УДК 004.056

Маг. И. В. Капелько
Науч. рук. доц. Д. В. Шиман
(кафедра программной инженерии, БГТУ)

АНАЛИЗ РАБОТОСПОСОБНОСТИ АЛГОРИТМА ТРЕХУРОВНЕВОЙ ТОКЕНИЗАЦИИ ДЛЯ АВТОМАТИЧЕСКОГО РЕФЕРИРОВАНИЯ ТЕКСТА

Автоматическое реферирование – это составление коротких изложений материалов, дайджестов, т.е. извлечение наиболее важных сведений из одного или нескольких документов и генерация на их основе лаконичных отчетов.

Для точного понимания термина реферата его можно сравнить с аннотацией. Реферат знакомит читателя с сутью содержания оригинала, используются формулировки и обобщения, заимствованные из текста оригинала. Аннотация даёт общее представление об оригинале. В аннотации отсутствует цитирование текста-источника. Основное содержание передаётся «своими словами», которые представляют собой высокую степень обобщения материала. Аннотация меньше информативного реферата в 3-4 раза. По объёму реферат всегда пространнее аннотации. Но рекомендации, касающиеся объёма обоих документов, в значительной степени варьируются. Так, для реферата разные авторы считают приемлемыми размеры от 200 до 1200 слов и сокращение текста в 3 или 8 и даже 10 раз. Реферату, состоящему из 10 – 120 слов (7 – 9 предложений), соответствуют оптимальные размеры аннотации от 40 до 60 слов (3 – 4 предложения).

По реферату можно составить мнение о содержании, сути излагаемого в оригинале содержания, аннотация даёт представление только о главной теме и о перечне вопросов, затрагиваемых в нём.

Описание работы алгоритма трёхуровневой токенизации текста. В основе алгоритма лежит метод TF-IDF (от англ. TF — termfrequency, IDF — inversedocumentfrequency) – статистическая мера, используемая для оценки важности слова в контексте документа, являющегося частью коллекции документов или корпуса.

Перед применением каких-либо методов реферирования необходимо использовать предварительную обработку текста: удаление стоп-слов, исправление грамматических ошибок.

Под трехуровневой токенизацией подразумевается, что токеном будет являться не только какое-то отдельное слово, но и предложение и абзац. Токеном первого уровня будет абзац текста, который включает в себя коллекцию токенов второго уровня – отдельные предложения этого абзаца. Токеном третьего уровня является слово. Для каж-

дого слова будет выделена его лемма – неизмененная, исходная форма слова, а для нее рассчитана TF-IDF. Где TF – отношение числа вхождений некоторой леммы к общему числу слов документа. Таким образом, оценивается важность леммы в пределах отдельного документа. IDF – это обратная частотность документов. Она измеряет непосредственно важность термина. В моем алгоритме ее можно выразить формулой:

$$IDF = \log_{10}(\frac{n}{w}),$$

где n – количество абзацев, w – количество абзацев в которых встречается токен.

После расчёта TF-IDF у каждого слова есть свое числовое значение, выражающее его вес в тексте. Для расчёта веса предложений необходимо сложить значения токенов (слов), которые входят в состав этого предложения и разделить на их количество. Далее ту же операцию необходимо применить и для абзацев: сложить значение весов предложений и разделить на количество этих предложений в абзаце.

Таким образом, можно выделить наиболее значимые слова, а на их основе выделить наиболее важные части текста.

Анализ результатов использования алгоритма трехуровневой токенизации. В ходе написания магистерской работы было создано программное средство для использования алгоритма трехуровневой токенизации текста. Приложение создано на языке C# при помощи технологии WinForms. Так же были подготовлены различные научные и художественные тексты. Все научные тексты имели отношение к IT сфере для удобства оценки качества реферата. Для проверки работоспособности алгоритма необходимо загрузить файл в формате txt в разработанное приложение. После чего начинается предварительная обработка текста и его разделение на токены, а также расчет их весов. В предварительную обработку текста также входит процесс удаление заголовков разделов, т.к. они в большинстве случаев содержат слова значение весов, которых достаточно велики. В результате на экран пользователю выводится таблица с результатами (рисунок 1).

На данном этапе можно посмотреть отсортированные списки токенов всех уровней. Далее необходимо выбрать процент реферата от основного текста и нажать кнопку показать результат. После этого происходит расчет необходимого числа параграфов, сортировка их по весам, а затем сортировка по номеру абзаца в тексте. Реферат выводится в отдельном окне.

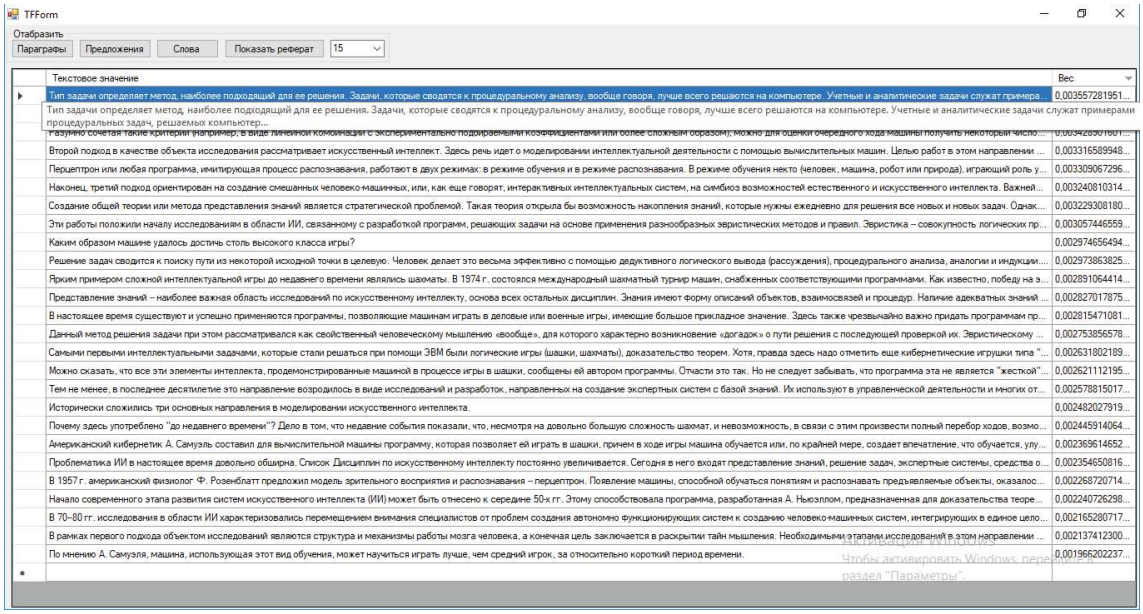

Рисунок 1 – Таблица весов токенов

В процессе проверки работоспособности алгоритма было выявлено, что наиболее информативный реферат получается только для научных текстов, т.к. у художественных текстов невозможно выделить только несколько основных абзацев. Оптимальный объем реферата от 15 до 20% от основного текста.

ЛИТЕРАТУРА

1. Miner. Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications. Academic Press, 2012 стр. 71–138.

2. TF-IDF с примерами кода: просто и понятно [Электронный ресурс] / NLPx – Режим доступа: http://nlpx.net/archives/57. – Дата доступа 10.04.2019.

3. Fayyad U., Piatetsky-Shapiro G., Smyth P. From Data Mining to Knowledge Discovery: an Overview // Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996. − стр. 1-34.