ДОДАТОК А

Апробація результатів роботи. Тези 1

УДК 004.85

ДОСЛІДЖЕННЯ МЕТОДІВ СКАНУВАННЯ ТА ІНДЕКСАЦІЇ ВЕБ-САЙТІВ ПОШУКОВОЮ СИСТЕМОЮ GOOGLE

Мартиненко А.О., здобувач вищої освіти *email: andrii.martynenko1@nure.ua*
Мельнікова Р.В., к.т.н. *email: roksana.melnikova@nure.ua*
Харківський національний університет радіоелектроніки

Актуальність та постановка проблеми. Спостерігаючи інтенсивну діджиталізацію усіх аспектів людської діяльності, можна прийти до висновку, що кожен бізнес обов'язково повинен мати власний веб-сайт для підвищення впізнаваності бренду та залучення вагомого потоку клієнтів. Метою дослідження є аналіз існуючих та проєктування ефективного рішення для аналізу проіндексованих веб-сайтів пошуковою системою Google, яке базується на головних та вагомих факторах ранжування [1], враховуючи індивідуальні особливості різних ніш веб-сайтів (медичної, юридичної, електронної комерції тощо) із залученням машинного навчання. Тому, дане дослідження є затребуваним та актуальним для будь-якого виду бізнесу, оскільки його результатами є покращення ранжування веб-сайтів з метою їх якісного та стрімкого просування до ТОП-3 у пошукової видачі Google.

Основні матеріали дослідження. Предметною областю даної роботи є оптимізація веб-сайтів для покращення ранжування в пошуковій системі Google з ціллю органічного просування, використовуючи засоби машинного навчання. А саме, збільшення органічного трафіку на веб-сайті шляхом підняття позицій в результатах пошуку за відповідними та релевантними до тематики пошуковими запитами.

SEO-оптимізація – це комплексний процес робіт, основні етапи якого: технічна оптимізація веб-сайту, створення якісного та унікального контенту, впровадження структурованих даних у форматі JSON та побудова стратегії зовнішніх посилань з релевантних ресурсів. Дані етапи узагальнені та розгалужуються на суттєву кількість підзадач, яка прямо-пропорційно залежить від першочергових факторів, таких як: вік домену, відсутність або наявність штрафних санкцій від пошукової системи Google, CMS або JavaScript-фреймворки за допомогою яких створено сайт, враховуючи технологію рендерингу сторінок: CSR або SSR, а також наявний посилальний профіль. Починаючи роботу з маленьких кроків та розробивши правильну стратегію просування можна отримати успішний результат з покращенням позицій в пошуковій видачі, що має значний вплив на конкурентоспроможність в умовах насиченого ринку.

На даний момент штучний інтелект та його підгалузь машинне навчання мають значний вплив на SEO кількома способами [2]. Пошукові системи використовують машинне навчання для аналізу отриманих даних на відповідність до контексту пошукових запитів, завдяки чому визначають рейтинг веб-сторінок. На рисунку 1, зображено схему роботи Google-бота [3] із урахуванням більш складного сканування JavaScript-сайтів. Бот відвідує лише відкриті для сканування сторінки в індивідуальному файлі robots.txt з каталогу, де розміщено сайт. До кожного сайту з наявною позначкою User-agent: Googlebot з дозволом сканування для Google-бота, або позначкою User-agent: *, для всіх можливих ботів, відбувається наступний процес сканування. На рисунку 1 наведено процес сканування сайтів, що створені засобами JavaScript-фреймворків та бібліотек.

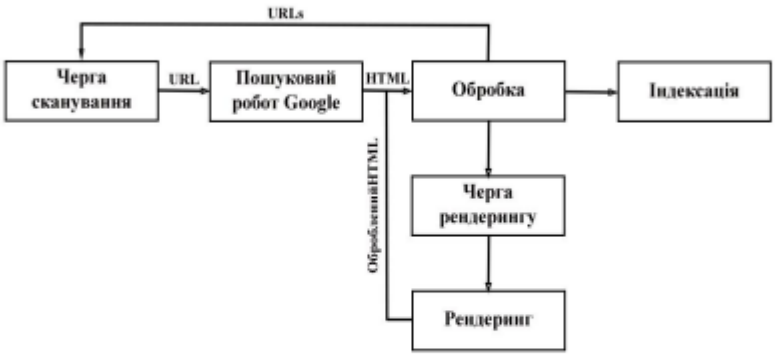


Рисунок 1 – Схема роботи Google-бота (при скануванні сайтів)

Основна відмінність від сайтів, створених з використанням CMS, таких як: WordPress або Shopify, полягає у типі рендерингу сторінок. На даний момент левова частка самописних сайтів створено без урахування рекомендованих умов від пошукових систем. До прикладу, веб-сайт створений з використанням стандартної бібліотеки для створення користувацьких інтерфейсів – React має технологію Client Side Rendering, в такому випадку відбувається двоетапний процес індексації. Спочатку Google-бот завантажує HTML-документ та програмний код JavaScript, після чого з використанням внутрішнього двигуна та середовища Chromium, виконує JavaScript-код, щоб отримати остаточний контент на сторінці. Така послідовність створює затримку між скануванням та індексацією веб-сайту.

Варто зазначити, що це не заперечує можливість сканування веб-сайту, але при такому сценарії витрачається в 2 рази більше краулінгового бюджету сайту [4], що призводить до зменшення ймовірності сканування, якомога більшої кількості сторінок.

Тому, перехід з наведеної бібліотеки React на JavaScript-фреймворк Next.js, який в основі використовує технологію Server Side Rendering та створений на базі React, дозволить зменшити кількість запитів до сайту та призведе до пришвидшення процесу сканування сторінок, оптимізувавши краулінговий бюджет, який визначається за двома факторами – краулінгова здатність (Crawl Capacity Limit), що відповідає за максимальну кількість запитів, яку Google-бот може здійснити на сервер сайту, не створюючи проблем для його продуктивності та краулінгова потреба (Crawl Demand), що базується на важливості та свіжості контенту на сторінках сайту.

Також зараз дуже широко залучається машинне навчання в інструментах, що дозволяють аналізувати пошуковий трафік та дані про залучення користувача. Що в свою чергу дуже спрощує роботу для SEO-фахівця, допомагаючи визначити та зрозуміти напрямки, які варто вдосконалити. Оскільки Google ретельно приховує роботу своїх алгоритмів та ніколи не повідомляє про їх оновлення або створення нових, перед SEO-спеціалістами постає задача методом досліджень, аналізу та практичних експериментів – визначити вплив нововведень на старі опрацювання та виявити нові тренди оцінювання пошуковою системою.

До прикладу, Ahrefs [5] – компанія, заснована українцем, яка надає потужний інструмент для дослідження великої кількості факторів оцінки веб-сайтів, що були досліджені за тривалий період часу та продовжують досліджуватись з постійними оновленнями пошукових алгоритмів Google. Розглянемо процес сканування пошукового бота від Ahrefs на рисунку 2.

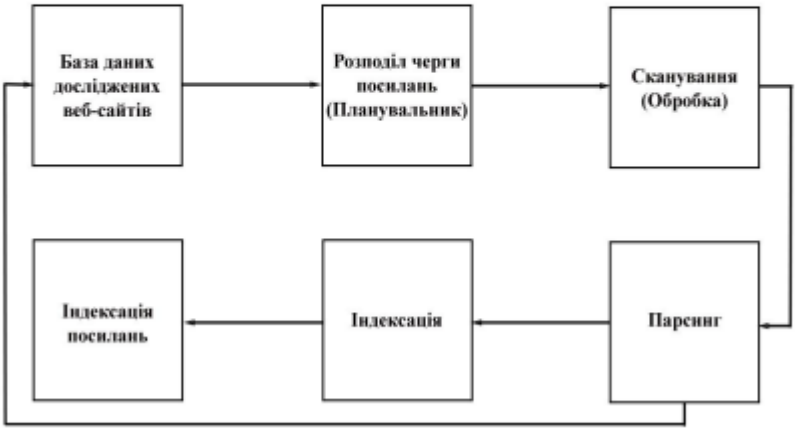


Рисунок 2 – Схема роботи бота Ahrefs (при скануванні сайтів)

Сканування обраного користувачем сайту починається з модифікації великої бази даних, що складається із вже відомих сервісу URL, адреси сторінок сайту перенаправляються до планувальника, який створює чергу сторінок, які мають бути просканованими. Коли URL-адреси сторінок будуть готові до аналізу, Ahrefs-бот надсилає їх до сканеру та завантажує вміст сторінок для обробки, суворо дотримуючись правил дозволу або заборони, встановлених у файлі robots.txt, який вже згадувався раніше. Робот-сканер доставляє необроблені дані до аналізатора, який робить парсинг сторінок та витягує посилання на цій сторінці, такі як: заголовок та інші відповідні метадані, після чого дані надсилаються на індексацію. В результаті чого, отримані дані додаються до індексу посилань і стають доступним у різних звітах інструменту Ahrefs.

