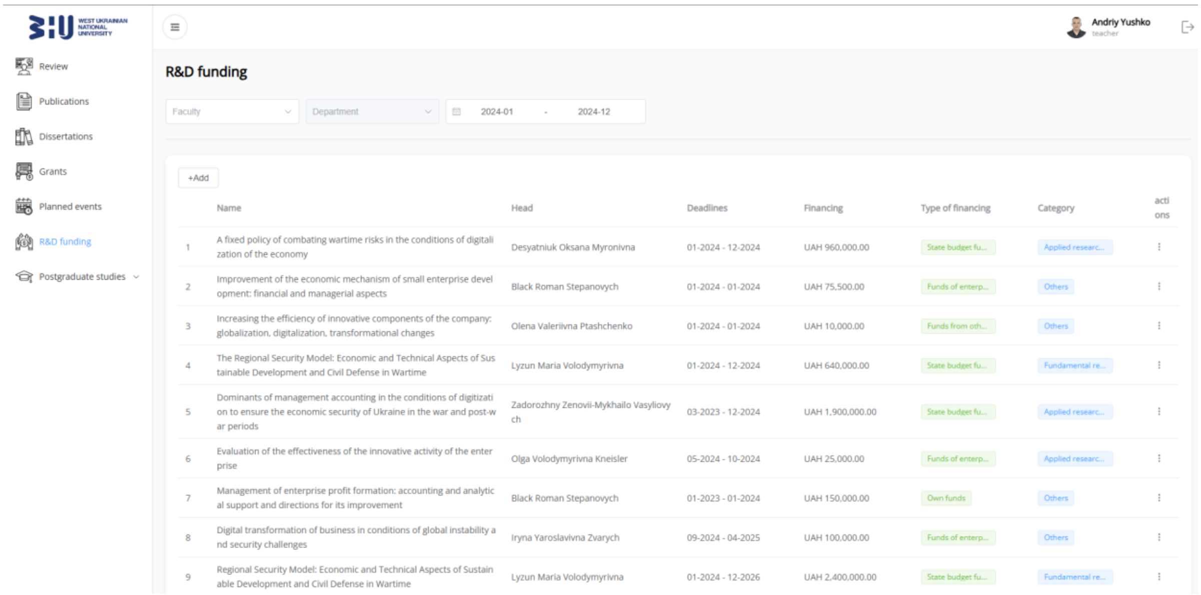
* Name

Middle name

Figure 21: Event planning form.

The R&D funding page displays a list of research and development (R&D) funded projects, including the name, manager, terms, amount of funding, type of funding, and category of each project (Fig. 22).

It is also possible to quickly search by faculty, department and deadline. As already mentioned earlier, the system provides for the possibility of automatic creation of a scientist's profile, which can be used in the future for the tasks of filtering publications.



