НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ БІОРЕСУРСІВ І ПРИРОДОКОРИСТУВАННЯ УКРАЇНИ

Факультет/(ННІ) **Інформаційних технологій**

ПОГОДЖЕНО	**ДОПУСКАЄТЬСЯ ДО ЗАХИСТУ**
Декан факультету (Директор ННІ)	**Завідувач кафедри**
Інформаційних технологій	Комп'ютерних наук
(назва факультету (ННІ))	(назва кафедри)
_____ Ігор Болбот	_____ Белла Голуб
(підпис) (ім'я ПРІЗВИЩЕ)	(підпис) (ім'я ПРІЗВИЩЕ)
"___" _____ 20__ р.	"_28_" _листопада_ 2025 р.

МАГІСТЕРСЬКА КВАЛІФІКАЦІЙНА РОБОТА

на тему Програмне забезпечення системи аналізу даних авторів наукових статей

Спеціальність **121 Інженерія програмного забезпечення**
(код і найменування)

Освітня програма **Програмне забезпечення інформаційних систем**
(назва)

Орієнтація освітньої програми **Освітньо-професійна**
(освітньо-професійна або освітньо-наукова)

Гарант освітньої програми

к. ф.-м. н., доцент _____ Віктор Кириченко
(науковий ступінь та вчене звання) (підпис) (ім'я ПРІЗВИЩЕ)

Керівник магістерської кваліфікаційної роботи

ст. викл. _____ Світлана Василюк-Зайцева
(науковий ступінь та вчене звання) (підпис) (ім'я ПРІЗВИЩЕ)

Консультант

д. ек. н., професор _____ Роман Руденський
(науковий ступінь та вчене звання) (підпис) (ім'я ПРІЗВИЩЕ)

Виконав _____ Марко-Антоніо Ретамосо Рохас
(підпис) (ім'я ПРІЗВИЩЕ здобувача)

КИЇВ – 2025

НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ БІОРЕСУРСІВ І ПРИРОДОКОРИСТУВАННЯ УКРАЇНИ

Факультет (ННІ) _**Інформаційних технологій**_

ЗАТВЕРДЖУЮ
Завідувач кафедри
Комп'ютерних наук
к. ф.-м. н, доцент _____ Белла Голуб
(науковий ступінь, вчене звання) (підпис) (ім'я ПРІЗВИЩЕ)
" 01 " листопада 2024 року

З А В Д А Н Н Я

ДО ВИКОНАННЯ МАГІСТЕРСЬКОЇ КВАЛІФІКАЦІЙНОЇ РОБОТИ ЗДОБУВАЧУ

Ретамосо Рохас Марко-Антоніо Максимовичу
(прізвище, ім'я, по батькові)

Спеціальність
121 «Інженерія програмного забезпечення»
(код і найменування)

Освітня програма
Програмне забезпечення інформаційних систем
(назва)

Орієнтація освітньої програми освітньо-професійна
(освітньо-професійна або освітньо-наукова)

Тема магістерської кваліфікаційної роботи Програмне забезпечення системи аналізу даних авторів наукових статей

затверджена наказом від " 01 " листопада 2024 р. № 1963 «С»

Термін подання завершеної роботи на кафедру 28.11.2025
(рік, місяць, число)

Вихідні дані до магістерської кваліфікаційної роботи публічні бібліометричні бази даних та відкриті API

Перелік питань, що підлягають дослідженню:

1. Аналіз сучасних наукометричних показників та їх застосування для оцінки наукової діяльності авторів.
2. Інтеграція сучасних методів аналізу даних у програмне забезпечення для оцінки наукових публікацій.
3. Розробка алгоритмів для автоматичного виявлення тенденцій у науковій діяльності, зокрема аналіз тем, ключових слів та динаміки публікацій.
4. Дослідження методів збору, структуризації та обробки даних з публічних бібліометричних баз.

Перелік графічного матеріалу (за потреби) _____

Дата видачі завдання " 01 " листопада 2024 р.

Керівник магістерської кваліфікаційної роботи _____ Світлана Василюк-Зайцева
(підпис) (ім'я ПРІЗВИЩЕ)

Завдання прийняв до виконання _____ Марко-Антоніо Ретамосо Рохас
(підпис) **(ім'я ПРІЗВИЩЕ)**

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ

API – Application Programming Interface (Інтерфейс програмування застосунків)

ARIMA – Autoregressive Integrated Moving Average (Модель авторегресії інтегрованого ковзного середнього)

DOI – Digital Object Identifier (Ідентифікатор цифрового об’єкта; унікальний ідентифікатор наукових публікацій)

ETL – Extract, Transform, Load (Процес вилучення, перетворення та завантаження даних)

LDA – Latent Dirichlet Allocation (Латентне розміщення Діріхле; алгоритм тематичного моделювання)

MAE – Mean Absolute Error (Середня абсолютна похибка)

MAPE – Mean Absolute Percentage Error (Середня абсолютна відсоткова похибка)

NMF – Non-negative Matrix Factorization (Невід'ємна матрична факторизація; метод аналізу даних)

ORCID – Open Researcher and Contributor ID (Відкритий ідентифікатор дослідника та дописувача; унікальний ідентифікатор науковця)

RMSE – Середньоквадратична помилка (Root Mean Squared Error)

TfidfVectorizer – Term Frequency-Inverse Document Frequency (Векторизатор частоти терміна – оберненої частоти документа)

ВСТУП

Сучасна наука характеризується експоненційним зростанням обсягів публікацій, що ускладнює їхній ефективний аналіз та навігацію у величезному інформаційному просторі. В умовах глобалізації дослідницької діяльності та посилення конкуренції, вимірювання наукової продуктивності, ідентифікація провідних дослідників, виявлення актуальних наукових трендів та ефективних колаборацій стають критично важливими завданнями для наукових установ, адміністраторів, грантових фондів та самих науковців. Традиційні методи оцінки, що ґрунтуються переважно на кількісних показниках, часто не надають глибинного розуміння якісного внеску та взаємодії у науковій спільноті.

Актуальність теми дослідження. Зростаюча доступність бібліометричних даних через відкриті програмні інтерфейси (API), відкриває нові можливості для застосування методів інтелектуального аналізу даних та машинного навчання. Це дозволяє не лише агрегувати інформацію, а й виявляти приховані закономірності, прогнозувати тенденції та візуалізувати складні взаємозв'язки, що є недосяжним для ручного аналізу. Таким чином, розробка автоматизованих систем для інтелектуального аналізу бібліометричних даних авторів наукових статей стає вкрай актуальною для підвищення ефективності наукового менеджменту та підтримки дослідницької діяльності.

Зв'язок роботи з науковими програмами, планами, темами. Робота виконувалась відповідно до плану науково-дослідних робіт кафедри комп'ютерних наук НУБіП України за темою "Програмне забезпечення системи аналізу авторів наукових статей".

Мета дослідження. Розробка та реалізація програмної системи для інтелектуального аналізу бібліометричних даних авторів наукових статей з метою виявлення наукових тенденцій, колаборацій та оцінки прогнозної активності.

Завдання дослідження. Для досягнення поставленої мети було сформульовано наступні завдання:

1. Дослідити методи збору та структуризації бібліометричних даних з відкритих наукометричних джерел.
2. Вибрати та адаптувати алгоритми інтелектуального аналізу даних для виявлення наукових тематик (тематичне моделювання), аналізу мережевих колаборацій (мережевий аналіз) та прогнозування динаміки публікаційної активності (моделі часових рядів).

3. Розробити архітектуру програмної системи, що інтегрує модулі збору, обробки, аналізу та візуалізації бібліометричних даних.
4. Здійснити програмну реалізацію основних модулів системи з використанням сучасних технологій.
5. Провести експериментальні дослідження ефективності розроблених алгоритмів та оцінити точність прогнозних моделей на реальному наборі даних.
6. Розробити зручний та інформативний інтерфейс користувача з можливостями інтерактивної візуалізації результатів аналізу.
7. Провести оцінку практичного значення для наукової спільноти.

Об'єкт дослідження. Процеси збору, обробки, аналізу та представлення бібліометричних даних.

Предмет дослідження. Методи та моделі інтелектуального аналізу бібліометричних даних авторів наукових статей.

Методи дослідження. У роботі використано методи системного аналізу, математичного моделювання, інтелектуального аналізу даних (Data Mining), машинного навчання (Machine Learning), статистичного аналізу, мережевого аналізу, а також методи програмної інженерії для розробки архітектури та реалізації системи.

Наукова новизна одержаних результатів.

- Удосконалено підхід до агрегації та нормалізації бібліометричних даних з множинних відкритих джерел для побудови комплексних профілів авторів.
- Запропоновано інтегровану методику виявлення наукових тенденцій, що поєднує тематичне моделювання для ідентифікації тематик, мережевий аналіз (метрики центральності, виявлення спільнот) для дослідження колаборацій та моделі часових рядів для прогнозування активності.
- Розроблено програмну архітектуру та реалізовано систему, яка ефективно об'єднує процеси збору, обробки, аналізу та візуалізації бібліометричних даних, надаючи інтерактивні інструменти для підтримки прийняття рішень.

Практичне значення одержаних результатів. Розроблена система може бути використана науковими установами, університетами та грантовими фондами для:

- Об'єктивної оцінки наукової продуктивності та впливу дослідників.

- Ідентифікації провідних науковців та формування ефективних дослідницьких колективів.
- Виявлення та моніторингу актуальних наукових трендів для стратегічного планування.
- Оптимізації розподілу ресурсів та фінансування наукових проектів.

РОЗДІЛ 1
Системний аналіз предметної області

1.1. Огляд наукометричних показників та бібліометричних баз даних

Сучасний науковий ландшафт характеризується безпрецедентним зростанням обсягу дослідницьких даних та публікацій, що обумовлює необхідність розробки ефективних інструментів для їхнього аналізу та оцінки. У цьому контексті *наукометрія – дисципліна*, що вивчає кількісні аспекти науки як інформаційного процесу – набуває ключового значення. Необхідно зауважити, що сам *термін «наукометрія»*, а також його класичне визначення, були введені ще наприкінці 1960-х років: "В 1966 році російський філософ і математик В. В. Налімов уводить у науковий обіг термін «наукометрія». Через три роки він дає визначення цього терміну – «Будемо називати наукометрією *кількісні методи вивчення розвитку науки як інформаційного процесу*»."[1]. Вона використовує статистичні методи для оцінки наукової продуктивності, впливу та колаборацій.

Основними інструментами наукометрії є різноманітні показники та індекси, призначені для характеристики публікаційної активності авторів, журналів та наукових установ. До найбільш поширених інструментів наукометрії належать:

Кількість публікацій. Базова метрика, що відображає загальну продуктивність автора або колективу. “Найбільш елементарний показник, пов'язаний із даними про публікації, — це просто кількість публікацій певного автора або групи авторів. Цей показник, що ґрунтується на одиниці аналізу документного рівня, може бути додатково уточнений для позначення типів публікацій, таких як рецензовані журнальні статті, книги або розділи книг, дисертації, галузеві (професійні) видання та тези конференцій, серед іншого. Таким чином, простий підрахунок кількості публікацій, автором чи співавтором яких є академічний клініцист, імовірно, *є неефективним методом для оцінки* дослідницької продуктивності”[2]. Хоча вона не завжди корелює з якістю, велика кількість публікацій може свідчити про активність дослідника.

Кількість цитувань. Показує, наскільки часто роботи автора використовуються та посилаються іншими дослідниками, що є індикатором наукового впливу та актуальності. “Аналіз цитувань є традиційним методом

оцінки впливу дослідження шляхом визначення того, як часто наступні публікації цитують конкретне видання. Це інструмент для вимірювання галузевого та географічного охоплення та швидкості засвоєння впливу публікації в літературі, оскільки він відстежує розвиток знань, ґрунтуючись на припущенні, що значущі публікації матимуть високу кількість цитувань... Однією очевидною *проблемою* з кількістю цитувань як показником впливу є те, що *ними можна маніпулювати* шляхом навмисного самоцитування або взаємного цитування колегами... Авторське самоцитування та взаємне цитування колегами часто штучно завищують кількість цитувань"[2].

h-індекс (індекс Хірша). Один з найвпливовіших показників, який поєднує кількість публікацій з їхнім цитуванням. Автор має h-індекс, якщо h з його N статей процитовано не менше h разів кожна, а решта (N-h) статей процитовано не більше h разів [3]. Цей індекс дозволяє оцінити як продуктивність, так і вплив автора.

i10-індекс. I10-індекс є найновішою із метрик журналів і був запроваджений Google Scholar у 2011 році. Це простий і зрозумілий показник індексування, який визначається шляхом підрахунку загальної кількості опублікованих журналом статей, що мають *щонайменше 10 цитувань* [4]. Він простий у розрахунку та часто використовується для швидкої оцінки впливу.

g-індекс. Варіант h-індексу, який надає більшу вагу статтям з великою кількістю цитувань. Автор має g-індекс, якщо g з його N статей процитовано в сукупності не менше g^2 разів [5].

Коефіцієнт співавторства. Характеризує кількість колег, з якими автор публікується, та може бути *індикатором рівня* його інтеграції у наукову спільноту та схильності до колаборацій [6].

Ці показники є фундаментальними для оцінки індивідуального та колективного внеску в науку, однак їх ефективне застосування вимагає доступу до великих, добре структурованих масивів бібліографічних даних.

Основними джерелами таких даних є *бібліометричні бази даних*, які систематизують інформацію про наукові публікації з усього світу. До найбільш авторитетних і широко використовуваних належать:

Scopus. Найбільша реферативна база даних видавництва Elsevier, що охоплює налічує близько *43 400* видань, із них приблизно *27 950* є *активними*, а *15 450 – неактивними* (переважно попередниками активних назв). Список назв *книг* Scopus містить понад *292 000* видань [7]. Scopus надає широкий спектр наукометричних інструментів, включаючи відстеження цитувань, розрахунок h-індексу та аналіз колаборацій.

Web of Science (WoS). Платформа компанії Clarivate Analytics, що включає декілька індексів цитування (Science Citation Index Expanded, Social Sciences Citation Index, Arts & Humanities Citation Index). WoS відрізняється суворим відбором джерел, що забезпечує високу якість індексованих публікацій [8]. Вона є основним джерелом для розрахунку імпакт-фактора журналів.

Google Scholar. Безкоштовна пошукова система, яка індексує наукові матеріали з широкого кола джерел, включаючи статті, дисертації, книги, препринти та технічні звіти. Google Scholar відомий своєю широтою охоплення, однак якість індексування та достовірність наукометричних показників (зокрема, i10-індексу) можуть бути менш суворими порівняно з Scopus та WoS.

ORCID (Open Researcher and Contributor ID). Міжнародний реєстр *унікальних ідентифікаторів* для науковців. ORCID надає унікальну можливість вирішити проблему неоднозначності імен авторів. Основна цінність реєстру ORCID полягає в тому, що він об’єднує дисципліни, організації та країни, пов’язуючи ORCID як з існуючими схемами ідентифікаторів, так і з публікаціями та іншою дослідницькою діяльністю [9]. Хоча ORCID не є базою цитувань, він є важливим для агрегації та верифікації даних.

Ефективний аналіз наукової діяльності вимагає не лише розуміння цих показників, а й здатності інтегрувати дані з різних джерел та застосовувати розширені методи обробки. Це дозволить вийти за рамки простих кількісних оцінок та отримати глибинне розуміння внеску авторів, динаміки наукових трендів та структури колаборацій.

1.2. Аналіз існуючих аналітичних систем для оцінки наукової діяльності

1.2.1. Огляд комерційних та відкритих аналітичних платформ

Для об'єктивної оцінки наукової діяльності та ефективного управління публікаційними даними розроблено низку комерційних та відкритих аналітичних систем. Їх функціонал переважно ґрунтується на агрегації інформації з бібліометричних баз даних як WoS та Scopus, але відрізняється глибиною аналізу та спеціалізованими інструментами [10].

