Міністерство освіти і науки України
НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «КИЄВО-МОГИЛЯНСЬКА АКАДЕМІЯ»
Кафедра мультимедійних систем факультету інформатики

ВИКОРИСТАННЯ НАУКОМЕТРИЧНИХ ПОКАЗНИКІВ ДЛЯ ОЦІНЮВАННЯ НАУКОВОЇ ДІЯЛЬНОСТІ

Текстова частина до дипломної роботи

Виконала студентка 2 курсу
МП «Інженерія програмного забезпечення»
Андрусів Соломія Ігорівна

Керівник дипломної роботи
доцент
Олецький Олексій Віталійович

Київ 2023

Міністерство освіти і науки України

НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «КИЄВО-МОГИЛЯНСЬКА АКАДЕМІЯ»

Кафедра мультимедійних систем факультету інформатики

ЗАТВЕРДЖУЮ

Зав. кафедри мультимедійних систем,

доцент, к. ф.-м. н. О. П. Жежерун

_____

(підпис)

“____”_____2023 р.

ІНДИВІДУАЛЬНЕ ЗАВДАННЯ

на дипломну роботу

студента Андрусів Соломії Ігорівни факультету інформатики 2 курсу

Тема: Використання наукометричних показників для оцінювання наукової діяльності

Зміст ТЧ до дипломної роботи:

Індивідуальне завдання

Календарний план

Зміст

Вступ

Розділ 1 Аналіз предметної області. Постановка завдання дипломної роботи

Розділ 2 Теоретичні відомості: наукові журнали та бібліографічні каталоги

Розділ 3 Теоретичні відомості: інтелектуальний аналіз даних

Розділ 4 Опис розробки програмного продукту

Висновки

Перелік використаних джерел

Дата видачі “___” _____ 2023 р. Керівник _____

(підпис)

Завдання отримав _____

(підпис)

Календарний план виконання дипломної роботи

Тема: Використання наукометричних показників для оцінювання наукової діяльності

Календарний план виконання роботи:

№ п/п	Назва етапу дипломного проекту (роботи)	Термін виконання етапу	Примітка
2.1.	Отримання завдання на дипломну роботу.	21.10.2022	
2.2.	Ознайомлення з існуючою інформацією по темі	22.10.2022-18.01.2023	
2.3.	Ознайомлення з існуючими системами-аналогами роботи	22.10.2022-18.01.2023	
2.4.	Початок створення практичної частини	16.01	
2.5.	Початок написання теоретичної частини	15.03	
2.6.	Подання проміжної версії практичної частини	27.03	
2.7.	Аналіз практичної частини; її коригування	15.04	
2.8.	Остаточне завершення написання теоретичної частини роботи та розробки практичної частини; коригування	15.04-05.06	
2.9.	Створення презентації	01.05-15.05	
2.10	Захист дипломної роботи	13.06	

Студент Андрусів С. І.

Керівник Олецький О. В.

“_____” _____

ЗМІСТ

Перелік термінів та умовних позначень

Data mining, або інтелектуальний аналіз даних – методика аналізу великого об’єму даних з використанням елементів машинного навчання.

Індекс Хірша – один з найпопулярніших індексів у наукометриці.

Імпакт-фактор – метрика виміру журналів.

Цитування – посилання на одну наукову роботу іншій, при цьому мається на увазі не згадка в тексті, а саме посилання у списку використаних джерел.

Квартиль – також метрика рейтингування журналів. Всього квартилів є 4, і позначаються вони відповідно Q1-Q4. Найбільш престижні журнали мають перший квартиль.

ВСТУП

Основною метою даної дипломної роботи є аналіз наукової діяльності авторів з використанням наукометричних показників та визначенням основних закономірностей на їх основі. У роботі використано такі основні метрики, як індекс Хірша[1], кількість цитувань, кількість робіт з кількістю цитувань більше 10(індекс i10)[2], імпакт-фактор[3] та кількість публікацій. Крім того, будуть проаналізовані основні метрики, які застосовуються в популярних наукових журналах, їх класифікації та ефективність.

Завданням дипломної роботи є інтелектуальний аналіз даних, що містять детальні відомості про авторів, їх публікації, та видавництва, у яких опубліковані дані наукові статті. Виокремлення певних закономірностей та аномалій, на основі статистичного та інтелектуального аналізу. Передбачення успіху чи невдачі публікації нової статті в залежності від галузі чи видавництва.

Об’єктом дослідження є наукові видавництва, в основному журнали та онлайн-ресурси типу Scopus та Google Scholar. Детально розглянуто їх класифікації, методики ранжування авторів і статей, основні наукометричні показники, що використовуються при цьому, в основному кількість цитувань, кількість публікацій, індекс Хірша тощо.

Предметом дослідження є закономірності, що виникають у результуючих наукометричних показниках авторів або конкретних наукових статей. У дослідженні буде враховано вплив таких факторів, як видавництво, галузь науки та інші показники на отримані результати, а також дослідження аномалій, якщо такі присутні.

Наукова новизна роботи полягає у використанні методів інтелектуального аналізу даних для виявлення закономірностей та аномалій у ранжуванні науковців та робіт в журналах за існуючими метриками.

Для виконання практичної частини роботи використано відкритий каталог науково-дослідницьких ресурсів OpenAlex[4], що надає API з даними про авторів, роботи та видавництва у вигляді неоднорідного графу зв'язаного посиланнями. Для підготовки датасету до подальшого аналізу було написано Python-скрипт, який здійснює мепінг та форматування необхідних полів, щоб зробити дані більш структурованими та зручними для подальшої обробки. Для інтелектуального аналізу даних та візуалізації результатів використано бібліотеку для машинного навчання з відкритим кодом Orange[5].

Текстова частина до дипломної роботи складається з вступу, чотирьох основних розділів та висновків.

У першому розділі «Аналіз предметної області. Постановка завдання дипломної роботи» розглядаються основні метрики формування рейтингу науковця чи його статті на момент дослідження та визначаються основні кроки для реалізації практичної частини.

У другому розділі «Теоретичні відомості: наукові журнали та бібліографічні каталоги» надано основну інформацію про основні категорії онлайн ресурсів. Також зроблено короткий огляд популярних онлайн-ресурсів та проаналізовано метрики, за якими ранжуються автори та їх публікації.

У третьому розділі «Теоретичні відомості: інтелектуальний аналіз даних» описано основні методи та принципи, що застосовуються для інтелектуального аналізу даних, з фокусом на визначенні тих методик, які найбільше підходять для нашого дослідження.

У четвертому розділі «Опис розробки програмного продукту» представлено детальний опис процесу реалізації практичної частини дослідження, включаючи аналіз набору даних, створення вибірки з початкового сету та її аналіз задопомогою засобів інтелектуальної обробки даних.

У висновках підбиваються підсумки проведеного дослідження, де було використано методи інтелектуального аналізу даних для проведення аналізу

формування рейтингів науковців та їх статей, та надаються пропозиції щодо удосконалення проведених процесів та усунення певних проблем, виявлених в ході проведення роботи.

Список використаної літератури містить посилання на наукові роботи та інші джерела інформації, використані в ході проведення дослідження. Всі матеріали, що доповнюють текст, зокрема діаграми, зображення та програмний код знаходяться у «Додатках».

РОЗДІЛ 1. АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ. ПОСТАНОВКА ЗАВДАННЯ ДИПЛОМНОЇ РОБОТИ

1.1. Аналіз сучасного стану питання та обґрунтування теми

Науковцям у сучасному діджиталізованому світі мало просто провести дослідження та написати статтю, навіть якщо результати дослідження містять значний прорив чи свіжий підхід до проблематики. Через перевантаження інформацією в Інтернеті, потрібно ще донести свою ідею до широкої аудиторії, що на ділі може бути складніше ніж саме дослідження. Саме цією задачею повинні займатися наукові журнали, каталоги та бібліографічні бази даних, які уже майже повністю перейшли у онлайн-формат.

Наукометрика - галузь загальної науки «про науку», що охоплює всі кількісні методи аналізу наукових праць, окремих дослідників та дослідницького процесу в цілому. Ці метрики у подальшому використовуються науковими ресурсами для пошуку найбільш релевантних статей на задану тематику, та від них, без перебільшень, зараз залежить кар’єра науковця.

Тому моя робота має на меті проаналізувати основні метрики, за якими онлайн-ресурси ранжують авторів та наукові роботи. Особливу увагу хочу звернути на актуальність індексу Хірша як найпопулярнішої метрики для ранжування науковців.

Також, на основі певних статистичних даних проаналізувати цитованість робіт у різних онлайн-ресурсах та фактори, які можуть впливати на таку оцінку. До прикладу, у тематичних журналах більша цитованість буде залежати від релевантності теми. У багатьох ресурсах зараз крім теми аналізується також анотація(abstract) та на основі

визначених штучним інтелектом ключових слів стаття пропонується як підходяща для певних пошукових запитів, навіть якщо вони не збігаються з назвою роботи. Багато електронних ресурсів прив'язані до університетів, наприклад, Oxford University Press, та можуть надавати роботам написаним своїми науковцями більший пріоритет.

У даній роботі проаналізовано тільки статті, написані англійською мовою, та опубліковані в електронному вигляді на одному з наукових онлайн-ресурсів.

1.2. Аналіз предметної області

Дані для аналізу, як найважливіший елемент дослідження, взято з відкритого каталогу OpenAlex, що агрегують та стандартизують дані, вже проіндексовані іншими великими каталогами, зокрема Misrosoft Academic[6], ORCID[7], Crossref[8], PubMed[9]. Вибірка зосереджена на авторах Стенфордського університету. Дані у датасеті поділяються на кілька сутностей: Автор, Робота, Установа, Концепт, Джерело та Видавництво. Усі сутності зв'язані між собою посиланнями та утворюють граф як на рис.1:

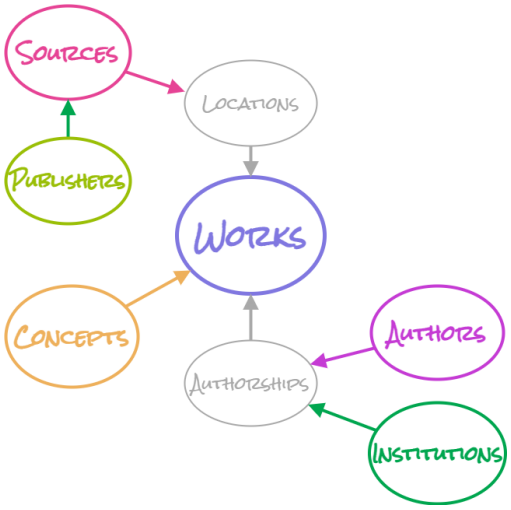


Рисунок 1: граф сутностей OpenAlex

Такий набір даних дозволяє проаналізувати зв’язки між авторами, їх науковими роботами та видавництвами, а сутність Concepts – також перевірити наскільки зрозумілою є анотація(abstract), адже ця сутність містить галузі, до яких можна віднести статтю, визначені автоматично на основі назви та анотації.

Сутність автора містить кілька уже обрахованих наукометричних показників:

- Кількість робіт автора.
- Кількість загальних цитувань усіх робіт автора.
- Кількість цитувань робіт автора щороку протягом останніх 10ти років.
- Індекс Хірша як базову метрику, що використовується у всіх системах ранжування.
- i-10 – індекс, що використовується Google Scholar та вказує на кількість робіт, що мають більше 10ти цитувань. З цього можна зробити висновок що 10 цитувань є першим умовним порогом для відсіювання взагалі неактуальних робіт.
- Імпакт-фактор, який зазвичай обчислюється для наукових журналів, та власне є річною середньою кількістю цитувань статей, опублікованих автором за останні два роки.