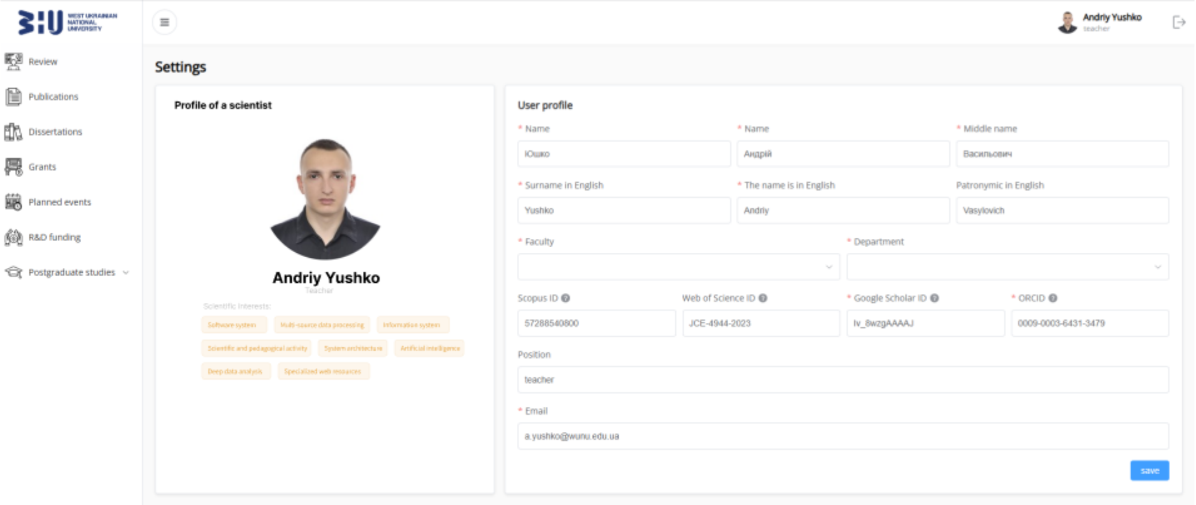
R&D funding

2024-01 - 2024-12

	Name	Head	Deadlines	Financing	Type of financing	Category	actions
1	A fixed policy of combating wartime risks in the conditions of digitalization of the economy	Desyatniuk Oksana Myronivna	01-2024 - 12-2024	UAH 960,000.00	State budget fu...	Applied researc...	⋮
2	Improvement of the economic mechanism of small enterprise development: financial and managerial aspects	Black Roman Stepanovych	01-2024 - 01-2024	UAH 75,500.00	Funds of enterp...	Others	⋮
3	Increasing the efficiency of innovative components of the company: globalization, digitalization, transformational changes	Olena Valeriivna Ptashchenko	01-2024 - 01-2024	UAH 10,000.00	Funds from oth...	Others	⋮
4	The Regional Security Model: Economic and Technical Aspects of Sustainable Development and Civil Defense in Wartime	Lyzun Maria Volodymyrivna	01-2024 - 12-2024	UAH 640,000.00	State budget fu...	Fundamental re...	⋮
5	Dominants of management accounting in the conditions of digitization to ensure the economic security of Ukraine in the war and post-war periods	Zadorozhny Zenovii-Mykhailo Vasyliovych	03-2023 - 12-2024	UAH 1,900,000.00	State budget fu...	Applied researc...	⋮
6	Evaluation of the effectiveness of the innovative activity of the enterprise	Olga Volodymyrivna Kneisler	05-2024 - 10-2024	UAH 25,000.00	Funds of enterp...	Applied researc...	⋮
7	Management of enterprise profit formation: accounting and analytical support and directions for its improvement	Black Roman Stepanovych	01-2023 - 01-2024	UAH 150,000.00	Own funds	Others	⋮
8	Digital transformation of business in conditions of global instability and security challenges	Iryna Yaroslavivna Zvarych	09-2024 - 04-2025	UAH 100,000.00	Funds of enterp...	Others	⋮
9	Regional Security Model: Economic and Technical Aspects of Sustainable Development and Civil Defense in Wartime	Lyzun Maria Volodymyrivna	01-2024 - 12-2026	UAH 2,400,000.00	State budget fu...	Fundamental re...	⋮

Figure 22: R&D funding review page.

Figure 23 shows the scientist's profile view page.



Settings

Profile of a scientist

Andriy Yushko
Teacher

Scientific interests:
Software system · Multi-source data processing · Information system · Scientific and pedagogical activity · System architecture · Artificial intelligence · Deep data analysis · Specialized web resources

User profile

* Name: Юшко
* Name: Андрій
* Middle name: Васильович
* Surname in English: Yushko
* The name is in English: Andriy
Patronymic in English: Vasylovich
* Faculty:
* Department:
Scopus ID: 57288540800
Web of Science ID: JCE-4944-2023
* Google Scholar ID: lv_8wzgAAAAJ
* ORCID: 0009-0003-6431-3479
Position: teacher
* Email: a.yushko@wunu.edu.ua

Figure 23: Scientist profile page.