Висновки. Досліджуючи методи сканування наведених ботів, можемо зробити висновок, що Google-бот постійно вдосконалюється та безпосередньо покладається на машинне навчання для вдосконалення обробки динамічного контенту, створеного JavaScript, обробляючи його у власному середовищі Chromium, завдяки чому може сканувати сайти побудовані на різних технологіях рендерингу, не лише статичні багатосторінкові, а також динамічні односторінкові (Single Page Application). Але на даний момент цей процес є ресурсозатратним для нього, через затримки між скануванням та індексацією сайтів.

В той же час, розглянутий бот від Ahrefs, працює як статичний краулер, що отримує дані лише із повністю завантаженої HTML-розмітки без можливості виконання динамічного JavaScript-коду. Проте він не витрачає додаткові ресурси на рендеринг, що дозволяє значно пришвидшити процес сканування, здатний аналізувати сайти, які використовують JavaScript, але тільки статичні повністю завантажені сторінки.

Таким чином, Google-бот підходить для сучасних JavaScript-орієнтованих сайтів, навіть, які створені з використанням CSR, завдяки можливості виконувати динамічний JavaScript-код, проте це затратно і повільно. Ahrefs-бот швидший і оптимізований для традиційного SEO-аналізу, але не обробляє динамічний контент, лише повністю завантажений HTML-документ.

Ahrefs-бот, хоч і не виконує динамічний JavaScript, але може використовувати машинне навчання для аналізу структур посилань, виявлення шаблонів поведінки посилань та прогнозування якості посилань на основі алгоритмів класифікації або кластеризації.

У підсумку, Google-бот є потужним інструментом для індексації сучасних веб-сайтів, побудованих на динамічних технологіях, а Ahrefs-бот — це швидке і ефективне рішення для аналізу посилань та SEO-оптимізації. Машинне навчання посилює можливості обох, хоча акценти у їх використанні суттєво різняться.

Для проектування ефективного рішення для аналізу проіндексованих веб-сайтів пошуковою системою Google варто враховувати специфіку роботи Google-бота, який виконує JavaScript для індексації динамічного контенту. Потрібно забезпечити оптимізацію рендеринга, використовуючи підходи, що дозволяють обробляти SSR (Server Side Rendering) та CSR (Client Side Rendering), щоб мінімізувати затримки та ресурсні витрати на краулінг. Важливо інтегрувати машинне навчання для семантичного аналізу контенту та пріоритизації важливих сторінок, адже Google-бот орієнтується на контекст і поведінкові сигнали. Аналіз посилальної структури також може бути підсилений через підходи, які використовуються в інструменті Ahrefs, такі як класифікація та кластеризація посилань за допомогою машинного навчання, що дозволить оцінювати якість посилань та метаданих для прогнозування їхнього впливу на ранжування в пошуковій системі.

Список використаних джерел:

1. Google Ranking Factors [Електронний ресурс] // Semrush. – 2024. – Режим доступу до ресурсу: https://go.semrush.com/Ranking-Factors.html#form.
2. The Impact of Machine Learning on SEO [Електронний ресурс] // Market Brew. – 2024. – Режим доступу до ресурсу: https://marketbrew.ai/the-impact-of-machine-learning-on-seo.
3. Understand the JavaScript SEO basics [Електронний ресурс] // Google. – 2024. – Режим доступу до ресурсу: https://developers.google.com/search/docs/crawling-indexing/javascript/javascript-seo-basics.
4. Crawl Budget [Електронний ресурс] // Ahrefs. – 2024. – Режим доступу до ресурсу: https://ahrefs.com/seo/glossary/crawl-budget.
5. How to use Ahrefs [Електронний ресурс] // Ahrefs. – 2024. – Режим доступу до ресурсу: https://ahrefs.com/academy/how-to-use-ahrefs.

ДОДАТОК Б

Апробація результатів роботи. Тези 2

УДК 004.85

КЛАСТЕРИЗАЦІЯ КОНТЕНТУ З ВИКОРИСТАННЯМ АЛГОРИТМІВ МАШИННОГО НАВЧАННЯ

Мартиненко А.О.
e-mail: andrii.martynenko1@nure.ua
Харківський національний університет радіоелектроніки, каф.ПІ
М. Харків, Україна

Examination of the use of machine learning for website content clustering. Clustering the semantic core allows for grouping similar keywords into distinct categories, thereby simplifying subsequent analysis and guiding the creation of relevant content. It is proposed to use K-means, Mini-batch K-means, Deep Embedded Clustering and Spectral Clustering to identify thematically similar groups of queries. The approach helps to uncover hidden structures and themes in large keyword sets, enabling more precise SEO strategies and an organized content architecture. Experimental evaluations highlight the effectiveness of each algorithm across various data volumes, noting differences in accuracy, computational demands, and interpretability.

В сучасних умовах стрімкого розвитку веб-технологій та загальної цифровізації якісна пошукова оптимізація (SEO) [1] стає одним із ключових чинників для успішного просування сайтів у пошукових системах. Конкуренція за високі позиції у видачі зростає з кожним днем, що зумовлює необхідність постійного вдосконалення методів та засобів оптимізації. Однією з фундаментальних складових оптимізації сайтів в пошукових системах є опрацювання семантичного ядра – це процес підбору та структурування ключових слів, що близькі за написанням та змістом, які відображають зміст та тематику веб-сайту за пошуковими намірами користувача.

Зазвичай семантичне ядро містить велику кількість ключових фраз, котрі можуть перетинатися за змістом, бути багатозначними або частково дублюватися. Внаслідок цього виникає проблема систематизації та ефективного групування ключових слів, що необхідно для побудови релевантних та тематичних посадкових сторінок, підготовки якісного контенту та визначення подальшої пріоритетності робіт з оптимізації сайту для успішного просування в результатах пошукової видачі.

Застосування ручного аналізу ключових запитів у таких умовах може бути надто трудомістким та містити високий ризик помилок чи пропуску важливих закономірностей.

Натомість методи машинного навчання, зокрема алгоритми кластеризації, дозволяють значно автоматизувати цей процес. Кластеризація дає змогу згрупувати велику кількість ключових слів за їх семантичною близькістю, а також виділити приховані теми чи підтеми в масиві даних.

Метою дослідження є порівняння ефективності застосування різних алгоритмів машинного навчання (K-means, Mini-batch K-means, Spectral Clustering та Deep Embedded Clustering) для кластеризації веб-контенту з метою вдосконалення SEO-стратегії та підвищення якості групування семантичного ядра сайтів.

Завдання дослідження полягає в аналізі принципів роботи алгоритмів K-means, Mini-batch K-means, Spectral Clustering та Deep Embedded Clustering, порівнянні одержаних результатів за обчислювальними витратами й часом виконання, а також визначенні найбільш доцільного алгоритму залежно від структури даних і доступних ресурсів.

Базовий алгоритм K-means [2], що часто є в основі інших алгоритмів, дозволяє групувати об'єкти за допомогою визначеної кількості кластерів. Суть алгоритму полягає у пошуку центра мас кожного кластеру та ітеративному переобчисленні їх положення доти, доки не буде досягнута збіжність.

Для опрацювання дуже великих наборів ключових фраз доцільно використовувати Mini-batch K-means, у якому обробка даних відбувається невеликими блоками. Такий підхід дає змогу суттєво пришвидшити конвергенцію алгоритму завдяки зменшенню обчислювальних витрат на кожному кроці, що особливо важливо, коли йдеться про десятки чи сотні тисяч ключових слів.

Алгоритм Spectral Clustering належить до спектральних методів, які спочатку будують матрицю суміжності або схожості між об'єктами, а потім перетворюють її у спектральний простір за допомогою власних векторів. У цьому просторі об'єкти стають лінійно роздільними, тож застосування традиційних методів K-means в основі алгоритму дозволяє успішно виявляти кластери. Такий підхід дає змогу краще враховувати нелінійні зв'язки між об'єктами та виявляти складнішу внутрішню структуру даних.

Алгоритм Deep Embedded Clustering (DEC) інтегрує спеціальну нейронну мережу для зменшення розмірності та кластеризацію за допомогою K-means. Спочатку автоенкодер навчається відображати вхідні дані, тобто текстові вектори ключових слів у простір меншої розмірності так, щоб зберегти головні особливості. Потім на стиснутих ознаках застосовується K-means для знаходження кластерів.

Дані для тестування моделей включали інформацію про сторінки веб-сайтів, таку як ключові слова. Ці дані оброблялися через TF-IDF для векторизації текстових характеристик, а поведінкові сигнали та метадані слугували основою для кластеризації. Програмні засоби, зокрема Python з бібліотеками scikit-learn та tensorflow, використовувалися для реалізації моделей. Метрики оцінки включали час виконання (середній час на одну ітерацію кластеризації), точність за допомогою Silhouette Score та Adjusted

Rand Index (ARI), а також масштабованість на наборах даних від 1000 до 100,000 сторінок. Результат наведено в таблиці 1.

