Министерство образования и науки Российской Федерации
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ
Н.Г.ЧЕРНЫШЕВСКОГО»

Кафедра дискретной математики и
информационных технологий

Информационные технологии анализа текстов

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 5 курса 521 группы
направления 09.03.01 «Информатика и вычислительная техника»
факультета компьютерных наук и информационных технологий
Калдихина Евгения Алексеевича

Научный руководитель
д. э. н., профессор _____ Л.В. Кальянов
подпись, дата

Зав. кафедрой
к. ф.-м.н., доцент _____ Л.Б. Тяпаев
подпись, дата

Саратов 2018

ВВЕДЕНИЕ

На современном этапе развития информационных технологий во всем мире актуальным является вопрос эффективной работы поисковых сервисов.

Актуальность данной работы обусловлена множеством ошибок и неточностей, допускаемых современными поисковыми системами. Не всегда, список ссылок и документов, выданных по поисковому запросу является точным, что вынуждает пользователя проводить дополнительную ручную классификацию полученных результатов и извлекать из списка самое необходимое. Это обусловлено низким качеством классификации документов, и слабой эффективностью распознавания текстовой информации на уровне вычислительной машины.

Целью данной работы является поиск и реализация наиболее эффективного решения задач аннотирования и классификации текстов. **Объектами исследования** являются текстовые документы различной тематики.

Задачами данной работы являются:

– рассмотрение основных проблем, решаемых при помощи интеллектуального анализа текста;

– рассмотрение основных задач интеллектуального анализа текста;

– обзор наиболее актуальных методов интеллектуального анализа текста;

– обзор программного обеспечения для решения задач интеллектуального анализа текста

– решение задачи аннотирования и классификации текстовых документов;

– проведение анализа полученных результатов.

Работа состоит из введения, трех глав, заключения, списка определений, обозначений и сокращений, и списка использованных источников.

Первая глава «Основные задачи интеллектуального анализа текста и методы их решения» описывает основные теоретические сведения, касающиеся процессов интеллектуального анализа текстов, а также приводятся примеры методов, способных решать современные задачи интеллектуального анализа

текста. Во второй главе «ПО для решения задач интеллектуального анализа текста» проведен обзор ПО, позволяющего проводить выстраивание аналитических алгоритмов и получать необходимые результаты. В третьей главе «Решение задач интеллектуального анализа текста средствами ПО KNIME» приведено решение задачи аннотирования текстовых документов, посредством «градиентного бустинга над решающими деревьями», и проведена оценка работы наиболее актуальных методов классификации, с выявлением самых эффективных и быстрых методов. В заключении приведены основные результаты и выводы по проделанной работе.

Основными проблемами, которые решает интеллектуальный анализ текста являются:

– автоматизация и увеличение качества анализа анкетных данных опросов и исследований;

– сортировка больших объемов информации, путем обработки текста и анализа его содержания;

– оптимизация результатов выдачи информации, поисковыми системами, на запрос пользователя;

– идентификация основных технических проблем, создание отчетов по данным анализа записей центра телефонных продаж;

– определение коренных причин возникновения проблем по данным сообщений о происшествиях;

– прогнозирование поведения и настроения потенциальных клиентов;

– обнаружение и визуализация взаимосвязей в использовании биомаркеров (биологических признаков);

– прогнозирование суброгационного потенциала, исходя из анализа заявлений о страховых компенсациях, поступающих в страховые компании.

На данный момент, отрасль анализа текстовой информации находится в постоянном развитии и совершенствовании, поэтому с каждым днем - одни проблемы будут терять актуальность, а на смену им – придут новые проблемы, которые окажутся решаемыми, посредством методов интеллектуального анализа текстов [1].

Исходя из того, какие проблемы может решать интеллектуальный анализ текста, был сформирован определенный ряд задач, к которым относятся:

– классификация текстов;

– кластеризация текстов;

– нахождение шаблонов данных;

– определение тематики или области знаний;

– аннотирование текстов;

– анализ тональности;

– задачи автоматической фильтрации контента;

– определение семантических связей.

Для реализации данных задач, потребуется выполнить ряд предварительных этапов. На рисунке 1 представлен процесс анализа текста.