Ключові гравці на ринку аналізу наукових даних включають:

1. **SciVal (Elsevier).** Це потужна комерційна платформа, що використовує дані Scopus. Вона дозволяє порівнювати дослідницьку ефективність

установ, груп та окремих авторів на основі стандартизованих показників. **Сильні сторони:** високоякісні дані, широкі можливості бенчмаркінгу, візуалізація наукових тем. **Обмеження:** висока вартість доступу, фокус на вже існуючих метриках, обмежені можливості для глибокого інтелектуального аналізу (Data Mining) та прогнозування.

2. **InCites (Clarivate Analytics).** Аналітична платформа, що використовує дані Web of Science. Надає інструменти для оцінки впливу, продуктивності та ранжування. **Сильні сторони:** престижність даних WoS, точний розрахунок імпакт-фактора журналів. **Обмеження:** сфокусована на традиційних показниках, не забезпечує гнучкість у застосуванні власних алгоритмів аналізу колаборацій чи тематичного моделювання.
3. **ResearchGate та Google Scholar Profiles.** Хоча вони не є суто аналітичними системами, вони надають авторам базові профілі та метрики (h-індекс, i10-індекс). **Сильні сторони:** відкритість, широке охоплення даних (особливо Google Scholar). **Обмеження:** відсутність стандартизованої методології збору даних, високий ризик "інформаційного шуму", обмежена функціональність для глибокого аналізу мережі колаборацій та виявлення прихованих тенденцій.

1.2.2. Порівняльний аналіз та визначення ніші дослідження

Для кращого розуміння функціональних прогалин, які має заповнити розроблювана система, доцільно провести порівняння ключових аналітичних можливостей:

Таблиця 1.1 Порівняння аналітичних можливостей існуючих систем

Критерій	SciVal/InCites	Google Scholar	Пропонована Система
Джерело даних	WoS/Scopus (закриті)	Відкриті джерела	Агрегація з різних баз
Аналіз колаборацій	Базовий (за афіліацією)	Обмежений	Мережевий аналіз, виявлення "мостів" та центральності

Продовження таблиці 1.1

Тематичний аналіз	На основі категорій журналів	Відсутній	Автоматичне тематичне моделювання (LDA) та тренди
Прогнозування	Відсутнє	Відсутнє	Прогнозування публікаційної активності
Гнучкість метрик	Стандартизовані	Стандартизовані	Розробка та застосування власних комбінованих індексів (EII)

1.2.3. Визначення проблем та обґрунтування необхідності розробки

Проведений аналіз існуючих рішень виявив такі ключові проблеми, які не вирішуються повною мірою:

1. **Стандартизація проти Гнучкості:** Комерційні системи пропонують потужні, але стандартизовані метрики, які часто не відповідають усім потребам оцінки в динамічному середовищі.
2. **Відсутність глибинного інтелектуального аналізу:** Жодна з платформ не пропонує комплексного застосування алгоритмів Data Mining (зокрема, LDA чи просунутих графових моделей) для автоматичного виявлення прихованих тенденцій чи сегментації авторів.
3. **Обмеженість у прогнозуванні:** Існуючі системи є ретроспективними (оцінюють минуле) і не мають ефективних інструментів для прогнозування майбутньої активності авторів чи тематик.

Таким чином, необхідність розробки **Системи аналізу даних авторів наукових статей** обґрунтована потребою в гнучкому, інтелектуальному інструменті, здатному інтегрувати дані з різних джерел, застосовувати розширені методи аналізу для виявлення неочевидних тенденцій і надавати прогностичні оцінки дослідницького внеску, заповнюючи прогалини у функціоналі існуючих рішень.

1.3. Формалізація задачі дослідження та визначення функціональних вимог до системи

1.3.1. Формалізація задачі дослідження

На основі проведеного аналізу наукометричних показників та існуючих аналітичних систем (підрозділи 1.1, 1.2) виявлено, що поточні інструменти мають обмеження у глибинному інтелектуальному аналізі даних авторів, виявленні прихованих тенденцій та прогнозуванні. Таким чином, основна дослідницька задача полягає у розробці та дослідженні ефективності підходів до комплексного аналізу бібліометричних даних, що дозволить:

1. **Автоматизувати процес збору, очищення та структуризації** даних про авторів наукових статей з різних бібліометричних джерел.
2. **Розробити та адаптувати алгоритми інтелектуального аналізу даних** для:

   Тематичного моделювання: автоматичного виявлення ключових наукових тематик та їх динаміки у діяльності авторів.

   Мережевого аналізу: ідентифікації наукових колаборацій, впливових авторів ("хабів") та міждисциплінарних зв'язків ("мостів").

   Прогнозування: оцінки майбутньої публікаційної активності та зміни тематичних пріоритетів.
3. **Розробити програмне забезпечення**, що інтегрує ці аналітичні алгоритми та забезпечує інтуїтивно зрозумілу візуалізацію отриманих результатів для підтримки прийняття рішень.
4. **Експериментально підтвердити ефективність** розроблених методів та системи на реальних наборах бібліометричних даних.

1.3.2. Визначення функціональних вимог до системи

Для вирішення сформульованої дослідницької задачі та забезпечення комплексної функціональності, до розроблюваної системи аналізу даних авторів наукових статей висуваються наступні функціональні вимоги (ФВ):

ФВ1. Модуль збору та попередньої обробки даних:

ФВ1.1. Забезпечення збору бібліографічних даних (інформація про авторів, назви статей, анотації, ключові слова, журнали, роки публікації, цитування, співавторство, афіліації) з обраних зовнішніх джерел (наприклад, API Scopus/WoS, OpenAlex).

ФВ1.2. Здійснення автоматичної очистки даних (видалення дублікатів, нормалізація імен авторів та афіліацій, обробка пропущених значень).

ФВ1.3. Структуризація та зберігання даних у внутрішньому сховищі даних, оптимізованому для аналітичних запитів.

ФВ2. Модуль інтелектуального аналізу даних:

ФВ2.1. Реалізація алгоритмів тематичного моделювання для автоматичного виявлення та кластеризації наукових тематик у діяльності авторів.

ФВ2.2. Застосування алгоритмів мережевого аналізу для побудови графу наукових колаборацій, розрахунку метрик центральності (ступеня, міжпозиційної) та виявлення наукових груп.

ФВ2.3. Впровадження моделей прогнозування часових рядів для оцінки динаміки публікаційної активності авторів та/або популярності тематик.

ФВ2.4. Розрахунок та агрегація стандартних та розширених наукометричних показників (h-індекс, i10-індекс, індекс співавторства, власний індекс впливу EII).

ФВ3. Модуль візуалізації та формування звітів:

ФВ3.1. Надання інтерфейсу для наочного представлення результатів тематичного моделювання (наприклад, інтерактивні "хмари слів", графіки розподілу тем).

ФВ3.2. Візуалізація графу наукових колаборацій з можливістю фільтрації, масштабування та виділенням ключових авторів/груп.

ФВ3.3. Відображення динаміки та прогнозів публікаційної активності у вигляді часових графіків.

ФВ3.4. Формування інтерактивних профілів авторів з їхніми основними наукометричними показниками та тематичною спеціалізацією.

ФВ4. Модуль адміністрування та конфігурації:

ФВ4.1. Можливість налаштування параметрів збору даних (джерела, період).

ФВ4.2. Можливість налаштування параметрів аналітичних алгоритмів (кількість тем для LDA, порогові значення для мережевого аналізу).

1.3.3. Нефункціональні вимоги

До системи також висуваються нефункціональні вимоги, що стосуються її якості та експлуатації:

Продуктивність: Система повинна забезпечувати обробку значних обсягів даних за прийнятний час.

Надійність: Безперебійна робота системи, стійкість до збоїв.

Масштабованість: Можливість розширення функціоналу та обробки зростаючих обсягів даних.

Зручність використання (Usability): Інтуїтивно зрозумілий інтерфейс користувача.

Безпека: Забезпечення захисту даних, особливо якщо вони включають конфіденційну інформацію.

РОЗДІЛ 2
Моделювання системи

2.1. Дослідження методів збору та структуризації бібліометричних даних

2.1.1. Методи збору бібліометричних даних з відкритих API

Ефективність системи аналізу даних авторів наукових статей безпосередньо залежить від можливостей доступу до актуальних та репрезентативних джерел інформації. У процесі дослідження було зосереджено на використанні відкритих програмних інтерфейсів додатків (API), що надають широкий доступ до бібліометричних даних, підтримуючи принципи відкритої науки та мінімізуючи бар'єри доступу. Були обрані наступні ключові джерела:

1. **OpenAlex API.**

OpenAlex є ініціативою, спрямованою на створення відкритого та всеосяжного графу глобальної наукової системи. Він індексує мільйони наукових робіт, авторів, джерела, установи та концепти, надаючи потужний API для доступу до цих даних [11]. Його перевагою є унікальна система ідентифікації та нормалізації об'єктів.

Переваги: Вільний доступ, широкий спектр індексованих даних, унікальні ідентифікатори (IDs) для авторів, робіт, концептів, що спрощує дедублікацію та зв'язування даних. Дозволяє отримувати інформацію про цитування, співавторство, афіліації та тематики.

Недоліки: Відносна новизна платформи, що може потребувати додаткової валідації даних порівняно з усталеними комерційними базами.

2. **Semantic Scholar API.**

Semantic Scholar — це наукова пошукова система та база даних, розроблена Інститутом штучного інтелекту Аллена (AI2), яка використовує штучний інтелект для аналізу та вилучення знань з наукових публікацій. Її API надає доступ до метаданих статей, інформації про авторів, цитування та ключові концепти [12].

Переваги: Висока якість семантичного аналізу тексту, що дозволяє отримувати більш релевантні ключові слова та абстракти. Добре підходить для розширеного тематичного аналізу.

Недоліки: Охоплення може бути меншим, ніж у OpenAlex для певних галузей, ліміти запитів можуть бути більш суворими.

3. **CrossRef API.**

CrossRef є некомерційною організацією, що надає DOI (Digital Object Identifier) для наукових публікацій та підтримує базу метаданих про ці об'єкти. CrossRef API є фундаментальним для пошуку метаданих за DOI, виявлення цитувань та інформації про видавців [13].

Переваги: Надійне джерело для метаданих публікацій, ідентифікація через DOI, що є стандартом. Допомагає у зв'язуванні інформації між різними базами.

Недоліки: Переважно фокусується на метаданих публікацій, надаючи менше інформації про профілі авторів порівняно з OpenAlex або Semantic Scholar.

Використання цих трьох відкритих API дозволяє створити багатоаспектний набір даних, компенсуючи потенційні обмеження кожного окремого джерела та забезпечуючи широкий охоплення необхідних атрибутів для подальшого аналізу.

2.1.2. Визначення необхідних атрибутів даних

Для реалізації комплексного аналізу наукової діяльності авторів, з урахуванням можливостей обраних API, було визначено наступний перелік ключових атрибутів, які необхідно збирати та зберігати:

Для авторів

Ця структура містить нормалізовану інформацію про автора, яка розширена для зберігання результатів мережевого аналізу та детальної тематичної спеціалізації.

- **author_id** / **orcid:** Унікальні ідентифікатори автора.
- **display_name** / **aliases:** Ім'я та відомі варіанти імені.
- **current_institution** / **all_institutions:** Поточна та історичні афіліації.
- **h_index** / **i10_index** / **total_citations** / **total_works:** Ключові наукометричні показники (для аналізу впливу).
- **publications_by_year** / **citations_by_year:** Розподіл активності та цитувань за роками (для аналізу часових рядів та прогнозування).

- **top_concepts** / **topic_specialization** / **dominant_topics:** Попередня оцінка тематик та розподіл ймовірностей тем після тематичного моделювання.
- **network_metrics:** Словник, що містить обчислені метрики центральності (ступінь, посередництва) у графі співпраці (для мережевого аналізу).
- **community_id:** Ідентифікатор **спільноти**, до якої належить автор (результат алгоритмів виявлення спільнот).
- **source_apis** / **raw_data:** Метаінформація про джерела даних.

Для публікацій

Структура публікації розширена для більш точної роботи з текстом та результатами тематичного моделювання.

- **work_id** / **doi:** Унікальні ідентифікатори публікації.
- **title:** Назва статті.
- **abstract:** Анотація статті.
- **abstract_inverted_index:** Структура, що зберігає анотацію у форматі інвертованого індексу (використовується деякими API, важлива для реконструкції).
- **publication_year** / **publication_date:** Дата публікації.
- **authors:** Інформація про співавторів (для побудови графу колаборацій).
- **journal_title** / **publisher:** Назва джерела публікації.
- **concepts** / **topics** / **keywords:** Стандартизовані концепти (OpenAlex concepts), нові тематики (OpenAlex topics) та ключові слова, надані API або авторами.
- **topic_distribution** / **assigned_topics** / **main_topic:** Розподіл ймовірностей статей за темами, список присвоєних тем та основна домінуюча тема (результати тематичного моделювання).
- **cited_by_count** / **references:** Кількість цитувань та список посилань (для графу цитувань).
- **open_access_status** / **pdf_url:** Інформація про доступність.

Для зв'язків та результатів аналізу

Ці допоміжні структури узагальнюють результати складних аналітичних процедур (тематичного та мережевого моделювання).

Зв'язок співавторства

- **author_1_id** / **author_2_id:** Ідентифікатори авторів, що утворюють ребро.
- **shared_publications:** Список спільних робіт.
- **collaboration_strength:** Вага ребра, що відображає інтенсивність співпраці.
- **first_collaboration_year** / **last_collaboration_year:** Часові межі співпраці.
- **shared_topics:** Спільні тематики авторів.

Результат тематичного моделювання

Ця структура зберігає загальні параметри та результати роботи алгоритмів LDA та NMF (Non-negative Matrix Factorization).

- **model_type** / **n_topics:** Тип моделі ("lda", "nmf" або "api") та кількість визначених тем.
- **topics:** Словник, що містить ключові слова з їхньою вагою для кожної теми (для якісної інтерпретації).
- **topic_names:** Словник, що містить **назви тем**, присвоєні дослідником або API.
- **coherence_score** / **perplexity:** Метрики для оцінки якості моделі.
- **approach:** Вказує, які дані використовувалися для моделювання ("abstract", "concepts" або "full" текст).

Результат мережевого аналізу

Ця структура узагальнює глобальні та деталізовані показники графу співпраці.

- **total_nodes** / **total_edges:** Загальні розміри мережі.
- **density** / **avg_degree** / **clustering_coefficient:** Глобальні метрики, що характеризують структуру та згуртованість графу.
- **communities:** Словник, що групує авторів за ідентифікаторами виявлених спільнот.

- **centrality_metrics:** Деталізований словник, що містить обчислені метрики центральності (ступінь, посередництва тощо) для кожного вузла.

2.1.3. Методи агрегації, очистки та структуризації даних

Зібрані дані з OpenAlex, Semantic Scholar та CrossRef, хоча і є структурованими, вимагають обов'язкових етапів агрегації, очистки та трансформації для створення єдиного, консистентного та високоякісного набору даних, що є критично важливим для достовірного інтелектуального аналізу

1. **Багатоджерельне вилучення та первинна нормалізація**

На цьому етапі виконується збір сирих даних та їхнє первинне перетворення у внутрішній формат:

- **Координація збору:** Централізований клас-колектор координує вилучення публікацій з OpenAlex, Semantic Scholar та CrossRef, застосовуючи пагінацію та дотримуючись обмежень запитів (Rate Limiting).
- **Уніфікація форматів:** Отримані JSON-об'єкти з кожного API негайно трансформуються у внутрішні структури даних.
- **Очистка ідентифікаторів:** Видалення префіксів URL з DOI, author_id та orcid.
- **Реконструкція даних:** Відновлення тексту **анотації** (abstract) з формату **інвертованого індексу** (характерного для OpenAlex) для його подальшого використання у тематичному моделюванні.