Імпакт-фактор як метрика для оцінки рейтингу автора не була в списку найбільш вживаних раніше, і, судячи з опису, розробники каталогу вирішили включити її експериментально. Тому у даній роботі зробимо акцент на імпакт-фактор як метрику в контексті оцінки автора та порівняння її з індексом Хірша. Зважаючи на те, що імпакт-фактор враховує тільки кількість цитувань за останні два роки, він відсіює неактивних авторів, що колись написали кілька непоганих робіт та «спочивають на лаврах», цим самим змушуючи дослідників публікувати нові статті, щоб підтримувати свій імпакт-фактор.

1.3. Постановка завдання

Основними завданнями моєї наукової роботи є:

1. Дослідження популярних журналів та їх методів оцінки наукових публікацій.
2. Дослідження датасету та виокремлення кількох показників, на основі яких у практичній частині роботи будуть перевірятися залежності.
3. Дослідження популярних методів інтелектуального аналізу даних та обрання тих, що найкраще підійдуть для вибірки даних із п.2.
4. Застосування методів інтелектуального аналізу до вибраного фрагменту даних та аналіз результатів. Пошук аномалій та моделювання нестандартних ситуацій.
5. Підбиття підсумків на основі отриманих результатів. Пояснення можливих аномалій.

РОЗДІЛ 2. ТЕОРЕТИЧНІ ВІДОМОСТІ: НАУКОВІ ЖУРНАЛИ ТА БІБЛІОГРАФІЧНІ КАТАЛОГИ

2.1. Класифікація наукових онлайн-ресурсів

Наукові статті, знаходячись на руках у автора, не мають ніякої цінності для суспільства. Щоб наукова спільнота отримала доступ до нового дослідження, і праця отримала певну вагу, потрібно її опублікувати. Наукові журнали зараз є передовими каналами для розповсюдження та популяризації наукових знань у формі описів процесів та результатів досліджень. Таким чином, вони є довідковим простором, у якому можна знайти найновішу інформацію у будь-якій науковій галузі. У сучасній ситуацій «опублікувати» це не тільки і не стільки додати статтю у печатний варіант журналу, як опублікувати електронний варіант на якомусь онлайн-ресурсі.

Існує кілька онлайн-типів ресурсів, на яких можна розміщувати наукові статті:

- Наукові журнали - це традиційний спосіб публікації наукових досліджень, тільки самі журнали тепер змінили або додали до друкованого формату ще електронний. Журнали можуть бути загальнодоступні або вимагати підписки, але в будь-якому випадку статті у них зазвичай проходять рецензування перед публікацією. Кілька прикладів наукових журналів:
    - Scientific American – охоплює всі аспекти науки та технологій, зосереджуючись на проривах і досягненнях. Найстаріший журнал США з тих, які досі випускаються.[10]

- National Geographic – відомий журнал про науку, природу та культуру з приголомшливими фотографіями та інформативними статтями.[11]
- Science - провідний рецензований журнал, що охоплює всі аспекти наукових досліджень.[12]
- Wired – охоплює науку, технології та культуру з акцентом на інновації та передові розробки.[13]

- Репозиторії - цифрові архіви, що містять повні тексти наукових публікацій, які можна безкоштовно або за оплату завантажувати та використовувати. Репозиторії зазвичай містять публікації, які надаються авторами або видавцями, а також можуть містити матеріали, які не були опубліковані в традиційних виданнях. Репозиторії можуть бути загальнодоступними або призначеними для певної спільноти.
  Основні приклади онлайн-репозиторіїв:
  - ArXiv - репозиторій для математичних, фізичних, інформатичних та інших наукових публікацій.[14]
  - PubMed Central - репозиторій для медичних публікацій.[9]
  - SSRN - репозиторій для соціальних наук, включаючи економіку, право та менеджмент.
  - RePEc - репозиторій для економічних публікацій та досліджень.
  - BioRxiv - репозиторій для наукових публікацій в галузі біології та пов'язаних з нею наук.
  - Scopus – також можна віднести до репозиторіїв, бібліографічна база даних з більше ніж 76 млн. записів наукових публікацій. Містить інформацію про наукові журнали, конференції, книги, тези доповідей та інші наукові джерела. Він не є архівним репозиторієм, але дозволяє

шукати та відстежувати наукові публікації та розробляти наукометричні показники.[15]

- Бібліографічні каталоги - це бази даних, що містять інформацію про наукові публікації, які були опубліковані в журналах, книгах та інших джерелах. Вони зазвичай містять метадані про статтю, такі як автор, заголовок, журнал або видавництво, рік публікації та інші релевантні дані. Каталоги не містять самі статті, а лише інформацію про них, що дозволяє знайти та отримати доступ до них. Вони є корисними для пошуку релевантних статей. Кілька прикладів широковживаних каталогів:
    - WorldCat - найбільший в світі бібліографічний каталог, який містить понад 2 мільярди записів про книги, періодичні видання та інші джерела інформації з усього світу.[16]
    - Library of Congress Online Catalog - містить записи про книги, періодичні видання, картографію, музичні записи, фільми та багато іншого, що належить до фондів Бібліотеки Конгресу США.
    - COPAC - який містить записи про книги, періодичні видання та інші джерела інформації з Великої Британії та Ірландії.
    - National Library of Medicine Catalog - містить записи про книги, періодичні видання та інші джерела інформації з медичної науки та здоров'я, що належать до фондів Національної медичної бібліотеки США.
    - European Library - містить записи про книги, періодичні видання та інші джерела інформації з Європейських країн, які партнери бібліотеки.

  - COPERNICUS - містить записи про публікації у галузі науки та технологій, зокрема з фізики, астрономії, математики та комп'ютерних наук.[17]
- Блоги та соціальні мережі - відносно новий спосіб публікації наукових досліджень. Науковці можуть писати про свої дослідження на власних веб-сайтах, в блогах або в соціальних мережах, де їх можуть прочитати колеги та широка громадськість.

2.2. Основні метрики наукових ресурсів

Наукові журнали та інші ресурси також бувають більш науково-важливими, та менш, і мають свої метрики, що ілюструють рейтинг журналу у наукометричній базі.

Квартилі для наукових журналів – це спосіб ранжування журналів на основі їх імпакт-фактору або кількості цитувань. У цій системі журнали поділяються на чотири квартилі, де Q1 представляє 25% кращих журналів за імпакт-фактором або кількістю цитувань, Q2 представляє наступні 25% і так далі. Це дозволяє дослідникам швидко визначити відносний вплив або престиж певного журналу в їхній галузі дослідження. Квартильний рейтинг можна використовувати як інструмент для оцінки якості досліджень, опублікованих у журналі, а також для прийняття рішень про те, куди подавати статті для публікації.

Квартилі можуть відрізнятися в різних галузях науки. Квартильні рейтинги визначаються на основі розподілу цитувань у певній галузі дослідження. Таким чином, квартиль для журналу в одній науковій галузі може не збігатися з квартилем для журналу в іншій науковій галузі. Крім того, квартильні рейтинги оновлюються щорічно, і тому можуть змінюватися з часом у міру зміни моделей цитування в певній галузі.

Квартильний рейтинг більш прозорий та зрозумілий для читачів, але, як ми бачимо, він базується на імпакт-факторі. Що ж таке імпакт-фактор для наукового журналу?

Імпакт-фактор — це показник, який використовується для вимірювання відносної важливості наукового журналу в галузі його досліджень. Він розраховується шляхом ділення кількості цитувань, отриманих статтями журналу за певний рік, на загальну кількість статей, опублікованих журналом за два попередні роки. Імпакт-фактор використовується як міра впливу журналу, при цьому більш високі імпакт-фактори зазвичай вказують на більш престижні журнали.

Для розрахунку імпакт-фактора використовується саме термін у два роки, оскільки вважається, що такий часовий проміжок забезпечує розумну оцінку поточного впливу журналу в науковому співтоваристві. Цей часовий проміжок дозволяє накопичити достатню кількість цитувань, відображаючи при цьому найновіші дослідження. Крім того, два роки є звичайним циклом публікацій для багатьох наукових галузей, що означає, що це достатній часовий проміжок для оцінки впливу дослідження, опублікованого в певному журналі.

Однак важливо зазначити, що імпакт-фактор піддавався критиці через низку причин, у тому числі через можливість маніпулювання цитуваннями і той факт, що він враховує лише цитування протягом певного періоду часу. Це також не міра якості окремих статей, опублікованих у журналі, а радше міра загального впливу журналу в цілому. Тому при оцінці якості та важливості наукового журналу важливо використовувати інші показники та фактори разом із імпакт-фактором. Деякі наукові ресурси, включаючи Google Scholar[18], навіть запропонували свої показники, якими можна виміряти якість публікацій. Наприклад, SCImago Journal запропонував використовувати термін у три роки замість двох, використаних для обчислення імпакт-фактору.

Так, топ-5 журналів за імпакт-фактором на кінець 2022 року[19]:

1. CA - A Cancer Journal for Clinicians(286.13)
2. Lancet(202.731)
3. New England Journal of Medicine(176.079)
4. JAMA - Journal of The American Medical Association(157.335)
5. Nature Reviews Molecular Cell Biology(113.915)

А згідно з Scimago Journal & Country Rank[20]:

1. CA - A Cancer Journal for Clinicians(56.204)
2. Nature Reviews Molecular Cell Biology(33.213)
3. Quarterly Journal of Economics(31.348)
4. Cell(25.716)
5. MMWR Recommendations and ReportsOpen Accessjournal(25.045)

Іноді кілька показників використовуються у комбінації, щоб домогтися найбільш релевантного результату. Відома бібліографічна база даних Scopus використовує комбінацію наступних методик для ранжування журналів[21]:

- SCImago Journal Rank (SJR) – вищезгадана метрика, що обраховує імпакт-фактор для трирічного часового проміжку.
- CiteScore – особлива метрика, розроблена Scopus, яка вимірює середню кількість цитувань на статтю, опубліковану в певному журналі за певний період часу (зазвичай один рік). CiteScore розроблений так, щоб бути більш повним і прозорим, ніж інші показники журналу, такі як імпакт-фактор, за рахонук включення у обчисленні всі типи документів, індексованих у Scopus, а не лише статті.
- Source-Normalized Impact per Paper (SNIP): метрика, розроблена Центром науково-технічних досліджень (CWTS), яка вимірює вплив журналу, беручи до уваги характеристики дослідницької

галузі, в якій він працює. SNIP використовує шаблони цитування журналів у галузі, щоб визначити «нормалізований у галузі» вплив цитування.

2.3. Рейтингування науковців у наукових журналах

На даний момент індекс Хірша все ще залишається одним із найпопулярніших методів ранжування авторів. Обчислення цього індексу доволі просте – він дорівнює максимальній кількості N робіт, цитування яких рівне або більше N. Незважаючи на те, що зараз все більше ресурсів використовують альтернативні методи ранжування, h-індекс залишається популярним і широко прийнятим показником впливу дослідження. Ось кілька прикладів журналів Q1, які все ще використовують h-індекс як головний показник рейтингу:

- Nature
- Science
- Cell
- The Lancet
- The Journal of the American Chemical Society (JACS)

Хоч індекс Хірша і залишається одним із основних показників, багато ресурсів комбінують показники. Розглянемо метрики, які використовуються у ранжуванні авторів на Scopus[15]:

- Індекс Хірша
- Відстеження огляду цитувань
- Інструменти візуального аналізу: загальна кількість цитованих документів, загальна кількість цитувань за рік та перелік документів із зазначенням номерів цитованих документів і посилань на цитовані документи за рік і за статтею.

