**Чумадевська Христина Василівна,**

Студентка, Західноукраїнський національний університет, м. Тернопіль

**Загородня Діана Іванівна**

к.т.н., доцент, Західноукраїнський національний університет, м. Тернопіль

Orcid ID: http://orcid.org/0000-0002-9764-3672

**МЕТОД КЛАСТЕРИЗАЦІЇ ДОКУМЕНТІВ ВДОСКОНАЛЕНИМ МЕТОДОМ K-СЕРЕДНІХ**

Системи управління документами (DMS) – це системи, які пропонують такі послуги, як зберігання, керування версіями, метаданими, безпека, а також можливості індексування та пошуку. Велика кількість документів може бути автоматично згрупована в класи документів, які містять схожу інформацію. Для цього застосовують методи кластеризації.

Елементами DMS для документів є: інтеграція, вилучення метаданих, збір, перевірка, індексування, зберігання, пошук, розповсюдження, безпека, робочий процес, співпраця, керування версіями, пошук, публікація та відтворення.

Процес кластеризації документів складається з декількох етапів. Спочатку виконується попередня обробка наборів даних, які надають набір токенів для моделі векторного простору (VSM). VSM – це процес пошуку, який працює за моделлю Tf-Idf. Для обчислення відстані між різними кластерами використовуються міри подібності.

У роботі для кластеризації документів використовується дерево подібності документів, яке виокремлює послідовність фраз і слів у документах. Для кластеризації сегментів на основі схожості використовується підхід побудови кластерів розташованих на максимально великій відстані [1]. Алгоритм генетичної кластеризації використовується для вирішення проблеми агрегації кластерів. Процес кластеризації зображений на рисунку 1.

Попередня обробка виконується над звичайними текстовими документами і генерує набір токенів для виводу в VSM. Ця методика забезпечує оптимальну якість кластерів. Основні етапи препроцесорної обробки полягають у наступному:

1. Фільтрація: для видалення розділових знаків і спеціальних символів.

2. Токенізація: для розбиття токенів на окремі слова та токени.

3. Зупинити видалення слів: слова, що не мають значення, видаляються.

4. Стеммінг: утворюється основна форма слів.

5. Обрізка: для видалення низькочастотних слів.