Таблиця 1 – Порівняння моделей

Модель	Час	Точність	Масштабованість
K-means	2.5 с	87%	Висока
Mini-batch K-means	1.8 с	85%	Дуже висока
Spectral Clustering	4.5 с	90%	Середня
DEC	6.0 с	92%	Висока

Кластеризація контенту на основі алгоритмів машинного навчання є важливим етапом для побудови якісної стратегії просування веб-сайтів в результатах пошукової видачі. Результати порівняння показали, що кожна модель має свої сильні та слабкі сторони залежно від обраного сценарію використання. Базовий K-means забезпечує швидке й точне групування сторінок, проте його ефективність знижується на великих обсягах даних. Mini-batch K-means демонструє оптимальні результати для великих наборів завдяки зменшенню обчислювальної складності, хоча його точність трохи нижча. Spectral Clustering підходить для даних із нелінійною структурою, але потребує більше ресурсів. Найкращі результати за точністю показала Deep Embedded Clustering (DEC), яка інтегрує нейронні мережі для зменшення розмірності та обробки багатовимірних даних, що робить її ідеальною для складних завдань, хоча й потребує значних обчислювальних ресурсів.

Під час дослідження алгоритми K-means, Mini-batch K-means, Spectral Clustering та Deep Embedded Clustering демонстрували різний рівень точності та швидкодії залежно від масштабу даних та ресурсів.

У результаті проведеного дослідження встановлено, що застосування алгоритмів машинного навчання для кластеризації контенту забезпечує істотні переваги в оптимізації процесів SEO. Використання кластеризації суттєво скорочує час ручного сортування та групування великих обсягів ключових слів, що є критично важливим за умови динамічного розширення семантичного ядра.

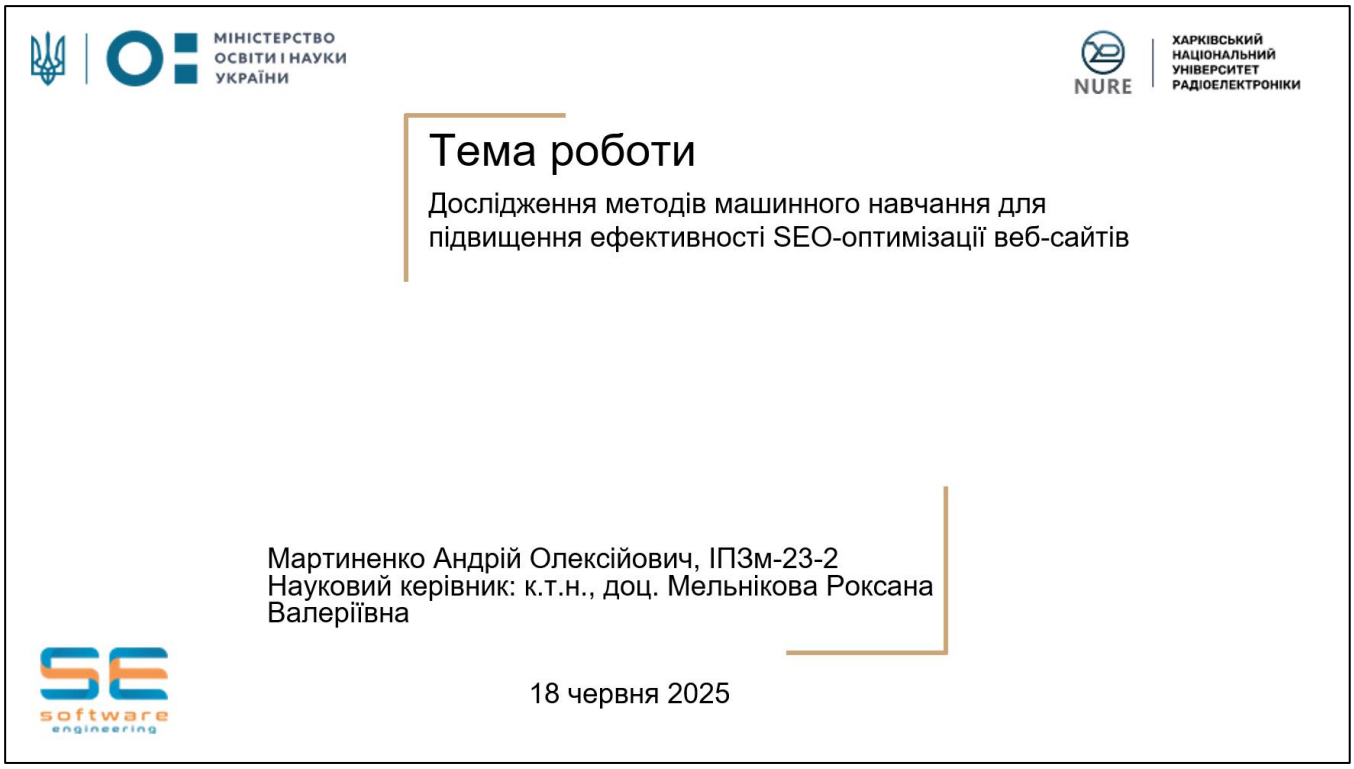
Список використаних джерел:

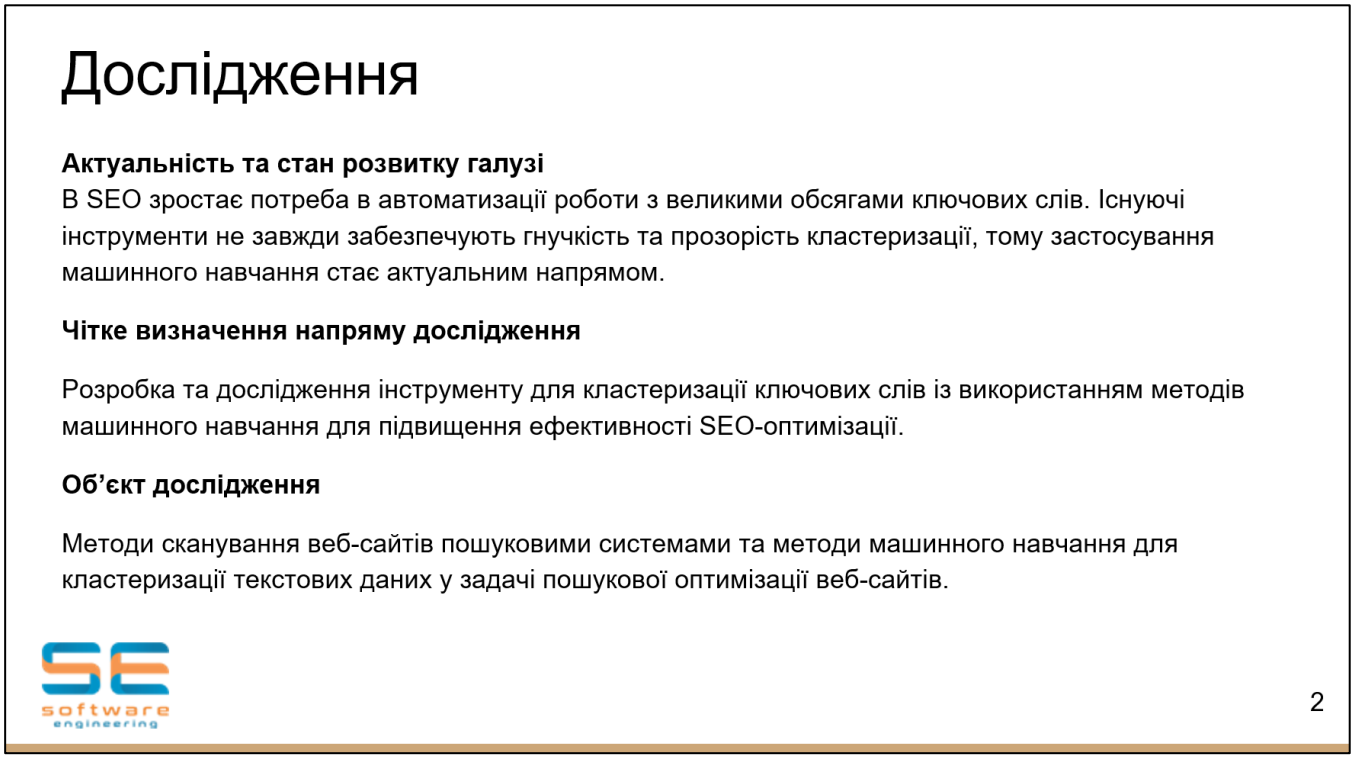
1. Enge E., Spencer S., Stricchiola J. The Art of SEO: Mastering Search Engine Optimization. O'Reilly Media, 2023. 925 p.