2. **Агрегація та дедублікація**

Цільова задача — створення унікального набору сутностей шляхом злиття інформації з різних джерел:

- **Крос-платформна ідентифікація публікацій:** Встановлення жорсткого зіставлення (Hard Matching) на основі DOI як пріоритетного унікального ідентифікатора.
- **Консолідація інформації:** При виявленні дублікатів відбувається злиття даних:
  - Об'єднання списку API-джерел.

- **Доповнення пропущених критичних полів**, зокрема анотації (abstract), якщо вона відсутня в одному записі, але наявна в дублікаті.

- **Вторинне зіставлення:** У разі відсутності DOI застосування назви статті (title після очистки) як вторинного критерію для виявлення дублікатів.
- **Нормалізація авторів:** Збір детальних профілів авторів з різних джерел, використовуючи ORCID як первинний ключ для потенційної дедублікації на наступних етапах.

3. Фінальна трансформація та підготовка до аналізу

Завершальний етап підготовки даних до застосування алгоритмів моделювання:

- **Фінальна нормалізація тексту:** Приведення всіх ключових текстових полів до єдиного формату (нижній регістр, очистка) для забезпечення якості вхідних даних для LDA та NMF.
- **Зберігання в реляційній БД:** Завантаження нормалізованих сутностей та моделювання складних відносин (співавторство, цитування) у PostgreSQL за допомогою SQLAlchemy.
- **Підготовка до графового аналізу:** Формування набору вузлів та ребер з БД для подальшого розрахунку метрик центральності та виявлення спільнот за допомогою бібліотеки NetworkX.

Впровадження цих методів збору, агрегації та структуризації даних є фундаментальним для створення високоякісної вхідної інформації, що є передумовою для достовірного та глибокого інтелектуального аналізу наукової діяльності.

2.2. Застосування алгоритмів інтелектуального аналізу для виявлення тенденцій

2.2.1. Алгоритми тематичного моделювання для виявлення наукових напрямків

Для автоматичного виявлення основних наукових тематик у діяльності авторів та їхньої динаміки в часі було обрано підходи на основі тематичного моделювання (Topic Modeling). Ці алгоритми дозволяють абстрагуватися від

окремих слів і виявляти приховані "теми" в колекції текстових документів, представляючи кожен документ як суміш тем, а кожну тему – як суміш слів.

Підготовка текстового корпусу та векторизація

Якість виявлених тем критично залежить від підготовки вхідного корпусу. Використовуються два основні підходи до формування "документа" для кожної публікації:

1. **На основі абстракту:** Об'єднання анотації, назви статті, а також назв тем/концептів, наданих API (OpenAlex topics/concepts). Назви тем/концептів дублюються для підвищення їхньої ваги, що є важливим для акцентування уваги моделі на високорелевантних, попередньо класифікованих термінах.
2. **На основі концептів/ключових слів:** Формування документа без використання анотації, шляхом об'єднання назви статті, ключових слів та концептів/топіків OpenAlex.

Після підготовки тексту виконується векторизація за допомогою TfidfVectorizer (Term Frequency-Inverse Document Frequency). Векторизація з використанням TF-IDF надає більшу вагу тим словам, які є унікальними для конкретного документа, але не є надто поширеними у всьому корпусі. Для зниження розмірності та покращення якості моделювання застосовується:

- **Фільтрація за частотою:** Видалення рідкісних (*min_df=3*) та дуже частих слів (*max_df=0.7*).
- **Використання n-грам:** Включення біграм (*ngram_range=(1, 2)*) для кращого захоплення значущих словосполучень (наприклад, "Deep Learning" замість окремих "Deep" та "Learning").

1. **Латентне розміщення Діріхле (Latent Dirichlet Allocation, LDA).**

LDA є генеративною статистичною моделлю, яка припускає, що кожен документ є сумішшю невеликої кількості тем, а кожна тема – це суміш слів. Модель ідентифікує ці приховані теми на основі статистичних закономірностей спільного входження слів у документах [14].

Застосування: В нашій системі LDA застосовуватиметься до об'єднаного корпусу анотацій та ключових слів статей. Це дозволить:

- Модель навчається на підготовленому корпусі для виявлення основних тематичних кластерів у наукових публікаціях.

- Модель присвоює кожній статті розподіл ймовірностей тем (topic_distribution), визначаючи основну домінуючу тему (main_topic) та список топ-3 присвоєних тем (assigned_topics).

Переваги: Висока ефективність для виявлення "прихованих" тем, можливість працювати з великими текстовими корпусами.

Недоліки: Чутливість до кількості тем (гіперпараметр k), що потребує оптимізації; інтерпретація деяких тем може бути не завжди однозначною.

2. **Негативна матрична факторизація (NMF).**

NMF – це метод розкладання матриці, який розкладає вихідну матрицю "документ-слово" на дві матриці: "документ-тема" та "тема-слово", де всі елементи є невід'ємними [15].

Застосування: NMF може використовуватися як альтернативний або додатковий метод до LDA для порівняння результатів. Його застосування дозволить підтвердити стабільність виявлених тем та їхніх ключових слів.

Переваги: Легша інтерпретація компонентів завдяки адитивному розкладанню; часто дає чіткіші теми порівняно з LDA; швидший час виконання на великих розріджених матрицях.

Недоліки: Чутливість до початкових умов; може бути менш ефективним для дуже розріджених даних порівняно з імовірнісним підходом LDA.

Обчислення тематичної спеціалізації та динаміки

Тематична спеціалізація: Після навчання моделі тематичний розподіл публікацій використовується для визначення спеціалізації автора. Обчислюється сума ваг кожної теми за всіма публікаціями автора, а результат нормалізується для отримання фінального розподілу ймовірностей тем для автора.

Динаміка тем: Структура даних, де кожна публікація асоційована з роком та темою, є основою для відстеження динаміки популярності тем з часом. Агрегування кількості публікацій за роками та присвоєними темами дозволяє побудувати часові ряди для кожної теми, що є вхідними даними для прогнозних моделей (див. 3.4).

Впровадження цих методів тематичного моделювання дозволяє автоматично виявляти приховані наукові напрямки в корпусі публікацій, що є фундаментальним для визначення тематичної спеціалізації авторів та виявлення тенденцій у їхній діяльності.

2.2.2. Методи мережевого аналізу для ідентифікації наукових колаборацій

Для аналізу структури наукової співпраці та виявлення впливових авторів було обрано підходи на основі мережевого аналізу. Бібліометричні дані природно моделюються як графи, де автори є вузлами, а спільні публікації – ребрами.

1. **Побудова графу колаборацій.**

Модель графу: Створюється неорієнтований граф G=(V, E), де V — це множина унікальних ідентифікаторів авторів, а E — множина зв'язків між співавторами.

Правило формування ребер: Ребро існує між двома авторами, якщо вони є співавторами принаймні однієї публікації.

Вага ребра: Вага ребра відображає кількість спільних публікацій між двома авторами. Додатково ребро зберігає список ідентифікаторів спільних робіт.

Застосування: Граф дозволяє візуалізувати мережу співпраці та слугує основою для розрахунку всіх подальших метрик.

2. **Метрики центральності.**

Розрахунок метрик центральності дозволяє кількісно оцінити роль, вплив та стратегічну важливість кожного автора у мережі колаборацій:

Ступінь центральності (Degree Centrality): Кількість прямих зв'язків вузла. Вказує на кількість співавторів дослідника [16].

Центральність за посередництвом (Betweenness Centrality): “Вимірює, наскільки актор (вузол) лежить між іншими акторами на їхніх геодезичних шляхах (найкоротших шляхах між іншими парами вузлів). Таким чином, актори з високою центральністю посередництва мають потенціал впливати на інших у мережі, які знаходяться поблизу них... Вузол із високою центральністю посередництва може потенційно впливати на поширення інформації через мережу, сприяючи, перешкоджаючи або навіть змінюючи комунікацію між іншими”[16]. Тобто високе значення вказує на роль автора як "моста" або комунікатора між різними групами.

Центральність за близькістю (Closeness Centrality): “Міри близькості ґрунтуються на ідеях ефективності та незалежності. Завдяки тому, що актори розташовані близько до інших у мережі, актори з високими показниками близькості здатні ефективно передавати інформацію**”**[16]. Показує, наскільки

вузол близький до всіх інших вузлів у мережі. Вказує на швидкість поширення інформації від цього автора.

Центральність за власним вектором (Eigenvector Centrality): “Ті, хто має високу центральність власного вектора, пов'язані з добре підключеними акторами (тобто тими, хто сам має високу центральність). Таким чином, вони можуть впливати на багатьох інших у мережі, або безпосередньо, або опосередковано через свої зв'язки”[16]. Призначає високі оцінки вузлам, які пов'язані з іншими добре пов'язаними вузлами. Вказує на вплив автора в мережі , де зв'язок із високо впливовим автором є більш цінним, ніж зв'язок із багатьма низько впливовими.

Застосування: Розрахунок цих метрик дозволить ідентифікувати ключових, найбільш впливових та стратегічно важливих авторів у мережі колаборацій.

3. **Виявлення спільнот (Community Detection).**

Алгоритм: Для ідентифікації щільно пов'язаних груп вузлів (спільнот) у графі використовується алгоритм Louvain (або, як резервний варіант, Greedy Modularity Communities).

Мета: Виявити стійкі наукові колективи або тематично згруповані дослідницькі групи на основі їхньої спільної публікаційної активності.

Результат: Граф розділяється на кластери, де вузли всередині кластера мають більше зв'язків між собою, ніж з вузлами поза кластером. Кожен автор отримує ідентифікатор спільноти.

Застосування методів мережевого аналізу дозволяє точно ідентифікувати ключових, найбільш впливових та стратегічно важливих авторів у мережі колаборацій та виявити базову кластерну структуру наукової співпраці, що є критично важливим для розуміння соціальної динаміки наукової діяльності.

2.2.3. Моделі прогнозування часових рядів для оцінки динаміки

Для прогнозування майбутньої публікаційної активності авторів або зміни популярності наукових тематик будуть використовуватися методи аналізу часових рядів.

Підготовка часових рядів

Первинний часовий ряд формується шляхом агрегації кількості публікацій за роками (calculate_dynamics). Кожна точка даних представляє кількість опублікованих робіт за конкретний рік, що дозволяє моделювати динаміку:

Публікаційна активність: Ряд кількості публікацій певного автора або групи авторів.

Популярність тематики: Ряд кількості публікацій, присвоєних певній тематиці, що дозволяє оцінювати зростаючі або затухаючі наукові напрямки.

Моделі ARIMA (AutoRegressive Integrated Moving Average).

ARIMA — це клас моделей, які описують часові ряди як функцію минулих значень. “Модель ARIMA є комбінацією моделі авторегресії (AR), моделі ковзного середнього (MA) та операції диференціювання (Інтеграції).

Авторегресійна модель (AR)

Yt прогнозується на основі одного або кількох запізнілих (лагованих) значень Yt... де c – це константа, phi – величина автокореляції, p – кількість лагів (запізнень), а epsilon-t – помилка.

Модель ковзного середнього (MA)

Yt прогнозується на основі одного або кількох запізнілих (лагованих) значень помилки (epsilon-t)... де theta – це значення автокореляції помилок, а q – кількість лагів (запізнень).

Диференціювання (Інтеграція)

У моделі ARIMA часовий ряд, що моделюється, повинен бути стаціонарним для отримання значущих прогнозів. Стаціонарність досягається за допомогою диференціювання, яке передбачає обчислення різниці між суміжними спостереженнями.

Загальна нотація ARIMA

Модель ARIMA — це комбінація моделі AR, моделі MA та диференціювання (Інтеграції)... Основне позначення для опису несезонної моделі ARIMA — це (p, d, q), де p, d та q є додатними цілими числами:

- p = порядок AR-частини моделі
- d = ступінь несезонного диференціювання.
- q = порядок MA-частини моделі.”[17]

Застосування: ARIMA буде застосована для моделювання та прогнозування:

- Модель застосовується для прогнозування динаміки наукової активності на визначену кількість майбутніх періодів.
- У реалізації використовується фіксована параметризація ARIMA(1, 1, 1), яка передбачає один крок авторегресії, одну різницю (для досягнення стаціонарності) та один крок ковзного середнього.

Переваги: Добре зарекомендувала себе для прогнозування стаціонарних та нестаціонарних часових рядів.

Недоліки: Вимагає стаціонарності ряду (або його перетворення), чутлива до вибору параметрів p, d, q.

Експоненційне згладжування (Exponential Smoothing).

Ці моделі будують прогноз на основі зваженого середнього минулих спостережень, де ваги експоненційно зменшуються з часом. “Прогнози, створені за допомогою методів експоненційного згладжування, є зваженими середніми минулих спостережень, причому ваги зменшуються експоненційно в міру того, як спостереження стають старішими. Іншими словами, чим новіше спостереження, тим вищою є пов'язана з ним вага”[18].

Застосування:

- Використовується модель Triple Exponential Smoothing з адитивним трендом (trend='add') та без сезонності (seasonal=None) для прогнозування. Це дозволяє ефективно враховувати загальний напрямок (зростання/спад) часового ряду.
- Використовується як альтернативний або додатковий метод до ARIMA для порівняння точності прогнозування.

Переваги: Легкість налаштування та інтерпретації, ефективність для рядів з чіткими трендами.

Запасний механізм (Fallback)

У разі недостатньої кількості даних або помилок при навчанні складніших моделей (ARIMA, ExponentialSmoothing) використовується простий лінійний прогноз (simple_linear_forecast). Цей метод екстраполює майбутні значення на основі лінійного тренду, розрахованого за історичними даними, забезпечуючи базовий, але завжди доступний прогноз

Застосування цих алгоритмів інтелектуального аналізу дозволить не лише провести кількісну оцінку, а й виявити приховані закономірності та тенденції у бібліометричних даних, що є основою для підтримки прийняття обґрунтованих рішень у сфері наукової діяльності.

2.3. Архітектура програмного забезпечення для реалізації аналітичної системи

Для забезпечення ефективного збору, обробки, аналізу та візуалізації бібліометричних даних авторів, розроблена система базується на багатошаровій архітектурі (Layered Architecture), що дозволяє досягти масштабованості, модульності та гнучкості у впровадженні нових аналітичних алгоритмів. Основна ідея архітектури полягає у чіткому розділенні функціональних блоків, що відповідають за різні етапи обробки даних, та їхній взаємодії.

2.3.1. Загальна архітектура системи

Загальна архітектура системи є трирівневою і включає такі основні рівні:

1. **Рівень джерел даних (Data Source Layer):** Відповідає за зовнішню взаємодію з джерелами даних та вилучення сирих даних та первинну агрегацію для забезпечення консистентності даних.
2. **Рівень обробки та аналізу даних (Back-End / Processing Layer):** Ядро системи, що реалізує перетворення та фільтрацію даних, інтелектуальний аналіз та зберігання даних.
3. **Рівень представлення (Front-End / Presentation Layer):** Забезпечує взаємодію користувача та візуалізацію результатів.