As can be seen from the figure above, the user has the opportunity not only to view his profile, but also to edit the necessary information.

10. Conclusion

This work emphasizes the need for automation of collection, processing and analysis of publication activity in the modern scientific environment. The increase in the volume of scientific information complicates the manual control and analysis of data, especially in large academic groups. The developed system described in this paper not only provides automated collection of information from scientometric databases such as Scopus and Web of Science, but also forms a profile of a scientist, which includes information about his scientific interests, publications, grants and participation in scientific events. This allows you to optimize the processes of managing scientific activities, making them more efficient and objective.

An important part of the work is the use of modern algorithms for automatic text analysis, such as TF-IDF, Named Entity Recognition (NER) and text vectorization, which contribute to the selection of keywords and the identification of scientific interests of researchers. The application of deep language models, such as BERT, GPT, as well as the capabilities of the Ollama platform for localized processing of big data, allows you to achieve high accuracy in text analysis, taking into account the semantic context and the relationship between terms.

In addition, a mathematical model of filtering irrelevant publications is built in the work, which is based on the profile of a scientist, which solves the problem of filtering the author's original works, thereby significantly increasing the accuracy of scientometric indicators.

Also, the use of vector representation of scientific interests and publications with the calculation of cosine similarity is proposed for the first time. This approach contributes to the objective assessment of scientific contributions, reducing the risk of inaccuracies due to random coincidence of surnames or errors in databases.

Another important component of this work is the integration of the Ollama platform into its own system, which allows the use of language models for accurate identification of scientific interests, as well as for automatic categorization and clustering of scientific materials. This greatly facilitates the formation of reports for scientific institutions, which allows you to quickly obtain generalized data on the activities of the university, faculties and departments.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT and Grammarly to check grammar and spelling, paraphrase, and reword the text. These tools help identify and correct grammatical errors, typos, and other writing mistakes, improving the clarity and professionalism of the text. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication’s content.

References

[1] D.-M. Petroșanu, A. Pîrjan, A. Tăbușcă, Tracing the influence of large language models across the most impactful scientific works, Electronics 12.24 (2023) 4957. https://doi.org/10.3390/electronics12244957.

[2] N. Lutsiv, T. Maksymyuk, M. Beshley, O. Lavriv, V. Andrushchak, A. Sachenko, L. Vokorokos, J. Gazda, Deep semisupervised learning-based network anomaly detection in heterogeneous information systems, Comput., Mater. & Contin. 70.1 (2022) 413–431. https://doi.org/10.32604/cmc.2022.018773.

[3] A. Sachenko, V. Kochan, V. Turchenko, V. Tymchyshyn and N. Vasylkiv, "Intelligent nodes for distributed sensor network," IMTC/99. Proceedings of the 16th IEEE Instrumentation and Measurement Technology Conference (Cat. No.99CH36309), Venice, Italy, 1999, pp. 1479-1484 vol. 3. https://doi.org/ 10.1109/IMTC.1999.776072

[4] V. Lytvyn, V. Vysotska, P. Pukach, Z. Nytrebych, I. Demkiv, A. Senyk, O. Malanchuk, S. Sachenko, R. Kovalchuk, N. Huzyk, Analysis of the developed quantitative method for automatic attribution of scientific and technical text content written in Ukrainian, Eastern-European J. Enterp. Technol. 6.2 (96) (2018) 19–31. https://doi.org/10.15587/1729-4061.2018.149596.

[5] A. Zaki Ahmed, M. Rodríguez Díaz, A methodology for machine-learning content analysis to define the key labels in the titles of online customer reviews with the rating evaluation, Sustainability 14.15 (2022) 9183. https://doi.org/10.3390/su14159183.

[6] Y. Wang, Research on the TF–IDF algorithm combined with semantics for automatic extraction of keywords from network news texts, J. Intell. Syst. 33.1 (2024). https://doi.org/10.1515/jisys-2023-0300.

