

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 трёхуровневой 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 – неизменная, 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:

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

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.

Текстовое значение	Вес
Тип задачи определяет метод, наиболее подходящий для ее решения. Задачи, которые сводятся к процедуральному анализу, вообще говоря, лучше всего решаются на компьютере. Учетные и аналитические задачи служат примерами процедуральных задач, решаемых компьютером...	0.003557281951...
Газуноу сочетая такие критерии (например, в виде линейной комбинации) с экспериментально подобранными коэффициентами или более сложным образом; можно для оценки очередного хода машины получить некоторый число...	0.002974655494...
Второй подход в качестве объекта исследования рассматривает искусственный интеллект. Здесь речь идет о моделировании интеллектуальной деятельности с помощью вычислительных машин. Целью работ в этом направлении...	0.003316589948...
Перцептрон или любая программа, имитирующая процесс распознавания, работают в двух режимах: в режиме обучения и в режиме распознавания. В режиме обучения некто (человек, машина, робот или природа), играющий роль у...	0.003309067296...
Наконец, третий подход ориентирован на создание смешанных человеко-машинных, или, как еще говорят, интерактивных интеллектуальных систем, на симбиоз возможностей естественного и искусственного интеллекта. Важней...	0.003240810314...
Создание общей теории или метода представления значений является стратегической проблемой. Такая теория открыла бы возможность накопления знаний, которые нужны ежедневно для решения все новых и новых задач. Однак...	0.003229308180...
Эти работы положили начало исследованиям в области ИИ, связанному с разработкой программы, решающих задачи на основе применения разнообразных эвристических методов и правил. Эвристика – совокупность логических пр...	0.003057446559...
Каким образом машине удалось достичь столь высокого класса игры?	0.002974655494...
Решение задач сводится к поиску пути из некоторой исходной точки в целевую. Человек делает это весьма эффективно с помощью дедуктивного логического вывода (рассуждения), процедурного анализа, аналогии и индукции...	0.002973853825...
Ярким примером сложной интеллектуальной игры до недавнего времени являлись шахматы. В 1974 г. состоялся международный шахматный турнир машин, снабженных соответствующими программами. Как известно, победу на э...	0.002891064414...
Представление знаний – наиболее важная область исследований по искусственному интеллекту, основа всех остальных дисциплин. Знания имеют форму описаний объектов, взаимосвязей и процедур. Наличие адекватных знаний ...	0.002827017875...
В настоящее время существуют и успешно применяются программы, позволяющие машинам играть в деловые или военные игры, имеющие большое прикладное значение. Здесь также чрезвычайно важно придать программам пр...	0.002815471081...
Данный метод решения задачи при этом рассматривался как свойственный человеческому мышлению «вообще», для которого характерно возникновение «догадок» о пути решения с последующей проверкой их. Эвристическому ...	0.002753856578...
Самыми первыми интеллектуальными задачами, которые стали решаться при помощи ЭВМ были логические игры (шахки, шахматы), доказательство теорем. Хотя, правда здесь надо отметить еще кибернетические игрышки типа "...	0.002631802189...
Можно сказать, что все эти элементы интеллекта, продемонстрированные машиной в процессе игры в шахки, сообщены ей автором программы. Странно это так. Но не следует забывать, что программа эта не является "жесткой"...	0.002621121195...
Тем не менее, в последнее десятилетие это направление возродилось в виде исследований и разработок, направленных на создание экспертных систем с базой знаний. Их используют в управленческой деятельности и многих от...	0.002578015017...
Исторически сложились три основных направления в моделировании искусственного интеллекта.	0.002482027919...
Почему здесь употреблено "до недавнего времени"? Дело в том, что недавние события показали, что, несмотря на довольно большую сложность шахмат, и невозможность, в связи с этим произвести полный перебор ходов, возмо...	0.002445914064...
Американский кибернетик А. Самуэль составил для вычислительной машины программу, которая позволяет ей играть в шахки, причем в ходе игры машина обучается или, по крайней мере, создает впечатление, что обучается, улу...	0.002369614652...
Проблематика ИИ в настоящее время довольно обширна. Список Дисциплин по искусственному интеллекту постоянно увеличивается. Сегодня в него входят представление знаний, решение задач, экспертные системы, средства о...	0.002354650816...
В 1957 г. американский физиолог Ф. Розенблат предложил модель зрительного восприятия и распознавания – перцептрон. Появление машины, способной обучаться понятиям и распознавать предъявляемые объекты, оказалось...	0.002268720714...
Начало современного этапа развития систем искусственного интеллекта (ИИ) может быть отнесено к середине 50-х гг. Этому способствовала программа, разработанная А. Ньюэллом, предназначенная для доказательства теоре...	0.002240726298...
В 70–80 гг. исследования в области ИИ характеризовались переключением внимания специалистов от проблем создания автономно функционирующих систем к созданию человеко-машинных систем, интегрирующих в единое цело...	0.002165280717...
В рамках первого подхода объектом исследований являются структура и механизмы работы мозга человека, а конечная цель заключается в раскрытии тайн мышления. Необходимыми этапами исследований в этом направлении ...	0.002137412300...
По мнению А. Самуэля, машина, использующая этот вид обучения, может научиться играть лучше, чем средний игрок, за относительно короткий период времени.	0.001966202237...

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://nlp.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.