2. Bishop C. M. Pattern Recognition and Machine Learning. Springer Science & Business Media, 2006. 758 p.

ДОДАТОК В

Слайди презентації





Аналіз предметної області та аналогів

Шляхом застосування методу багатокритеріального вибору вдалося визначити найкращі SEO-інструменти, орієнтуючись на ключові критерії: ефективність сканування веб-сайтів та інтеграцію методів машинного навчання.

- Google
- Ahrefs
- Serpstat
- Screaming Frog

Тому для подальшого дослідження було обрано саме такі SEO-інструменти.

Постановка задачі

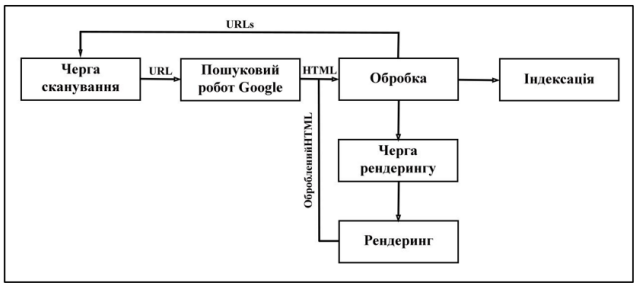
Метою дослідження є вирішення наступних задач:

- оцінити функціональні можливості SEO-інструментів. Провести детальний аналіз функціональних можливостей інструменту Ahrefs та Google-бота для сканування та аналізу веб-сайтів;
- проаналізувати переваги та недоліки інструментів;
- розглянути методи машинного навчання, що використовуються для кластеризації контенту;
- на основі отриманих даних розробити програмне забезпечення з інтеграцією методів машинного навчання для автоматизації рутинних процесів аналізу даних при веденні SEO-оптимізації;
- застосувати отримані результати кластеризації для стратегії просування веб-сайту в пошукових системах.

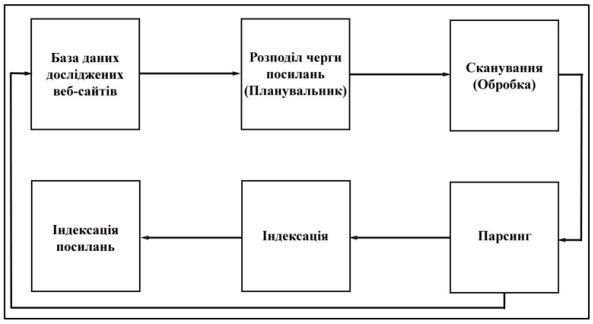
Автоматизація процесу групування ключових слів дає змогу значно заощадити час, який зазвичай витрачається на ручну класифікацію семантики, що особливо важливо при роботі з великими обсягами даних.

Порівняння методів сканування

Google-Бот



Ahrefs-Бот



Порівняння методів

Метод	Час виконання	Точність (Silhouette Score)	Масштабованість	Примітки
K-means	2.5 с	87%	Висока	Базовий метод для кластеризації
Mini-batch K-means	1.8 с	85%	Дуже висока	Оптимізований для великих наборів даних
Spectral Clustering	4.5 с	90%	Середня	Для нелінійно роздільних даних
Deep Embedded Clustering	6.0 с	92%	Висока	Найкраща точність серед методів

Методологія

Опис використаних методів дослідження

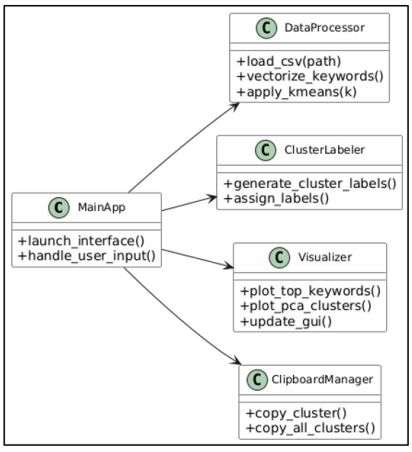
У роботі використано методи машинного навчання без вчителя, зокрема кластеризацію алгоритмом K-means. Також застосовано векторизацію тексту методом TF-IDF і зменшення розмірності даних за допомогою PCA для візуалізації результатів.

Інструментарій та технології, використані в роботі

Розробку реалізовано мовою Python із використанням бібліотек для обробки даних, машинного навчання, побудови графіків та створення графічного інтерфейсу: Pandas, scikit-learn, Matplotlib, CustomTkinter.

Архітектура системи для проведення експериментального дослідження

Діаграма сценаріїв використання



Діаграма класів

Опис програмного забезпечення, що було використано у дослідженні

Опис процесу розробки

Розробка включала аналіз вхідних даних, вибір методу кластеризації, програмну реалізацію обробки тексту, створення графічного інтерфейсу, інтеграцію візуалізацій та тестування з реальними SEO-запитами.

Вибрані мови програмування та фреймворки

Мова програмування: Python. Для реалізації функціоналу використано фреймворки та бібліотеки:
- **scikit-learn:** машинне навчання;
- **Pandas, NumPy:** обробка даних;
- **Matplotlib:** побудова графіків;
- **CustomTkinter:** графічний інтерфейс.

Зміст проведеного експерименту

Методи

– кластеризація ключових слів за допомогою алгоритму K-means;
– векторизація тексту методом TF-IDF;
– візуалізація результатів за допомогою PCA та гістограм.

Вхідні дані

CSV-файл, що містить ключові слова та їх пошуковою частотність, за якими можна ранжуватись в результатах пошукових систем.

Критерії

– релевантність кластерів (однорідність пошукового наміру);
– частотність запитів у межах кластера.

Зміст проведеного експерименту

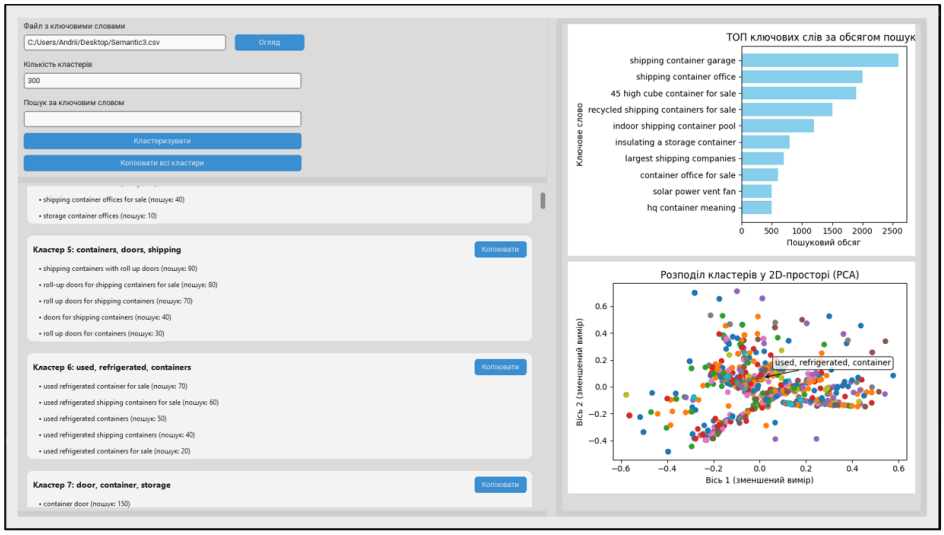
Послідовність

Завантаження даних, обробка, кластеризація, побудова графіків, ручна перевірка точності кластеризації в SERP.

Вимірювання

Вимірювання ефективності, кінцевий результат приросту органічного трафіку та показів в результатах пошуку після оптимізації веб-сайту використовуючи програмний продукт.

Результати експерименту



Результати експерименту

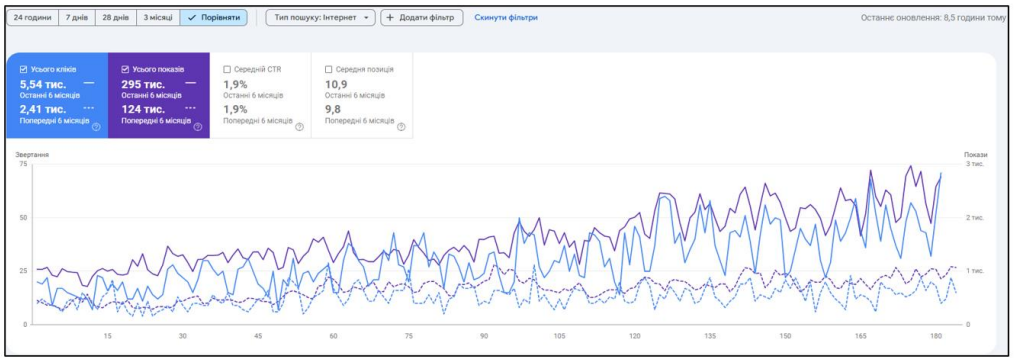
Кластер 40: opening, shipping, container

- side opening containers for sale (пошук: 100)
- side opening shipping containers (пошук: 80)
- shipping container side opening (пошук: 50)
- side opening storage container (пошук: 20)

Запит	Мета-теги Title в результатах пошуку (ТОП-5)
side opening container for sale	Buy a 20ft Open Side Container
	20FT Open Side New (One Trip) Shipping Container
	Open-Side Shipping Containers for Sale
	Open sided shipping containers for sale
	Open Side Containers for Sale or Lease
side opening containers	Buy a 20ft Open Side Container
	Open-Side Shipping Containers for Sale
	Open sided shipping containers for sale
	Side Opening Shipping Containers
	20FT Open Side New (One Trip) Shipping Container
side opening storage container	Open sided shipping containers for sale
	Buy a 20ft Open Side Container
	Open-Side Shipping Containers for Sale
	Modular Open-Sided Shipping Containers Features | MMPS

Результати експерименту

Результати експерименту



За період проведення оптимізації сайту, кількість переходів на сайт з результатів органічного пошуку зросла на **130%** з 2,41 тисячі до 5,54 тисяч.
Водночас загальна
кількість показів в пошуковій видачі зросла на **138%** з 124 тисяч до 295 тисяч.

