2 Torrey L. Transfer Learning. Handbook of Research on Machine Learning Applications / L.Torrey, J.Shavlik // IGI Global 2009, 22 c.

3 Zeiler M.D. Dept. of Computer Science / M.D.Zeiler, R.Fergus. // New York University, USA 2013, 16 c.

О.В. Антошина

ПРИМЕНЕНИЕ АЛГОРИТМОВ СЕМАНТИЧЕСКОГО АНАЛИЗА ТЕКСТОВОЙ ИНФОРМАЦИИ ДЛЯ ОПРЕДЕЛЕНИЯ ЗАИМСТВОВАНИЙ

(Самарский университет)

Введение

Свободный доступ к огромным массивам информации в сети Интернет способствует развитию плагиата в различных сферах человеческого общества. Под плагиатом понимают незаконное использование чужого изобретения или произведения без указания источника заимствования [1]. Для выявления заимствований и оценки самостоятельности автора применяют специализированные системы [2]. Главным недостатком некоторых автоматизированных систем поиска плагиата является отсутствие методов лингвистического анализа, в том числе семантического и морфологического.

Методы анализа текстовой информации

Задачи семантического и морфологического анализа способствуют классификации текстов, определению предметной области, а также более точному обнаружению плагиата.

Одной из задач морфологического анализа является определение нормальной формы искомого слова (процесс лемматизации), от которого наследуется данная словоформа, и набора параметров этой словоформы [3]. Таким образом, морфологический анализ позволяет избежать дополнительной обработки всех словоформ.

Другой тип морфологического анализа построен на основании правил с использованием процедурного подхода и позволяет привести словоформу к единой форме с помощью стемминга – выделения основы слова. Рассмотрим оба подхода [4].

1. Лемматизация – процесс приведения словоформ с одинаковым понятием к единой нормальной форме для увеличения релевантности поиска [5] и уменьшения количества анализируемых слов. Операция лемматизации может быть представлена в виде отображения (1):

$$T \rightarrow L, \quad (1)$$

где T – множество всех терминов, L – множество всех лемм. Данное преобразование позволяет уменьшить размер индексной информации и ускорить обработку текстового документа.

2. Стемминг – приближенный эвристический процесс, который находит основу слова [6]. В отличие от лемматизации стеммер не требует наличия

словаря (хранилища данных) и основан на правилах морфологии. Процесс стемминга осуществляется с помощью специального анализатора, разбивающего входные данные на наборы (последовательность) токенов и затем пропускающего токены через группу фильтров. В основе стемминга лежит алгоритм Портера [6], приводящий словоформы к единой форме.

Методы морфологического анализа реализованы библиотекой полнофункционального текстового поиска Apache Lucene [7], которая реализована в собственной системе проверки на текстовые заимствования.

Методы семантического анализа направлены на определение сущностей исходного текста, свойств и отношений между ними. Основными задачами семантического анализа являются возможность дальнейшего представления текста, его оценка, разбиение на составные части, анализ, в частности возможность применения специализированных методов для обнаружения текстовых заимствований.

Среди методов семантических поисковых моделей выделяют:

1) Метод накопительных сумм и его модификации для отслеживания изменений отклонений стилометрических характеристик. Основные этапы данного метода [2]:
   a. выбор пары характеристик – функций предложения и подсчет предложений для каждой пары-функции,
   b. вычисление средних значений для каждой пары-функции,
   c. построение накопительной суммы отклонений,
   d. построение графика зависимости предложений и кумулятивных сумм,
   e. сопоставление текстов по виду графиков (сравнение текстов).
2) Метод «шинглов» и его модификации. Основные этапы метода «шинглов» [2]:
   a. канонизация текста, включающая этапы морфологического анализа и этап очищения от стоп-слов.
   b. разбиение на шинглы внахлест (шингл – последовательность слов),
   c. вычисление хеш-значений (хеш-функций) шинглов.
   d. случайная выборка восьмидесяти четырех значений контрольных сумм.
   e. сравнение и определение результатов.
3) I-Match-метод, основанный на лексических принципах [8], его модификации и другие методы.

Апробация программы

Для выявления текстовых заимствований путем сравнения пары текстовых документов была реализована программа на языке программирования Java (среда – Intellij Idea), использована библиотека Lucene версии 7.1.0. и алгоритм «шинглов».

Сравниваемые тексты:

- Эталонные тексты:
  1. отрывок из статьи О.Р. Демидова «Плагиат: норма или аномалия?»,

2. текст из открытого тренировочного сборника задач (ФИПИ) по С.И. Каширину «Лётчик-парашютист»,
3. текст из открытого тренировочного сборника задач (ФИПИ) по Л. Соболеву «Морские спасатели».

- Сравниваемый текст – переведенный Яндекс-переводчиком на английский и обратно на русский эталонный текст.

