УДК 004.912

РОЗРОБКА ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ ДЛЯ ВИЯВЛЕННЯ АКАДЕМІЧНОГО ПЛАГІАТУ НА ОСНОВІ СТАТИСТИЧНОГО АНАЛІЗУ ТЕКСТУ

В. А. Голубенко[1], *Ю. І. Дорофєєв[2]*

[1] магістрант кафедри САІТ, НТУ «ХПІ», Харків, Україна
[2] завідувач кафедри САІТ, докт. техн. наук, НТУ «ХПІ», Харків, Україна
holubenko.vlad@gmail.com

Тема використання людиною інтелектуальних напрацювань інших людей є актуальною ще зі стародавніх часів. Приблизно у 80 році нашої ери римський поет Марціал використав слово «плагіарус», з якого пішло слово «плагіат», до іншого поета Фідентінуса, який видавав його твори за свої [1]. Після Середньовіччя багато всесвітньо відомих авторів також займалися запозиченням чужих творів та видаванням їх під своїми іменами. Так, наприклад, виявили, що один з найбільш відомих рисунків Леонардо да Вінчі «Вітрувіанська людина» є копією роботи його друга Джакомо Андреа.

Застосування сучасних ІТ технологій, зокрема, у сфері вищої освіти несе багато користі, але разом з тим з'являються певні негативні наслідки, такі як плагіат. Плагіатом вважається привласнення чужої інтелектуальної власності, в даному випадку тексту, без зазначення відповідного посилання. Отже, стає очевидним, що для виявлення авторства певної публікації необхідний інструмент, який дозволяє визначити випадки текстових запозичень. Існує достатньо подібних програм, наприклад Unicheck або AntiPlagiarism, але більшість з них є платними.

Метою даної роботи є розробка програмного продукту для автоматичного пошуку академічного плагіату на основі використання методів статистичного аналізу текстів. Для досягнення зазначеної мети необхідно вирішити наступні задачі: 1) визначити джерела та способи пошуку текстів для порівняння; 2) розробити способи зіставлення та обробки текстів; 3) програмно реалізувати процес попередньої обробки тексту; 4) програмно реалізувати обрані алгоритми зіставлення тексту; 5) здійснити тестування коректності порівняння текстів.

В результаті роботи було створено програмне забезпечення із застосуванням мови програмування Python. Для пошуку частин тексту, який перевіряється на наявність запозичень, використано адаптивний варіант алгоритму шинглів (англ. shingles - лусочки). Попередня обробка тексту включає токенізацію (розбиття на слова), що було реалізовано за допомогою пакету бібліотек та програм NLTK. З метою оцінювання міри збігання як між текстами, так і окремими реченнями застосовано метод «мішка слів» [2, 3], при застосуванні якого текст описується у вигляді мультимножини слів, не беручи до уваги граматику та послідовність слів. Оскільки у мультимножині зберігається лише основа слова, для реалізації стемінгу (англ. stemming - процес скорочення слова до основи шляхом відкидання допоміжних частин, таких як закінчення чи суфікс) були використані бібліотеки Snowball та uk_stemmer.

Список літератури:

1. The World's First "Plagiarism" Case [Електрон. ресурс]. – Режим доступу: https://www.plagiarismtoday.com/2011/10/04/the-world%E2%80%99s-first-plagiarism-case/.

2. *Маннинг К.Д.* Введение в информационный поиск/ *Маннинг К.Д., Рагхаван П., Шютце Х.* // «И. Д. Вильямс». – 2011. – 528 с.

3. *Ali A. M. E. T.* Survey of Plagiarism Detection Methods */ A. M. E. T. Ali, H. M. D. Abdulla, V. Snasel* // Proceedings of Fifth Asia Modelling Symposium. – 2011. – С. 39 – 42.