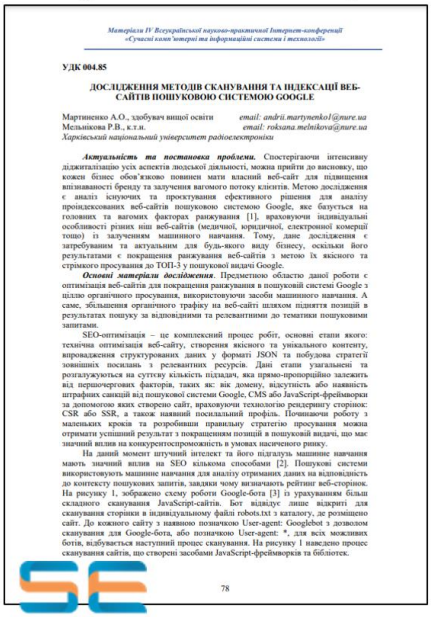
Аналіз отриманих результатів

Отримані результати повністю відповідають поставленим цілям дослідження: в результаті дсолідження вдалося підвищити ефективність просування веб-сайту в пошуковій системі Google з використанням методів машинного навчання.

Аналіз показав високу релевантність сформованих кластерів і практичну придатність створеної програми для ведення стратегії Programmatic SEO.

Результати підтверджують, що застосування машинного навчання в якості допоміжного інструменту дозволяє добре структурувати семантику та охоплювати ширший спектр пошукових намірів.

Публікація результатів



Матеріали IV Всеукраїнської науково-практичної Інтернет-конференції «Сучасні комп'ютерні та інформаційні системи і технології»

УДК 004.85

ДОСЛІДЖЕННЯ МЕТОДІВ СКАНУВАННЯ ТА ІНДЕКСАЦІЇ ВЕБ-САЙТІВ ПОШУКОВОЮ СИСТЕМОЮ GOOGLE

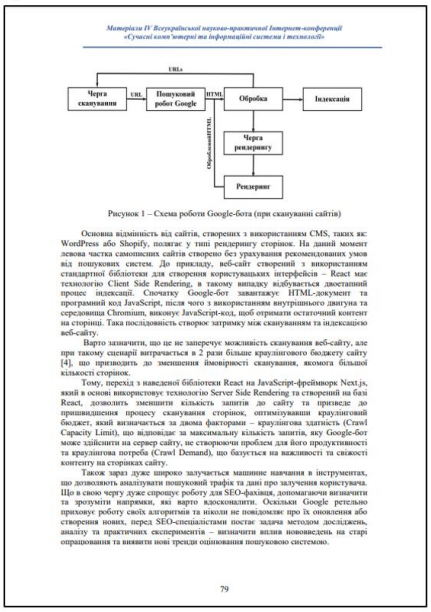
Мартиненко А.О., здобувач вищої освіти *email: andrii.martynenko1@nure.ua*
Мельникова Р.В., к.т.н. *email: roksana.melnikova@nure.ua*
Харківський національний університет радіоелектроніки

Актуальність та постановка проблеми. Спостерігаючи інтенсивну діджиталізацію усіх аспектів людської діяльності, можна прийти до висновку, що кожен бізнес обов'язково повинен мати власний веб-сайт для підвищення впізнаваності бренду та залучення вагомого потоку клієнтів. Метою дослідження є аналіз існуючих та пропонування ефективного рішення для аналізу проіндексованих веб-сайтів пошуковою системою Google, яке базується на головних та вагомих факторах ранжування [1], враховуючи індивідуальні особливості різних ніш веб-сайтів (медичної, юридичної, електронної комерції тощо) із залученням машинного навчання. Тому, дане дослідження є затребуваним та актуальним для будь-якого виду бізнесу, оскільки його результатами є покращення ранжування веб-сайтів з метою їх якісного та стрімкого просування до ТОП-3 у пошуковій видачі Google.

Основні матеріали дослідження. Предметною областю даної роботи є оптимізація веб-сайтів для покращення ранжування в пошуковій системі Google з цілями органічного просування, використовуючи засоби машинного навчання. А саме, збільшення органічного трафіку на веб-сайті шляхом підняття позицій в результатах пошуку за відповідними та релевантними до тематики пошуковими запитами.

SEO-оптимізація – це комплексний процес робіт, основні етапи якого: технічна оптимізація веб-сайту, створення якісного та унікального контенту, впровадження структурованих даних у форматі JSON та побудова стратегії зовнішніх посилань з релевантних ресурсів. Дані етапи узагальнені та розглядуються на сутому кількість підзадач, яка прямо-пропорційно залежить від першочергових факторів, таких як: вік домену, відсутність або наявність штрафних санкцій від пошукової системи Google, CMS або JavaScript-фреймворки за допомогою яких створено сайт, враховуючи технологію рендерингу сторінок: CSR або SSR, а також наявний посилальний профіль. Починаючи роботу з маленьких кроків та розробивши правильну стратегію просування можна отримати успішний результат з покращенням позицій в пошуковій видачі, що має значний вплив на конкурентоспроможність в умовах насиченого ринку.

На даний момент штучний інтелект та його підгалузь машинне навчання мають значний вплив на SEO кількома способами [2]. Пошукові системи використовують машинне навчання для аналізу отриманих даних на відповідність до контексту пошукових запитів, завдяки чому визначають рейтинг веб-сторінок. На рисунку 1, зображено схему роботи Google-бота [3] із урахуванням більш складного сканування JavaScript-сайтів. Бот відвідує лише відкриті для сканування сторінки в індивідуальному файлі robots.txt з каталогу, де розміщено сайт. До кожного сайту з наявною позначкою User-agent: Googlebot з дозволом сканування для Google-бота, або позначкою User-agent: *, для всіх можливих ботів, відбувається наступний процес сканування. На рисунку 1 наведено процес сканування сайтів, що створені засобами JavaScript-фреймворків та бібліотек.



Матеріали IV Всеукраїнської науково-практичної Інтернет-конференції «Сучасні комп'ютерні та інформаційні системи і технології»

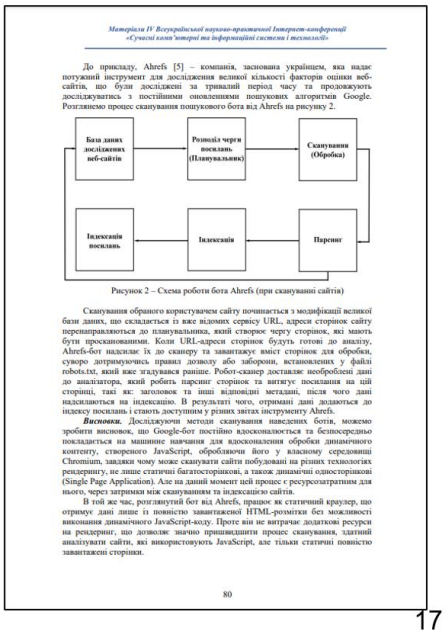
Рисунок 1 – Схема роботи Google-бота (при скануванні сайтів)

Основна відмінність від сайтів, створених з використанням CMS, таких як: WordPress або Shopify, полягає у типі рендерингу сторінок. На даний момент левова частка самописних сайтів створено без урахування рекомендованих умов від пошукових систем. До прикладу, веб-сайт створений з використанням стандартної бібліотеки для створення користувацьких інтерфейсів – React має технологію Client Side Rendering, в такому випадку відбувається двоетапний процес індексації. Спочатку Google-бот завантажує HTML-документ та програмний код JavaScript, після чого з використанням внутрішнього двигуна та середовища Chromium, виконує JavaScript-код, щоб отримати остаточний контент на сторінці. Така послідовність створює затримку між скануванням та індексацією веб-сайту.