Як бачимо, Scopus також поки ще спирається на індекс Хірша як основний метод ранжування авторів. Проте, на сторінці з інформацією про метрики документа можна знайти багато нових метрик:

- загальна кількість цитувань за діапазоном дат за вибором користувача;
- цитування за рік для діапазону;
- percentile - порівняльний аналіз цитувань, відсоток авторів, які мають вищий рейтинг певного автора на основі вибраного показника;
- зважене цитування, в залежності від галузі;
- також нещодавно додано метрику кількості переглядів документа на Scopus;
- PlumX Metrics: п’ять комплексних показників на рівні предметів, які дають зрозуміти, як люди взаємодіють з окремими фрагментами результатів досліджень (статтями, матеріалами конференцій, розділами книг та багатьма іншими) в онлайн-середовищі(рис. 2):

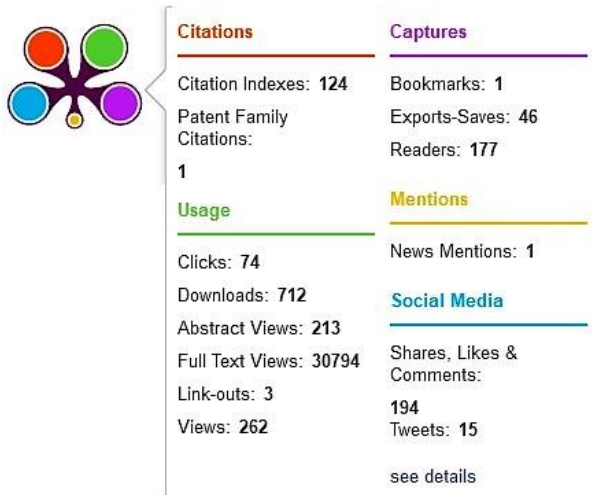


Рисунок 2: Scopus PlumX Metrix

Кількість переглядів, та для деяких ресурсів кількість завантажень наукових робіт – відносно нові метрики, що підходять тільки для оцінки електронних наукових ресурсів. У той час як цитування зазвичай вважають

індикатором академічного впливу, завантаження відображають радше рівень привабливості чи популярності публікації для користувачів веб-сервісу(які не завжди є науковцями).

Також популярності зараз набуває нова метрика Altmetrics, що також враховує популярність наукових робіт серед широкої Інтернет-аудиторії.

Альтметрика — це відносно новий спосіб вимірювання впливу досліджень, який виходить за рамки традиційних показників на основі цитування, таких як h-індекс і імпакт-фактор. Оцінка Altmetrics — це показник, який вимірює увагу, яку дослідницька стаття привернула в Інтернеті, включаючи соціальні мережі, засоби масової інформації, блоги, політичні документи та інші онлайн-джерела. Це дає можливість оцінити безпосередній і ширший вплив дослідження за межами академічної спільноти. Altmetrics обчислюється за допомогою алгоритму, який враховує різні онлайн-джерела та призначає оцінку кожному джерелу на основі його впливу та релевантності. У описі алгоритму не зазначено напряму, але зважені оцінки онлайн-ресурсів та перебір можливих поосилань роблять його схожим на вдосконалену версію алгоритму Page Rank, який був детально розглянутий у моїй попередній роботі з наукометрики.[27]

Altmetrics зараз використовується багатьма науковими журналами, репозиторіями та платформами, щоб забезпечити альтернативний спосіб вимірювання впливу та охоплення досліджень. Ось кілька прикладів журналів і платформ, які використовують Altmetrics:

1. PLOS ONE BIOLOGY – рецензований журнал із відкритим доступом, який публікує дослідження в усіх галузях науки та медицини. Даний журнал був одним із перших, який прийняв Altmetrics як спосіб вимірювання впливу своїх статей.

2. SpringerLink — платформа, що надає доступ до мільйонів наукових статей і розділів книг від Springer та її філій.
3. Frontiers — платформа, яка публікує журнали з відкритим доступом у різноманітних галузях, включаючи науку, здоров’я та техніку.
4. Figshare — це сховище даних, яке дозволяє дослідникам ділитися своїми дослідницькими даними та керувати ними.
5. Mendeley — це довідковий менеджер і соціальна мережа для дослідників.

Метрика Altmetrics може бути корисною для дослідників, щоб визначити, які з їхніх статей мають найбільший вплив за межами академічного середовища, чи допомогти спонсорам і політикам відстежувати ширший вплив інвестицій у дослідження. Дана метрика також може надати ранню ознаку потенційного впливу нового відкриття чи результату дослідження. Однак важливо зазначити, що показник Altmetrics не слід використовувати як заміну традиційним показникам на основі цитувань, оскільки він не обов’язково відображає якість або наукову значимість дослідження. Натомість він надає додатковий погляд на вплив дослідження, який враховує його ширший охоплення та вплив за межами академічної спільноти.

РОЗДІЛ 3. ТЕОРЕТИЧНІ ВІДОМОСТІ: ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ДАНИХ

3.1. Загальна інформація про інтелектуальний аналіз даних

Інтелектуальний аналіз даних(data mining), також відомий як виявлення знань у базах даних (Knowledge Discovery in Databases, KDD) — це процес пошуку та вилучення корисної інформації та знань із великих і складних наборів даних. Це підгалузь інформатики та штучного інтелекту, яка передбачає використання алгоритмів, статистичних моделей та інших обчислювальних методів для аналізу та виявлення закономірностей, взаємозв’язків і розуміння даних.

Інтелектуальний аналіз даних використовується в широкому діапазоні програм, включаючи бізнес, фінанси, маркетинг, охорону здоров’я та наукові дослідження. Процес інтелектуального аналізу даних починається з надання певних вхідних даних інструментам інтелектуального аналізу, які в свою чергу використовують статистичні дані й алгоритми для відображення знайдених шаблонів. Результати аналізу можна візуалізувати за допомогою цих інструментів, проаналізувати вручну та надалі застосовувати для модифікації та вдосконалення бізнесу(рис 3).

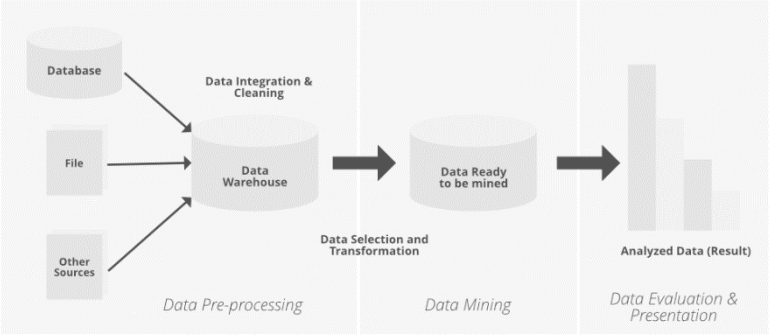


Рисунок 3: Інтелектуальний аналіз даних

Детальний флоу виявлення знань із даних складається з наступних кроків:

1. Очищення даних (Data cleaning) - видалення шуму та нерелевантних даних.
2. Інтеграція даних(Data integration) - поєднання кількох джерел очищених даних.
3. Вибір даних (Data selection) - дані, що стосуються завдання аналізу, витягуються з бази.
4. Трансформація даних (Data transformation) - дані перетворюються у формат, що підходить для інтелектуального аналізу, шляхом виконання функцій зведення або агрегації.
5. Інтелектуальний аналіз даних (Data mining) – найважливіший процес, у якому застосовуються інтелектуальні методи для пошуку шаблонів даних.
6. Оцінка зразків (pattern evaluation) - виявлення найпопулярніших шаблонів.
7. Презентація знань (knowledge presentation) – отримані дані візуалізуються та показуються користувачеві.

Як можна побачити, інтелектуальний аналіз даних без попередніх кроків неможливий, не менш важливим є процес вибірки та підготовки даних до аналізу та представлення їх у формі, необхідній для обраних інструментів аналізу даних.

Найпопулярніші формати даних для інтелектуального аналізу даних можуть відрізнятися залежно від конкретної програми та типу даних, що аналізуються. Однак є формати даних, які використовуються для інтелектуального аналізу частіше за інші:

- CSV (значення, розділені комами) - простий текстовий формат, у якому кожен рядок представляє запис даних, а окремі значення розділені комою, або іншим розділювальним символом.
- JSON(JavaScript Object Notation) - легкий формат обміну даними, який зазвичай використовується для передачі даних між веб-додатками.
- SQL(мова структурованих запитів) - це стандартна мова, яка використовується для керування реляційними базами даних.
- XML(розширена мова розмітки) - мова розмітки, яка використовується для зберігання та передачі даних.
- HDF5 (ієрархічний формат даних) - це формат файлу, призначений для зберігання та керування великими та складними наборами даних.

Загалом, CSV і SQL широко використовуються в інтелектуальному аналізі даних завдяки своїй простоті та сумісності з більшістю програмного забезпечення для аналізу даних. JSON і XML зазвичай використовуються у веб-додатках та для обміну даними між різними системами. HDF5 частіше можна зустріти в наукових і інженерних програмах, які включають великі та складні набори даних.

3.2. Основні техніки інтелектуального аналізу

Існує багато методик та алгоритмів для інтелектуального аналізу даних, які зазвичай використовуються для вилучення необхідних знань із великих і складних наборів даних. Ось деякі з найпопулярніших технік:

1. Асоціативні правила (association rules) – техніка виявлення зв’язків між різними атрибутами даних або елементами. Її також

іноді називають правилами ринкового кошика. Інтелектуальний аналіз правил асоціації має кілька застосувань і зазвичай використовується для кореляції продажів у наборах даних або медичних даних. Алгоритм працює так, що у вас є різні дані, наприклад, список продуктів, які ви купували за останні шість місяців. Він обчислює відсоток товарів, які купуються разом. Маючи цю інформацію, магазини можуть планувати акції, рекламувати групи товарів та прогнозувати.

2. Кластеризація (clustering) - поділ даних на кластери на основі пов’язаних характеристик. Іншими словами, ми можемо сказати, що кластерний аналіз — це техніка для ідентифікації подібних даних. Об’єкти кластеризуються за принципом максимізації внутрішньокласової подібності та мінімізації міжкласової подібності. Тобто кластери об’єктів створюються так, що об’єкти всередині кластера мають високу схожість на відміну від інших, але є різними об’єктами в інших кластерах.
3. Класифікація (classification) – велика група технік, що базуються на процесі пошуку набору моделей (або функцій), які описують і розрізняють класи даних або концепції, з метою використання моделі для прогнозування класу об’єктів, мітка класу яких невідома. Це робиться за допомогою таких алгоритмів, як дерева рішень, k-найближчий сусід або логістична регресія. Класифікація даних — це двоетапний процес:
    a. Етап навчання: тут будується модель. Попередньо визначений алгоритм застосовується до даних для аналізу з наданою міткою класу, і будуються правила класифікації.
    b. Етап класифікації: модель використовується для прогнозування міток класу для заданих даних. Точність правил класифікації оцінюється тестовими даними, які,

якщо визнаються точними, використовуються для класифікації нових кортежів даних.

4. Дерева рішень (decision trees) - використовуються для класифікації або прогнозування результату на основі встановленого списку критеріїв або рішень. Дерево рішень використовується для введення ряду каскадних запитань, які сортують набір даних на основі наданих відповідей. Кожен вузол дерева представляє перевірку значення атрибута, кожна гілка позначає результат перевірки, а листя дерева представляють класи або розподіл класів. Дерева рішень популярні, оскільки не вимагають жодних знань предметної області, можуть представляти багатовимірні дані та легко перетворюються на правила класифікації.
5. Випадковий ліс (random forest) - метод поєднує кілька дерев рішень для створення більш точної та стабільної моделі. Основна ідея випадкового лісу полягає у створенні безлічі дерев рішень, кожне з яких базується на різній підмножині навчальних даних і випадковому виборі функцій. Кожне дерево навчено передбачати цільову змінну на основі вибраних ознак. Під час прогнозування кінцевий результат визначається більшістю голосів прогнозів від усіх дерев у лісі. Зазвичай є більш точним ніж звичайне дерево рішень.
6. Регресія (regression) - статистичний метод, який використовується в інтелектуальному аналізі даних для моделювання зв’язку між залежною змінною та однією або кількома незалежними змінними. Це методика навчання під наглядом, що означає, що для розробки прогнозної моделі потрібні позначені навчальні дані. Мета регресійного аналізу — знайти таку математичну функцію, яка може точно передбачити значення залежної змінної на основі значень незалежних змінних. Ця функція зазвичай

представляється у вигляді лінійного рівняння, але також може приймати інші форми, такі як поліноміальна, експоненціальна або логарифмічна функції.