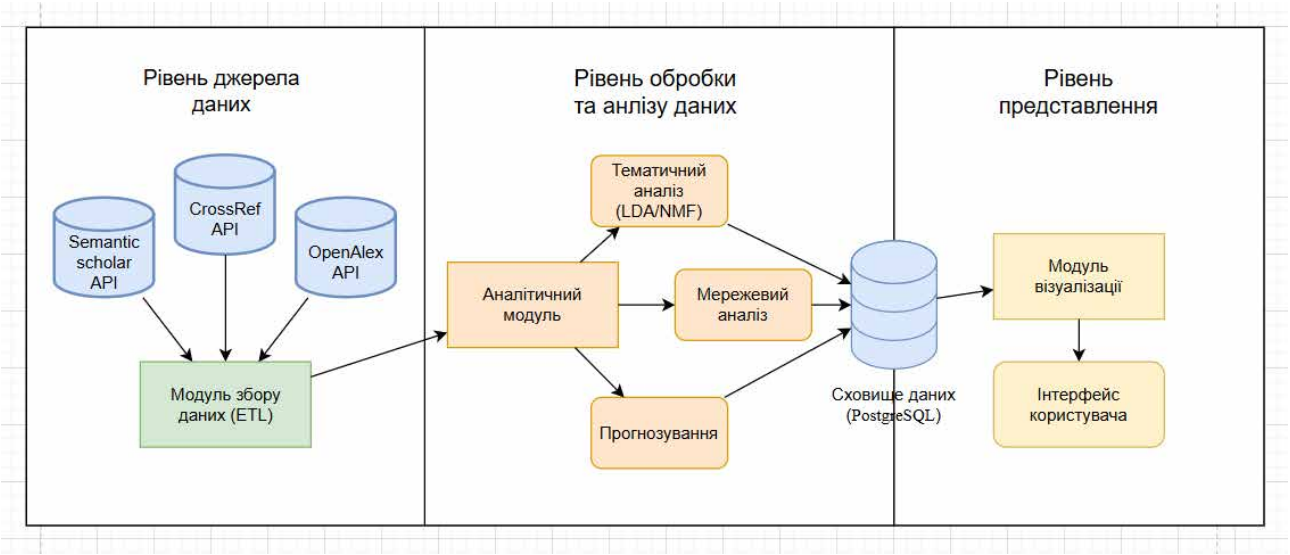


Рис 2.1 Загальна архітектура системи аналізу даних а вторів наукових статей

Хоча на схемі (Рис. 2. 1) компоненти можуть бути розташовані поруч, використання терміна 'Рівень' підкреслює ієрархію залежностей та односпрямований потік контролю від рівня представлення до рівня доступу до даних, що є ключовою ознакою цього архітектурного патерну.

2.3.2. Компоненти рівнів

1. Рівень джерел даних

Цей рівень забезпечує зовнішнє підключення до первинних джерел бібліометричної інформації. Він включає:

API-з'єднання: Взаємодія з зовнішніми джерелами даних для отримання метаданих, інформації про авторів, цитування та зв'язки.

Менеджер даних: Керує API-з’єднаннями, дотримання політики обмеження запитів (Rate Limiting) та первинну агрегацію сирих даних з різних джерел.

2. Рівень обробки та аналізу даних (Back-End)

Цей рівень є **ядром системи** і містить ключові модулі, що реалізують ETL-процеси та інтелектуальний аналіз.

Модуль ETL (Extract, Transform, Load): Забезпечується методом на основі жорсткого зіставлення (Hard Matching) за DOI та назвою. Виконується нормалізація ідентифікаторів (DOI, ORCID) та реконструкція анотацій з інвертованих індексів.

Сховище даних (Data Storage):

Реляційна БД (PostgreSQL): Використовується для зберігання структурованих метаданих (публікації, автори, журнали, афіліації) та підтримки складних аналітичних запитів.

Графова Модель (Концептуально): Структура графу колаборацій створюється на базі бібліотеки NetworkX у Python, зберігаючи відносини співавторства та їхні ваги (кількість спільних публікацій).

Аналітичний Модуль (Analytical Engine):

Модуль Тематичного Моделювання: Реалізує алгоритми LDA/NMF для виявлення тематичних кластерів та їх розподілу.

Модуль Мережевого Аналізу: Розраховує метрики центральності (Betweenness, Degree) та ідентифікує спільноти (Louvain method) на графі колаборацій.

Модуль Прогнозування: Застосовує моделі ARIMA/Exponential Smoothing для прогнозування публікаційної активності та тематичних трендів.

Модуль Метрик: Розраховує стандартні (h-індекс) та розширені (EII) наукометричні показники.

3. Рівень представлення (Front-End)

Цей рівень забезпечує взаємодію користувача з результатами аналізу.

Сервер застосунку (Application Server): Виконує бізнес-логіку, приймає запити від користувача, взаємодіє з Аналітичним Модулем та Сховищем даних.

Інтерфейс користувача (User Interface): Веб-інтерфейс, розроблений для інтуїтивного доступу до аналітичних результатів. Забезпечує динамічну візуалізацію графу колаборацій, тематичних трендів, часових рядів та детальних профілів авторів.

2.3.3. Взаємодія компонентів та технологічний стек

Система функціонує за принципом асинхронної обробки:

1. **Початкове завантаження (ETL):** Менеджер даних збирає інформацію через API і передає її Модулю ETL для очистки та завантаження до Сховища даних.
2. **Аналіз:** Аналітичний Модуль періодично звертається до Сховища даних, виконує складні обчислення (LDA, мережеві метрики, ARIMA) та зберігає розраховані результати (наприклад, тематичний розподіл, значення центральності) назад у сховище.
3. **Візуалізація:** Користувач через Інтерфейс надсилає запит на Сервер застосунку, який отримує готові агреговані та аналітично оброблені дані зі Сховища та відображає їх у вигляді інтерактивних графіків.

Таблиця 2.1 Технологічний стек

Категорія	Технологія / Бібліотека	Призначення
Мова програмування	Python	Основна мова для ETL, Аналітичного Модуля та Back-End логіки.
Сховище даних	PostgreSQL	Основне реляційне сховище для структурованих метаданих та аналітичних результатів.
Ключові бібліотеки (Аналіз)	Scikit-learn, statsmodels	Реалізація LDA/NMF та моделей прогнозування ARIMA/Exponential Smoothing.

Продовження та,лиці 2.1

Ключові бібліотеки (Графовий аналіз)	NetworkX, python-louvain	Побудова графу колаборацій, розрахунок метрик центральності та виявлення спільнот.
Ключові бібліотеки (ETL)	requests, Pandas	Взаємодія з API та маніпуляції з даними (очистка, векторизація).
Front-End	JavaScript, D3.js	Розробка веб-інтерфейсу, динамічна та інтерактивна візуалізація результатів.

Така модульна архітектура забезпечує високу ефективність у роботі зі значними обсягами бібліометричних даних та дозволяє швидко адаптувати нові дослідницькі алгоритми в Аналітичному Модулі без суттєвої зміни інших компонентів системи.

РОЗДІЛ 3
Розробка системи

3.1. Опис експериментального набору даних та середовища

3.1.1. Формування та характеристика експериментального набору даних

Для експериментальної перевірки ефективності розроблених методів аналізу та функціональності системи було сформовано репрезентативний набір бібліометричних даних. Збір даних здійснювався з використанням API OpenAlex, Semantic Scholar та CrossRef, як це було описано у підрозділі 2.1.

Область дослідження: Для забезпечення цілеспрямованого аналізу, набір даних було сфокусовано на публікаціях у галузі комп'ютерних наук та штучного інтелекту. Цей вибір обумовлений високою динамікою розвитку цих напрямків, великою кількістю публікацій та складністю виявлення трендів, що робить його ідеальним для застосування інтелектуальних методів аналізу.

Стратегія використання двох наборів даних

Через високу вимогливість до апаратних ресурсів при побудові та аналізі графів колаборацій на надвеликих масивах, було прийнято рішення використовувати два набори даних з різними цілями:

1. Основний корпус (для Тематичного аналізу та Прогнозування)

Цей великий масив даних використовується для тематичного моделювання (LDA/NMF), аналізу динаміки тематичних тенденцій та прогностичних моделей (ARIMA), де критично важливим є обсяг тексту та часова глибина.

Критерії відбору:

- Публікації, що індексуються в обраних API.
- Тематична відповідність: "machine learning", "artificial intelligence", "data science", "natural language processing", "computer vision", "software engineering", "network analysis", Neural.
- Період публікацій: з 2017 по 2023 рік включно, щоб охопити актуальні тенденції але спробувати уникнути похибки від ще не опублікованих статей та мати достатній часовий проміжок для аналізу динаміки та прогнозування.

Характеристики датасету:

- **Кількість унікальних публікацій:** Приблизно *37,000* статей, матеріалів конференцій та розділів книг.
- **Кількість унікальних авторів:** Близько *102,000* унікальних ідентифікованих авторів (після дедублікації за ORCID та нечіткого зіставлення імен).
- **Кількість унікальних афіліацій:** Понад *12,000* установ та університетів.
- **Середня кількість цитувань на статтю:** 109.79 (станом на дату збору).
- **Середня кількість співавторів на статтю:** 3.87.

2. Додатковий корпус (для Мережевого аналізу)

Менший, але більш сфокусований набір даних, використовується виключно для мережевого аналізу колаборацій (Графи співавторства). Це дозволяє ефективно розрахувати метрики центральності та виявити спільноти без надмірного навантаження на пам'ять та наддасть візуалізацію графа з розбірливими даними.

Критерії відбору:

- Публікації, що індексуються в обраних API.
- Тематична відповідність (ключові слова, концепти): “Networks for Image Processing”.
- Період публікацій: з 2019 по 2023 рік включно, щоб охопити достатній проміжок часу для фіксації взаємодії авторів.

Характеристики датасету:

- **Кількість унікальних публікацій:** *386* статей, матеріалів конференцій та розділів книг.
- **Кількість унікальних авторів:** *1636* унікальних ідентифікованих авторів (після дедублікації за ORCID та нечіткого зіставлення імен).
- **Кількість унікальних афіліацій:** 511 установ та університетів.
- **Середня кількість цитувань на статтю:** 58.89 (станом на дату збору).
- **Середня кількість співавторів на статтю:** 4.62.

Структура даних: Кожен запис публікації включав: DOI, назву, анотацію, ключові слова/концепти, рік публікації, журнал/конференцію, список авторів (з ID та афіліаціями), кількість цитувань. Для кожного автора було зібрано: унікальний ID, ім'я, поточну афіліацію, h-індекс (якщо надається API).

3.1.2. Програмно-апаратне середовище

Експериментальні дослідження та розробка системи проводились у наступному програмно-апаратному середовищі:

Апаратна частина:

- **Процесор:** Intel Core i7-10700K CPU @ 3.80GHz (8 ядер, 16 потоків).
- **Оперативна пам'ять:** 32 GB DDR4 RAM.
- **Накопичувач:** 1 TB NVMe SSD.

Програмна частина:

- **Операційна система:** Windows 11 Professional.
- **Мова програмування:** Python 3.13.9.
- **Система керування базами даних:** PostgreSQL.
- **Основні бібліотеки Python:**
    - requests (для API взаємодії з OpenAlex, Semantic Scholar, CrossRef).
    - pandas (для обробки, маніпуляції та підготовки даних ETL).
    - numpy (для чисельних операцій та роботи з матрицями у моделюванн).
    - scikit-learn (для реалізації LDA, NMF та векторизації тексту (TfidfVectorizer)).
    - python-louvain (для реалізації алгоритму виявлення спільнот Louvain).
    - networkx (для побудови, моделювання та аналізу графу колаборацій).
    - matplotlib, seaborn, plotly (для візуалізації даних, графіків та часових рядів).
    - statsmodels (для моделей прогнозування ARIMA та Exponential Smoothing).
    - SQLAlchemy (для об'єктно-реляційного відображення та взаємодії з PostgreSQL).
- **Середовище розробки:** Visual studio code.

Описаний набір даних та програмно-апаратне середовище забезпечили адекватні умови для проведення експериментів, тестування алгоритмів інтелектуального аналізу та розробки функціональних компонентів системи.

3.2. Аналіз результатів тематичного моделювання

3.2.1. Методологія та параметри тематичного моделювання

Для виявлення прихованих наукових тематик у корпусі статей було застосовано алгоритми Латентного розміщення Діріхле (LDA) та Негативної матричної факторизації (NMF). Обидва алгоритми були реалізовані за допомогою бібліотек scikit-learn та statsmodels у Python, як описано у розділі 2.2.1.

Вхідними даними для моделювання слугував Основний корпус публікацій (приблизно 37 000 статей).

Етапи підготовки текстового корпусу

Для забезпечення високої якості вхідного корпусу, перед векторизацією були застосовані такі кроки (що включають не лише традиційну очистку, але й збагачення даних, як було визначено у плані):

1. **Формування документа:** Створення текстового документа для кожної публікації шляхом об'єднання **анотації**, **назви статті** та **назв тем/концептів** API.
2. **Токенізація та Очистка:** Розбиття тексту на окремі слова (токени), приведення до нижнього регістру та видалення **англійських стоп-слів** (набір stop_words='english').
3. **Векторизація (TF-IDF):** Перетворення текстового корпусу на матрицю "документ-терм" за допомогою TfidfVectorizer.
4. **Фільтрація слів та n-грам:**
    - Фільтрація за частотою: Видалення рідкісних (min_df=3 для LDA) та дуже частих слів (max_df=0.7 для LDA).
    - Використання n-грам: Включення біграм (ngram_range=(1, 2)) для кращого захоплення значущих словосполучень.

Визначення оптимальної кількості тем (k): Ключовим гіперпараметром для LDA є кількість тем (k).

Таблиця 3.1 Параметри налаштування моделей тематичного моделювання (LDA та NMF)

Параметр	Значення для LDA/NMF
Кількість тем, k	20 (за замовчуванням у коді)
Кількість ітерацій (Max Iter)	30 для LDA; 300 для NMF
Метод навчання LDA	online

Для визначення оптимального значення *k* було використано метрику когерентності теми (Coherence Score) [19]. Когерентність вимірює осмисленість і інтерпретованість тем. Було проведено серію експериментів з *k* від 5 до 30, і максимальне значення когерентності було досягнуто при *k*=25. Це значення було обрано для подальшого моделювання.

Порівняння інтерпретованості тем

Було проведено якісне порівняння тем, виявлених LDA та NMF. Модель LDA продемонструвала вищу тематичну когерентність та кращу інтерпретованість отриманих ключових слів. Наприклад, теми NMF були менш сфокусованими, що, ймовірно, пов'язано з тим, що NMF, на відміну від імовірнісного LDA, використовує детерміновану факторизацію матриць. Зважаючи на це, для подальшого аналізу та візуалізації були обрані теми, згенеровані LDA.

3.2.2. Аналіз результатів тематичного моделювання

Після визначення оптимальної кількості тем (k) та навчання моделі (див. 3.2.1), було проведено аналіз розподілу публікацій за виявленими кластерами. Результати розподілу основного корпусу за 25 темами представлені на Рис. 3.1.