Варто зазначити, що це не заперечує можливість сканування веб-сайту, але при такому сценарії витрачається в 2 рази більше краулінгового бюджету сайту [4], що призводить до зменшення ймовірності сканування, якомога більшої кількості сторінок.

Тому, перехід з наведеної бібліотеки React на JavaScript-фреймворк Next.js, який в основі використовує технологію Server Side Rendering та створений на базі React, дозволить зменшити кількість запитів до сайту та призведе до пришвидшення процесу сканування сторінок, оптимізуванням краулінгового бюджет, який визначається за двома факторами – краулінгова здатність (Crawl Capacity Limit), що відповідає за максимальну кількість запитів, яку Google-бот може здійснити на сервер сайту, не створюючи проблем для його продуктивності та краулінгова потреба (Crawl Demand), що базується на важливості та свіжості контенту на сторінках сайту.

Також зараз дуже широко залучається машинне навчання в інструментах, що дозволяють аналізувати пошуковий трафік та дані про залучення користувача. Що в свою чергу дуже спрощує роботу для SEO-фахівця, допомагаючи визначити та зрозуміти напрямки, які варто вдосконалити. Оскільки Google ретельно приховує роботу своїх алгоритмів та ніколи не повідомляє про їх оновлення або створення нових, перед SEO-спеціалістами постає задача методом досліджень, аналізу та практичних експериментів – визначити вплив нововведень на старі опрацювання та виявити нові тренди оцінювання пошукової системою.



Матеріали IV Всеукраїнської науково-практичної Інтернет-конференції «Сучасні комп'ютерні та інформаційні системи і технології»

До прикладу, Ahrefs [5] – компанія, заснована українцем, яка надає потужний інструмент для дослідження великої кількості факторів оцінки веб-сайтів, що були досліджені за тривалий період часу та продовжують досліджуватись з постійними оновленнями пошукових алгоритмів Google. Розглянемо процес сканування пошукового бота від Ahrefs на рисунку 2.

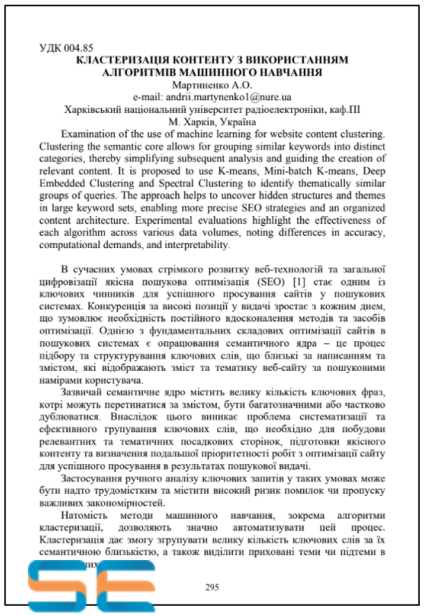
Рисунок 2 – Схема роботи бота Ahrefs (при скануванні сайтів)

Сканування обраного користувачем сайту починається з модифікації великої бази даних, що складається із вже відомих сервісу URL, адреси сторінок сайту перенаправляються до планувальника, який створює чергу сторінок, які мають бути просканованими. Коли URL-адреси сторінок будуть готові до аналізу, Ahrefs-бот надсилає їх до сканеру та завантажує вміст сторінок для обробки, суворо дотримуючись правил дозволу або заборони, встановлених у файлі robots.txt, який вже згадувався раніше. Робот-сканер доставляє необроблені дані до аналізатора, який робить парсинг сторінок та витягує посилання на цій сторінці, такі як: заголовок та інші відповідні метадані, після чого дані надсилаються на індексацію. В результаті чого, отримані дані додаються до індексу посилань і стають доступним у різних звітах інструменту Ahrefs.

Висновки. Досліджуючи методи сканування наведених ботів, можемо зробити висновок, що Google-бот постійно вдосконалюється та безпосередньо покладається на машинне навчання для вдосконалення обробки динамічного контенту, створеного JavaScript, обробляючи його у власному середовищі Chromium, завдяки чому може сканувати сайти побудовані на різних технологіях рендерингу, не лише статичні багатосторінкові, а також динамічні односторінкові (Single Page Application). Але на даний момент цей процес є ресурсозатратним для нього, через затримки між скануванням та індексацією сайтів.

В той же час, розглянутий бот від Ahrefs, працює як статичний краулер, що отримує дані лише із повністю завантаженої HTML-розмітки без можливості виконання динамічного JavaScript-коду. Проте він не витрачає додаткові ресурси на рендеринг, що дозволяє значно пришвидшити процес сканування, здатний аналізувати сайти, які використовують JavaScript, але тільки статичні повністю завантажені сторінки.

Публікація результатів



УДК 004.85

КЛАСТЕРИЗАЦІЯ КОНТЕНТУ З ВИКОРИСТАННЯМ АЛГОРИТМІВ МАШИННОГО НАВЧАННЯ

Мартиненко А.О.
e-mail: andrii.martynenko1@nure.ua
Харківський національний університет радіоелектроніки, каф.ПІ
м. Харків, Україна

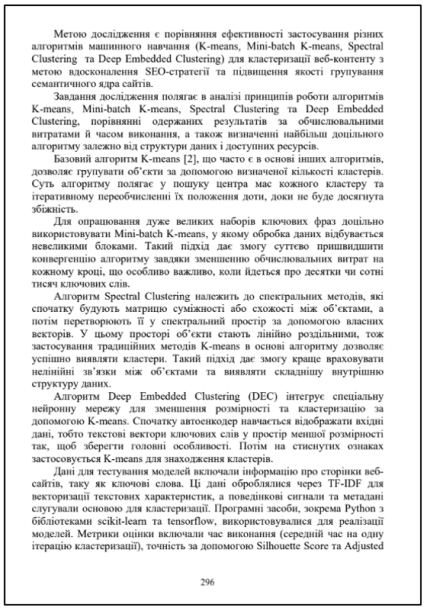
Examination of the use of machine learning for website content clustering. Clustering the semantic core allows for grouping similar keywords into distinct categories, thereby simplifying subsequent analysis and guiding the creation of relevant content. It is proposed to use K-means, Mini-batch K-means, Deep Embedded Clustering and Spectral Clustering to identify thematically similar groups of queries. The approach helps to uncover hidden structures and themes in large keyword sets, enabling more precise SEO strategies and an organized content architecture. Experimental evaluations highlight the effectiveness of each algorithm across various data volumes, noting differences in accuracy, computational demands, and interpretability.

В сучасних умовах стрімкого розвитку веб-технологій та загальної цифровізації якісна пошукова оптимізація (SEO) [1] стає одним із ключових чинників для успішного просування сайтів у пошукових системах. Конкуренція за високі позиції у видачі зростає з кожним днем, що зумовлює необхідність постійного вдосконалення методів та засобів оптимізації. Однією з фундаментальних складових оптимізації сайтів в пошукових системах є опрацювання семантичного ядра – це процес підбору та структурування ключових слів, що близькі за написанням та змістом, які відображають зміст та тематику веб-сайту за пошуковими намірами користувача.

Зазвичай семантичне ядро містить велику кількість ключових фраз, котрі можуть перетинатися за змістом, бути багатозначними або частково дублюватися. Внаслідок цього виникає проблема систематизації та ефективного групування ключових слів, що необхідно для побудови релевантних та тематичних посадкових сторінок, підготовки якісного контенту та визначення подальшої пріоритетності робіт з оптимізації сайту для успішного просування в результатах пошукової видачі.

Застосування ручного аналізу ключових запитів у таких умовах може бути надто трудомістким та містити високий ризик помилок чи пропуску важливих закономірностей.

Натомість методи машинного навчання, зокрема алгоритми кластеризації, дозволяють значно автоматизувати цей процес. Кластеризація дає змогу згрупувати велику кількість ключових слів за їх семантичною близькістю, а також виділити приховані теми чи підтеми в



Метою дослідження є порівняння ефективності застосування різних алгоритмів машинного навчання (K-means, Mini-batch K-means, Spectral Clustering та Deep Embedded Clustering) для кластеризації веб-контенту з метою вдосконалення SEO-стратегії та підвищення якості групування семантичного ядра сайтів.

Завдання дослідження полягає в аналізі принципів роботи алгоритмів K-means, Mini-batch K-means, Spectral Clustering та Deep Embedded Clustering, порівнянні одержаних результатів за обчислювальними витратами й часом виконання, а також визначенні найбільш доцільного алгоритму залежно від структури даних і доступних ресурсів.