7. Перехресна перевірка (cross-validation) — це техніка інтелектуального аналізу даних, яка використовується для оцінки продуктивності прогнозної моделі. Основна ідея перехресної перевірки полягає в тому, щоб використовувати частину доступних даних для навчання моделі, а решту даних — для перевірки продуктивності моделі. Найпоширенішою формою перехресної перевірки є k-кратна перехресна перевірка, коли дані поділяються на k частиин однакового розміру. Потім модель тренується на k-1 з цих частин і тестується на тій, що залишилася. Цей процес повторюється k разів, причому кожна частина даних використовується один раз як дані тестування. Потім знаходиться середня продуктивність моделі для всіх k частин, яка і є точністю прогнозування моделі. Перехресну перевірку можна застосовувати до широкого діапазону прогнозних моделей, включаючи регресію, класифікацію та алгоритми кластеризації.
8. Виявлення аномалій (anomaly detection) - техніка інтелектуального аналізу даних, яка використовується для визначення точок даних або спостережень, які значно відрізняються від більшості даних. Мета виявлення аномалій полягає в тому, щоб ідентифікувати точки даних з відхиленнями і або видалити їх із набору даних, або провести їх подальше дослідження, щоб визначити, чи вони мають значення. Для виявлення аномалій використовується кілька методів, зокрема статистичні методи, методи кластеризації та алгоритми машинного навчання.
9. Прогнозування (prediction) - двоетапний процес, подібний до процесу класифікації даних. Він використовує комбінацію інших

методів інтелектуального аналізу даних, таких як тенденції, кластеризація, класифікація тощо. Він аналізує минулі події або випадки в правильній послідовності, щоб передбачити майбутню подію. Прогноз можна розглядати як побудову та використання моделі для оцінки класу непозначеного об’єкта або для оцінки значення чи діапазонів значень атрибута, який, ймовірно, матиме даний об’єкт. Прогнозування можна використовувати для побудови моделі для оцінки класу непозначеного об’єкта або для оцінки значення чи діапазонів значень атрибута, які може мати даний об’єкт.

3.3. Data mining та наукометрика

Українські науковці Олеся Мриглод та Юрій Головач з Національної Академії Наук України у співпраці з британським науковцем Ральфом Кенна із університету Ковентрі у 2018 році було проведено порівняльне дослідження впливу цитувань та завантажень на оцінку академічної публікації з використанням інтелектуального аналізу даних на основі статистичних даних, наданих європейським журналом «Europhysics Letters» (EPL).[22]

У результаті дослідження було проілюстровано залежність кількості цитувань та завантажень для європейського журналу EPL на діаграмах(рис.4). З наданих діаграм видно що кількість завантажень та цитувань у більшості випадків лінійно залежна, але є винятки, та на ці показники також впливає вік самої публікації, дисципліна, вид публікації та особливості конкретного журналу.

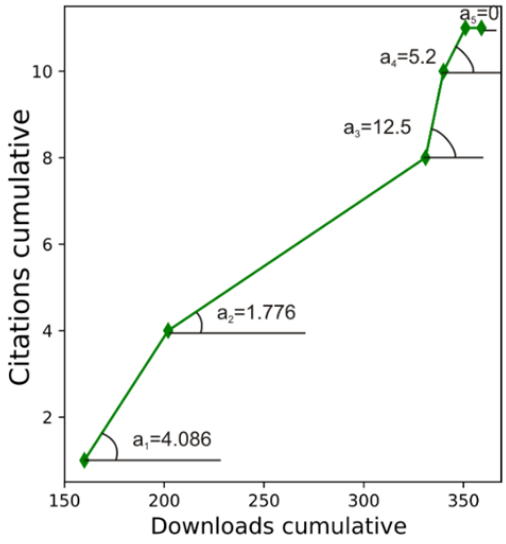


Рисунок 4: залежність завантажень та цитувань

Проаналізувавши всі доступні наукометричні показники, вирішено обрати такі основні метрики оцінок наукових публікацій та журналів для подальшого інтелектуального аналізу:

- індекс Хірша автора;
- імпакт-фактор автора;
- кількість цитувань автора;
- кількість цитувань публікації;
- вік публікації(в роках)
- імпакт-фактор журналу
- квартиль журналу.

Для інтелектуального аналізу даних на основі вищезазначених наукометричних показників буде змодельовано такі основні ситуації:

1. Визначити квартиль журналу на основі усіх інших наявних метрик журналу.
2. Визначити квартиль на основі даних про публікації. Зробити висновки про залежність популярності наукової статті від журналу у якому вона опублікована.

3. Можливо також зробити вибірку із кількох журналів одного квартиля та методом класифікації визначити журнал на основі даних публікацій.
4. Пошук аномалій серед авторів та публікацій методом кластеризації.
5. Визначити вік публікації на основі її цитувань.
6. Визначити стаж автора на основі індексу Хірша, цитувань та імпакт-фактора.
7. Визначити метрики автора на основі всіх інших наявних метрик. Порівняти результати.
8. Визначити метрики журнала на основі всіх інших наявних метрик. Порівняти результати.

РОЗДІЛ 4. ОПИС РОЗРОБКИ ПРОГРАМНОГО ПРОДУКТУ

4.1. Аналіз сутностей даних

Для подальшого інтелектуального аналізу обрано датасет, наданий відкритим каталогом OpenAlex.[4] Дані у каталозі представляються у вигляді графу, що складається з таких основних сутностей:

- Source
- Author
- Work
- Institution
- Concept
- Publisher
- Geo

В даному дослідженні ми зупинимося на детальному аналізі перших трьох сутностей та їх атрибутів.

- Source – сутність, що описує електронний ресурс, у якому опублікована стаття. Для дослідження можна виділити такі корисні атрибути даної сутності:
  - display_name: назва ресурсу;
  - type: тип електронного видання. Всього є 4 типи: журнал, репозиторій, конференція, та платформа для електронних книг;
  - works_count: загальна кількість робіт, опублікованих у даному ресурсі.
  - cited_by_count: загальна кількість робіт, які цитують роботи, розміщені на цьому ресурсі;

  - counts_by_year: кількість нових робіт та доданих цитувань за кожний з останніх десяти років, розділених за роками. Для аналізу було зроблено вибірку з трьох останніх років(2021-2023);
  - summary_stats.2yr_mean_citedness: імпакт-фактор для даного електронного ресурсу;
  - summary_stats.h_index – індекс Хірша;
  - summary_stats.i10_index – кількість робіт, у яких кількість цитувань більше 10-ти(метрика Google Scholar);
  - даний репозиторій не містить інформації про квартиль, до якого відноситься ресурс, тому при необхідності ця інформація буде додаватися вручну.
- Author – сутність, що описує автора наукових робіт. Має такі важливі нам атрибути:
  - works_count: загальна кількість робіт, написаних даним автором.
  - cited_by_count: загальна кількість робіт, які цитують роботи цього автора;
  - counts_by_year: кількість нових робіт автора та доданих цитувань за кожний з останніх десяти років, розділених за роками;
  - summary_stats.2yr_mean_citedness: імпакт-фактор для даного автора;
  - summary_stats.h_index – індекс Хірша;
  - summary_stats.i10_index – кількість робіт, у яких кількість цитувань більше 10-ти(метрика Google Scholar);
- Work – сутність наукової роботи. Має такі автрибути:
  - cited_by_count: загальна кількість робіт, які цитують дану роботу;
  - publication_year: рік, в який опублікована робота;

- sources: посилання на джерела, у яких опублікована дана робота;
- authorships: посилання на усіх авторів роботи;
- cited_by_api_url: посилання на всі роботи що цитують дану роботу.

4.2. Процес формування вибірки даних

Для аналізу даних було зроблено кілька вибірок з використанням Python-бібліотеки pyalex[23], що реалізує зручний доступ до OpenAlex API. Зокрема, було зроблено такі вибірки:

1. Для перевірки зв’язку між віком публікації та її цитуваннями було вибрано 3000 випадкових робіт. Вік робіт було обчислено у роках на основі дати публікації.
2. Для визначення стажу автора на основі індексу Хірша, цитувань та імпакт-фактора було вибрано 5000 випадкових авторів з усіма їх метриками. Цей же датасет використовувався для визначення залежностей метрик одна від одної. Стаж автора було обчислено у роках на основі дати створення акаунту.
3. Для визначення індексу Хірша журналу на основі всіх його інших метрик було обрано 1000 випадкових журналів.
4. Для визначеннння квартилю журналу на основі усіх інших наявних метрик журналу було обрано по 10 журналів для кожного квартилю.
5. Для визначення квартилю журналу на основі метрик опублікованих робіт було взято ті самі 40 журналів, по 10 на квартиль, та по 25 робіт для журналу. Загалом – 1000 робіт.
6. Для кластеризації журналів певного квартилю було проведено 4 експерименти, в кожному з яких використано частину даних з п.4.

4.3. Інтелектуальний аналіз даних: Orange

Для виконання інтелектуального аналізу було використано програмний пакет візуального програмування Orange[5]. Програма дозволяє застосовувати ряд алгоритмів інтелектуального аналізу даних до заданої вибірки даних та формують передбачення на основі перехресної перевірки. Наприклад, модель для аналізу даних про наукові журнали на основі їх поділу на квартилі виглядає як на рис.5.

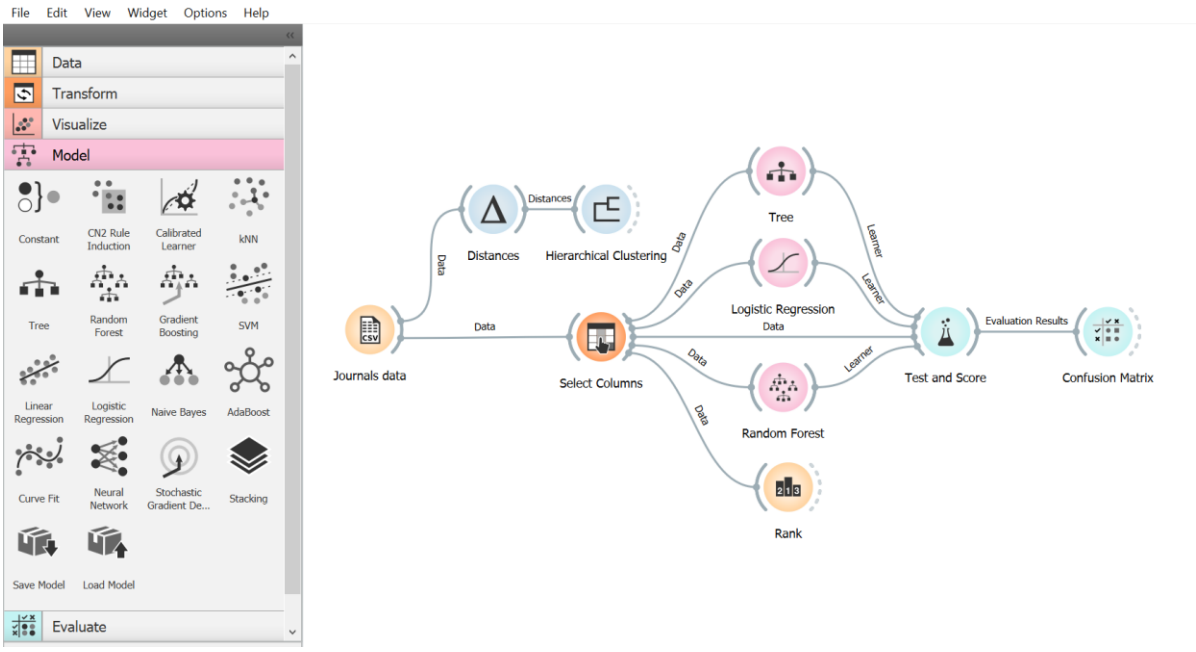


Рисунок 5: Orange Journals Data Mining

Для моделювання та перевірки ситуацій, описаних у п.3.3 було застосовано такі техніки інтелектуального аналізу даних:

1. Для перевірки зв’язку між віком публікації та її цитуваннями було порівняно результати виконання таких трьох методів як випадковий ліс, лінійна регресія та дерево рішень з навчанням методом перехресної перевірки.