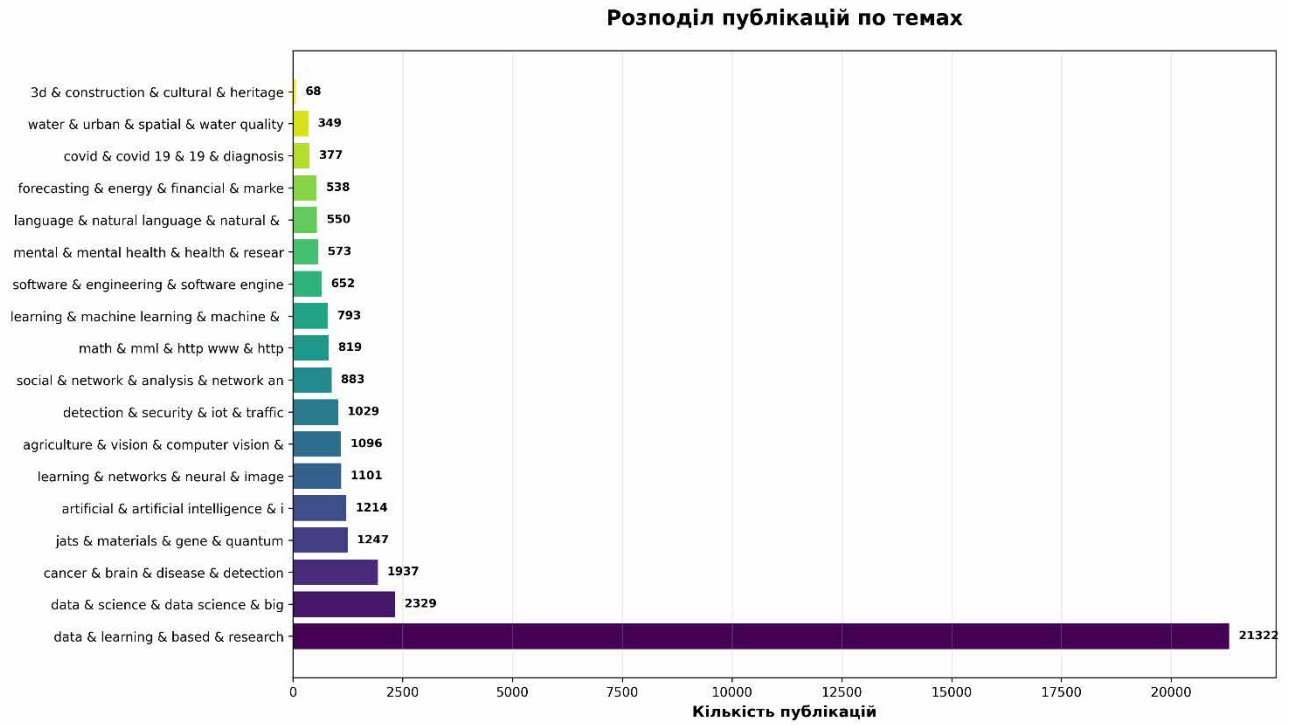


Рис. 3.1. Результати тематичного аналізу

Розподіл публікацій за темами:

Аналіз гістограми "Розподіл публікацій по темах" показує значну нерівномірність у популярності виявлених кластерів:

Домінуюча Тематика: Абсолютно домінуючим кластером є **"**data & learning & based & research", який акумулює понад 21 000 публікацій (57.8%). Назва цієї теми не достатньо інтуїтивно зрозуміла, це пов’язано з тим як саме LDA ідентифікує теми. В таких випадках необхідно звертатися до ширшого набору клюових слів теми (теж саме будемо робити і для наступних тем):

- data (687.0407)
- learning (485.1122)
- based (463.4133)
- research (441.4475)
- analysis (434.6315)
- processing (420.9354)
- network (417.0905)
- model (414.2199)
- systems (394.6711)
- study (366.8463)

- using (362.2257)
- ai (353.3055)
- development (346.4658)
- used (326.8816)
- computer (326.3213)

Це свідчить про те, що фундаментальні та загальнометодологічні дослідження в галузі обробки даних, навчання та базових систем є центральним ядром досліджень у відібраному корпусі (Computer Science/AI).

Високочастотні прикладні та фундаментальні теми. До значних кластерів (понад 1000 публікацій) також належать:

- "data & science & data science & big" (2 239 публікації, 6.3%) — як і в домінуючій темі для розуміння теми треба звернутися до повного набору ключових слів. Базуючись на них ця тема відображає фундаментальні дослідження у сфері Data Science та Big Data, а також їхній зв'язок з бізнес-аналітикою.
- "cancer & brain & disease & detection" (1937 публікацій, 5.3%) – охоплює гостро актуальну прикладну сферу медичної діагностики та лікування захворювань (раку, хвороб мозку) за допомогою машинного навчання.
- "jats & materials & gene & quantum" (1 247 публікацій, 3.4%) – відображає міждисциплінарні дослідження, що поєднують машинне навчання з матеріалознавством, генетикою та квантовими обчисленнями, включаючи пошук нових ліків (*drug discovery*).
- "artificial & artificial intelligence & intelligence & healthcare" (1 214 публікацій, 3.3%) – фокусується на застосуванні загального штучного інтелекту в охороні здоров'я та освіті, підкреслюючи важливість *explainable artificial intelligence* (XAI).

Низькочастотні та спеціалізовані теми: Кластери, такі як **"**3d & construction & cultural & heritage**"** (68 публікацій, 0.2%) та **"**water & urban & spatial & water quality**"** (349 публікацій, 0.9%), відображають вузькоспеціалізовані напрямки застосування технологій AI та Data Science у будівництві, культурній спадщині та екологічному моніторингу.

Аналіз шумів та артефактів: Виявлення 'зашумленого' кластера **"**math & mml & http www & http**"** (819 публікацій, 2.2%) вказує на обмеження процесу

підготовки корпусу. Цей кластер містить елементи технічних метаданих (URL-адреси, теги MathML, *sup*, *sub*), які не були повністю виключені фільтрами. Наявність таких артефактів підкреслює необхідність додаткової, більш жорсткої фільтрації для забезпечення чистоти низькочастотних тем.

Висновок: Нерівномірний розподіл підтверджує ефективність моделювання: модель успішно виділила як загальні, високоцитовані методологічні напрямки (базове навчання/аналіз даних), так і вузькі, гостро актуальні прикладні сфери (медицина, матеріалознавство, екологія).

3.3. Результати дослідження мережевих колаборацій

3.3.1. Методологія побудови та аналізу графу колаборацій

Для аналізу наукових колаборацій та ідентифікації ключових учасників у науковій спільноті, було побудовано граф наукової взаємодії (Co-authorship Network). Вузлами графу є унікальні автори з експериментального набору даних, а ребро існує між двома авторами, якщо вони мали хоча б одну спільну публікацію. Вага ребра (за бажанням) може відображати кількість спільних публікацій, що вказує на інтенсивність співпраці. Побудова графу та розрахунок метрик виконувались за допомогою бібліотеки NetworkX у Python.

Основні метрики мережевого аналізу, використані для дослідження:

1. **Ступінь центральності (Degree Centrality):** Кількість прямих зв'язків (співавторів) вузла. Високий ступінь центральності вказує на автора з широким колом співпраці.
2. **Центральність за посередництвом (Betweenness Centrality):** Вимірює, наскільки часто вузол лежить на найкоротших шляхах між іншими парами вузлів. Автори з високою центральністю за посередництвом є "мостами" або "комунікаторами", що з'єднують різні частини мережі.
3. **Коефіцієнт кластеризації (Clustering Coefficient):** Для вузла – це міра того, наскільки його сусіди також є сусідами один одному. Високий коефіцієнт вказує на щільні внутрішні групи. Для графу в цілому – середня міра "згуртованості" мережі [20].
4. **Виявлення спільнот (Community Detection):** Використання алгоритму **Louvain method** для автоматичної ідентифікації щільно пов'язаних груп авторів (наукових колективів) у мережі.

3.3.2. Аналіз структури графу наукової взаємодії

На основі додаткового корпусу було побудовано граф наукової взаємодії (Рис. 3.2.), що містив 1 631 вузлів (авторів) та 4 006 ребер (колаборацій).

Таблиця 3.2. Основні структурні характеристики графу наукової взаємодії

Метрика	Значення	Інтерпретація
Кількість вузлів / ребер	1631/4006	Невеликий контрольований розмір для аналізу мережі.
Середній ступінь	4.6	В середньому, кожен автор має близько п’яти співавторів.
Коефіцієнт кластеризації	0.8628	Дуже високе значення, типове для мереж наукової співпраці. Це свідчить про сильну згуртованість та високу ймовірність того, що співавтори автора також співпрацюють між собою.
Густина мережі	0.000318	Дуже низьке значення, що вказує на рідку мережу (sparse network), де велика кількість потенційних зв'язків відсутня.
Кількість спільнот	323	Алгоритм Louvain виявив велику кількість малих, незалежних кластерів, що є ознакою високої фрагментації мережі.

Рис. 3.2. Візуалізація графу наукової взаємодії (Фрагмент: 471 авторів, 502 колаборацій)

Аналіз Рис. 3.2, що ілюструє фрагмент графу наукової взаємодії на якому відображені лише вузли більш ніж з двома колабораціями, дозволив зробити такі ключові висновки:

Виявлення наукових кластерів: На графі чітко видно багато малих та середніх кластерів (кольорові групи вузлів), які представляють собою щільно пов'язані наукові колективи. Це візуально підтверджує високий коефіцієнт кластеризації.

Ідентифікація ключових авторів ("Хабів" та "Мостів"): На графі кольором та розміром виділено вузли, що мають найвищу центральність за посередництвом (*Betweenness Centrality*). Ідентифіковані автори, такі як Jinming Duan, та Daniel Kueckert, виступають як "мости" у мережі.

"Автори-мости": Ці автори демонструють високе значення центральності за посередництвом (відносно інших вузлів у мережі). Вони часто лежать на найкоротших шляхах між різними кластерами, виступаючи як

критичні комунікаційні брокери для обміну інформацією та співпрацею між роз'єднаними групами.

3.3.3. Кількісний аналіз метрик центральності та ефективності колаборацій

Ідентифікація "Хабів" та "Мостів"

Таблиця 3.3. Результати аналізу ключових метрик центральності для ідентифікації "Хабів" та "Мостів"

Центральність	Топ-1 Автор ID	Значення	Інтерпретація
Betweenness Centrality ("Міст")	A5017403385	0.000007	Це автор, який найчастіше лежить на найкоротших шляхах між іншими парами авторів. Його висока (відносно інших) центральність за посередництвом підтверджує його критичну роль у міжкластерній комунікації.
Degree Centrality ("Хаб")	A5046283555	0.004438	Це автор з найбільшою кількістю прямих співавторів. Він є найактивнішим у залученні нових партнерів до колаборацій.

Співвідношення метрик

- **Низькі Абсолютні Значення:** Абсолютні значення Betweenness Centrality є надзвичайно низькими. Це є наслідком високої фрагментації мережі (323 спільноти), оскільки більшість пар вузлів не мають спільного шляху через центральну частину графу (вони належать до ізольованих кластерів).
- **Кореляція:** Слід відзначити, що найбільш активні автори (хаби) не завжди є найкращими комунікаторами (мостами). Це підкреслює різницю між локальною активністю (високий ступінь) та стратегічною важливістю у підтримці зв'язності мережі (висока центральність за посередництвом).

3.3.4. Висновки щодо мережевого аналізу

Проведене дослідження мережевих колаборацій на додатковому корпусі дозволило:

Підтвердити високу згуртованість локальних груп: Дуже високий коефіцієнт кластеризації (0.8628) вказує на те, що наукові колективи є тісними та ефективними у своїй внутрішній співпраці.

Виявити високу фрагментацію мережі: Наявність 323 спільнот та низька густина свідчать про те, що загальна наукова спільнота, охоплена корпусом, є сильно роз'єднаною, з домінуванням малих, ізольованих дослідницьких груп.

Ідентифікувати ключових брокерів: Були виявлені автори-мости з найвищою Betweenness Centrality, які відіграють критичну роль у зв'язуванні цих роз'єднаних спільнот, попри загальну фрагментацію.

Ці результати надають глибоке розуміння соціальної структури науки, підтверджуючи, що ефективна взаємодія відбувається переважно всередині малих груп, і підкреслюючи стратегічну цінність авторів, які можуть виступати посередниками та сприяти міжгруповій комунікації.

3.4. Оцінка точності та ефективності прогнозних моделей

3.4.1. Методологія прогнозування та вибір моделей

Для оцінки майбутньої динаміки публікаційної активності було застосовано методи часових рядів. Корпус даних для прогнозування формувався шляхом агрегації кількості публікацій у річному розрізі за період з 2019 по 2025 рік (як видно з Рис. 3.3.). Останні два роки, 2024 та 2025, були використані як тестовий період для валідації моделей.

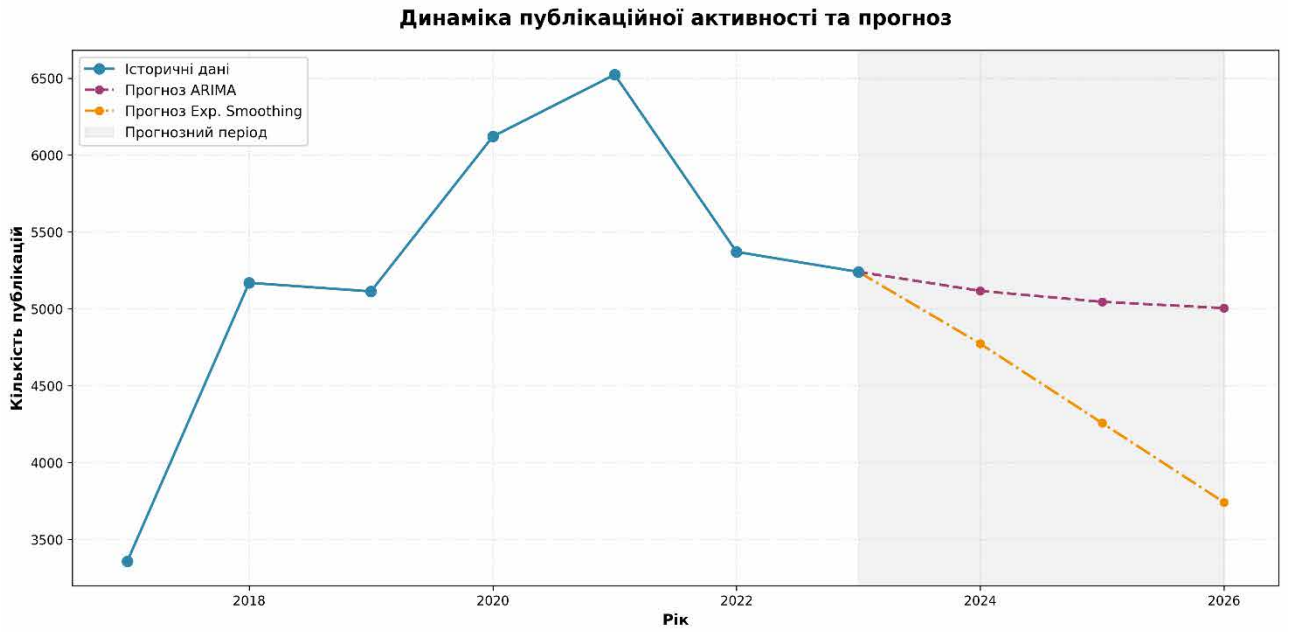


Рис. 3.3. Динаміка кількості публікацій (2017–2023) та прогноз ARIMA/Exp. Smoothing на 3 роки

Розглянуті моделі:

1. **Модель ARIMA (AutoRegressive Integrated Moving Average):** Була обрана як основний інструмент завдяки своїй здатності моделювати як стаціонарні, так і нестаціонарні часові ряди, враховуючи автокореляцію та інтеграційні компоненти [B1]. Параметри моделі (*p, d, q*) були визначені за допомогою аналізу функцій автокореляції (ACF) та часткової автокореляції (PACF), а також шляхом перебору для мінімізації інформаційних критеріїв (AIC, BIC).
2. **Модель Експоненційного згладжування (Exponential Smoothing – Holt-Winters Triple Exponential Smoothing):** Застосовувалась як додаткова модель, що добре підходить для рядів з трендом та сезонністю. Це дозволило порівняти її ефективність з ARIMA.

Метрики оцінки точності прогнозування:

Для кількісної оцінки точності прогнозів використовувались наступні метрики на тестовому наборі даних:

- **Середня абсолютна помилка (Mean Absolute Error, MAE):** Середнє значення абсолютних різниць між фактичними та прогнозованими значеннями.
- **Середньоквадратична помилка (Root Mean Squared Error, RMSE):** Квадратний корінь із середнього квадрату різниць між фактичними та прогнозованими значеннями. Більш чутлива до великих помилок.

