РАЗРАБОТКА ПРОГРАММ, ПРЕДНАЗНЧЕННЫХ ДЛЯ СРАВНЕНИЯ ДВУХ ТЕКСТОВ С ИСПОЛЬЗОВАНИЕМ МЕТОДА «ДАКТИЛОСКОПИИ» И АЛГОРИТМА «ШИНГЛОВ»

Н.В. Тимошин, Г.И. Шкатова
Томский политехнический университет
xaoma@mail.ru

Введение

В современном информационном мире существование плагиата является серьезной проблемой, требующей большого внимания. Понятие плагиата не имеет вполне определённого содержания, и в частных случаях не всегда возможно однозначно отделить его от сопредельных понятий: подражания, заимствования, соавторства и других подобных случаев сходства произведений. О какой бы задаче из вышеперечисленных не шла речь, общим для них является вопрос, связанный с установлением схожести между собой разных текстов. Задача сравнения текстов стала особенно актуальной с появлением интернета, когда соблюдать авторское право становится всё труднее и даже невозможно.

К настоящему моменту времени разработано много методов оценки схожести текстов. Методы характеризуются по типу оценки сходства. Глобальная оценка использует большие части текста или документа для нахождения сходства в целом, в то время как локальные методы на входе проверяют ограниченный сегмент текста.

К группе локальных методов можно отнести методы, основанные на проверке документа «дословным перекрытием». В таких методах используются классические алгоритмы сравнения строк: алгоритм Кнута -Морриса - Пратта, алгоритм Рабина-Карпа и др.

Проверка подозрительных документов в этой ситуации требует расчёта и хранения эффективно сопоставимые представления всех документов в справочной коллекции, которые сравниваются попарно. Однако сопоставление подстроки является нежизнеспособным решением для проверки больших коллекций документов (алгоритм отрабатывает в среднем 2h сравнений, где h — длина строки, в которой ведётся поиск).

Анализ "множества слов" является упрощением представления, используемого в обработке естественного языка и поиска информации. В этой модели текст представлен как неупорядоченный набор слов. Документы представлены в виде одного или нескольких векторов, которые используются для попарного вычисления сходства.

Цитирование — компьютерный метод выявления плагиата, предназначенный для использования в научных документах, позволяющий использовать цитаты и справочный материал. Определяет общие цитаты двух научных работ.

Шаблон цитат представляет собой подпоследовательности, содержащие не только общие цитаты для двух документов, но и подобный порядок и близость цитат в тексте, являющихся основными критериями для определения шаблона цитат.

Стилометрия или изучение языковых стилей — это статистический метод для выявления авторства анонимных документов и для компьютерной проверки на плагиат.

Строятся стилометрические модели для различных сегментов текста, отрывков, которые стилистически отличаются от других. И путём сравнения моделей можно обнаружить плагиат.

В настоящее время наиболее распространённым является метод, получивший название «Дактилоскопия». Из ряда документов выбирается набор из нескольких подстрок, которые и являются «отпечатками». Рассматриваемый документ будет сравниваться с «отпечатками» для всех документов коллекции. Найденные соответствия с другими документами указывают на общие сегменты текста. [1]

В представленной работе реализован метод с элементами «Дактилоскопии», получивший название «Алгоритм Шинглов». [2]

Логика метода сравнения текстов

Логика метода представлена на рис. 1. Исходными данными (блоки «Ввести проверяемый текст» и «Ввести эталонный текст») для анализа являются два текста. Текст «эталонный» - текст с подтвержденным авторством. Второй текст, назовем его «проверяемый» - текст с подозрением на плагиат. По логике алгоритма, проверяемый и эталонный тексты подвергаются канонизации. На этапе канонизации все слова преобразуются к нижнему регистру, убираются знаки пунктуации, вводные слова, служебные части речи. Канонизированный текст разбивается на шинглы, для которых создаются таблицы идентификаторов. Под шинглами понимаются последовательные наборы слов заданной длины. Мера схожести определяется процентом совпадающих шинглов.