2. Для аналізу метрик автора також використали випадковий ліс, лінійну регресію та дерева рішень з навчанням методом перехресної перевірки.
3. Для перевірки розбиття наукових журналів на квартилі, порівняно результати виконання таких трьох методів як випадковий ліс, логічна регресія та дерево рішень, а також застосовано кластерний аналіз та рангування метрик журналу за їх впливом на результат класифікації.

4.4. Огляд результатів

Orange застосовує різні метрики для оцінки прогнозованих результатів для числових значень, що коливаються у певному проміжку дійсних чисел, та класових значень, тобто таких, що мають кілька можливих варіантів для передбачення.

Для класових значень маємо наступні метрики:

- AUC(area under curve) – метрика, що використовується для оцінки ефективності моделі бінарної класифікації. AUC представляє ступінь роздільності класів моделі, тобто наскільки добре модель здатна розрізняти позитивні та негативні зразки. Загалом показник AUC вище 0,8 вважається хорошим результатом.
- CA(classification accuracy) – метрика, що показує, наскільки добре модель класифікації здатна правильно передбачити мітки класів нових екземплярів. Значення знаходяться в межах від 0 до 1, та часто представляються у відсотках.
- Precision – міра частки правильних позитивних передбачень до усіх випадків, які прогнозовано будуть позитивними.

- Recall - міра частки правильно передбачених позитивних серед усіх справді позитивних випадків.

Для числових значень:

- MAE(mean absolute error) – середня абсолютна похибка. Чим нижче значення похибки, тим краще. Загалом, значення MAE менше 10% від діапазону цільової змінної можна вважати хорошим результатом.
- RMSE (root mean squared error) – середньоквадратична помилка. Значення RMSE менше 20% від діапазону цільової змінної можна вважати хорошим результатом.
- R2(R-squared) – коефіцієнт детермінації, представляє частку дисперсії залежної змінної, яку можна передбачити на основі незалежних змінних. Значення R-квадрат варіюються від 0 до 1, при цьому вищі значення вказують на кращу відповідність моделі регресії даним. Загалом, значення R-квадрат більше 0,7 можна вважати хорошим результатом.

Отже, розглянемо результати застосування інтелектуального аналізу до змодельованих випадків.

1. Індекс Хірша журналу на основі інших наявних метрик

Індекс Хірша журналу також гарно визначається на основі наявних метрик журналу:

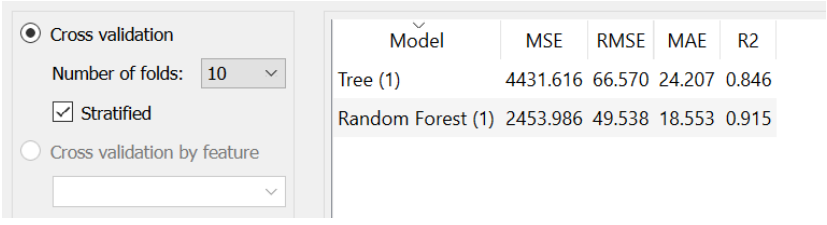


Рисунок 6: прогнозування індексу Хірша журналу

Метрик журналу більше, ніж у автора, тому у нас є можливість проранжувати їх та вияснити, які з метрик були найбільш вагомими при обчисленні індексу Хірша:

		#	Univar. reg.	RReliefF
1	cited_by_count_2021		173.375	0.334
2	cited_by_count_2022		168.075	0.332
3	cited_by_count_2023		157.491	0.325
4	cited_by_count		143.306	0.322
5	i10_index		82.695	0.283
6	works_count_2021		30.789	0.264
7	works_count_2023		26.565	0.260
8	works_count_2022		25.137	0.257
9	works_count		49.296	0.246
10	2yr_mean_citedness		2.765	0.115

Рисунок 7: ранжування метрик журналу

Як бачимо, наш імпакт-фактор(2year mean citedness) відіграє найменшу роль у визначенні індексу Хірша, що логічно, адже індекс Хірша не залежить від часового проміжку. В той же час найбільш вагомими є кількості цитувань.

2. Квартиль журналу на основі інших наявних метрик журналу

Спробували визначити квартиль журналу на основі інших наявних метрик. Модель навчилася визначати квартиль журналу з точністю до 70%:

Model	AUC	CA	F1	Precision	Recall
Tree	0.780	0.625	0.608	0.624	0.625
Random Forest	0.876	0.725	0.716	0.715	0.725
Logistic Regression	0.876	0.725	0.718	0.722	0.725

Рисунок 8: прогнозування квартилю журналу

Глянемо на confuse matrix щоб побачити, у яких кейсах наша модель не змогла правильно класифікувати науковий журнал:

Actual \ Predicted	Q1	Q2	Q3	Q4	Σ
Q1	10	0	0	0	10
Q2	0	4	5	1	10
Q3	0	3	6	1	10
Q4	0	0	1	9	10
Σ	10	7	12	11	40

Рисунок 9: матриця невідповідностей(confusion matrix) для визначення квартилю

Як бачимо, найкраще визначаються «граничні» випадки, тобто найвищий та найнижчий квартиль. А найтонга межа переходу між другим та третій квартилями. При цьому межа між першим та другим квартилем досить однозначна, модель жодного разу не «підняла» журнал другого квартилю до першого.

Застосуємо також кластерний аналіз:

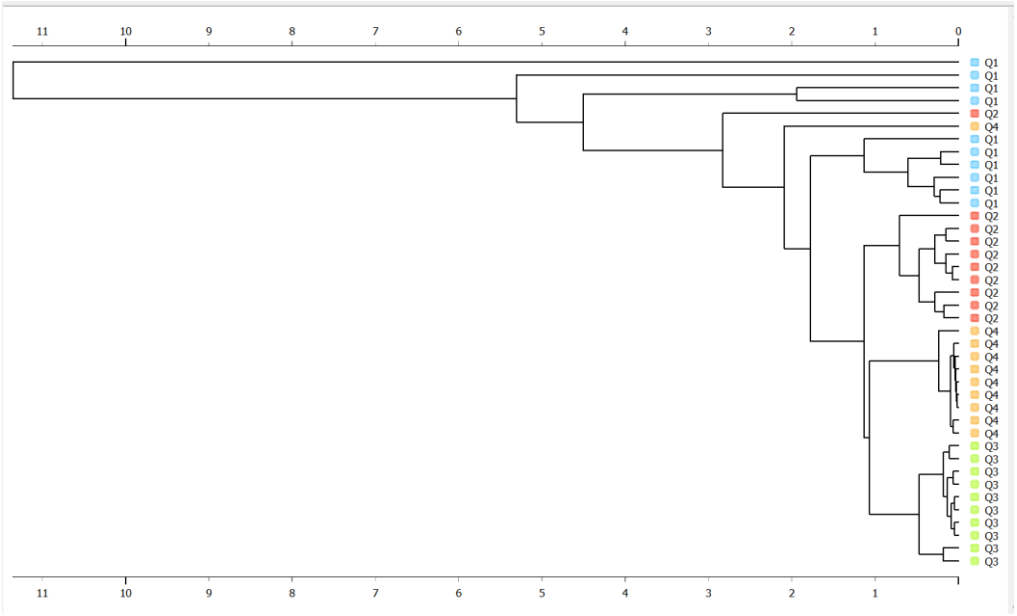
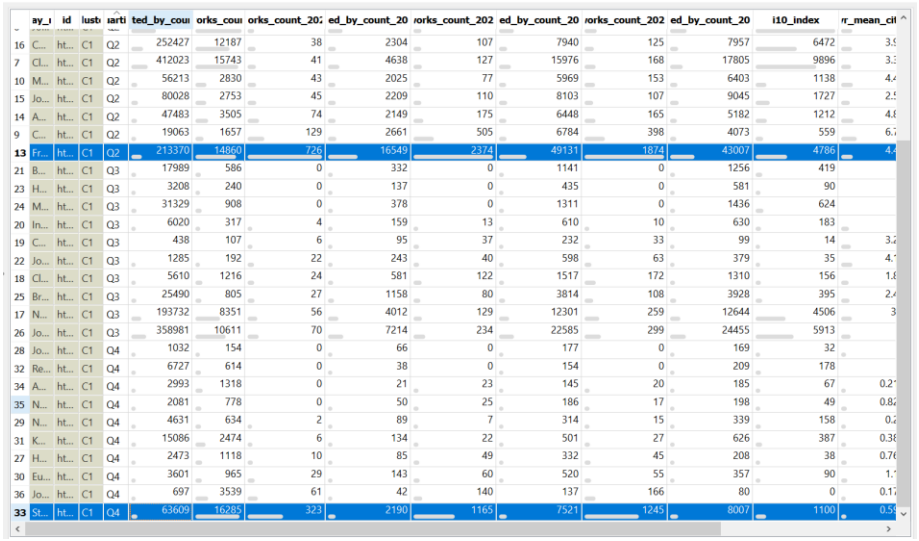


Рисунок 10: кластерний аналіз по квартилю

На основі кластерного аналізу бачимо тільки одну «змішану» зону. При детальному розгляді знайшли один журнал другого

квартилю та один четвертого з досить високими показниками метрик у порівнянні з іншими показниками свого класу:



	ay_i	id	lust	arti	ted_by_cou	orks_cou	orks_count_20	ed_by_count_20	orks_count_202	ed_by_count_20	orks_count_202	ed_by_count_20	i10_index	r_mean_cit
16	C...	ht...	C1	Q2	252427	12187	38	2304	107	7940	125	7957	6472	3.5
7	Cl...	ht...	C1	Q2	412023	15743	41	4638	127	15976	168	17805	9896	3.3
10	M...	ht...	C1	Q2	56213	2830	43	2025	77	5969	153	6403	1138	4.4
15	Jo...	ht...	C1	Q2	80028	2753	45	2209	110	8103	107	9045	1727	2.5
14	A...	ht...	C1	Q2	47483	3505	74	2149	175	6448	165	5182	1212	4.8
9	C...	ht...	C1	Q2	19063	1657	129	2661	505	6784	398	4073	559	6.7
13	Fr...	ht...	C1	Q2	213370	14860	726	16549	2374	49131	1874	43007	4786	4.4
21	B...	ht...	C1	Q3	17989	586	0	332	0	1141	0	1256	419	
23	H...	ht...	C1	Q3	3208	240	0	137	0	435	0	581	90	
24	M...	ht...	C1	Q3	31329	908	0	378	0	1311	0	1436	624	
20	In...	ht...	C1	Q3	6020	317	4	159	13	610	10	630	183	
19	C...	ht...	C1	Q3	438	107	6	95	37	232	33	99	14	3.2
22	Jo...	ht...	C1	Q3	1285	192	22	243	40	598	63	379	35	4.1
18	Cl...	ht...	C1	Q3	5610	1216	24	581	122	1517	172	1310	156	1.8
25	Br...	ht...	C1	Q3	25490	805	27	1158	80	3814	108	3928	395	2.4
17	N...	ht...	C1	Q3	193732	8351	56	4012	129	12301	259	12644	4506	3
26	Jo...	ht...	C1	Q3	358981	10611	70	7214	234	22585	299	24455	5913	
28	Jo...	ht...	C1	Q4	1032	154	0	66	0	177	0	169	32	
32	Re...	ht...	C1	Q4	6727	614	0	38	0	154	0	209	178	
34	A...	ht...	C1	Q4	2993	1318	0	21	23	145	20	185	67	0.2
35	N...	ht...	C1	Q4	2081	778	0	50	25	186	17	198	49	0.8
29	N...	ht...	C1	Q4	4631	634	2	89	7	314	15	339	158	0.2
31	K...	ht...	C1	Q4	15086	2474	6	134	22	501	27	626	387	0.38
27	H...	ht...	C1	Q4	2473	1118	10	85	49	332	45	208	38	0.76
30	Eu...	ht...	C1	Q4	3601	965	29	143	60	520	55	357	90	1.1
36	Jo...	ht...	C1	Q4	697	3539	61	42	140	137	166	80	0	0.17
33	St...	ht...	C1	Q4	63609	16285	323	2190	1165	7521	1245	8007	1100	0.5

Рисунок 11: аномалії кластерного аналізу

Журнал «Frontiers in Neuroscience»[24] належить до другого квартилю, але при цьому є провідним журналом в галузі неврології, що й забезпечило йому такі високі показники.