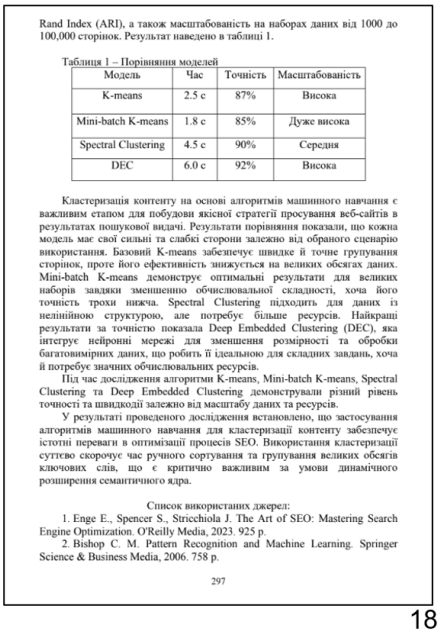
Базовий алгоритм K-means [2], що часто є в основі інших алгоритмів, дозволяє групувати об'єкти за допомогою визначеної кількості кластерів. Суть алгоритму полягає у пошуку центра має кожного кластеру та ітеративному переобчисленні їх положення доти, доки не буде досягнута збіжність.

Для опрацювання дуже великих наборів ключових фраз доцільно використовувати Mini-batch K-means, у якому обробка даних відбувається невеликими блоками. Такий підхід дає змогу суттєво пришвидшити конвергенцію алгоритму завдяки зменшенню обчислювальних витрат на кожному кроці, що особливо важливо, коли йдеться про десятки чи сотні тисяч ключових слів.

Алгоритм Spectral Clustering належить до спектральних методів, які спочатку будують матрицю суміжності або схожості між об'єктами, а потім перетворюють її у спектральний простір за допомогою власних векторів. У цьому просторі об'єкти стають лінійно роздільними, тож застосування традиційних методів K-means в основі алгоритму дозволяє успішно виявляти кластери. Такий підхід дає змогу краще враховувати нелінійні зв'язки між об'єктами та виявляти складнішу внутрішню структуру даних.

Алгоритм Deep Embedded Clustering (DEC) інтегрує спеціальну нейронну мережу для зменшення розмірності та кластеризацію за допомогою K-means. Спочатку автоенкодер навчається відображати вхідні дані, тобто текстові вектори ключових слів у простір меншої розмірності так, щоб зберегти головні особливості. Потім на стиснутих ознаках застосовується K-means для знаходження кластерів.

Дані для тестування моделей включали інформацію про сторінки веб-сайтів, таку як ключові слова. Ці дані оброблялися через TF-IDF для векторизації текстових характеристик, а поведінкові сигнали та метадані слугували основою для кластеризації. Програмні засоби, зокрема Python з бібліотеками scikit-learn та tensorflow, використовувалися для реалізації моделей. Метрики оцінки включали час виконання (середній час на одну ітерацію кластеризації), точність за допомогою Silhouette Score та Adjusted



Rand Index (ARI), а також масштабованість на наборах даних від 1000 до 100,000 сторінок. Результат наведено в таблиці 1.

Таблиця 1 – Порівняння моделей

Модель	Час	Точність	Масштабованість
K-means	2.5 с	87%	Висока
Mini-batch K-means	1.8 с	85%	Дуже висока
Spectral Clustering	4.5 с	90%	Середня
DEC	6.0 с	92%	Висока

Кластеризація контенту на основі алгоритмів машинного навчання є важливим етапом для побудови якісної стратегії просування веб-сайтів в результатах пошукової видачі. Результати порівняння показали, що кожна модель має свої сильні та слабкі сторони залежно від обраного сценарію використання. Базовий K-means забезпечує швидке й точне групування сторінок, проте його ефективність знижується на великих обсягах даних. Mini-batch K-means демонструє оптимальні результати для великих наборів завдяки зменшенню обчислювальної складності, хоча його точність трохи нижча. Spectral Clustering підходить для даних із нелінійною структурою, але потребує більше ресурсів. Найкращі результати за точністю показала Deep Embedded Clustering (DEC), яка інтегрує нейронні мережі для зменшення розмірності та обробки багатовимірних даних, що робить її ідеальною для складних завдань, хоча й потребує значних обчислювальних ресурсів.

Під час дослідження алгоритми K-means, Mini-batch K-means, Spectral Clustering та Deep Embedded Clustering демонстрували різний рівень точності та швидкодії залежно від масштабу даних та ресурсів.

У результаті проведеного дослідження встановлено, що застосування алгоритмів машинного навчання для кластеризації контенту забезпечує істотні переваги в оптимізації процесів SEO. Використання кластеризації суттєво скорочує час ручного сортування та групування великих обсягів ключових слів, що є критично важливим за умови динамічного розширення семантичного ядра.

Список використаних джерел:

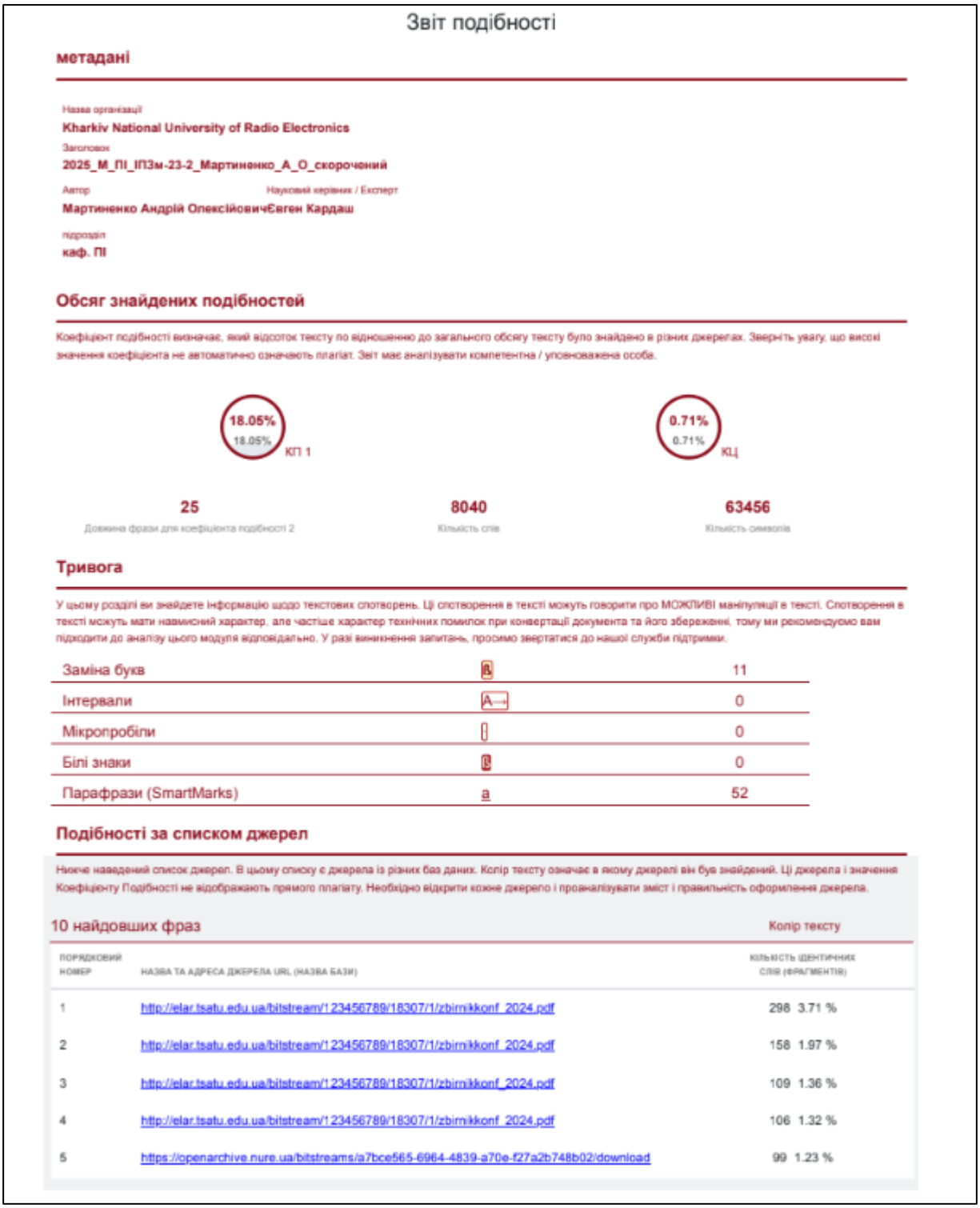
1. Enge E., Spencer S., Stricchiola J. The Art of SEO: Mastering Search Engine Optimization. O'Reilly Media, 2023. 925 p.
2. Bishop C. M. Pattern Recognition and Machine Learning. Springer Science & Business Media, 2006. 758 p.