Класс Plagiarism

Основу программы составляет класс Plagiarism (Плагиат). Данный класс реализует все описанные выше функции. Полями класса являются объекты структуры text_0. Структура хранит в себе текст и соответствующую ему хэш-таблицу.

Рис.1. Концептуальная схема работы алгоритма

Апробация алгоритма

Для проверки работы алгоритма в качестве эталонного текста использовался отрывок из работы «Применение кластеризации, при моделировании искусственных иммунных сетей» с международной научно-практической конференции студентов, аспирантов и молодых ученых 2010 года.

Проверяемый текст построен на базе эталонного путем внесения следующих изменений: замена слов, добавление новых слов, добавление лишних знаков препинания, изменения регистров. Фрагменты текстов представлены на рисунке 2. На данном рисунке красным цветом отмечены измененные части текста.

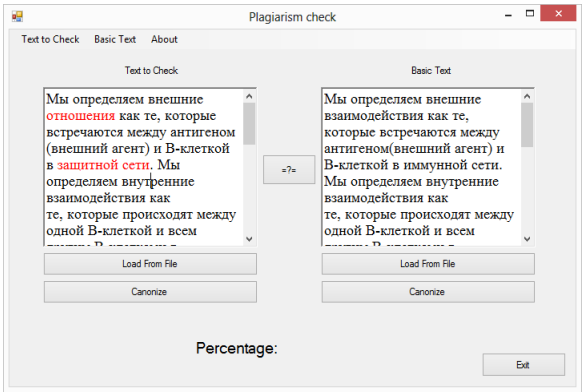

Рис. 2. Отображение эталонного и проверяемого текстов

Канонизированные тексты представлены на рисунке 3. После канонизации текстов для них создаются хэш-таблицы, которые позволяют произвести сравнительный анализ текстов. На рисунке 3 так же представлен результат сравнения текстов.

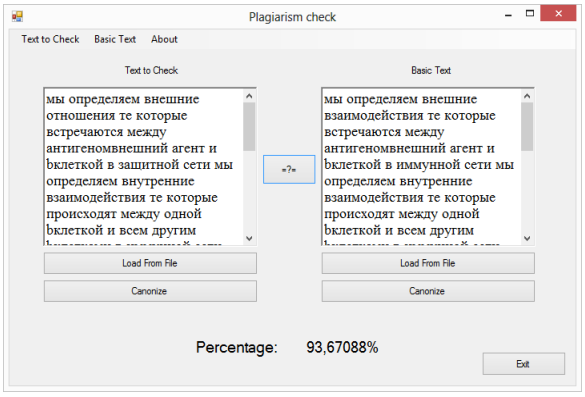

Рис. 3. Канонизированные тексты после сравнения

Результат сравнения текстов показал схожесть на 93,67%, что является показателем высокой схожести текстов.

Заключение

1. Путем проверки результатов работы метода на разных тестовых данных, установлено что разработанный класс может использоваться для грубой оценки меры схожести текстов.
2. Метод позволяет перейти к решению более глобальной задачи - выявлению плагиата.
3. Возможно улучшение работы алгоритма за счет
4. «Классовая» разработка программы упрощает внесение изменений в алгоритмы и добавление новых методов

Литература

1. [1] Определение плагиата [Электронный ресурс] // Википедия – свободная энциклопедия. URL: http://ru.wikipedia.org/wiki/Определение_плагиата (Дата обращения 22.05.2014)
2. [2] Алгоритм шинглов [Электронный ресурс] // Википедия – свободная энциклопедия. URL: http://ru.wikipedia.org/wiki/Алгоритм_шинглов (Дата обращения 22.05.2014)
3. Поиск плагиата методом шинглов [Электронный ресурс] // Кафедра АСОИУ ОмГТУ вики – страница. URL: http://wiki.asoiu.com/index.php/Поиск_плагиата_методом_шинглов (Дата обращения 22.05.2014)
4. Python: Алгоритм шинглов – поиск нечетких дубликатов текста [Электронный ресурс] // Code is art. URL: http://www.codeisart.ru/python-shingles-algorithm/ (Дата обращения 22.05.2014)