

ПІДХІД ДО АНАЛІЗУ УНІКАЛЬНОСТІ ПРАКТИЧНИХ РОБІТ СТУДЕНТІВ

Крепич С.Я.¹⁾, Глухов О.В.²⁾, Співак І.Я.³⁾

Західноукраїнський національний університет

¹⁾ к.т.н., доцент; ²⁾ магістрант; ³⁾ к.т.н., доцент

I. Постановка проблеми

В університетському середовищі проведення лабораторних робіт є важливою складовою академічного процесу. Проте, однією з основних проблем, з якою зіштовхуються викладачі, є виявлення плагіату та копіювання студентами результатів робіт. У сучасному освітньому середовищі для перевірки унікальності випускових кваліфікаційних робіт часто використовуються спеціалізовані сервіси, такі як Turnitin, Unicheck та інші. Ці сервіси пропонують платні плагіат-детектори, які автоматично перевіряють текстові документи на унікальність, порівнюючи їх з великою базою даних наукових статей, книг та Інтернет-ресурсів. Проте, вони зазвичай використовуються лише для архівних робіт, таких як випускові кваліфікаційні роботи, і не завжди доступні для перевірки лабораторних або практичних завдань. У той час, як унікальність лабораторних чи практичних робіт зазвичай перевіряється викладачем самостійно, що може бути часо- та працевитратним процесом. Викладачеві доводиться аналізувати та порівнювати кожен студентський проект з іншими текстами, що може бути важко і неефективно, особливо у великих групах. Для вирішення цієї проблеми необхідно розробити ефективний додаток, який забезпечить можливість виявлення збігів у лабораторних роботах та інших завданнях. Для цієї цілі найбільше підходить розробка ПЗ на комп'ютер, на базі операційної системи Windows.

II. Мета роботи

Мета роботи – створення програмного забезпечення, яке дозволить виявляти збіги у лабораторних роботах та інших завданнях, що надасть викладачам засоби для виявлення плагіату та копіювання в академічному середовищі.

III. Огляд методів

Метод TF-IDF (Term Frequency-Inverse Document Frequency): Цей метод використовується для оцінки важливості кожного слова у документі, порівнюючи його частоту в документі з його частотою в корпусі документів. Він може бути використаний для визначення ступеня схожості між документами на основі їхнього змісту.

	m	e	i	l	e	n	s	t	e	i	n	
l	0	1	2	3	4	5	6	7	8	9	10	11
e	1	1	2	3	3	4	5	6	7	8	9	10
v	2	2	1	2	3	3	4	5	6	7	8	9
e	3	3	2	2	3	4	4	5	6	7	8	9
n	4	4	3	3	3	3	4	5	6	6	7	8
s	5	5	4	4	4	4	3	4	5	6	7	7
h	6	6	5	5	5	5	4	3	4	5	6	7
t	7	7	6	6	6	6	5	4	4	5	6	7
e	8	8	7	7	7	7	6	5	4	5	6	7
i	9	9	8	8	8	7	7	6	5	4	5	6
n	10	10	9	8	9	8	8	7	6	5	4	5
	11	11	10	9	9	9	8	8	7	6	5	4

Алгоритм Левенштейна: Цей алгоритм використовується для визначення відстані між двома рядками шляхом підрахунку мінімальної кількості операцій (вставка, видалення, заміна символів), необхідних для перетворення одного рядка на інший. Він може бути використаний для визначення рівня схожості між текстовими документами.

Метод N-грам: Цей метод використовується для аналізу тексту шляхом розбиття його на

Рисунок 1 - Алгоритм Левенштейна послідовності з n символів (зазвичай n = 2 або 3) і порівняння цих послідовностей між двома документами. Він може виявляти схожість між документами навіть у випадках, коли слова переставлені або змінено порядок слів.

