

Поповський О.І.
провідний інженер,
Інститут інформаційних технологій і засобів навчання
Національної академії наук України,
popovsky44@gmail.com

ОГЛЯД ПРОГРАМ ПОРІВНЯЛЬНОГО АНАЛІЗУ НА ЗБІГ

В даний час глобальні інформаційні мережі являють собою невід'ємне від життя суспільства складне соціальне, інформаційне і правове явище. В Інтернеті зосереджена колосальна кількість найрізноманітнішої інформації. Цю інформацію можна знайти і використовувати в тих або інших цілях. Захист результатів творчості, авторських прав безпосередньо пов'язане із захистом свободи особистості, прав людини взагалі. Найчастіше плагіат знаходить своє вираження у привласненні авторства на чужі результати інтелектуальної праці шляхом публікації їх під своїм ім'ям. Плагіат можливий і в частковому використанні чужого твору або цитуванні без посилання на джерело.

У складі НАПН України працюють близько 20 наукових установ у галузі психології та педагогіки. Результатом їх роботи є статті, дисертації, звіти. Одним із кроків по підвищенню якості робіт і захисту авторських прав може бути впровадження в практику перевірки на плагіат. У рамках цієї роботи була проведена технологічна перевірка роботи трьох програм аналізу текстів на збіг.

1. Комп'ютерна система порівняльного аналізу електронних текстів, призначена для виявлення збігів у текстових документах Національного авіаційного університету (НАУ України).
2. Програма Etxt Антиплагіат ресурсу www.etxt.ru (біржа копірайту).
3. Електронна система Anti-Plagiarist для перевірки на плагіат документів Хмельницького національного університету (ХНУ України).

1. Комп'ютерна система порівняльного аналізу електронних текстів

Національного авіаційного університету (НАУ України)

Ця система розроблена і застосовується в НАУ України для перевірки курсових і дипломних робіт. Всі документи перекладаються в формат *.txt* і вводяться в базу даних. База даних на MySQL. Програма порівнює лише текстову частину документа з базою даних НАУ України.

Результат роботи:

1. Показуються відсоток збігу і з яким документом (або з декількома).
2. Складається звіт.
3. Відображення на монітор співпадаючих частин тексту.

2. Програма Etxt Антиплагиат ресурсу www.etxt.ru

Програма Etxt Антиплагиат призначена для перевірки документів на оригінальність в Інтернеті. Документ перекладається в канонізований текст у форматі *.txt* і перевіряється. Перевірка виконується по шинглах. Шингл - фрагмент тексту, що складається з послідовності декількох слів. Пошук в Інтернеті здійснюється декількома пошуковими системами. В результаті показується відсоток оригінальності тексту та список сайтів з відсотком збігу у відповідному кольорі. Робота програми перевірялася на двох документах, складених з даних, узятих в основному з Інтернету. Програма знайшла всі джерела даних. При виявленні 100% збігу показується джерело плагіату і програма закінчує роботу.

3. Електронна система Anti-Plagiarist

Електронна система Anti-Plagiarist для перевірки на плагіат документів Хмельницького національного університету (ХНУ України). Перевірка виконується на сервері ХНУ. Користувачеві системи надається логін і пароль для доступу до системи на підставі договору. Документи перевіряються по базі даних ХНУ і можуть перевірятися по контенту сайтів в Інтернеті. Система дозволяє перевіряти документи у форматах *.doc*, *.txt*, *.pdf*. Найбільш переважний формат *.doc*. Після входу в систему документ вводиться в базу даних (і залишається в ній), виконується аналіз на порівняння по базі даних і видається звіт. Перевірка виконується по лексемах. Лексема — це окреме слово з усією сукупністю властивих йому форм словозміни й значень у різних контекстах. У звіті вказується кількість і відсоток збігу символів і лексем і джерело плагіату. У звіті також вказується у відсотках кількість помилок в документі. Це скорочення і терміни, які відсутні в словнику.

Виконано введення в базу даних одного і того ж документа під різними найменуваннями. Система відразу ж визначила повний збіг.

Якщо вводити в систему документи у форматі *.pdf*, то в деяких з них кодування кирилиці не розпізнається, що призводить до некоректної роботи програми. Ці документи вимагають додаткового перетворення в інший формат *.pdf*. Також відмічено, що при введенні в форматі *.pdf* деякі символи розпізнаються невірно, що призводить до помилок. Перевірка в Інтернеті виконується по лексемах. Формується файл звіту, в якому під номерами вказуються адреси джерел збігу. Також вказується відсоток збігу з кожним джерелом, найбільший відсоток збігу і сумарний відсоток збігу. За заявою розробника перевищення двоцентного збігу вважається плагіатом і ця частина документа виділяється червоним кольором. Перевірка по лексемах займає багато часу і триває кілька годин.

Проведено теоретико-експериментальні дослідження, досвід практичної роботи, розробка та впровадження у навчальний процес демонстраційної версії експертної системи.

Транспорт, що використовує простий пошук та метод співставлення зі зразком, на базі відкритого коду мови логічного програмування ТУРБО-ПРОЛОГ 2.

Висновок

Для впровадження в НАПН України системи перевірки на плагіат найкраще використовувати програму Anti-Plagiarist Хмельницького національного університету. Для цього треба купити сервер, встановити на ньому необхідне програмне забезпечення і створити базу даних НАПН України. Для створення необхідної бази даних потрібно, щоб усі роботи подавались на електронних носіях в форматі *.doc* і для цього потрібно багато часу і зусиль. Сервер потрібно встановити в приміщенні з обмеженим доступом.

Для перевірки робіт в Інтернеті краще використовувати програму Etxt Антиплагиат ресурсу *www.etxt.ru*, так як вона використовує пошук по шинглах, що набагато швидше ніж пошук по лексемах.

Список використаних джерел:

1. Плагіат и антиплагиат- что это такое и как проверить тексты сайта на уникальность онлайн или с помощью программ, как быть с плагіаторами. URL: <http://ktonanovenkogo.ru/seo/plagiat-antiplagiat-chto-eto-takoe-kak-proverit-teksty-sajta-unikalnost-onlajn-programm-plagiator.html>.
2. EtxtАнтиплагиат. URL: <http://etxt.ru>.
3. Anti-Plagiarism . URL: <http://ikc2.tup.km.ua>.