Підсумки

- проведено аналіз існуючих програмних систем для SEO-оптимізації;
- розглянуто сучасні виклики SEO, порівняно процес сканування еталонного Google-бота та конкурента Ahrefs-бота;
- порівняно методи кластеризації;
- розроблено застосунок з інтеграцією методів кластеризації та виконано ручну перевірку вихідного результату в результатах пошуку Google;
- отримані вихідні дані застосовано для впровадження стратегії Programmatic SEO на веб-сайті та порівняно ефективність просування до та після початку оптимізації;
- опубліковано тези доповіді «Кластеризація контенту з використанням алгоритмів машинного навчання» на двадцять дев'ятий міжнародний молодіжний форум «РАДІОЕЛЕКТРОНІКА ТА МОЛОДЬ В XXI ст.»;
- опубліковано тези доповіді «Дослідження методів сканування та індексації веб-сайтів пошуковою системою Google» на IV Всеукраїнську науково-практичну Інтернет-конференцію «Сучасні комп'ютерні та інформаційні системи і технології»

ДОДАТОК Г

Звіт з результатами перевірки на унікальність тексту в базі ХНУРЕ



Звіт подібності

метадані

Назва організації
Kharkiv National University of Radio Electronics
Заголовок
2025_M_ПІ_ІПЗм-23-2_Мартиненко_А_О_скорочений
Автор
Мартиненко Андрій Олексійович
Науковий керівник / Експерт
Євген Кардаш
підрозділ
каф. ПІ

Обсяг знайдених подібностей

Коефіцієнт подібності визначає, який відсоток тексту по відношенню до загального обсягу тексту було знайдено в різних джерелах. Зверніть увагу, що високі значення коефіцієнта не автоматично означають плагіат. Звіт має аналізувати компетентна / уповноважена особа.

18.05% 18.05% КП 1

0.71% 0.71% КЦ

25	8040	63456
Довжина фрази для коефіцієнта подібності 2	Кількість слів	Кількість символів

Тривога

У цьому розділі ви знайдете інформацію щодо текстових спотворень. Ці спотворення в тексті можуть говорити про МОЖЛИВІ маніпуляції в тексті. Спотворення в тексті можуть мати навмисний характер, але частіше характер технічних помилок при конвертації документа та його збереженні, тому ми рекомендуємо вам підходити до аналізу цього модуля відповідально. У разі виникнення запитань, просимо звертатися до нашої служби підтримки.

Заміна букв		11
Інтервали		0
Мікропробіли		0
Білі знаки		0
Парафрази (SmartMarks)		52

Подібності за списком джерел

Нижче наведений список джерел. В цьому списку є джерела із різних баз даних. Колір тексту означає в якому джерелі він був знайдений. Ці джерела і значення Коефіцієнту Подібності не відображають прямого плагіату. Необхідно відкрити кожне джерело і проаналізувати зміст і правильність оформлення джерела.

10 найдовших фраз — Колір тексту

ПОРЯДКОВИЙ НОМЕР	НАЗВА ТА АДРЕСА ДЖЕРЕЛА URL (НАЗВА БАЗИ)	КІЛЬКІСТЬ ІДЕНТИЧНИХ СЛІВ (ФРАГМЕНТІВ)
1	http://elar.tsatu.edu.ua/bitstream/123456789/18307/1/zbirnikkonf_2024.pdf	298 3.71 %
2	http://elar.tsatu.edu.ua/bitstream/123456789/18307/1/zbirnikkonf_2024.pdf	158 1.97 %
3	http://elar.tsatu.edu.ua/bitstream/123456789/18307/1/zbirnikkonf_2024.pdf	109 1.36 %
4	http://elar.tsatu.edu.ua/bitstream/123456789/18307/1/zbirnikkonf_2024.pdf	106 1.32 %
5	https://openarchive.nure.ua/bitstreams/a7bce565-6964-4839-a70e-f27a2b748b02/download	99 1.23 %

6	https://openarchive.nure.ua/bitstreams/a7bce565-6964-4839-a70e-f27a2b748b02/download	83 1.03 %
7	http://elar.tsatu.edu.ua/bitstream/123456789/18307/1/zbirnikkonf_2024.pdf	66 0.82 %
8	https://openarchive.nure.ua/bitstreams/d8dd7e38-abf1-4639-80de-456037263681/download	63 0.78 %
9	http://elar.tsatu.edu.ua/bitstream/123456789/18307/1/zbirnikkonf_2024.pdf	41 0.51 %
10	https://github.com/lain530/force-directed-layout-algorithms/blob/master/forcelayout/draw.py	36 0.45 %

з бази даних RefBooks (0.00 %)

ПОРЯДКОВИЙ НОМЕР	ЗАГОЛОВОК	КІЛЬКІСТЬ ІДЕНТИЧНИХ СЛІВ (ФРАГМЕНТІВ)

з домашньої бази даних (0.00 %)

ПОРЯДКОВИЙ НОМЕР	ЗАГОЛОВОК	КІЛЬКІСТЬ ІДЕНТИЧНИХ СЛІВ (ФРАГМЕНТІВ)

з програми обміну базами даних (0.00 %)

ПОРЯДКОВИЙ НОМЕР	ЗАГОЛОВОК	КІЛЬКІСТЬ ІДЕНТИЧНИХ СЛІВ (ФРАГМЕНТІВ)

з Інтернету (18.05 %)

ПОРЯДКОВИЙ НОМЕР	ДЖЕРЕЛО URL	КІЛЬКІСТЬ ІДЕНТИЧНИХ СЛІВ (ФРАГМЕНТІВ)
1	http://elar.tsatu.edu.ua/bitstream/123456789/18307/1/zbirnikkonf_2024.pdf	783 (7) 9.74 %
2	https://openarchive.nure.ua/bitstreams/d8dd7e38-abf1-4639-80de-456037263681/download	259 (14) 3.22 %
3	https://openarchive.nure.ua/bitstreams/a7bce565-6964-4839-a70e-f27a2b748b02/download	215 (7) 2.67 %
4	https://dev59.com/Krbna4cB1Zd3GeqPbXrr	38 (3) 0.47 %
5	https://github.com/lain530/force-directed-layout-algorithms/blob/master/forcelayout/draw.py	36 (1) 0.45 %
6	https://www.alibaba.com/showroom/open-sided-shipping-container_6.html	23 (4) 0.29 %
7	https://note.com/yeku_sub/n/nb71efbfa8340	23 (2) 0.29 %
8	https://www.conexdepot.com/product/20ft-open-side-new-one-trip-shipping-container/	22 (2) 0.27 %
9	https://www.fprefabhouse.com/showroom/20-open-side-shipping-container/	22 (3) 0.27 %
10	https://openarchive.nure.ua/bitstreams/d6adea9d-b281-4e0d-a432-aec52df9427c/download	20 (1) 0.25 %
11	https://how2matplotlib.com/matplotlib-annotate.html	10 (1) 0.12 %

Список прийнятих фрагментів (немає прийнятих фрагментів)

ПОРЯДКОВИЙ НОМЕР	ЗМІСТ	КІЛЬКІСТЬ ОДНАКОВИХ СЛІВ (ФРАГМЕНТІВ)

ВСТУП

Світ сучасних технологій невпинно розвивається, і важливе місце в цьому розвитку займають інструменти для сканування та аналізу веб-сайтів. У сфері SEO (Search Engine Optimization) такі інструменти, як Ahrefs, Serpstat, Seranking, Majestic, MOZ, Screaming Frog, а також методи сканування, які використовує Google, є невід'ємною частиною процесу оптимізації. Вони дозволяють виявляти технічні та контентні помилки, аналізувати структуру веб-сайту, а також надавати рекомендації щодо їхнього усунення. Це дозволяє підвищувати якість сайтів, їхню швидкість

ДОДАТОК Ґ

Експертний висновок нормоконтроль

Експертний висновок результатів перевірки кваліфікаційної роботи

студент (посада) | програмної інженерії (кафедра) | ІПЗм-23-2 (група)

Андрій МАРТИНЕНКО
(прізвище, ім'я, по батькові)

Зауваження

Пункт ДСТУ 3008-2015	Зміст пункту	Сторінка кваліфікаційної роботи
1	2	3
	7.1 Загальні положення	
7.1.25	Не дозволено розміщувати назву розділу, підрозділу, а також пункту й підпункту на останньому рядку сторінки.	19
	7.3 Нумерація сторінок звіту	
	7.5 Рисунки	
	7.6 Таблиці	
	7.7 Переліки	
	7.8 Примітки	
	7.9 Виноски	
	7.10 Формули та рівняння	
	7.11 Посилання	
	7.13 Список авторів	
	7.14 Скорочення та умовні познаки	
	7.15 Додатки	
Методичні вказівки до виконання кваліфікаційної роботи магістра… **ЗАТВЕРДЖЕНО** кафедрою ПІ протокол № 5 від 13.11.2023р. 3.2 Оформлення пояснювальної записки згідно з ДСТУ 3008:2015 Звіти у сфері науки і техніки. Структура та правила оформлювання. **Шаблон** затверджений засіданням кафедри №3 від 16.10.2023.	Кількість сторінок (рисунків, таблиць, джерел) заявлених в рефераті повинна співпадати з кількістю сторінок (рисунків, таблиць, джерел) в записці.	4

Експерт _____ (підпис) Вадим НЕЧВОЛОД (прізвище, ініціали)

13.06.2025