ВИКОРИСТАННЯ ПРОГРАМ ПЕРЕКЛАДАЦЬКОЇ ПАМ'ЯТІ ДЛЯ ВИЯВЛЕННЯ ПЛАГІАТУ В ТЕКСТАХ

Курсін А.І., Дженюк Н.В., Смишляєв А.В.
НТУ «ХПІ», м. Харків

Актуальність виявлення плагіату в текстах зростає в наш час з розвитком електронних методів зберігання, пошуку та копіювання інформації. З юридичної точки зору, об'єктом плагіату можуть бути не ідеї, думки або концепції, а лише їх оформлення, зовнішня оболонка. У випадку текстів – це фрагменти, цілком або частково позичені у інших авторів без належним чином оформлених посилань. Виявлення випадків плагіату вручну є практично неможливим. Серед існуючих засобів автоматизації цього процесу можна назвати, наприклад, Інтернет-сервіси antiplagiat.ru, copyscape.com та ін., але в них відсутні можливості пошуку у власному корпусі текстів користувача – лише в Інтернеті; не гарантується і конфіденційність документа, що перевіряється. Ці недоліки частково вирішуються такими програмами, як PlagiatInform, що мають, однак, високу ціну.

Ми хочемо звернути увагу на можливість використання для пошуку плагіату наявні програми, що належать до класу систем перекладацької пам'яті (ПП), такі як TRADOS, DejaVu та ін. Їхня функція – пропонувати можливий переклад певного речення на основі пошуку схожих з ним речень у корпусі вже переведених текстів. Аналіз задачі пошуку плагіату показує, що, попри свою складність та інтелектуальність, це є, по суті, лише пошук текстових фрагментів у великому масиві текстової інформації. Системи ПП мають в своєму складі потрібний механізм пошуку речень з повним і, що важливо, частковим збігом.

Розроблена нами методика полягає у внесені в БД програми ПП корпусу текстів, у яких буде проводитися пошук. Оскільки до БД програми ПП вносяться лише пари текстів оригінал-переклад, кожен документ нашого корпусу вноситься в парі сам з собою з певними «хитрощами», щоб обійти обмеження програм на введення пар текстів, що належать до однієї мови. Подальша обробка тексту, що перевіряється, проводиться згідно з процедурою перекладу нового тексту в програмі ПП. В результаті програма виділяє в тексті речення, для якого найдений збіг у корпусі, надає поруч нього «переклад», який завдяки внесенню в БД пар текстів «сам з собою» є оригінальним реченням з корпусу документів, та відсоток збігу цих речень. Врахування неповного збігу речень дає методиці стійкість до можливих спроб перехитрити її за допомогою перестановки та заміни окремих слів.