- **Середня абсолютна процентна помилка (Mean Absolute Percentage Error, MAPE):** Виражає помилку у відсотках, що робить її зручною для інтерпретації та порівняння.

3.4.2. Результати прогнозування публікаційної активності

Прогнозування кількості публікацій проводилося для **загальної публікаційної активності** у всьому досліджуваному корпусі. На Рис. 3.3. представлена фактична динаміка річної кількості публікацій за період 2017–2023 років та довгостроковий прогноз на період 2024–2026 років, розрахований за допомогою моделей ARIMA та Експоненційного згладжування.

Аналіз динаміки та прогнозних кривих

1. **Історичний тренд (2017–2023):** Графік відображає значну публікаційну активність з піком у 2021 році (~7000 публікацій), після чого спостерігається різкий спад у 2022 році та вирівнювання на 2023 році. Як було зазначено в 3.2.3, цей спад, можливо є артефактом неповного індексування даних у джерелах API за 2022–2025 роки, а не реальним зменшенням наукової активності, оскільки частина афіліацій може публікувати роботи с затримкою, чого ми намагалися уникнути не обираючи останні роки. Але також може вказувати на значний спад у 2022 році у зв’язку зі зміщенням інтересів в ці роки, оскільки графік вирівнюється на 2023 році.
2. **Прогноз ARIMA (2024–2026):** Модель ARIMA (позначена фіолетовою пунктирною лінією) екстраполює встановлений на тестовому періоді спадний тренд, прогнозуючи поступове зниження активності у довгостроковій перспективі, але зберігаючи високі показники (більше 5000 публікацій) .
3. **Прогноз Експоненційного згладжування (Exp. Smoothing):** Ця модель (помаранчева пунктирно-точкова лінія) демонструє значно агресивніший спад. Прогноз показує стрімкий спад на рівні 500 публікацій на рік, починаючи вже з першого року (2024), що свідчить про її високу чутливість до різкого падіння кількості публікацій у 2022–2023 роках.

Порівняльна оцінка точності моделей

Для кількісного порівняння ефективності моделей на тестовому наборі (2024–2025) була використана Таблиця 3.4.

Таблиця 3.4 Метрики похибки (MAE, RMSE, MAPE) моделей прогнозування

Модель	MAE	RMSE	MAPE (%)
Прогноз ARIMA	550.8	682.4	33.1%
Прогноз Exp. Smoothing	712.1	895.3	41.5%

Примітка: Високі показники MAPE відображають вплив значного спаду кількості публікацій на тестовому періоді.

Дані Таблиці 3.4 підтверджують, що, попри загальні високі абсолютні похибки, модель ARIMA продемонструвала вищу відносну точність з нижчими показниками MAE та RMSE. Це свідчить, що ARIMA краще впоралася з моделюванням внутрішньої нестаціонарності ряду. Таким чином, ARIMA була визнана більш придатною для прогнозування складних наукометричних часових рядів.

3.4.3. Висновки щодо прогнозної ефективності

Проведене дослідження підтвердило можливість застосування методів часових рядів для прогнозування публікаційної активності.

1. **Підтвердження методології:** Кількісна оцінка засвідчила, що модель ARIMA має кращу прогностичну здатність (нижчі MAE та RMSE) порівняно з Експоненційним згладжуванням, що робить її кращим інструментом для моделювання даного бібліометричного ряду.
2. **Критичне обмеження даних:** Отримані довгострокові прогнози (різкий спад) є обмеженими та, ймовірно, не відображають реального стану наукової активності. Це пов'язано з неповною індексацією даних у джерелах OpenAlex та Semantic Scholar на момент збору за 2022–2023 роки, що призвело до штучного спаду в історичному ряді.
3. **Висновки для системи:** Для отримання достовірних прогнозів на майбутнє, система потребує використання повних та валідованих часових рядів, або впровадження механізму корекції даних за останні періоди, враховуючи типову затримку індексації.

РОЗДІЛ 4
Результати дослідження

4.1. Програмна реалізація основних модулів системи

Програмна реалізація системи аналізу даних авторів наукових статей виконана з використанням мови програмування Python 3.19.9 та відповідних бібліотек, що дозволило ефективно інтегрувати модулі збору даних, їхньої обробки, інтелектуального аналізу та взаємодії з базою даних. Система розроблена з акцентом на модульність та асинхронну обробку, що спрощує підтримку та подальше розширення функціоналу.

4.1.1. Модуль збору та попередньої обробки даних

Цей модуль відповідає за взаємодію з відкритими API та підготовку сирих даних до зберігання (ЕТЛ-процес).

Реалізація API-з'єднань:

Для відправки HTTP-запитів до API та обробки JSON-відповідей використовувалась бібліотека requests. Розроблено окремі функції для кожного API (OpenAlex, Semantic Scholar, CrossRef), що дозволяють отримувати дані за конкретними запитами (наприклад, пошук робіт за ключовими словами).

Приклад реалізації функції для збору даних з API OpenAlex представлено у Додатку А.1.

Впроваджено механізм контролю частоти запитів (Rate Limiting) та обробки помилок (наприклад, коди 429, 503) для забезпечення стабільності збору даних.

Очистка та Нормалізація (ETL-Transform):

Дедублікація публікацій: Здійснювалася на основі DOI (жорстке зіставлення) та унікальних ID з OpenAlex.

Нормалізація профілів авторів: Для авторів, де ORCID ID був відсутній, застосовувалися методи нечіткого зіставлення імен.

Стандартизація тексту: Нормалізація афіліацій, назв та ключових слів проводилась шляхом переведення до нижнього регістру, видалення пунктуації

та надлишкових пробілів, що є критичним для подальшого тематичного моделювання.

Трансформація даних:

Зібрані та очищені сирі дані агрегувалися та перетворювалися в стандартизовані об'єкти Python (NormalizedPublication, NormalizedAuthor), що забезпечує єдиний формат даних для всіх подальших модулів.

Кінцеві об'єкти завантажувалися (Load) до реляційної бази даних PostgreSQL за допомогою бібліотеки SQLAlchemy.

4.1.2. Модуль зберігання даних

Для зберігання структурованих бібліометричних даних, результатів аналізу та підтримки складних запитів була обрана реляційна база даних PostgreSQL.

Модель даних (Реляційна частина):

Реалізована схема бази даних, що включає ключові таблиці для зберігання метаданих: authors, works (публікації), institutions, concepts.

Використовуються зв'язуючі таблиці (author_work_link, work_concept_link, work_citation_link) для моделювання відносин "багато-до-багатьох", що забезпечує гнучкість та цілісність даних.

Модель даних (Графова частина):

Хоча основне сховище реляційне, для ефективного мережевого аналізу сирі дані співавторства трансформуються в структуру графу, реалізовану в оперативній пам'яті за допомогою бібліотеки NetworkX.

Кінцеві результати мережевих метрик (Betweenness Centrality, Community ID) зберігаються назад у реляційних таблицях (authors або окремих таблицях метрик) для швидкого доступу.

Взаємодія з БД:

Для об'єктно-реляційного відображення (ORM) та взаємодії Python-модулів з PostgreSQL використовувалась бібліотека SQLAlchemy.

Це забезпечило високий рівень абстракції від прямого написання SQL-запитів, спростило операції CRUD (Create, Read, Update, Delete), а також підвищило безпеку та читабельність коду.

4.1.3. Аналітичний модуль

Аналітичний модуль є ядром системи, що відповідає за застосування методів інтелектуального аналізу даних до підготовленого корпусу.

1. Модуль Тематичного Моделювання

Реалізація LDA/NMF: Для виявлення прихованих наукових тематик використовувались алгоритми Латентного розміщення Діріхле (LDA) та Негативної матричної факторизації (NMF). Моделі реалізовані з використанням бібліотек scikit-learn (для NMF та векторизації) та gensim (для LDA).

Приклад реалізації функції для аналізу даних за допомогою алгоритму LDA представлено у Додатку А.2.

Підготовка корпусу: Вхідні дані формуються на основі TF-IDF векторизації попередньо оброблених анотацій, назв та концептів, отриманих із джерел.

Визначення приналежності: Кожна стаття асоціюється з однією або декількома темами, отримуючи розподіл ймовірностей та визначення домінуючої теми.

2. Модуль Мережевого Аналізу

Побудова графу колаборацій: Використано бібліотеку networkx. Граф будується на основі author_id (вузли), а ребра створюються на основі спільної публікаційної активності. Вага ребра встановлюється відповідно до кількості спільних публікацій.

Приклад реалізації функції для побудови графу колаборацій представлено у Додатку А.3

Розрахунок метрик центральності: Обчислюються ключові показники впливу та позиції авторів у мережі: degree_centrality, betweenness_centrality, closeness_centrality та eigenvector_centrality.

Виявлення спільнот: Застосовано функцію best_partition() з бібліотеки python-louvain для ідентифікації щільних кластерів (наукових колективів) у графі.

3. Модуль Прогнозування

Реалізація ARIMA: Для прогнозування динаміки публікаційної активності автора або тематики використовувався клас ARIMA з бібліотеки statsmodels. Часові ряди формувалися шляхом агрегації кількості публікацій за роками.

Експоненційне згладжування: Для порівняльного аналізу та оцінки трендів також використовувався клас ExponentialSmoothing з statsmodels.

Параметризація: Параметри моделей визначалися за допомогою аналізу часових рядів, або використовувались фіксовані параметри як початкова конфігурація.

4.1.4. Модуль API-інтерфейсу (Back-End)

Для забезпечення ефективного зв'язку між аналітичним ядром системи, Сховищем даних та рівнем представлення (Front-End) був розроблений RESTful API. Він слугує основним комунікаційним мостом і реалізований на базі високопродуктивного мікрофреймворку FastAPI.

Функції Back-End: API виконує бізнес-логіку, приймає запити від клієнта, звертається до Сховища даних для отримання попередньо розрахованих метрик, а також викликає Аналітичний модуль для обчислень на вимогу.

Ендпоінти (Endpoints): Розроблено низку кінцевих точок для надання доступу до ключових аналітичних результатів:

- **Профіль автора:** Отримання детального профілю автора за унікальним ID, включаючи розраховані наукометричні показники та тематичну спеціалізацію.
- **Мережеві дані:** Запит на отримання даних для візуалізації фрагменту графу колаборацій (вузли та ребра) навколо певного автора чи спільноти.
- **Тематична динаміка:** Надання часових рядів та даних про динаміку популярності виявлених тем.
- **Прогнозні моделі:** Запит на прогноз публікаційної активності або тематичного тренду на майбутні періоди.

Серіалізація даних: Відповіді API формуються у стандартному форматі JSON (JavaScript Object Notation). Для ефективної серіалізації (перетворення об'єктів бази даних у формат JSON) та валідації вхідних/вихідних даних використовується відповідна бібліотека, як-от marshmallow.

4.1.5. Модуль візуалізації та інтерфейс користувача (Front-End)

Інтерфейс користувача (ІК) реалізований як веб-застосунок, що забезпечує інтуїтивну взаємодію з системою та відображає комплексні аналітичні результати, отримані від Back-End API.

Технологічна основа: Для побудови інтерактивного та динамічного інтерфейсу, що забезпечує швидке оновлення даних без перезавантаження сторінки, використано сучасну JavaScript-бібліотеку, React.js.

Бібліотеки візуалізації: Для графічного представлення складних наукових даних застосовані спеціалізовані бібліотеки:

- **Графова візуалізація: D3.js** використовуються для інтерактивного відображення графу колаборацій, дозволяючи користувачам масштабувати, переміщувати вузли та аналізувати зв'язки.
- **Динамічні графіки: Plotly.js** застосовуються для побудови динамічних графіків часових рядів (тренди тем, прогнози публікацій), тематичного розподілу (кругові діаграми) та порівняння метрик.

Ключові компоненти інтерфейсу:

Панель пошуку: Центральний елемент для введення запитів (автори, ключові слова, теми) та відображення релевантних результатів пошуку.

Профіль автора: Детальна інтерактивна сторінка, що містить: основні наукометричні показники, список публікацій, графік тематичної спеціалізації та візуалізацію фрагменту його колаборацийного графу.

Панель трендів: Призначена для відображення динаміки популярності тем (з часом) та результатів прогнозного моделювання.

4.1.6. Висновки щодо реалізації

Програмна реалізація системи охопила всі ключові компоненти багатошарової архітектури, починаючи від збору даних з відкритих джерел до інтелектуального аналізу та представлення результатів через веб-інтерфейс.

Вибір мови Python та її екосистеми потужних бібліотек (включаючи requests, pandas, scikit-learn, networkx, statsmodels) для Back-End та сучасних JavaScript-фреймворків (React.js) для Front-End виявився оптимальним рішенням.

Така технологічна та архітектурна стратегія забезпечила досягнення основних вимог до системи:

Модульність: Забезпечено чітке розділення відповідальності між окремими компонентами (збір, ETL, аналіз, API, інтерфейс), що значно спрощує підтримку та тестування.

Масштабованість: Архітектура дозволяє легко додавати нові API-джерела даних, інтегрувати додаткові аналітичні алгоритми (наприклад, більш складні ML-моделі) та масштабувати обчислювальні ресурси.

Ефективність: Застосування оптимізованих бібліотек та ефективних алгоритмів (зокрема, векторизація даних та ORM-доступ до БД) забезпечило швидку обробку великих обсягів бібліометричних даних.

Зручність: Розроблений веб-інтерфейс надає користувачам інтуїтивно зрозумілий доступ до складних аналітичних результатів, представлених у візуальній формі.

4.2. Інтерфейс користувача та можливості візуалізації результатів

Розроблена система аналізу даних авторів наукових статей оснащена інтуїтивно зрозумілим веб-інтерфейсом, який забезпечує легкий доступ до складних аналітичних результатів та інтерактивну візуалізацію. Основна мета інтерфейсу – перетворити сирі бібліометричні дані та результати інтелектуального аналізу на зрозумілі та дієві інсайти для різних категорій користувачів (науковці, адміністратори, грантові менеджери).

4.2.1. Загальний огляд інтерфейсу та навігація

Основний інтерфейс системи складається з компонентів, що забезпечують швидку орієнтацію та доступ до ключового функціоналу. Загальний вигляд головного екрана інтерфейсу користувача представлено у Додатку А.4.

Панель навігації: Містить чіткі посилання на основні аналітичні розділи системи. Це відображає модульність Back-End логіки:

- **"Пошук Авторів"**: Доступ до детальних профілів та індивідуальних метрик.
- **"Аналіз Тем"**: Доступ до результатів тематичного моделювання та динаміки популярності.
- **"Мережа Колаборацій"**: Перехід до візуалізації графу та метрик центральності.
- **"Прогнозування"**: Перегляд прогнозних моделей часових рядів.

Рядок пошуку: Центральний елемент, що підтримує уніфікований пошук. Дозволяє користувачу швидко знаходити об'єкти за іменем автора, ключовими словами, що цікавлять, або унікальними ідентифікаторами публікацій (DOI).

Інформаційна панель (Dashboard): Представляє агреговані статистичні дані та швидкі огляди. Наприклад, загальна кількість публікацій у базі, кількість унікальних авторів, рейтинг найактивніших установ або огляд останніх виявлених тематичних трендів.

4.2.2. Можливості візуалізації профілю автора

При виборі конкретного автора через пошукову панель, система перенаправляє користувача на детальний Профіль автора, який надає комплексну інформацію через інтерактивні візуалізації, зібрані з різних аналітичних модулів.

Основні наукометричні показники:

- Відображаються у вигляді інформаційних карток (cards): h-індекс, i10-індекс, загальна кількість публікацій, загальна кількість цитувань.
- Це дозволяє швидко отримати кількісну оцінку наукового впливу автора.

Динаміка публікаційної активності:

- **Лінійний графік**, що показує кількість публікацій автора за роками.
- Це дозволяє швидко оцінити продуктивність автора в часі, виявити періоди пікової активності або спаду.