Ресурс четвертого квартилю «Studies in computational intelligence» має невелику кількість цитувань, відповідно, індекс Хірша всього 66, але при цьому має досить велику кількість робіт (більше тисячі для 2021 та 2022 років). Такі показники можуть бути пов’язані з типом ресурсу, так як це серія книг, а не науковий журнал, а також з тим, що вони стараються бути першими, хто публікує нові розробки та досягнення в різних сферах обчислювального інтелекту, відповідно генеруючи велику кількість контенту.

3. Квартиль на основі даних про публікації журналу

На основі даних про публікації квартиль журналу визначається з найкращою точністю 69% з використанням методу випадкового лісу:

Model	AUC	CA	F1	Precision	Recall
Random Forest	0.888	0.690	0.690	0.691	0.690
Logistic Regression	0.844	0.587	0.591	0.627	0.587
Decision Tree	0.808	0.641	0.641	0.645	0.641

Рисунок 12: передбачення квартилю на основі публікацій

Матриця невідповідностей виглядає наступним чином:

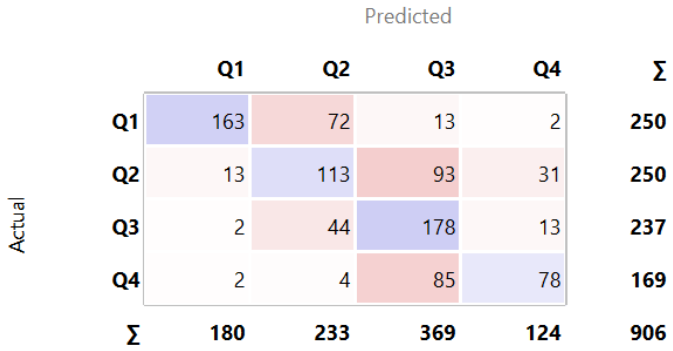


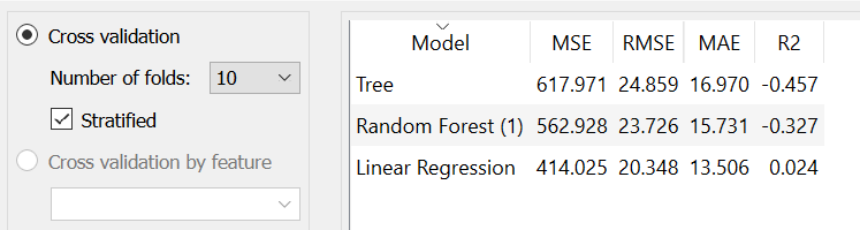
Рисунок 13: confusion matrix для квартилю на основі публікацій

З матриці видно, що у контексті робіт, межі між сусідніми квартилями журналів дещо змилис. Враховуючи метрики, які використовувалися для аналізу, а саме кількість цитувань роботи та вік публікації, можна зробити висновок, що статистика журналів 3 та 4 квартилів уже майже зрівнялася.

Кластерний аналіз не показав конфліктних зон.

4. Визначення віку публікації на основі її цитувань.

В результаті застосування трьох методів інтелектуального аналізу до даних про публікації, бачимо, що жодна з моделей не підходить для прогнозування віку публікації в залежності від кількості цитувань(рис.6).



Model	MSE	RMSE	MAE	R2
Tree	617.971	24.859	16.970	-0.457
Random Forest (1)	562.928	23.726	15.731	-0.327
Linear Regression	414.025	20.348	13.506	0.024

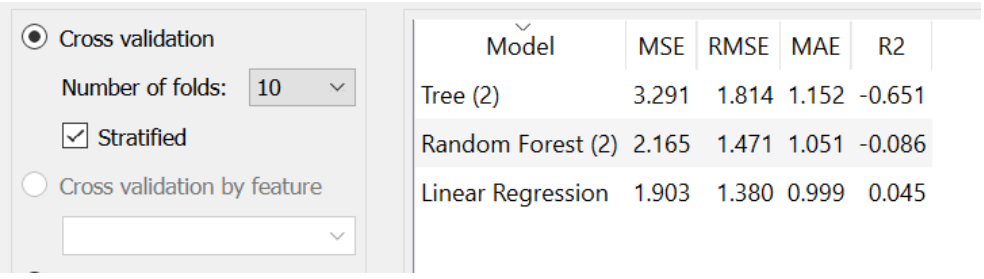
Рисунок 14: Прогнозування віку публікації

З цього можемо зробити висновок, що модель дещо контрінтуїтивна та не може знайти залежність між кількістю цитувань та віком публікації. Отже, ці дві метрики не залежать одна від одної.

У загальному, статистичні метрики для опису наукової роботи є дуже обмеженими, оскільки наявних метрик недостатньо для вагомих висновків про роботу. Для коретного опису роботи потрібно використовувати інвертовані індекси з часто вживаними термінами, які можна отримати, провівши детальний аналіз анотації, а іноді навіть усього тексту роботи. Це може бути темою для окремої наукової роботи.

5. Визначення стажу автора

Об’єкт автора містить більше метрик для певного аналізу, але визначенням стажу на основі цих метрик картина схожа до попередньої:



Model	MSE	RMSE	MAE	R2
Tree (2)	3.291	1.814	1.152	-0.651
Random Forest (2)	2.165	1.471	1.051	-0.086
Linear Regression	1.903	1.380	0.999	0.045

Рисунок 15: прогнозування стажу автора

В результаті модель показує що між стажем автора та його наукометричними показниками немає чіткої залежності, або ж взаємозв'язок між цими показниками потребує подальших досліджень.

Однією з ситуацій, коли стаж автора не залежить від його наукометричних показників, можна змоделювати таку в якій науковець щороку пише роботи, які ніхто не цитує, при цьому маючи високий показник кількості робіт, і маленький цитувань. Також варто згадати архетип основоположника, коли науковець першим публікує свіжу ідею у певній сфері, та вона починає активно цитуватися.

А той же час метрики можна легко передбачити на основі одна одної:

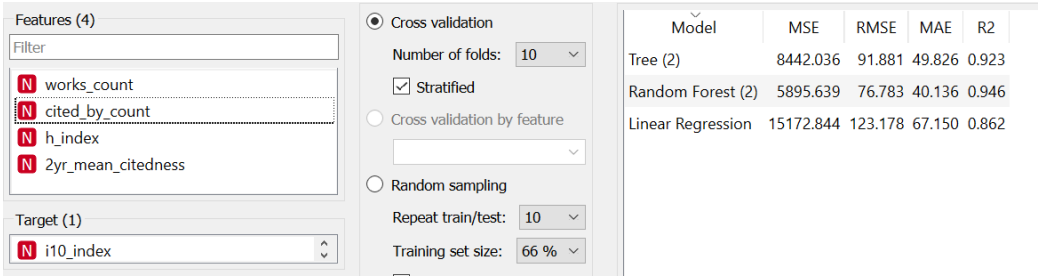


Рисунок 16: прогнозування індексу i10 автора

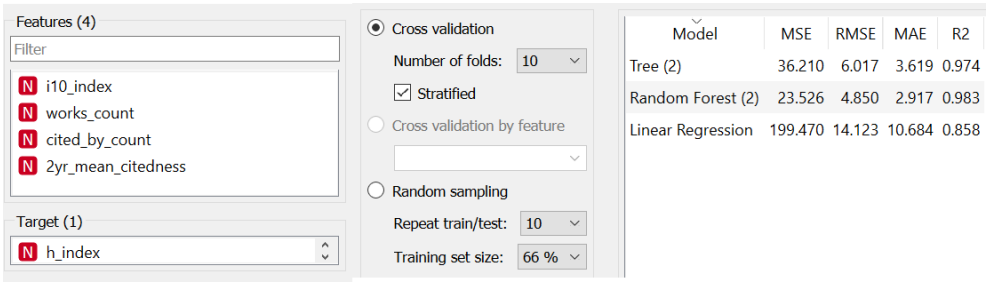


Рисунок 17: прогнозування індексу Хірша автора

6. Визначення основної галузі досліджень науковця методом кластеризації

Для проведення цього дослідження було використано інше джерело даних, а саме датасет з даними про 100000 науковців, що був використаний для дослідження, проілюстрованого журналом PlosBiology[31]. Використовуючи такі показники, як рік першої та останньої публікацій, кількість публікацій, кількість цитувань індекс Хірша, кластерний аналіз у 76,2% правильно розподілив авторів по кластерам відносно їх поля досліджень, в основному у близьких по сенсу кластерах:

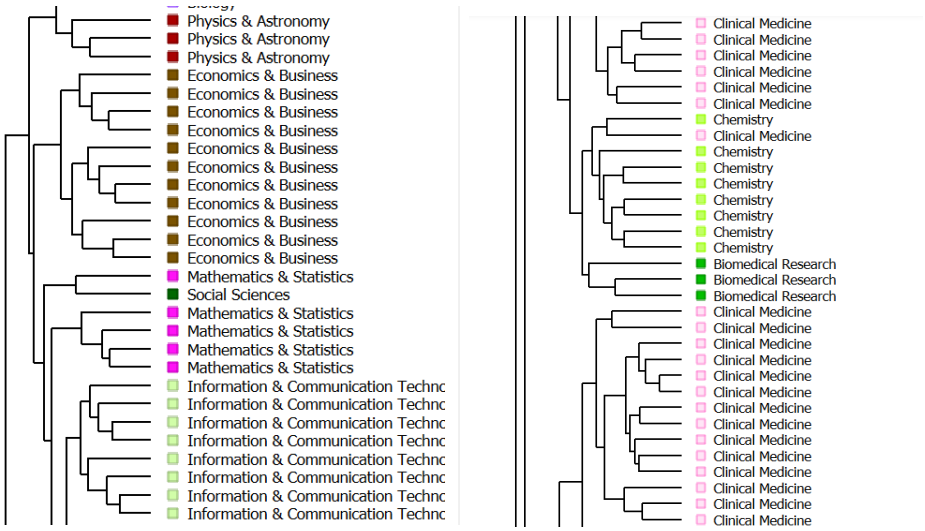


Рисунок 18: кластерний аналіз за галузями

ВИСНОВКИ

У даній дипломній роботі детально проаналізовано наукометричні показники для наукових робіт, авторів та наукових журналів, та зроблено спроби визначити або спростувати певні закономірності на їх основі. Основний акцент у роботі було зроблено на індекс Хірша, i10, імпакт-фактор, кількість цитувань, вік робіт та кількість публікацій. Крім цього, було детально проаналізовано квартильну систему ранжування журналів та вплив належності журналу до певного квартилю на показники його публікацій.

Було змодельовано кілька закономірнісних ситуацій залежності метрик журналів, авторів та публікацій та на основі даних бібліографічного каталогу OpenAlex зроблено вибірки даних необхідні для їх перевірки. З використанням методів інтелектуального аналізу даних, а саме дерева рішень, лінійної та логістичної регресії та випадкового лісу, було проаналізовано вибірки даних та спрощено чи підтверджено вищезазначені закономірності. У ході кластерного аналізу було виявлено та проаналізовано кілька аномальних ситуацій.

У даному дослідженні досить великий акцент було зроблено саме на науковий журнал як сутність, що поєднує собою авторів, публікації та інші журнали, публікації яких цитують дані. Результати дослідження можуть бути покращені із вдосконаленням збору даних, змінами в методах аналізу, додаванням нових метрик та відкриттям нових шляхів для дослідження.

