УДК 004.91

Д. С. Чорний
С. М. Захарченко

АВТОМАТИЗОВАНА СИСТЕМА КОНТРОЛЮ КОНТЕНТУ СТУДЕНТСЬКИХ РОБІТ

Вінницький національний технічний університет

Анотація
У роботі проаналізовано методи побудови автоматизованої системи контролю контенту студентських робіт.
Ключові слова : токенізація,нормалізація, n-грам, TF–IDF.

Abstract
The methods of construction for automated system of control of content of student's works are analyzed in the paper.
Keywords: tokenization,normalization, n-gram, TF–IDF.

Вступ

Інформатизація суспільства, стрімкий розвиток технологій та вільний доступ до продуктів інтелектуальної власності полегшують процес використання та розповсюдження інформації.

Перевірка контенту документів є поширеною задачею, вона може застосовуватися у різних сферах діяльності людини [1]. Тому розробка автоматизованої системи контролю контенту студентських робіт є дуже актуальним завданням. Вона може включати у себе: перевірку джерел походження – джерела плагіату, також перевірку окремих компонент – пошук варіанту завдання, IP адрес, інших слів, речень.

Результати дослідження

Задачу інтелектуальної перевірки можна розбити на окремі етапи (рис.1).

Перший етап вирішення проблеми автоматичної перевірки текстів є перетворення документів. На даному етапі документи у вигляді послідовності символів перетворюються у форму, придатну для алгоритмів машинного навчання відповідно до поставленої задачі. Важливо здійснити на даному етапі токенізацію та нормалізацію [2]. Одиницею тексту є слово. Токенізація означає поділ тексту на окремі одиниці слів, які називаються лексемами. В результаті токенізації оригінальний рядок тексту перетворюється на список слів, з яких він складався. Токенізація тексту виконується в кілька етапів:

1. Приведення вхідного тексту до нижнього регістру.
2. Видалення усіх розділових знаків та заміна їх на пробіли.
3. Оголошення слів окремими лексемами.

Другий етап - побудова класифікаційної функції. Якість класифікації залежить як від того, як документи будуть перетворені у векторне подання, так і від алгоритму, який буде застосовано на другому етапі [3]. Важливо зазначити, що методи перетворення тексту у вектор є специфічними для завдання класифікації текстів і можуть залежати від колекції документів, типу тексту (простого, структурованого) та мови документа. Методи машинного навчання, що використовуються на другому етапі, не є специфічними для проблеми класифікації тексту, а також використовуються в інших областях, наприклад, для розпізнавання шаблонів [4].

На наступному етапі здійснюється порівняння документа запиту з іншими. Паралельно можна здійснювати пошук окремих елементів – це може бути пошук обов’язкових компонентів, або щось інше.

У кінці формується звіт про роботу системи.