[7] A. Singh, A. Garg, Named entity recognition (NER) and relation extraction in scientific publications, Int. J. Recent Technol. Eng. (IJRTE) 12.2 (2023) 110–113. https://doi.org/10.35940/ijrte.b7846.0712223.

[8] L. M. Pham, H. C. The, LNLF-BERT: transformer for long document classification with multiple attention levels, IEEE Access (2024) 1. https://doi.org/10.1109/access.2024.3492102.

[9] H. D. Abubakar, M. Umar, Sentiment classification: review of text vectorization methods: bag of words, tf-idf, word2vec and doc2vec, SLU J. Sci. Technol. 4.1&2 (2022) 27–33. https://doi.org/10.56471/slujst.v4i.266.

[10] H.-S. Lee, H.-S. Shim, Implementation of generative AI using metaverse-based LLM, Korea Ind. Technol. Converg. Soc. 29.2 (2024) 123–132. https://doi.org/10.29279/jitr.2024.29.2.123.

[11] M. Brown, A. Gruen, G. Maldoff, S. Messing, Z. Sanderson, M. Zimmer, Web scraping for research: legal, ethical, institutional, and scientific considerations, 2024. https://doi.org/10.48550/arXiv.2410.23432.

[12] D. P. Pau, F. M. Aymone, Forward learning of large language models by consumer devices, Electronics 13.2 (2024) 402. https://doi.org/10.3390/electronics13020402.

[13] C.-N. Hang, P.-D. Yu, R. Morabito, C.-W. Tan, Large language models meet next-generation networking technologies: A review, Future Internet 16.10 (2024) 365. https://doi.org/10.3390/fi16100365.

[14] M. Dyvak, P. Stakhiv, A. Pukas, Algorithms of parallel calculations in task of tolerance ellipsoidal estimation of interval model parameters, Bull. Pol. Acad. Sci. 60.1 (2012). https://doi.org/10.2478/v10175-012-0022-9.

[15] M. Dyvak, I. Voytyuk, N. Porplytsya, A. Pukas, Modeling the process of air pollution by harmful emissions from vehicles, in: 2018 14th international conference on advanced trends in radioelecrtronics, telecommunications and computer engineering (TCSET), 2018, pp. 1272–1276. https://doi.org/10.1109/TCSET.2018.8336426.

[16] N. Ocheretnyuk, I. Voytyuk, M. Dyvak, Y. Martsenyuk, Features of structure identification the macromodels for nonstationary fields of air pollutions from vehicles, in: Proceedings of international conference on modern problem of radio engineering, telecommunications and computer science, 2012, pp. 444–444.

[17] M. Dyvak, Parameters identification method of interval discrete dynamic models of air pollution based on artificial bee colony algorithm, in: 2020 10th international conference on advanced computer information technologies (ACIT), 2020, pp. 130–135. https://doi.org/10.1109/ACIT49673.2020.9208972.

[18] M. Dyvak, A. Pukas, I. Oliynyk, A. Melnyk, Selection the "saturated" block from interval system of linear algebraic equations for recurrent laryngeal nerve identification, in: 2018 IEEE second international conference on data stream mining & processing (DSMP), 2018, pp. 444–448. https://doi.org/10.1109/DSMP.2018.8478528.

[19] M. Pirnau, M. A. Botezatu, I. Priescu, A. Hosszu, A. Tabusca, C. Coculescu, I. Oncioiu, Content analysis using specific natural language processing methods for big data, Electronics 13.3 (2024) 584. https://doi.org/10.3390/electronics13030584.

[20] M. Gkevrou, D. Stamovlasis, Illustration of a software-aided content analysis methodology applied to educational research, Educ. Sci. 12.5 (2022) 328. https://doi.org/10.3390/educsci12050328.

[21] N. Le, D. Tran, R. Sturgill, Content analysis of three-dimensional model technologies and applications for construction: current trends and future directions, Sensors 24.12 (2024) 3838. https://doi.org/10.3390/s24123838.