Можна також продовжити аналіз в ключі наукових публікацій. Зокрема, можна розглянути релевантність наукових статей заданому пошуковому запиту. Одними із шляхів є аналіз наукових праць з використанням інвертованого індексу частовживаних термінів для оцінки релевантності робіт для певного запиту, чи на відповідність певній галузі знань. Це можна зробити задопомогою інтелектуального аналізу анотації, вступу чи навіть всього тексту роботи.

Список використаної літератури

1. "H-index" *Wikipedia*.[Електронний ресурс]. Режим доступу: https://en.wikipedia.org/wiki/H-index
2. "Measuring your research impact: i10-Index" *Cornell University Library*.[Електронний ресурс]. Режим доступу: https://guides.library.cornell.edu/impact/author-impact-10
3. "Impact factor" *Wikipedia*.[Електронний ресурс]. Режим доступу: https://en.wikipedia.org/wiki/Impact_factor
4. "OpenAlex API Documentation".[Електронний ресурс]. Режим доступу: https://docs.openalex.org/
5. "Orange Data Mining Tool".[Електронний ресурс]. Режим доступу: https://orangedatamining.com/workflows/
6. *Microsoft Academic* February 22, 2016.[Електронний ресурс]. Режим доступу: https://www.microsoft.com/en-us/research/project/academic/
7. *ORCID*.[Електронний ресурс]. Режим доступу: https://orcid.org/
8. *Crossref*.[Електронний ресурс]. Режим доступу: https://www.crossref.org/
9. *PubMed*.[Електронний ресурс]. Режим доступу: https://pubmed.ncbi.nlm.nih.gov/
10. *Scientific American*.[Електронний ресурс]. Режим доступу: https://www.scientificamerican.com/
11. *National Geographic*.[Електронний ресурс]. Режим доступу: https://www.nationalgeographic.com/
12. *Science*.[Електронний ресурс]. Режим доступу: https://www.science.org/
13. *Wired*.[Електронний ресурс]. Режим доступу: https://www.wired.com
14. *arXiv*.[Електронний ресурс]. Режим доступу: https://arxiv.org/
15. *Scopus*.[Електронний ресурс]. Режим доступу: https://www.scopus.com/home.uri

16. *WorldCat*.[Електронний ресурс]. Режим доступу: https://www.worldcat.org/

17. *Copernicus*.[Електронний ресурс]. Режим доступу: https://www.copernicus.eu/en

18. "Best publications by citations" *Google Scholar*.[Електронний ресурс]. Режим доступу: https://scholar.google.com/citations?view_op=top_venues

19. "Top 100 Highest Impact Factor Journals of 2022" *Journal impact factor* January 14, 2023.[Електронний ресурс]. Режим доступу: https://impactfactorforjournal.com/highest-impact-factor-journals/

20. "Journal Rank" *Scimago Journal & Country Rank*.[Електронний ресурс]. Режим доступу: https://www.scimagojr.com/journalrank.php

21. "Metrics to show journal, article & author influence" *Scopus*.[Електронний ресурс]. Режим доступу: https://www.elsevier.com/solutions/scopus/how-scopus-works/metrics

22. Olesya Mryglod, Yurij Holovatch, Ralph Kenna "Data Mining in Scientometrics: usage analysis for academic publications" *2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP)* (241-246) August 21, 2018.[Електронний ресурс]. Режим доступу: https://arxiv.org/pdf/1807.03353.pdf

23. PyAlex python library.[Електронний ресурс]. Режим доступу: https://pypi.org/project/pyalex/

24. *Frontiers in Neuroscience*.[Електронний ресурс]. Режим доступу: https://www.frontiersin.org/journals/neuroscience

25. "Data Mining Techniques".[Електронний ресурс]. Режим доступу: https://www.javatpoint.com/data-mining-techniques

26. "Data Mining Examples: Most Common Applications of Data Mining 2023" March 20, 2023.[Електронний ресурс]. Режим доступу: https://www.softwaretestinghelp.com/data-mining-examples

27."How is the Altmetric Attention Score calculated?" Sep 21, 2021.[Електронний ресурс]. Режим доступу: https://help.altmetric.com/support/solutions/articles/6000233311-how-is-the-altmetric-attention-score-calculated-

28. David B. Resnik, JD., PhD, Elizabeth Wager, PhD, and Grace E. Kissling, PhD "Retraction policies of top scientific journals ranked by impact factor" *PubMed* July 2015.[Електронний ресурс]. Режим доступу: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4511053/

29. "What Is Data Mining? How It Works, Benefits, Techniques, and Examples" *Investopedia* April 15, 2023.[Електронний ресурс]. Режим доступу: https://www.investopedia.com/terms/d/datamining.asp

30."List of Q4 journals" *List of Journals*.[Електронний ресурс]. Режим доступу: https://listofjournals.com/q4.php

31.Ioannidis JPA, Baas J, Klavans R, Boyack KW "A standardized citation metrics author database annotated for scientific field" *PLoS Biol* 17(8): e3000384 August 12, 2019.[Електронний ресурс]. Режим доступу: https://doi.org/10.1371/journal.pbio.3000384

ДОДАТКИ

Додаток 1: Діаграми і дані

(довідниковий)

	1	2	3	4	5	6	7	8
1	DISPLAY_NA...	CR...	CITED_BY_COUNT	2YR_MEAN_CITEDNESS	EXP_YEARS	H_INDEX	I10_INDEX	WORKS_COUNT
2	Solomon H. Snyder	2016	167 187 works	6.45	7 years	212	893	1 207
3	Tim D. Spector	2016	140 841 works	21.52	7 years	186	937	1 567
4	Kazuo Shinozaki	2016	124 157 works	8.35	7 years	175	646	1 208
5	Yury Gogotsi	2016	148 796 works	28.66	7 years	175	585	977
6	Cornelia M. van Duijn	2016	129 501 works	7.80	7 years	174	770	1 290
7	Zhuang Liu	2016	125 726 works	9.69	7 years	171	712	1 306
8	Klaus Müllen	2016	152 839 works	8.46	7 years	169	1 712	2 446
9	Terutaro Nakamura	2016	175 915 works	2.72	7 years	166	3 048	11 400
10	Gregg W. Stone	2016	116 189 works	17.78	7 years	164	1 046	2 029
11	Chad A. Mirkin	2016	132 215 works	1.59	7 years	164	759	1 329
12	Simon Baron-Cohen	2016	100 945 works	4.11	7 years	164	581	1 207
13	Kenneth S. Kendler	2016	129 490 works	4.20	7 years	163	854	1 400
14	Marc Ladanyi	2016	146 524 works	13.31	7 years	163	505	894
15	Lei Li	2020	222 561 works	3.94	3 years	161	4 820	13 775
16	Aviv Regev	2016	140 058 works	21.70	7 years	161	431	939
17	S. J. Chen	2016	166 206 works	2.23	7 years	159	2 733	9 316
18	Stefan D. Anker	2016	135 039 works	22.81	7 years	158	810	1 426
19	John H. Seinfeld	2016	95 063 works	5.88	7 years	158	721	1 321
20	T. Kobayashi	2018	174 347 works	2.74	5 years	157	2 852	10 263
21	C. D. Marsden	2016	100 490 works	0.00	7 years	157	824	1 243
22	G. Wang	2020	151 996 works	4.16	3 years	155	1 877	6 611
23	Frederik Barkhof	2016	108 126 works	4.92	7 years	154	1 041	2 245

Рисунок 19: Показники науковців

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	DISPLAY_NAME	QUAR	2YR_MEAN_CITEDNESS	H_INDEX	I10_INDEX	WORKS_COUNT	CITED_BY_COUNT	CITED_BY_COUNT_2021	CITED_BY_COUNT_2022	CITED_BY_COUNT_2023	WORKS_COUNT_2021	WORKS_COUNT_2022	WORKS_COUNT_2023
2	Nature	Q1	20.43	1 584	101 294	435 147	20 759 447	1 172 506	1 111 700	348 866	3 776	4 103	1 234
3	Cell	Q1	46.68	1 086	20 116	25 868	7 231 523	431 296	406 923	121 793	610	431	157
4	CA: A Cancer Journal for Clinicians	Q1	248.45	213	1 012	4 748	525 123	62 751	71 368	22 219	56	54	17
5	Nature Reviews Molecular Cell Biology	Q1	45.23	517	1 577	4 820	819 093	74 908	76 221	25 071	123	125	39
6	Quarterly Journal of Economics	Q1	11.23	470	3 037	6 677	905 006	51 378	43 179	13 834	48	44	20
7	The New England Journal of Medicine	Q1	43.32	1 290	39 371	127 930	9 505 467	568 615	527 961	156 533	1 146	1 123	345
8	Nature Medicine	Q1	37.48	646	6 844	14 205	1 826 096	180 367	158 966	49 179	610	571	193
9	Nature Reviews Genetics	Q1	19.79	445	1 609	4 649	651 502	56 804	53 649	16 434	121	123	33
10	Reviews of Modern Physics	Q1	41.84	541	2 856	3 551	1 142 588	69 370	60 770	18 021	33	38	10
11	The American Economic Review	Q1	9.71	475	7 269	10 759	1 118 481	69 442	57 155	19 182	129	131	36
12	Clinical and Experimental Immunology	Q2	5.37	188	9 898	15 743	412 023	17 805	16 976	4 638	168	127	41
13	Journal of General Virology	Q2	4.07	225	13 216	17 102	689 453	26 003	22 274	6 734	196	113	23
14	Cell death discovery	Q2	6.79	51	559	1 657	19 063	4 073	6 784	2 661	388	505	129
15	Mitochondrion	Q2	4.45	106	1 138	2 830	56 213	6 403	5 969	2 025	153	77	43
16	Molecular Brain Research	Q2	0.00	152	3 866	5 269	203 403	4 059	3 529	1 039	0	0	0
17	Plasmid	Q2	2.08	112	1 574	2 642	74 393	2 461	2 186	638	36	23	7
18	Frontiers in Neuroscience	Q2	4.44	146	4 786	14 860	213 370	43 007	49 131	16 549	1 874	2 374	726
19	Advances in Atmospheric Sciences	Q2	4.84	79	1 212	3 595	47 483	5 152	6 448	2 149	165	175	74
20	Journal of Applied Meteorology and Clima...	Q2	2.60	112	1 727	2 753	80 028	9 045	8 103	2 209	107	110	45
21	Cellular Immunology	Q2	3.94	144	6 472	12 187	252 427	7 957	7 940	2 304	125	107	38
22	Neurochemistry International	Q3	3.84	148	4 506	8 351	193 732	12 644	12 301	4 012	259	129	56
23	Clinical liver disease	Q3	1.89	28	106	1 216	9 610	1 310	1 517	581	172	122	24
24	Current research in structural biology	Q3	3.24	11	14	107	438	99	232	95	33	37	6
25	International journal of hepatology	Q3	2.08	40	183	317	6 020	630	610	159	10	13	4
26	BMC Structural Biology	Q3	0.00	64	419	586	17 989	1 256	1 111	332	0	0	0
27	Journal of translational autoimmunity	Q3	4.18	16	35	192	1 285	379	598	243	83	40	22
28	Human gene therapy. Clinical development	Q3	0.00	34	90	240	3 208	581	435	137	0	0	0
29	Molecular Membrane Biology	Q3	0.00	86	824	908	31 329	1 436	1 311	378	0	0	0
30	Brain connectivity	Q3	2.45	66	395	895	25 490	3 928	3 814	1 108	108	80	27
31	Journal of Neuroscience Methods	Q3	3.12	260	5 913	10 611	358 981	24 455	22 585	7 214	299	234	70
32	Hellenic Journal of Nuclear Medicine	Q4	0.76	15	38	1 118	2 473	208	332	85	45	49	10
33	Journal of relationships research	Q4	0.35	16	32	104	1 032	109	177	66	0	0	0
34	Neues Jahrbuch Fur Mineralogie-abhandl...	Q4	0.26	26	158	834	4 831	339	314	89	15	7	2
35	European annals of allergy and clinical im...	Q4	1.15	19	90	965	3 601	357	520	143	55	60	29
36	Kodai Mathematical Journal	Q4	0.38	41	387	2 474	15 086	626	501	134	27	22	6
37	Reliable Computing	Q4	0.00	36	178	614	6 777	269	154	38	0	0	0
38	Studies in computational intelligence	Q4	0.60	66	1 100	16 285	63 609	8 007	7 521	2 190	1 249	1 165	323
39	Acta Arachnologica	Q4	0.22	18	67	1 318	2 993	185	145	21	20	23	0
40	Norsk epidemiologi	Q4	0.82	19	49	778	2 081	198	186	60	17	25	0
	Double-click to edit Optics	Q4	0.18	5	0	3 539	697	80	137	42	166	140	61