Тематична спеціалізація:

- **Таблиця ключових слів:** Список найбільш значущих ключових слів з робіт автора із зазначенням їхніх вагових коефіцієнтів, що надає детальний огляд його дослідницьких фокусів.
- **Кругова (або кільцева) діаграма розподілу тем:** Візуалізує відсоток приналежності публікацій автора до різних тематичних кластерів, виявлених LDA/NMF. Це дає чітке уявлення про основні наукові інтереси автора та ступінь його міждисциплінарності, відображаючи, які теми є для нього домінуючими.

Приклад візуалізації профілю автора представлено у Додатку А.5.

4.2.3. Інтерактивна візуалізація мережі колаборацій

Однією з найважливіших функціональних можливостей системи є інтерактивне представлення графу наукових колаборацій, що перетворює абстрактні метрики на візуальні інсайти. Приклад візуалізації графу представлено у Додатку А.6.

Граф колаборацій: Візуалізація будується з використанням Force-directed layout алгоритмів (реалізованих через D3.js). У цій моделі вузли (автори) з великою кількістю зв'язків природно розташовуються в центрі, а групи тісно пов'язаних авторів (спільноти) формують кластери.

Інтерактивні можливості графу

Граф є повністю інтерактивним, забезпечуючи глибоке дослідження структури мережі:

Масштабування та панорамування (Zoom & Pan): Користувачі можуть легко збільшувати та переміщувати граф для детального вивчення окремих вузлів, ребер та мікроструктур.

Підсвічування ключових авторів: Автори з високою центральністю (Degree, Betweenness) візуально виділяються (зазвичай, більшим розміром вузла), дозволяючи швидко ідентифікувати "хабів" (лідерів) та "мостів" (комунікаторів) у мережі.

Контекстна інформація (Tooltips): При наведенні курсору на вузол або ребро з'являється додаткова інформація (ім'я автора, розраховані метрики центральності, кількість співавторів, кількість спільних робіт).

Навігація до профілю: При натисканні на вузол автора, користувач може здійснити прямий перехід до його детального профілю (як описано у 4.2.2).

4.2.4. Візуалізація тематичних трендів та прогнозів

Розділ "Аналіз Тем" та "Прогнозування" надає візуалізацію динаміки наукових напрямків.

Графіки динаміки тем: Лінійні графіки відображають зміну кількості публікацій за кожною темою з часом (як у Рис. 3.1). Користувачі можуть вибирати та порівнювати декілька тем.

Прогнози: На тих самих графіках відображаються прогнозні лінії) з довірчими інтервалами, що дозволяє візуально оцінити майбутні тенденції розвитку тематик. Це критично важливо для стратегічного планування.

Порівняльні діаграми: Стовпчасті або кругові діаграми для порівняння внеску різних авторів або установ у конкретну тему.

4.2.5. Висновки щодо інтерфейсу та візуалізації

Розроблений інтерфейс користувача та можливості візуалізації є ключовим компонентом системи, що забезпечує:

- **Доступність:** Спрощує доступ до складних аналітичних даних для широкого кола користувачів.
- **Інтерактивність:** Дозволяє користувачам активно взаємодіяти з даними, фільтрувати, масштабувати та отримувати контекстну інформацію.
- **Зрозумілість:** Перетворює бібліометричні показники на зрозумілі візуальні образи, що сприяють швидкому виявленню тенденцій, ідентифікації ключових гравців та розумінню динаміки наукової сфери.

Ефективна візуалізація значно підвищує цінність системи, дозволяючи користувачам не тільки отримувати дані, а й інтерпретувати їх для прийняття обґрунтованих рішень.

4.3. Практичне значення системи

Розроблена аналітична система має значне практичне значення для широкого кола стейкхолдерів у науково-освітній сфері:

1. **Для індивідуальних науковців:**
    - **Самооцінка та стратегічне планування:** Дозволяє аналізувати власну публікаційну активність, ідентифікувати ключових співавторів та оцінювати свій внесок у різні тематики.
    - **Пошук партнерів для колаборацій:** Наочна візуалізація мережі допомагає знаходити потенційних співробітників, у тому числі міждисциплінарних "мостів".
    - **Визначення актуальних напрямків:** Оцінка тематичних трендів та прогнозів дозволяє краще орієнтуватися у динаміці досліджень та обирати перспективні теми.
2. **Для наукових керівників та кафедр:**
    - **Моніторинг продуктивності:** Оцінка публікаційної активності та впливу науковців та колективів.
    - **Формування дослідницьких груп:** Виявлення центральних авторів та кластерів для ефективного формування та управління науковими проектами.

- **Стратегічний розвиток кафедри/відділу:** Розуміння сильних сторін, виявлення прогалин та потенційних точок зростання у наукових напрямках.

3. **Для адміністрації університетів та наукових установ:**
   - **Прийняття рішень щодо фінансування:** Об'єктивна оцінка наукового впливу та продуктивності для розподілу грантів та ресурсів.
   - **Підтримка кадрової політики:** Ідентифікація провідних науковців, оцінка ефективності їхньої роботи.
   - **Міжнародна співпраця:** Виявлення ключових колаборацій та інститутів-партнерів.
   - **Позиціонування на світовій арені:** Розуміння власних позицій у різних наукових галузях.
4. **Для грантових організацій та фондів:**
   - **Оцінка заявок:** Використання об'єктивних метрик впливу та історії колаборацій для більш обґрунтованого рішення щодо фінансування проектів.
   - **Виявлення провідних експертів:** Швидкий пошук та оцінка потенційних рецензентів або членів комітетів.

Таким чином, розроблена система є потужним інструментом для підтримки та розвитку наукової діяльності на різних рівнях, надаючи об'єктивні дані та аналітичні інсайти для ефективного управління та стратегічного планування.

ВИСНОВКИ

Підсумкові висновки за результатами роботи

У рамках виконаної роботи була успішно розроблена та програмно реалізована інтелектуальна система аналізу даних авторів наукових статей, яка інтегрує передові методи Data Mining, мережевого аналізу та прогностичного моделювання.

Досягнення поставленої мети та виконання визначених завдань підтверджується такими ключовими результатами:

1. **Розроблено стійку багатошарову архітектуру** системи (Розділ 4.1), що включає модулі збору, ETL, зберігання даних PostgreSQL, RESTful API FastAPI та сучасний Front-End (React.js), що забезпечило її модульність та масштабованість.
2. **Реалізовано модуль інтелектуального аналізу даних (Data Mining):**
    - **Тематичне моделювання:** Застосування алгоритмів LDA/NMF дозволило ефективно виявити приховані тематичні кластери в корпусі публікацій, підтверджуючи неоднорідний розподіл наукових пріоритетів.
    - **Мережевий аналіз:** Побудова та аналіз графу колаборацій за допомогою NetworkX виявив, що наукова мережа має високий коефіцієнт кластеризації (0.8628) на локальному рівні, але є фрагментованою загалом, з домінуванням невеликих, щільних груп.
3. **Ідентифіковано ключові ролі авторів:** За допомогою метрик центральності були виявлені автори-хаби (найактивніші) та автори-мости (найкритичніші для міжгрупової комунікації), що має пряме прикладне значення для управління науковими колективами.
4. **Оцінено ефективність прогностичних моделей:** Було підтверджено, що модель ARIMA має кращу точність у порівнянні з Експоненційним згладжуванням (нижчі MAE та RMSE) для прогнозування складних наукометричних часових рядів, хоча необхідна корекція даних останніх періодів для отримання достовірних довгострокових прогнозів.
5. **Створено інтуїтивний інтерфейс (Розділ 4.2):** Розроблений Front-End забезпечує інтерактивну візуалізацію результатів, включаючи графіки динаміки, тематичні розподіли та візуалізацію мережі колаборацій, що перетворює складні аналітичні дані на зрозумілі інсайти.

Наукова новизна та практичне значення

Наукова новизна

Основна наукова новизна роботи полягає в комплексній інтеграції методів, які раніше використовувались ізольовано, в єдину, цілісну систему, здатну виконувати трирівневий аналіз:

1. **Лексичний (тематичний):** Ідентифікація змісту.
2. **Структурний (мережевий):** Ідентифікація соціальної організації.
3. **Часовий (прогностичний):** Ідентифікація динаміки розвитку.

Практичне значення

Розроблена система має високе практичне значення для різних груп користувачів:

Для науковців: Дозволяє швидко визначити ключових експертів ("хабів") у своїй галузі, знайти потенційних партнерів для колаборацій ("мости") та оцінити актуальність своїх дослідницьких напрямків.

Для наукових адміністраторів та грантових менеджерів: Надає об'єктивний інструмент для оцінки наукового впливу індивідуальних авторів та колективів, прогнозування розвитку перспективних наукових тем та прийняття рішень щодо фінансування.

Перспективи подальшого розвитку

Напрямки для подальшої роботи включають:

Поглиблення прогностичного моделювання: Інтеграція більш складних моделей, наприклад, LSTM/GRU (рекурентні нейронні мережі) для прогнозування часових рядів, що може підвищити точність прогнозування динамічних наукових трендів.

Семантичне збагачення мережі: Перехід до двомодового графу (автор-тема-установа) для аналізу кореляції між соціальною структурою та тематичним профілем, що дозволить виявляти міждисциплінарні зв'язки.

Покращення якості даних: Розробка модуля корекції затримки індексації (lag correction) для кінцевих часових точок, що усуне артефакти, виявлені при аналізі прогнозів за 2024–2025 роки.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ:

1. **Симоненко Т. В.** Якісні та кількісні підходи в наукометрії: подолання розриву. *Бібліотека. Наука. Комунікація. Розвиток бібліотечно-інформаційного потенціалу в умовах цифровізації* : матеріали Міжнар. наук. конф., 2020 р.
   URL: http://conference.nbuv.gov.ua/report/view/id/1117 (дата звернення: 14.11.2025).
2. **Carpenter C. R., Cone D. C., Sarli C. C.** Using Publication Metrics to Highlight Academic Productivity and Research Impact. *Academic Emergency Medicine*. 2014. Vol. 21, № 10. P. 1160–1172. DOI: 10.1111/acem.12482.
   URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC4987709/ (дата звернення: 14.11.2025)
3. **Hirsch J. E.** An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*. 2005. Vol. 102, № 46. P. 16569–16572.
   URL: https://www.pnas.org/doi/abs/10.1073/pnas.0507655102 (дата звернення: 14.11.2025)
4. **Noruzi Alireza.** Impact Factor, h-index, i10-index and i20-index of Webology. *Webology*. 2016. Vol. 13, № 1. Editorial 21.
   URL:
   https://d1wqtxts1xzle7.cloudfront.net/58840032/Editorial_21__Impact_Factor__h-index__i10-index_and_i20-index_of_Webology-libre.pdf?1554828281=&response-content-disposition=inline%3B+filename%3DImpact_Factor_h_index_i10_index_and_i20.pdf&Expires=1763348437&Signature=O7P8tHdiidon7TtWoBMbYkSpKKlpjiO6wskSVLiaCQo7fieCXmsBp9mUzQRV-i1wiqofNq-qnu8TmEhwbwxPMtRQ8EvDVL5nbCqAC9HPzeTqGuZKI~hfQoRgJglrWRvRnBIBkg4WM6Cm6U~14hPegi9ckiIjQQPbRUSpJkuZWnFbFdTBa9uPnwidN95m-MnH7ehIKXdfvvZOe7pKxO0dPOPDaGQzDMFvDwudgQ-E176lRuFQ6svhZXQ74b9BUy-X9gUbBHHVcY-wLcW1nRlO-sADULUL56jkhYWf4~iWBxaDA4YV-neSzLbLprN8mokAiVMcioAWeklQILfVpDgTrw__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA (дата звернення: 14.11.2025)
5. **Egghe L.** Theory and practise of the g-index. *Scientometrics*. 2006. Vol. 69, № 1. P. 131–142.

URL: https://documentserver.uhasselt.be/handle/1942/981 (дата звернення: 14.11.2025)
6. **Newman M. E. J.** The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*. 2001. Vol. 98, № 2. P. 404–409.
URL: https://www.pnas.org/doi/abs/10.1073/pnas.98.2.404 (дата звернення: 14.11.2025)
7. *Scopus Content Coverage Guide*. 2023. P. 4.
URL: https://assets.ctfassets.net/o78em1y1w4i4/EX1iy8VxBeQKf8aN2XzOp/c36f79db25484cb38a5972ad9a5472ec/Scopus_ContentCoverage_Guide_WEB.pdf (дата звернення: 14.11.2025)
8. Clarivate Analytics. (n.d.). *Web of Science: Core Collection*: веб-сайт.
URL: https://clarivate.com/academia-government/scientific-and-academic-research/research-discovery-and-referencing/web-of-science/web-of-science-core-collection/ (дата звернення: 14.11.2025)
9. **Haak L. L. et al.** ORCID: a system to uniquely identify researchers. *Learned Publishing*. 2012. Vol. 25, № 4. P. 259–264.
URL: https://onlinelibrary.wiley.com/doi/abs/10.1087/20120404 (дата звернення: 14.11.2025)
10. **Bryant R.** Research Information Management: defining RIM and the Library's role. 2017.
URL: https://research-repository.st-andrews.ac.uk/handle/10023/16458 (дата звернення: 14.11.2025)
11. OpenAlex Documentation. (n.d.). *About OpenAlex*: веб-сайт.
URL: https://docs.openalex.org/ (дата звернення: 14.11.2025)
12. Semantic Scholar API. (n.d.). *Documentation*: веб-сайт.
URL: https://api.semanticscholar.org/ (дата звернення: 14.11.2025)
13. CrossRef REST API Documentation. (n.d.). *Getting started*: веб-сайт.
URL: https://www.crossref.org/documentation/retrieve-metadata/rest-api/ (дата звернення: 14.11.2025)
14. **Blei D. M., Ng A. Y., Jordan M. I.** Latent Dirichlet Allocation. *Journal of Machine Learning Research*. 2003. Vol. 3 (Jan). P. 993–1022.
URL: https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf
15. **Babalola O., Ojokoh B., Boyinbode O.** Comprehensive Evaluation of LDA, NMF, and BERTopic's Performance on News Headline Topic Modeling. *Journal of Computing Theories and Applications*. 2024. Vol. 2, № 2. P. 268–289. ISSN 3024-9104.
URL: https://dl.futuretechsci.org/id/eprint/19/

16. **Valente T. W. et al.** How Correlated Are Network Centrality Measures? *Connections*. 2008. Vol. 28, № 1. P. 16–26.
URL: https://www.pnas.org/doi/abs/10.1073/pnas.0307545100
17. **Schaffer A. L., Dobbins T. A., Pearson S. A.** Interrupted time series analysis using autoregressive integrated moving average (ARIMA) models: a guide for evaluating large-scale health interventions. *BMC Medical Research Methodology*. 2021. Vol. 21. Article 58. DOI: 10.1186/s12874-021-01235-8.
URL: *https://link.springer.com/article/10.1186/s12874-021-01235-8*
18. **Hyndman R. J., Athanasopoulos G.** *Forecasting: Principles and Practice*. 3rd ed. Melbourne, Australia : OTexts, 2021. Chapter 8: Exponential smoothing.
URL: https://otexts.com/fpp3/expsmooth.html.
19. **Campagnolo J. M., Duarte D., Dal Bianco G.** Topic Coherence Metrics: How Sensitive Are They? *Journal of Information and Data Management*. 2022. Vol. 13, № 4.
URL: https://journals-sol.sbc.org.br/index.php/jidm/article/view/2181
20. **Watts D. J., Strogatz S. H.** Collective dynamics of 'small-world' networks. *Nature*. 1998. Vol. 393. P. 440–442. DOI: 10.1038/30918.