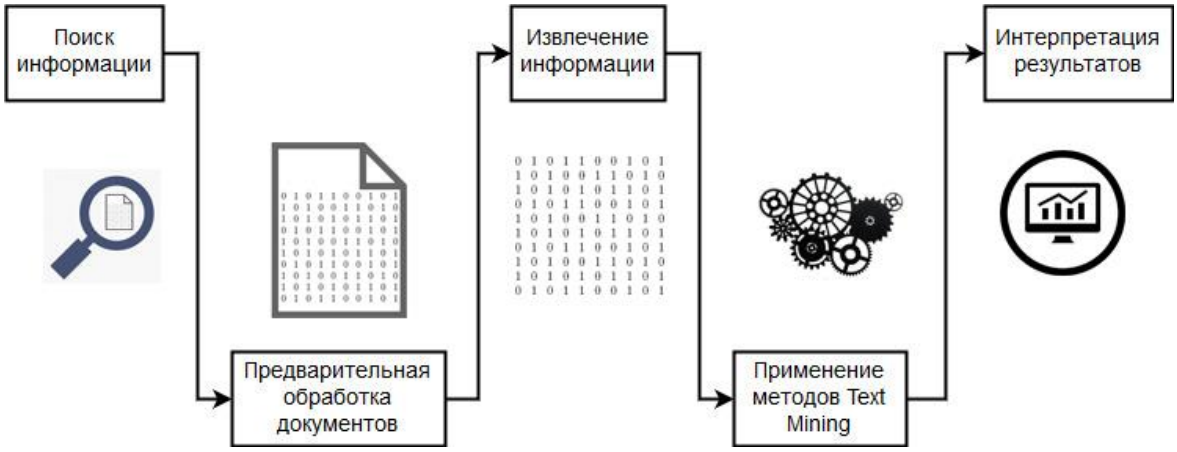

Рисунок 1 – Процесс анализа текста.

Наиболее эффективными и сохраняющими актуальность методами кластеризации и классификации на данный момент являются:

– скрытый семантический анализ или индексирование;

– кластеризация методом суффиксного дерева;

– метод K-средних;

– наивный Байесовский классификатор;

– метод концептуальной индексации;

– метод «самоорганизующейся сети Кохонена»;

– дерево принятия решений;

– метод «опорных векторов»;

– метод «ближайшего соседа»;

– метод «градиентного бустинга над решающими деревьями» [2, 3].

Прежде чем приступить к решению основных задач, были рассмотрены три крупных программных продукта – KNIME, RapidMiner и Statistica. После обзора этих программных продуктов, была составлена таблица с преимуществами и недостатками каждого из продуктов, оцененных по десяти-

бальной шкале, где 0 – самая низкая оценка и 10 – самая высокая. Результаты представлены в таблице 1.

Таблица 1 – Сравнение продуктов для решения задач интеллектуального анализа текстов.

Продукт	Доступность	Функциональность	Удобство	Общая оценка
Statistica	7	9	8	8
RapidMiner Studio	8	9	9	9
KNIME	10	9	9	9

После выбора программного обеспечения, были реализованы алгоритмы, решающие задачи аннотирования и классификации. В виду того, что полное представление алгоритма аннотирования занимает большое количество места, в дальнейшем будут приведены последовательные части этого алгоритма.

На рисунке 2 – приведен универсальный алгоритм извлечения текстовой информации из текстовых документов, работающий со всеми известными текстовыми файлами.

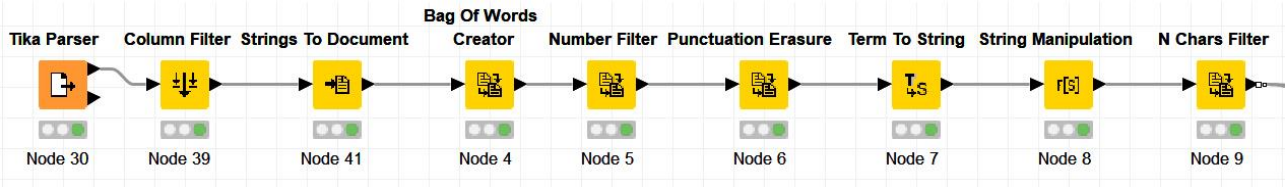

Рисунок 2 – Алгоритм извлечения текстовой информации из текстовых документов и ее последующая обработка.

На рисунке 3 – изображен алгоритм по извлечению статистических данных из текста, таких как – частота слов и n-грамм.

