Поповський О.І.
провідний інженер відділу електронних інформаційних ресурсів і мережних технологій
Інституту інформаційних технологій і засобів навчання НАПН України

ОГЛЯД РОБОТИ ПРОГРАМ ПЕРЕВІРКИ ТЕКСТІВ НА ПЛАГІАТ

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. Важливим методом захисту від плагіату є перевірка робіт програмами порівняльного аналізу текстів. У цих цілях було виконано дослідження роботи трьох програм:

1. Комп'ютерна система порівняльного аналізу електронних текстів, призначена для виявлення збігів у текстових документах Національного авіаційного університету (НАУ України).
2. Програма Etxt Антиплагиат 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 (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. Можливий вивід 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. 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.

Приклад частини звіту програми:

[10:30:04] Запрос к поисковой системе №83 [1] (500 миллисек.): Rambler

[10:30:04] Запрос к поисковой системе №70 [1] (1109 миллисек.): Qip

[10:30:04] Загружена страница из запроса №83-3 (343 миллисек.):
http://studentam.net.ua/content/view/7560/

[10:30:04] Загружена страница из запроса №70-3 (156 миллисек.):
http://uareferats.com/index.php/referat/details/25610

[10:30:05] Загружена страница из запроса №70-2 (281 миллисек.):
http://murzim.ru/nauka/psihologija/vvedenie-v-psihologiju/26645-sklonnosti-i-zadatki.html

[10:30:05] Запрос к поисковой системе №86 [1] (1140 миллисек.): Google

[10:30:05] Отсутствуют результаты запроса №86 к поисковой системе

[10:30:05] Запрос к поисковой системе №89 [1] (609 миллисек.): Bing

[10:30:05] Загружена страница из запроса №84-3 (2453 миллисек.):
http://baeva.org.ua/?q=node%2F1&paged=10

[10:30:05] Загружена страница из запроса №70-1 (437 миллисек.):
http://www.br.com.ua/kurs/Pravo/104122.htm

[10:30:09] Ra Найдено 100% совпадений (25562 миллисек.) по адресу:
http://lib.iitta.gov.ua/527/1/Dis_Spirin.doc (468 миллисек.)

[10:30:29] ПС[ош.]:0, ПС[неож. ош.]:0, ПС[закач.]:82, ПС[таймаут]:0, ОС[ош.]:0, ОС[неож. ош.]:0, ОС[закач.]:100, ОС[таймаут]:0

[10:30:29] **Уникальность текста равна 0%**

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. 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.

Приклад частини звіту:

Anti-Plagiarism v-08.245

Максимальное совпадение с одним документом 99.0%

Словари проверки: en_US, ru_RU, ua_UA. **Ошибок в документах: 10%**

ID: 6830 Название: Диференційований пидхід у вивченні основ Добавлено в БД: 2013-01-03 Авторы: Спірін О.М. Руководители: Консультанты: Опоненты:	Документ		Суммарное совпадение по Базе Данных	
	Символы	Лексемы	Символы	Лексемы
	348812	1969	346639 (99%)	1969 (100%)

Источник плагиата

ID	Описание	Наличие плагиата в документе	
		Символы	Лексемы
6725	Название: диференційований підхід у вивченні основ штучного інтелекту Добавлено в БД: 2012-12-27 Авторы: Спірін О.М. Руководители: Консультанты: Опоненты:	346639 (99.0%)	2034 (100.0%)

При введенні в систему документів 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 двопроцентного збігу вважається плагіатом і ця частина документа виділяється червоним кольором.

Приклад частини звіту:

Anti-Plagiarism v-08.245

ID: 6725

Название: диференційований підхід у вивченні основ штучного інтелекту
Добавлено в БД: 2012-12-27
Авторы: Спірін О.М.
Руководители:
Консультанты:
Опоненты:

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 [54] [55] [56] [57].***

D	Совпадение в документе	Источник совпадения
56	0.6723093%	https://twitter.com/search%3Fq%3D%25232
55	0.6723093%	https://twitter.com/2
57	0.6723093%	http://www.youtube.com/watch%3Fv%3DQFCSXr6qnv4
54	0.6723093%	http://en.wikipedia.org/wiki/2_(number)
59	0.60095394%	http://www.urbandictionary.com/define.php%3Fterm%3D%253C3
5	0.072367504%	http://www.cnn.com/
337	0.070090204%	http://en.wikipedia.org/wiki/Members_of_the_Australian_House_of_Representatives,_1996%25E2%2580%25931998
335	0.070090204%	http://www.imdb.com/title/tt0115279/
338	0.070090204%	http://www.modularfords.com/f5/

Максимальное совпадение с одним источником: 0.6723093%

https://twitter.com/search%3Fq%3D%25232

Сумма совпадений: 8.287345%

Рекомендації:

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 бази даних необхідно, щоб усі роботи подавались на електронних носіях в форматі *.doc* і для цього потрібно багато часу. Сервер потрібно встановити в приміщенні з обмеженим доступом.

Для перевірки робот 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.