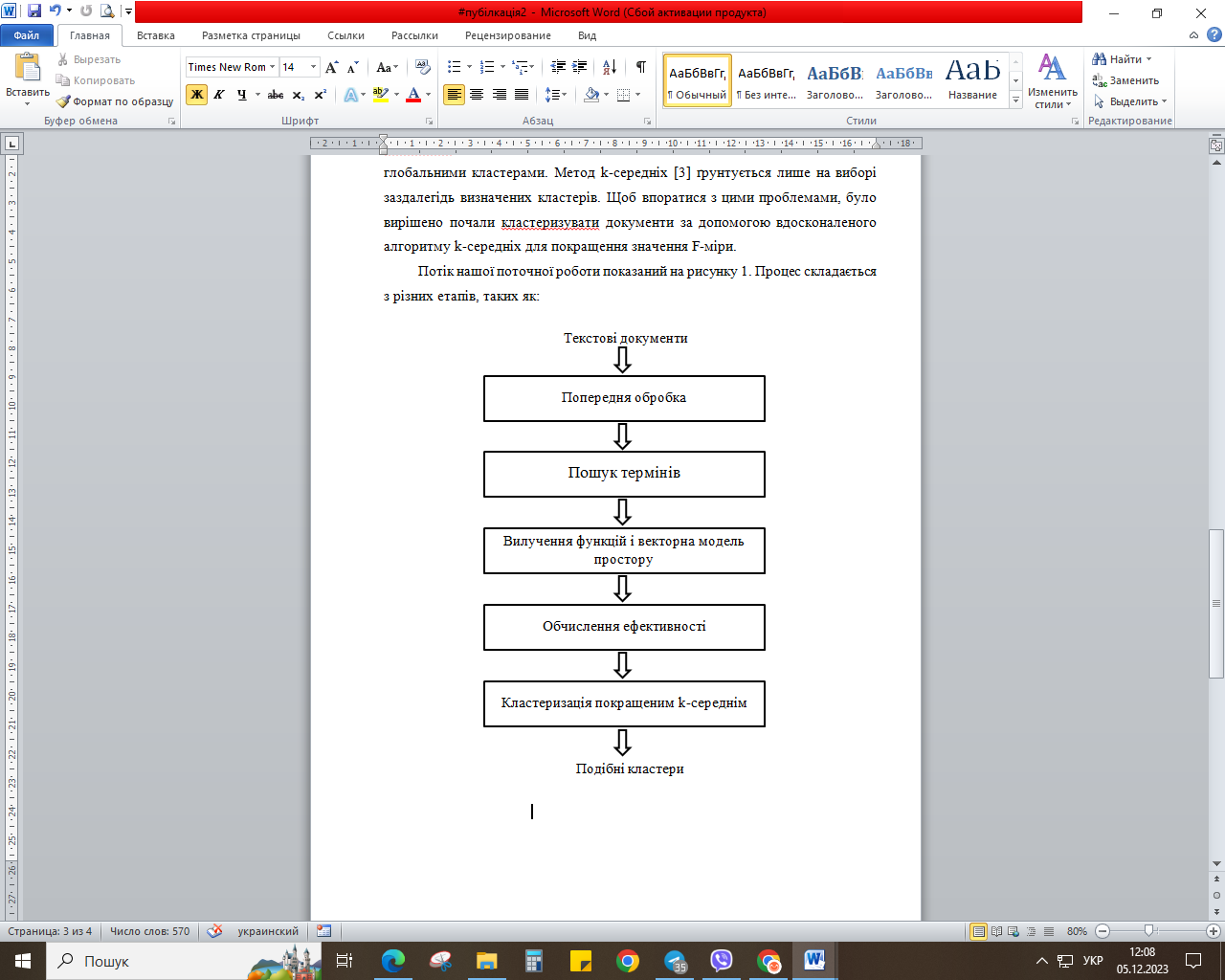


Рисунок 1 – Процес кластеризації документів

Пошук термінів знаходить ексклюзивні терміни з кожної доступної категорії. Кожному терміну присвоюється порогове значення як вага. Частота терміна tf (*i, j*) – це кількість разів, коли термін *i* зустрічається в документі *j*. Якщо частота терміна tf більша за порогове значення, то значення додається, інакше відхиляється.

Вилучення функцій використовується для видалення набору ключових слів з документів. VSM – це техніка пошуку в інтелектуальному аналізі даних, також відома як модель «частота терміна - частота документа». Це стандартна алгебраїчна модель представлення тексту. Кожен документ представляється у вигляді n-вимірного вектора за допомогою вектора ознак. Значення кожного елемента вектора відображає важливість відповідної ознаки в документі. За допомогою цієї моделі схожість між документами вимірюється шляхом обчислення відстані між векторами документів. Якщо документи містять однакові ключові слова, вони вважаються схожими. Частота терміна нормалізується відносно максимальної частоти всіх термінів, що зустрічаються в документі.

Також обчислюються евклідова схожість, оцінка ефективності кластеризації (вимірюється за допомогою F-міри), точність (визначається як відношення кількості позитивних результатів до кількості позитивних результатів плюс кількість хибних результатів), відгук (визначається як відношення кількості істинно позитивних спрацьовувань до кількості істинно позитивних спрацьовувань плюс кількість помилково негативних результатів).

Наступний крок – це покращений метод k-середніх. Удосконалена кластеризація за методом k-середніх використовує алгоритм на основі розбиття. Одним з таких алгоритмів є Bisecting K Means Methods [2], який починає з розбиття всієї множини точок на два кластери, вибирає один із них, поділяє його, а потім повторює цей процес, поки не створить k кластерів. Гібридна бісектриса k-середніх використовує комбінацію бісектриси k-середніх та ієрархічного алгоритму розбиття для отримання оптимальних кластерів. Цей удосконалениий алгоритм спрямований на автоматичну кластеризацію та усунення недоліків методу K-Means [3].

Робота виконується на наборі даних mini\_newsgroups [4]. Для порівняння класифікація виконувалась методом k-середніх та покращеним методом k-середніх на 300 документах з mini\_Newsgroup. Результати представлені на рисунку 2.