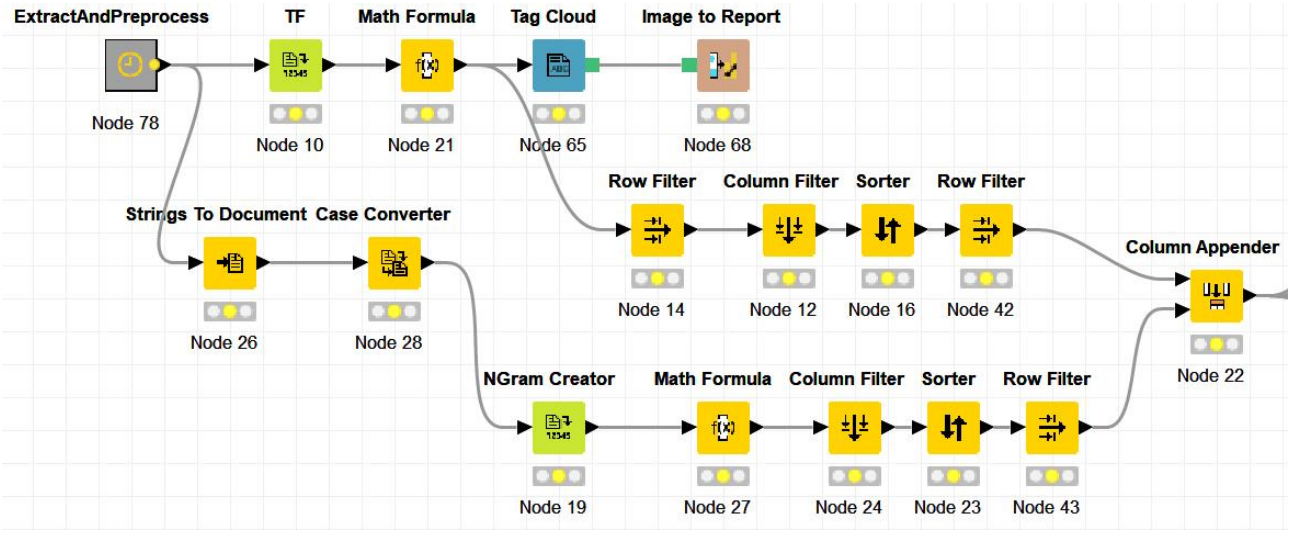

Рисунок 3 – Алгоритм, подсчитывающий количество слов и n-грамм, после предварительной обработки текста.

Далее, алгоритм подсчитывающий количество слов и n-грамм осуществляет дополнительную фильтрацию слов, оставляя около 20 значений для слов и n-грамм.

После прохождения этих этапов происходит следующий этап, основанный на техниках машинного обучения с использованием «градиентного бустинга над решающими деревьями» [4]. Идея бустинга заключается в комбинации функций с невысокой обобщающей способностью, которые выстраиваются в ходе итеративного процесса, и на каждом последующем шаге новая модель обучается с использованием данных об ошибках предыдущих.

Далее осуществляется сортировка полученных значений по их наибольшей значимости, результат работы алгоритма для текста с тематикой «системное программирование», можно увидеть на рисунке 4.

Row ID	S Prediction ()	D Confidence
Row1	вычислительных систем	1
Row4	развитии аппаратуры	1
Row5	системные программы	1
Row6	системных программ	1
Row7	успехи развитии	1

Рисунок 4 – Данные, полученные в результате применения метода «градиентного бустинга над решающими деревьями».

Для решения задачи классификации, были использованы четыре метода – «дерево принятия решений», метод «опорных векторов», метод «k-ближайших соседей» и «градиентный бустинг над решающими деревьями». Для лучшей наглядности результата, выборка документов была расширена до 20 документов. Данная выборка была разделена на 4 группы по 5 документов со следующими тематиками:

1) информационные технологии;
2) физика;
3) медицина;
4) экономика.

Точность общей классификации документов методом «дерево принятия решений» составила около 60%, методом «опорных векторов» - приблизительно 45%, методом «k-ближайших соседей» – 85% и методом «градиентного бустинга над решающими деревьями» - 85%.

Наиболее эффективным среди четырех методов, оказался метод «градиентного бустинга над решающими деревьями». Его эффективность равна 85%, как и у метода «k-ближайших соседей», но точность определения принадлежности классу оказалась выше чем у других методов.

ЗАКЛЮЧЕНИЕ

В ходе выполнения выпускной квалификационной работы были получены следующие результаты:

– рассмотрены основные проблемы интеллектуального анализа текста;

– рассмотрены основные задачи интеллектуального анализа текста;

– проведен обзор наиболее актуальных методов интеллектуального анализа текста;

– проведен обзор программного обеспечения для решения задач интеллектуального анализа текста;

– решена задача аннотирования и классификации текста, с выявлением наиболее эффективного метода классификации.

В результате выполненной работы, были исследованы новые эффективные способы по извлечению текста из документов и его дальнейшей обработке, которые позволили обнаружить новые закономерности и подходы к решению задач аннотирования и классификации текста. Разработанные методики позволяют более эффективно производить аннотирование документов и классифицировать их, в дальнейшем это позволит повысить эффективность и скорость алгоритмов поисковых сервисов.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 Barker K. Cornacchia N. Using Noun Phrase Heads to Extract Document Keyphrases. Advances in Artificial Intelligence. 2000, vol. 1822, pp. 40–52.

2 Rapid MineR Data Mining Use Cases and Business Analytics Applications / editors, Markus Hofmann, Ralf Klinkenberg, includes bibliographical references and index. ISBN-13: 978-1-4822-0550-3.

3 Christopher D. Manning, Hinrich Schutze. Foundation of Statistical Language Processing. MIT Press.

4 S.E. Robertson. K. Spak Jones., 1976. Relevance weighting of search terms. Journal of the American Society for Information Science May-June 1976.