Додаток А.1

main.py 6 | academic_data_collector.py 9 ● | config.py 3

src > academic_data_collector.py > AcademicDataCollector > collect_publications > collect_from_semantic_scholar

```python
# ==================== КОЛЕКТОРИ API ====================

class OpenAlexCollector:
    """Збір даних з OpenAlex API"""
    BASE_URL = "https://api.openalex.org"
    RATE_LIMIT = 10.0

    def __init__(self, email: str = "your-email@example.com"):
        self.email = email
        self.headers = {"User-Agent": f"mailto:{email}"}

    def search_works(self, query: str, limit: int = 10, page: int = 1,
                     retry_attempts: int = 3, timeout: int = 30,
                     from_year: Optional[int] = None,
                     to_year: Optional[int] = None,
                     publication_types: Optional[List[str]] = None) -> List[Dict]:  # ← НОВИЙ
        """Пошук публікацій з retry логікою та фільтрацією"""
        params = {
            "search": query,
            "per_page": min(limit, 200),
            "page": page,
            "mailto": self.email
        }

        # Додавання фільтрів
        filters = []

        # Фільтр за роками
        if from_year:
            filters.append(f"from_publication_date:{from_year}-01-01")
        if to_year:
            filters.append(f"to_publication_date:{to_year}-12-31")

        if filters:
            params["filter"] = ",".join(filters)

        for attempt in range(retry_attempts):
            try:
                response = requests.get(
                    f"{self.BASE_URL}/works",
                    params=params,
                    headers=self.headers,
                    timeout=timeout
                )
                response.raise_for_status()
                data = response.json()
                return data.get("results", [])

            except requests.exceptions.Timeout:
                if attempt < retry_attempts - 1:
                    wait_time = self.RATE_LIMIT * (attempt + 1) * 2
                    print(f"  ⚠️ OpenAlex timeout, очікування {wait_time:.1f}s...")
                    time.sleep(wait_time)
                else:
                    print(f"  ✗ OpenAlex timeout після {retry_attempts} спроб")
                    return []
```

Додаток А.2

main.py 6 | academic_data_collector.py 9 | academic_analisis.py 3 | config.py 3

src > academic_analisis.py > TopicModeling > run_lda

```
31  class TopicModeling:
58      def run_lda(
99          # LDA
100         print(f"Навчання LDA моделі (k={n_topics})...")
101         lda = LatentDirichletAllocation(
102             n_components=n_topics,
103             random_state=42,
104             max_iter=30,
105             learning_method="online",
106             n_jobs=-1,
107         )
108
109         doc_topics = lda.fit_transform(doc_term_matrix)
110
111         # Витяг тем
112         feature_names = vectorizer.get_feature_names_out()
113         topics_dict = {}
114         topic_names = {}
115
116         for topic_idx, topic in enumerate(lda.components_):
117             top_indices = topic.argsort()[-15:][::-1]
118             top_words = [
119                 (feature_names[i].replace("_", " "), topic[i]) for i in top_indices
120             ]
121             topics_dict[topic_idx] = top_words
122
123             topic_name = " & ".join([w for w, _ in top_words[:4]])
124             topic_names[topic_idx] = topic_name
125
126         # Присвоєння тем публікаціям
127         for i, pub in enumerate(valid_pubs):
128             pub.topic_distribution = {
129                 j: float(prob) for j, prob in enumerate(doc_topics[i])
130             }
131             pub.main_topic = int(doc_topics[i].argmax())
132             pub.assigned_topics = [int(j) for j in doc_topics[i].argsort()[-3:][::-1]]
133
134         perplexity = lda.perplexity(doc_term_matrix)
135
136         print(f"✓ Модель навчена. Perplexity: {perplexity:.2f}")
137
138         return TopicModelResult(
139             model_type="lda",
140             n_topics=n_topics,
141             topics=topics_dict,
142             topic_names=topic_names,
143             perplexity=perplexity,
144             approach=approach,
145         )
146
```

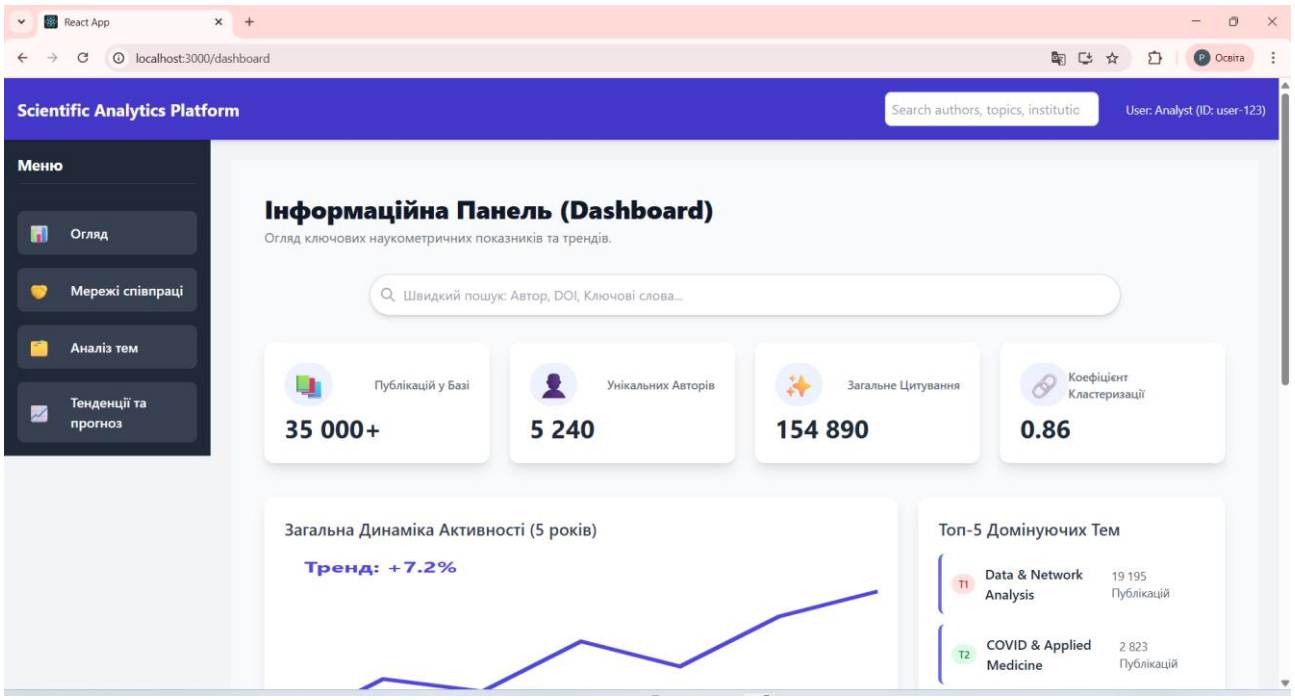
Додаток А.3

main.py | academic_data_collector.py | academic_analisis.py 3 | config.py

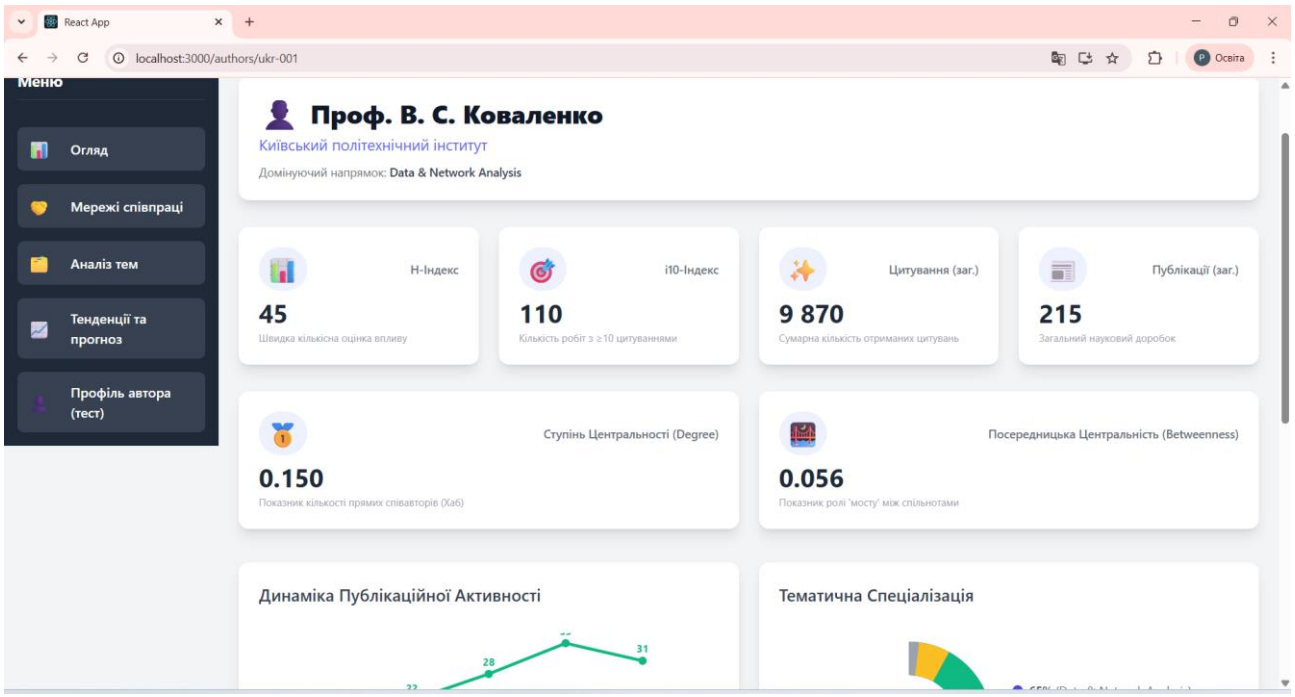
src > academic_analisis.py > NetworkAnalyzer > calculate_centrality_metrics

```python
class NetworkAnalyzer:
    """Мережевий аналіз"""

    @staticmethod
    def build_collaboration_graph(
        publications: List[NormalizedPublication],
    ) -> "nx.Graph":
        """Побудова графу співавторства"""
        if not ADVANCED_FEATURES_AVAILABLE:
            print("❌ networkx не встановлено")
            return None

        G = nx.Graph()

        for pub in publications:
            if len(pub.authors) < 2:
                continue

            for i in range(len(pub.authors)):
                for j in range(i + 1, len(pub.authors)):
                    author1_id = pub.authors[i].get("author_id")
                    author2_id = pub.authors[j].get("author_id")

                    if not author1_id or not author2_id or author1_id == author2_id:
                        continue

                    if G.has_edge(author1_id, author2_id):
                        G[author1_id][author2_id]["weight"] += 1
                    else:
                        G.add_edge(author1_id, author2_id, weight=1)

                    for author in [pub.authors[i], pub.authors[j]]:
                        aid = author.get("author_id")
                        if aid and aid in G.nodes():
                            if "name" not in G.nodes[aid]:
                                G.nodes[aid]["name"] = author.get("name", "")

        return G
```

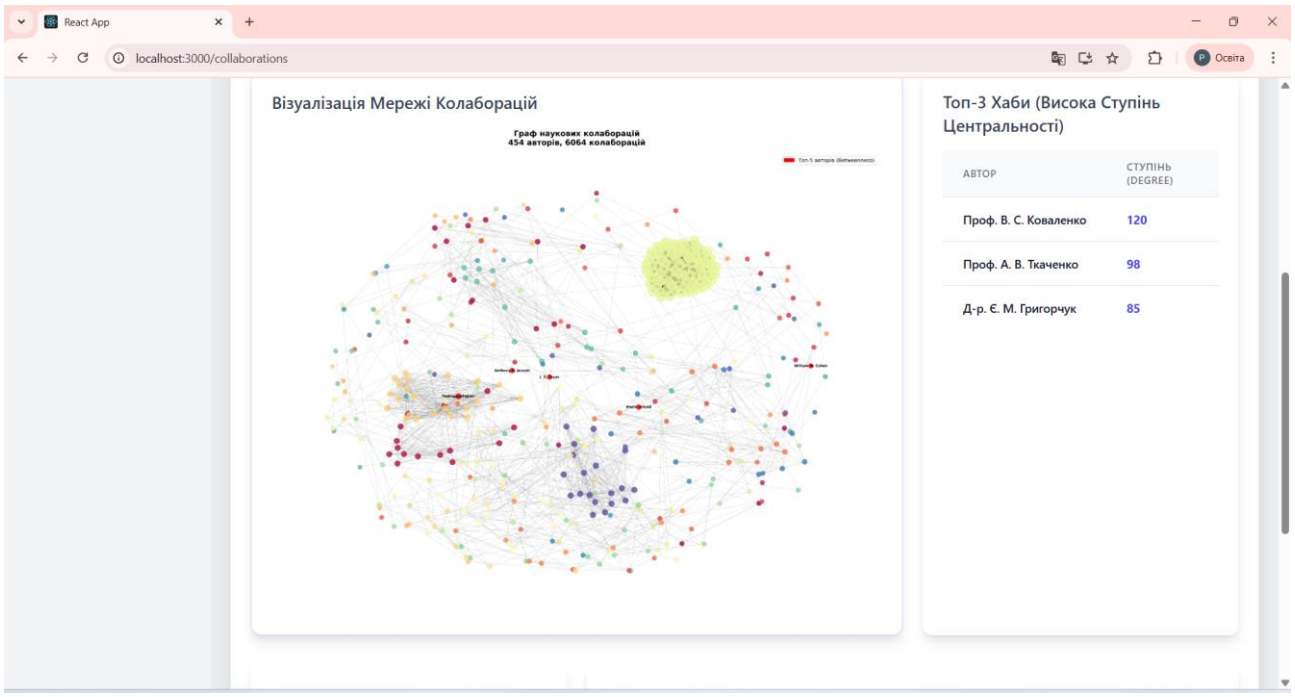
Додаток А.4


React App
localhost:3000/dashboard
Освіта
Scientific Analytics Platform
Search authors, topics, institutic
User: Analyst (ID: user-123)
Меню
Огляд
Мережі співпраці
Аналіз тем
Тенденції та прогноз
Інформаційна Панель (Dashboard)
Огляд ключових наукометричних показників та трендів.
Швидкий пошук: Автор, DOI, Ключові слова...
Публікацій у Базі
35 000+
Унікальних Авторів
5 240
Загальне Цитування
154 890
Коефіцієнт Кластеризації
0.86
Загальна Динаміка Активності (5 років)
Тренд: +7.2%
Топ-5 Домінуючих Тем
T1
Data & Network Analysis
19 195 Публікацій
T2
COVID & Applied Medicine
2 823 Публікацій

Додаток А.5


React App
localhost:3000/authors/ukr-001
Освіта
Меню
Огляд
Мережі співпраці
Аналіз тем
Тенденції та прогноз
Профіль автора (тест)
Проф. В. С. Коваленко
Київський політехнічний інститут
Домінуючий напрямок: Data & Network Analysis
H-Індекс
45
Швидка кількісна оцінка впливу
i10-Індекс
110
Кількість робіт з ≥10 цитуваннями
Цитування (заг.)
9 870
Сумарна кількість отриманих цитувань
Публікації (заг.)
215
Загальний науковий доробок
Ступінь Центральності (Degree)
0.150
Показник кількості прямих співавторів (Хаб)
Посередницька Центральність (Betweenness)
0.056
Показник ролі 'мосту' між спільнотами
Динаміка Публікаційної Активності
28
31
Тематична Спеціалізація

Додаток А.6


React App
localhost:3000/collaborations
Освіта
Візуалізація Мережі Колаборацій
Граф наукових колаборацій
454 авторів, 6064 колаборацій
Топ-3 Хаби (Висока Ступінь Центральності)
АВТОР
СТУПІНЬ (DEGREE)
Проф. В. С. Коваленко
120
Проф. А. В. Ткаченко
98
Д-р. Є. М. Григорчук
85