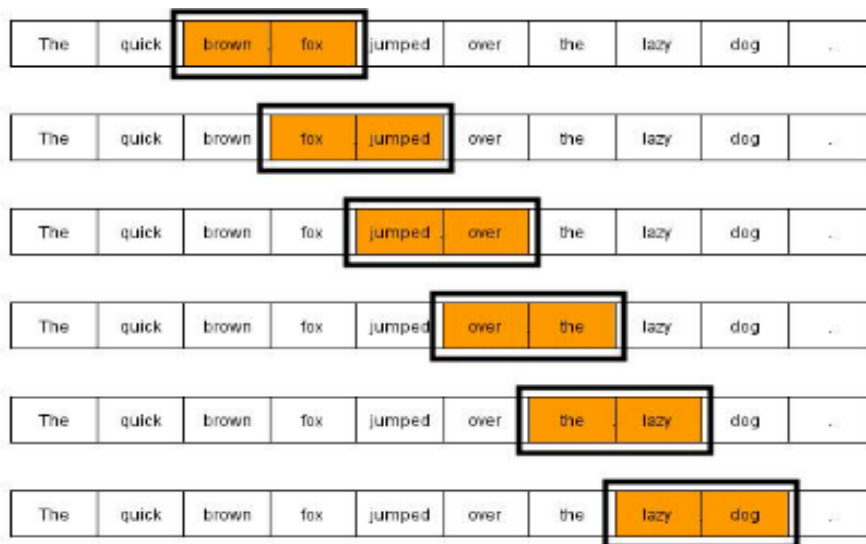


Рисунок 2 - Метод N-грам

Машинне навчання: Методи машинного навчання, такі як нейронні мережі, науково-статистичні методи та класифікаційні алгоритми, можуть бути використані для автоматичного виявлення збігів у лабораторних роботах на основі великої кількості навчальних даних.

У таблиці 1 представлені дані експерименту із застосування методу N-грам для оцінки подібності текстів між собою.

Таблиця 1

Порівняння застосування методу N-грам до текстів різної схожості

Текст№1	Текст№2	Збіжність тексту
Мовлення як вид людської діяльності завжди зорієнтоване на виконання певного комунікативного завдання. Висловлюючи думки і почуття, людина ставить конкретну мету – щось повідомити, про щось переконати тощо. Існує багато визначень тексту. Наведемо окремі з них.	Мовлення як вид людської діяльності завжди зорієнтоване на виконання певного комунікативного завдання. Висловлюючи думки і почуття, людина ставить конкретну мету – щось повідомити, про щось переконати тощо. Існує багато визначень тексту. Наведемо окремі з них.	98.9%
«Текст – це витвір мовленнєвого процесу, що відзначається завершеністю, об'єктивований у вигляді письмового документа, літературно опрацьований відповідно до типу документа.	«Текст – певна, з функціонально-сміслового погляду упорядкована, група речень або їх аналогів, які являють собою, завдяки семантичним і функціональним взаємовідношенням елементів, завершену смислово єдність»	2.44%
Мовлення як вид людської діяльності завжди призначена на виконання якое комунікативного завдання. Висловлюючи думки і почуття, людина ставить певну мету – щось повідомити, про щось переконати . Існує багато визначень тексту. Наведемо окремі з них.	Мовлення як вид людської діяльності завжди зорієнтоване на виконання певного комунікативного завдання. Висловлюючи думки і почуття, людина ставить конкретну мету – щось повідомити, про щось переконати тощо. Існує багато визначень тексту. Наведемо окремі з них.	73.2%

Висновок

У роботі було проведено дослідження існуючих методів оцінювання унікальності текстів. Метод TF-IDF враховує важливість слова в документі та базі, але не враховує порядок слів; алгоритм Левенштейна ефективний для визначення відстані між текстовими рядками, однак не враховує семантичні зв'язки; метод N-грам добре виявляє схожість певним чином впорядкованих слів. Проведений експеримент показав адекватність методу N-грам для застосування при перевірці практичних робіт студентів на унікальність змісту.

Список використаних джерел

1. ntchosting <https://www.ntchosting.com/encyclopedia/frameworks/symfony/> (дата звернення: 10.04.2024)
2. Turnitin <https://www.turnitin.com/infographics/the-plagiarism-spectrum> (дата звернення: 15.04.2024)
3. Herrera F. Multilabel classification: problem analysis, metrics and techniques / F. Herrera, F. Charte, A. J. Rivera, M. J. del Jesus // Springer International Publishing Switzerland, 2016. — P. 65-78.
4. Taylor J. An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements / J. Taylor. — University Science Books, 1999. — P. 128-129.