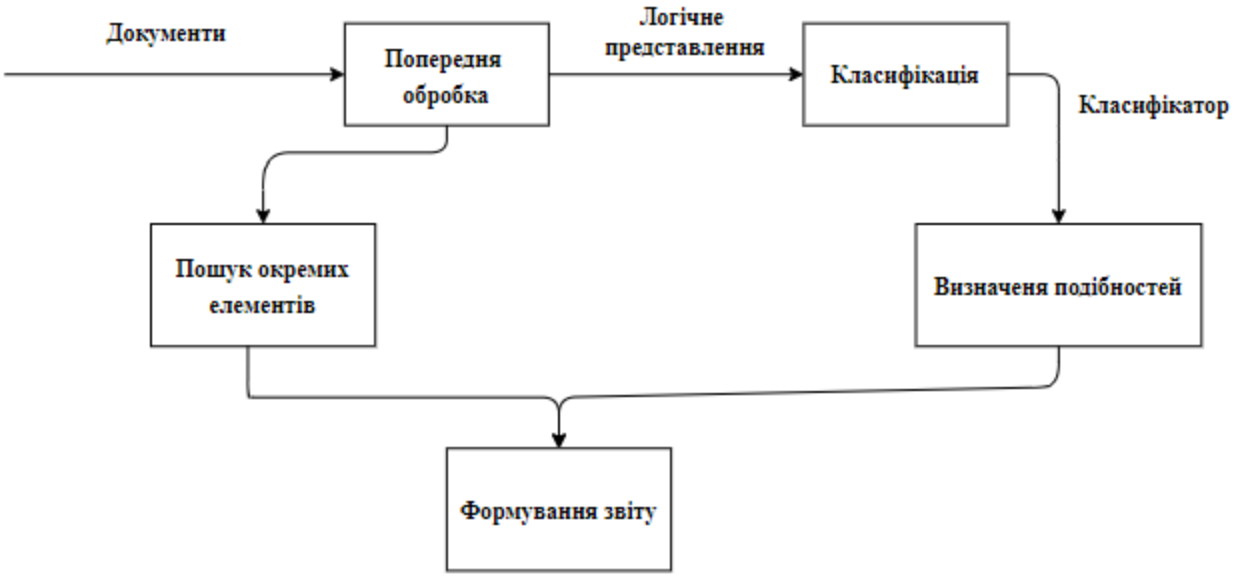

Рис.1. – основні етапи інтелектуальної перевірки документів

Для представлення документів було обрано модель векторно простору. Модель векторного простору представляє документи и як вектори у багатовимірному просторі, розміри яких є термінами, що використовуються для побудови індексу для представлення документів. Для покращення роботи використовуються n-грами. n-грами - це послідовності елементів, як вони з'являються в документі. Буква n вказує, скільки елементів слід врахувати. Наприклад, існують біграми (2 грами), триграми (3 грами), 4 грами, 5 грамів або інші.

У якості статистичної величини використовується TF–IDF. TF–IDF – посідає провідне місце серед схем зважування термінів сьогодні. Це числова статистична величина, показкує, наскільки важливим є вага слова для документа чи колекції документів.

Щоб знайти подібність векторів, використовуємо косинусну міру кута між векторами: чим гостріший кут, тим більший косинус.

Висновки

Проведено аналіз побудови автоматизованої системи контролю студентських робіт. Визначено основні етапи її створення.

Розглянуто розроблену теоретико-матиматичну модель вирішення поставленого завдання.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Михайловський Ю.Б. Система Anti-Plagiarism як інструмент запобіганняплагіату в навчальній та науковій діяльності / Ю.Б. Михайловський, Н.А.Длугунович // Вісник Хмельницького національного університету. Технічнінауки. – 2013. – № 3.— С. 162—168.
2. Moses S. Charikar. Similarity estimation techniques from rounding algorithms. Proceedings of the 34th Annual ACM Symposium on Theory of Computing, 2002, p. 380
3. Шинкаренко В. І. Система контролю плагіату в студентських роботах[Електронний ресурс] / В. І. Шинкаренко, О. С. Куроп' ятник // СхідноЄвропейський журнал передових технологій. – 2012. – Том. 4. – № 2(58). – С.32-36.
4. Bolilyi V. O. Check the Uniqueness of the Text in the Assessment of Student Work Creative or Exploratory / V.O. Bolilyi, V. V. Kopotii // Naukovi zapysky NDU im. M. Hoholia. – 2011. – № 7 (34).— P. 134—145

Чорний Денис Сергійович, студент групи 1КІ-19м факультет інформаційних технологій та комп'ютерної інженерії, Вінницький національний технічний університет, м. Вінниця, 1ki15b.chornyi@gmail.com.

Захарченко Сергій Михайлович – кандидат технічних наук, доцент кафедри обчислювальної техніки, Вінницький національний технічний університет, Вінниця, e-mail: zahar@vntu.net;

Chornyi Denys S, student, Faculty of information Technologies and Computer Engineering, Vinnytsa

National Technical University, Vinnytsia, 1ki15b.chornyi@gmail.com.

Zaharchenko Sergiy M. – PhD, Assistant Professor of the Computer Techniques Chair, Vinnitsa National Technical University. Vinnitsa, e-mail: zahar@vntu.net.