Рисунок 20: Дані вибірки журналів із квартилями

	1	2	3	4	5	6	7	8	9	10	11	12
1	FIRSTYR	H18...	HM18...	LASTYR	NAME1	NAME2	NAME22	NPCIT...	NC9618 ...	NCSFL...	NPCIT...	NPSFL
2	1967	134	59.75829833	2019	Surgery	Oncology & Carcinogenesis	Clinical Medicine	48688	69123	35368	46776	620
3	1979	120	51.58072329	2019	Surgery	Gastroenterology & Hepatology	Clinical Medicine	52252	70521	33764	48778	1057
4	1973	113	45.99184205	2007	Surgery	Orthopedics	Clinical Medicine	39617	42901	16701	38472	259
5	1957	105	44.2500071	2018	Surgery	Gastroenterology & Hepatology	Clinical Medicine	33033	49705	34004	31568	1368
6	1968	88	44.15633364	2019	Surgery	Dentistry	Clinical Medicine	18565	30090	12903	17547	258
7	1982	102	43.95161211	2014	Surgery	Gastroenterology & Hepatology	Clinical Medicine	31096	41295	13550	29780	374
8	1977	100	43.92215122	2019	Surgery	Emergency & Critical Care Medicine	Clinical Medicine	30135	43564	9027	28618	309
9	1982	66	42.69484127	2019	Surgery	Orthopedics	Clinical Medicine	10503	17093	10445	10001	372
10	1984	103	42.49836381	2017	Surgery	Immunology	Clinical Medicine	32124	41177	16149	30375	240
11	1968	49	42.44722222	2000	Surgery	Oncology & Carcinogenesis	Clinical Medicine	6581	9438	7983	6500	158
12	1974	90	41.81812806	2016	Surgery	Gastroenterology & Hepatology	Clinical Medicine	18869	28563	14965	18042	353
13	1982	87	40.96676675	2019	Surgery	Medical Informatics	Clinical Medicine	26661	31801	15148	24894	548
14	1962	115	40.89166674	2019	Surgery	Oncology & Carcinogenesis	Clinical Medicine	34367	49510	15699	32815	332
15	1988	99	40.72063161	2019	Surgery	Gastroenterology & Hepatology	Clinical Medicine	36015	44005	8463	33136	334
16	1982	71	39.72233356	2018	Surgery	Microbiology	Clinical Medicine	14417	19294	13615	13937	277
17	1966	94	38.97996335	2019	Surgery	Gastroenterology & Hepatology	Clinical Medicine	23901	34055	15060	22855	359
18	1973	84	38.96598663	2019	Surgery	Gastroenterology & Hepatology	Clinical Medicine	23308	28497	10203	21918	448
19	1970	62	38.84920635	2017	Surgery	Orthopedics	Clinical Medicine	9191	13636	8912	8872	217
20	1987	84	38.82321975	2019	Surgery	Dentistry	Clinical Medicine	24717	31129	16496	23142	512
21	1986	92	38.77138898	2019	Surgery	Gastroenterology & Hepatology	Clinical Medicine	33801	41702	33307	32472	492
22	1987	101	38.26166415	2019	Surgery	Oncology & Carcinogenesis	Clinical Medicine	28454	40818	15762	26932	326
23	1978	94	38.08676456	2019	Surgery	Gastroenterology & Hepatology	Clinical Medicine	31513	44057	22774	29762	713

Рисунок 21: Дані із галузями(із PLOS Biology)

Додаток 2: Програмний код

(обов'язковий)

```python
import csv
from datetime import datetime
from pyalex import Works, Authors, Sources, Institutions, Concepts, Publishers

# Q1-Q4 journals
q1_names = ['Nature', "Cell", "CA: A Cancer Journal for Clinicians",
            "Nature Reviews Molecular Cell Biology", "Quarterly Journal of Economics",
            "The New England Journal of Medicine", "Nature Medicine",
            "Nature Reviews Genetics", "Reviews of Modern Physics", "The American Economic Review"]
q2_names = ["Clinical and Experimental Immunology", "Journal of General Virology",
            "Cell death discovery", "Mitochondrion", "Molecular Brain Research",
            "Plasmid", "Frontiers in Neuroscience", "Advances in Atmospheric Sciences",
            "Journal of Applied Meteorology and Climatology", "Cellular Immunology"]
q3_names = ["Neurochemistry International", "Clinical liver disease",
            "Current research in structural biology", "International journal of hepatology",
            "BMC Structural Biology", "Journal of translational autoimmunity",
            "Human gene therapy. Clinical development", "Molecular Membrane Biology",
            "Brain connectivity", "Journal of Neuroscience Methods"]
q4_names = ["Hellenic Journal of Nuclear Medicine", "Journal of relationships research",
            "Neues Jahrbuch Fur Mineralogie-abhandlungen",
            "European annals of allergy and clinical immunology", "Kodai Mathematical Journal",
            "Reliable Computing", "Studies in computational intelligence",
            "Acta Arachnologica", "Norsk epidemiologi =", "Journal of Applied Optics"]


def print_to_csv(values, filename):
    with open(filename, 'w', encoding='utf8', newline='') as output_file:
        fc = csv.DictWriter(output_file, fieldnames=values[0].keys())
        fc.writeheader()
        fc.writerows(values)


def find_sources_detailed(names, q):
    journals = []
    for n in names:
        j = Sources().filter(display_name=n) \
            .select(['display_name', 'id', "cited_by_count", "works_count", "counts_by_year",
                     "summary_stats", "type"]) \
            .get()[0]
        # print(n, j)
        if q:
            j['quartile'] = "Q" + (str)(q)
        j["works_count_2023"] = j['counts_by_year'][0]['works_count']
        j["cited_by_count_2023"] = j['counts_by_year'][0]['cited_by_count']
        j["works_count_2022"] = j['counts_by_year'][1]['works_count']
        j["cited_by_count_2022"] = j['counts_by_year'][1]['cited_by_count']
        j["works_count_2021"] = j['counts_by_year'][2]['works_count']
        j["cited_by_count_2021"] = j['counts_by_year'][2]['cited_by_count']
        del j['counts_by_year']
        j["h_index"] = j["summary_stats"]["h_index"]
        j["i10_index"] = j["summary_stats"]["i10_index"]
        j["2yr_mean_citedness"] = j["summary_stats"]["2yr_mean_citedness"]
        del j['summary_stats']
        journals.append(j)
    return journals
```

```python
def find_sources(names, q):
    journals = []
    for n in names:
        j = Sources().filter(display_name=n) \
            .select(['display_name', 'id', 'homepage_url']) \
            .get()[0]
        print(n, j)
        j['quartile'] = "Q" + (str)(q)
        journals.append(j)
    return journals


def find_papers_by_sources(journals):
    total_works = []

    for j in journals:
        works = Works().filter(best_oa_location={"source": {"id": j['id']}},
                               publication_year=">2010") \
            .select(['title', 'publication_year', 'cited_by_count', 'type']) \
            .sort(cited_by_count="desc").get(per_page=200)
        # 'concepts',  'ngrams_url'
        for w in works:
            w['source'] = j['display_name']
            w['quartile'] = j['quartile']
            w['age_year'] = 2023 - w['publication_year']
        total_works += works

    return total_works


def get_detailed_works_quartile():
    journals = find_sources(q1_names, 1)
    total_works = []
    for j in journals:
        pages = Works().filter(best_oa_location={"source": {"id": j['id']}},
                               publication_year=">2000") \
            .select(['title', 'publication_year', 'cited_by_count', 'type']) \
            .sort(cited_by_count="desc").paginate(per_page=200, n_max=1000)
        # 'concepts',  'ngrams_url'
        for works in pages:
            for w in works:
                w['source'] = j['display_name']
                w['age_year'] = 2023 - w['publication_year']
            total_works += works
    print_to_csv(total_works, "q1-papers-smaller.csv")

```

```
104
105 def generate_pages_authors():
106     pager = Authors() \
107         .filter(from_created_date="<2023-01-01", cited_by_count=">0",
108                 works_count=">0") \
109         .select(['cited_by_count', 'display_name',
110                  "works_count", "created_date", "summary_stats"]) \
111         .paginate(per_page=200, n_max=5000)  # 5000
112
113     authors = []
114     for page in pager:
115         for a in page:
116             a["exp_years"] = 2023 - datetime.strptime(a["created_date"], '%Y-%m-%d').year
117             a["h_index"] = a["summary_stats"]["h_index"]
118             a["i10_index"] = a["summary_stats"]["i10_index"]
119             a["2yr_mean_citedness"] = a["summary_stats"]["2yr_mean_citedness"]
120             del a['summary_stats']
121             authors.append(a)
122
123     print_to_csv(authors, "authors_stats.csv")
124     print("successfully saved to the authors_stats.csv")
125
126
127 def generate_pages_works():
128     pager = Works().select(['display_name', 'id', "cited_by_count", "publication_year", "type"]) \
129         .paginate(per_page=200, n_max=3000)
130
131     works = []
132     for page in pager:
133         for w in page:
134             w["age"] = 2023 - w["publication_year"]
135             works.append(w)
136     print_to_csv(works, "works_stats.csv")
137     print("successfully saved to the works_stats.csv")
138
139
140 def generate_pages_sources():
141     pager = Sources().select(['display_name', 'id', "cited_by_count", "works_count",
142                               "summary_stats", "type"]) \
143         .paginate(per_page=200, n_max=1000)
144
145     journals = []
146     for page in pager:
147         for j in page:
148             j["h_index"] = j["summary_stats"]["h_index"]
149             j["i10_index"] = j["summary_stats"]["i10_index"]
150             j["2yr_mean_citedness"] = j["summary_stats"]["2yr_mean_citedness"]
151             del j['summary_stats']
152             journals.append(j)
153     print_to_csv(journals, "journals_stats.csv")
154     print("successfully saved to the journals_stats.csv")
155
```

```
156
157 def get_journals_detailed_data():
158     journals = []
159     journals += find_sources_detailed(q1_names, 1)
160     journals += find_sources_detailed(q2_names, 2)
161     journals += find_sources_detailed(q3_names, 3)
162     journals += find_sources_detailed(q4_names, 4)
163     print_to_csv(journals, "journals_stats_with_quartiles.csv")
164     print("successfully saved to the journals_stats_with_quartiles.csv")
165
166
167 def get_journals_papers_data():
168     journals_l = []
169     journals_l += find_sources(q1_names, 1)
170     journals_l += find_sources(q2_names, 2)
171     journals_l += find_sources(q3_names, 3)
172     journals_l += find_sources(q4_names, 4)
173
174     journals_t = []
175     journals_t += find_sources([test_names[0]], 1)
176     journals_t += find_sources([test_names[1]], 2)
177     journals_t += find_sources([test_names[2]], 3)
178     journals_t += find_sources([test_names[3]], 4)
179
180     learning_data = find_papers_by_sources(journals_l)
181     testing_data = find_papers_by_sources(journals_t)
182
183     print_to_csv(learning_data, "quartiles_data_scatter.csv")
184     # print_to_csv(testing_data, "quartiles_test_data_scatter.csv")
185
186
187 if __name__ == '__main__':
188     get_detailed_works_quartile()
189
```