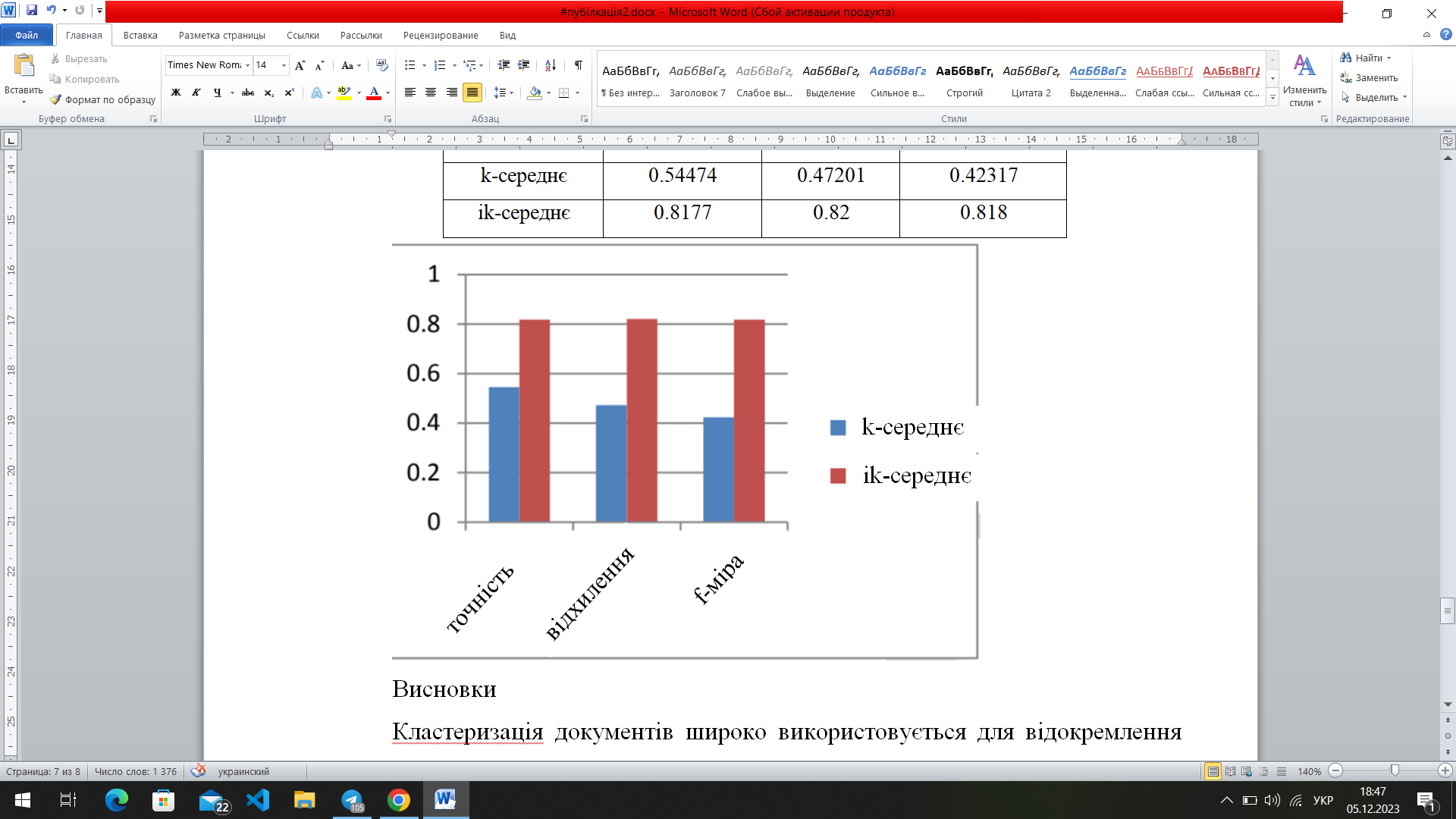


Рисунок 2 – Результати класифікації документів методом k-середніх та удосконаленим методом

F-міра має більше значення для запропонованого алгоритму порівняно з існуючим алгоритмом. Також значення точності та відхилення є кращими для запропонованого алгоритму порівняно з існуючим алгоритмом.

На рисунку 3 представлено порівняння цих методів за часом. Існуючий метод потребує більшого часу виконання, ніж запропонований.