Форма ввода данных представлена на рисунке 1.

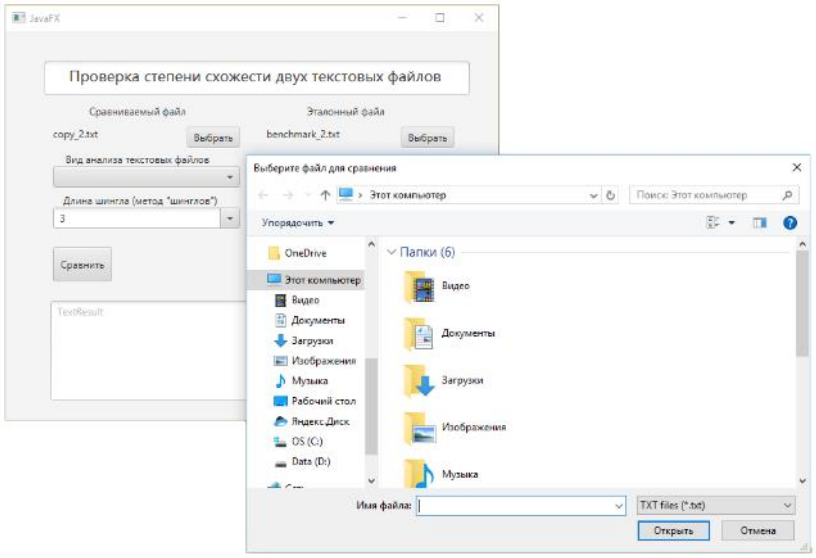

Рисунок 1 – Форма ввода данных

В таблице 2 представлены результаты работы собственной программы и существующих онлайн-сервисов, в таблице 1 – результаты сравнения текстов в случаях применения следующих морфологических методов:

1) лемматизации,
2) стемминга,
3) лемматизации и стемминга.

Таблица 1. Результаты работы программы Java (длина шингла – 3)

Морфологические методы	Процент заимствования первого текста	Процент заимствования второго текста	Процент заимствования третьего текста
Лемматизация	34,806	12,204	16,35
Стемминг	36,314	11,417	17,11
Лемматизация и стемминг	36,530	12,60	17,87

Таблица 2. Результаты работы программы Java и онлайн-сервисов

Сервис/Программа	Процент заимствования первого текста	Процент заимствования второго текста	Процент заимствования третьего текста
Java-программа	35,68	12,60	17,87
Онлайн сервис Back Links Manager	6,00	0,00	0,00
Онлайн сервис www.majento.ru	0,00	1,00	0,00
Онлайн сервис ciox.ru	41,79	30,56	35,52

Заключение

В статье были рассмотрены морфологические методы анализа, стемминг Портера и лемматизация, а также семантический метод «шинглов», определяющий процент заимствования текстовой информации. Были приведены результаты работы программы с применением этих методов (лемматизация и стемминг отдельно и вместе), а также сравнение наиболее точного способа нахождения (с применением лемматизации и стемминга) с некоторыми существующими бесплатными системами поиска заимствований. В дальнейшем в системе будут учтены синонимы и перестановка слов.

Литература

[1] Толково-энциклопедический словарь русского языка [Электронный ресурс]. – Режим доступа: https://slovar.cc/rus/tolk-enc/1462480.html (24.12.2016).

[2] Мошина, О.В. Применение методов семантического анализа текстовой информации [Текст] / О.В. Мошина, О.А. Гордеева // Труды Международного симпозиума «Надежность и качество». –2017. – Т. 1. – С. 371-375.

[3] Словарь терминов [Электронный ресурс]. – Режим доступа: https://seopult.ru/library/Словоформа (01.10.2017).

[4] Большакова, Е.И. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учебное пособие / Е.И. Большакова, Э.С. Клышинский, Д.В. Ландэ, А.А. Носков, О.В. Пескова, Е.В. Ягунова // – М.: МИЭМ. – 2011. – 272 с.

[5] Пересторонин П. Стеммер Портера для русского языка [Электронный ресурс]. – Режим доступа: https://medium.com/@eigenein/стеммер-портера-для-русского-языка-d41c38b2d340 (14.06.2017).

[6] Жердева, М.В. Стемминг и лемматизация в Lucene.NET / М.В. Жердева, В.М. Артюшенко // Лесной вестник. – 2016, №3. – С. 131-133.

[7] Apache Lucene Core [Электронный ресурс]. – Режим доступа: http://lucene.apache.org/core/ (02.10.2017).

[8] Астапова О.П. Исследование и разработка методов поиска плагиата в многоязычных корпусах текстов / О.П. Астапова [Электронный ресурс]. – Режим доступа: http://seminar.at.ispras.ru/wp-content/uploads/2012/07/Astapova-thesis1.pdf (24.12.2016).