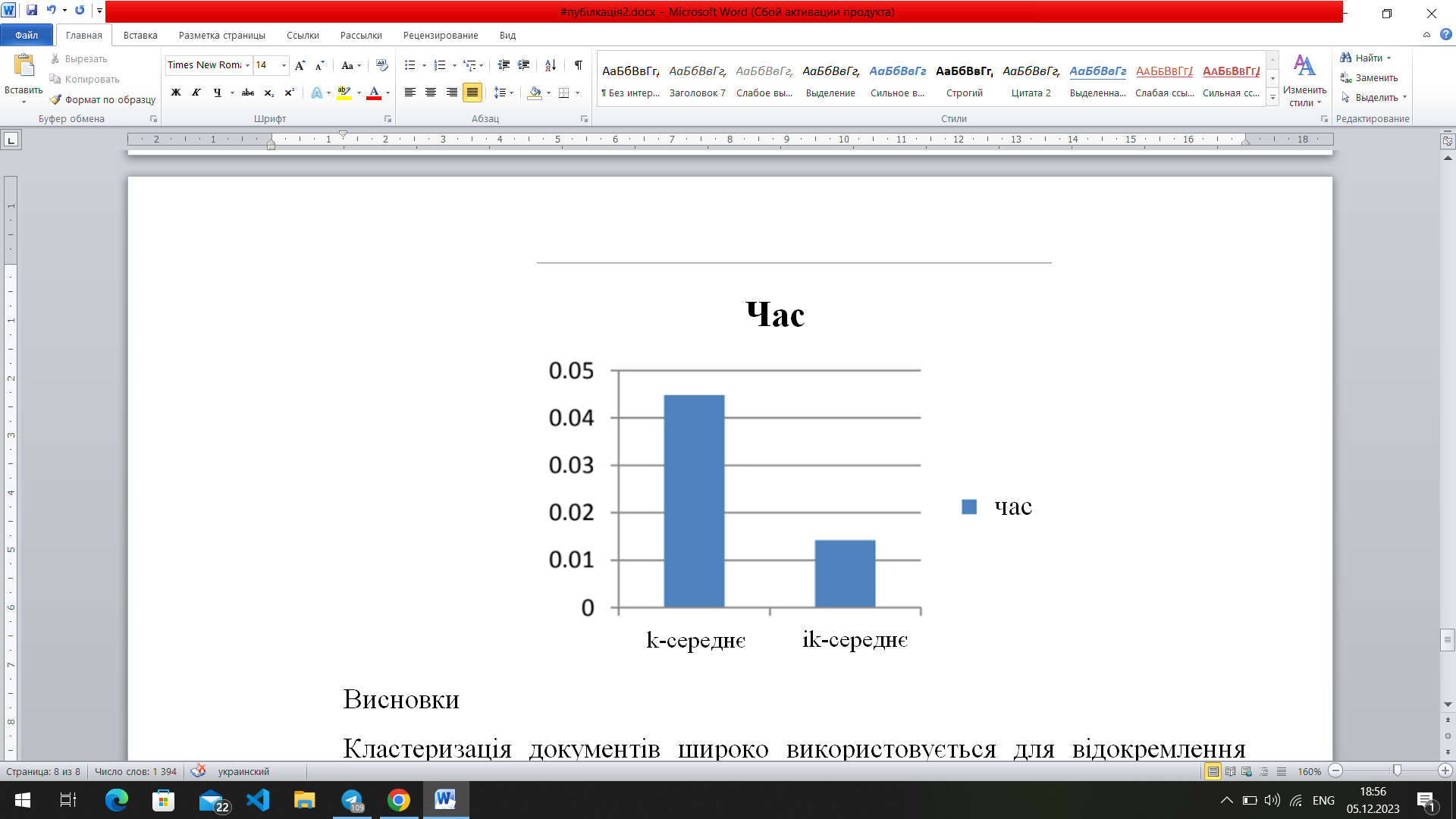


Рисунок 3 – Порівняння методів за часом виконання

В умовах інтенсивної генерації документів, кластеризація стає необхідним інструментом для структурування, управління та зручного доступу до інформації, а також може служити основою для впровадження різноманітних технологій аналізу та автоматизації обробки даних. Традиційний алгоритм k-середніх добре працює з певними документами, а центроїди обираються випадково. У запропонованому алгоритмі центроїди прогнозуються вручну. Експериментальні результати показали, що покращений алгоритм k-середніх працює краще, ніж існуючий алгоритм з точки зору точності, f- міри та часу.

**Література**

1. Rupesh Kumar Mishra, Knika Sain, Sakshi Bagri, «Text Document Clustering On The Basis Of Inter Passage Approach By Using K-Means», International Conference On Computing, Communication And Automation,(ICCCA- 2015), may 15-16, pp:110-113,IEEE, 2018.
2. Pradeep Rai. Shubha Singh, «A Survey Of Clustering Techniques», International Journal Of Computer Applications, volume 7,pp:1-5, 2020.
3. Improvement of K-means Cluster Quality by Post Processing Resulted Clusters. URL: https://www.sciencedirect.com/science/article/pii/S1877050922000096, 2022 .
4. UC Irvine Machine Learning Repository. URL: https://archive.ics.uci.edu/.