Министерство образования Республики Беларусь
Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК 004.852

Чугаинов
Кирилл Викторович

Машинное обучение с использованием облачной платформы IBM Bluemix

АВТОРЕФЕРАТ

на соискание степени магистра информатики и вычислительной техники по специальности 1-40 81 04 «Обработка больших объемов информации»

Научный руководитель
Пилецкий Иван Иванович
доцент, к. ф.-м. н

Минск, 2017

КРАТКОЕ ВВЕДЕНИЕ

В конце XX века возник взрывообразный рост различных публикаций в сети Интернет. Увеличение доступности и снижение порога вхождения, привели к тому, что, и в настоящее время, количество информации растёт экспоненциально. Согласно статистике, объём цифровой информации удваивается каждые восемнадцать месяцев. Социальные сети, блоги, новостные порталы и развлекательные ресурсы ежедневно заполняются большими объемами контента. По большей части этот поток состоит из неструктурированных данных, представленных текстами на естественных языках.

С точки зрения обычных пользователей, возникает необходимость поиска нужной информации среди всего информационного «шума». Эта необходимость породила для аналитиков и разработчиков программного обеспечения ряд задач, среди которых можно выделить автоматическую обработку текстов с целью получения структурированных данных. Потом эти данные можно использовать для решения множества других проблем, таких как поиск, получение фактических данных, анализ количественных и качественных характеристик и т.п.

В данной работе будут рассмотрены различные способы обработки текстов на естественном языке и классификация полученных результатов с помощью алгоритмов машинного обучения. В качестве примера использования данных алгоритмов спроектировано и разработано приложение для кластеризации новостей из различных источников по ключевым словам.

Незатронутыми остаются вопросы лингвистического обоснования выбранных методов построения моделей, которые, в силу своей специфики и множественности подходов, могут стать темой отдельной диссертации.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность

Для решения многих современных проблем используется машинное обучение. Машинное обучение – актуальная и интенсивно развивающаяся область научного знания и передовая технологическая дисциплина, которая позволяет упростить и автоматизировать обработку больших объемов данных. Изучением и созданием алгоритмов машинного обучения занимаются самые крупные и передовые корпорации, но этот механизм позволяет решать и ряд более мелких задач.

В условиях современного развития технологий и сети Интернет, многие новостные издательства перешли на онлайн формат. Небольшой размер публикаций и скорость, с которой можно донести новость пользователям, породили проблему новостного избытка. Количество статей и их тематик значительно превышают ожидаемый уровень, что усложняет поиск только тех тем, в которых пользователь заинтересован сильнее прочих. Поэтому задача кластеризации новостного потока стоит достаточно остро для обычного пользователя.

Цели и задачи исследования

Целью данной работы является исследование методов машинного обучения и создание приложения для анализа с использованием алгоритмов машинного обучения.

Для достижения поставленной цели необходимо решить следующие задачи:

1. Провести анализ особенностей структуры текста интернет-новостей.
2. Изучить подход классификации и особенности существующих алгоритмов кластеризации.
3. Провести сравнительный анализ нескольких алгоритмов, наиболее подходящих к обработке новостного потока.
4. На основе полученных данных сделать вывод о наиболее подходящем методе обработки и кластеризации новостного потока.
5. Разработать приложение «News Analyzer», реализующее выбранный подход.

Объектом являются алгоритмы и методы машинного обучения.

Предметом является приложение, использующее алгоритмы машинного обучения, для решения задачи.

Основной гипотезой, положенной в основу данной работы является возможность использования алгоритмов машинного обучения для обработки и кластеризации текстов интернет-новостей, написанных на естественном языке.

Личный вклад соискателя

Результаты, приведенные в диссертации, получены соискателем лично. Вклад научного руководителя И. И. Пилецкого заключается в формулировке целей и задач и руководстве ходом работы.

Апробация результатов диссертации

Основные положения диссертационной работы докладывались на международной научно-практической конференции «Актуальные вопросы современных исследований» № 2 (Омск, май 2017) и № 3 (Омск, июнь 2017).

Опубликованность результатов диссертации

По теме диссертации опубликовано 2 печатные работы в сборниках трудов и материалов конференций.

Структура и объём диссертации

Диссертация состоит из общей характеристики работы, введения, четырех глав, заключения, списка использованных источников, приложений.

Общий объём работы составляет 63 страницы, 9 рисунков и 1 таблицы, список использованных источников из 20 наименований на 2 страницах и 1 приложение на 5 страницах.

Во введении обозначена проблематика информационного переизбытка в сети Интернет.

В первой главе рассмотрена предметная область и ее проблемы, определены основные цели и задачи исследования.

Во второй главе проведен анализ существующих решений, производится обоснование выбора метода кластеризации новостного потока.

В третьей главе анализируются платформы и сервисы, позволяющие реализовать выбранный в третьей главе подход.

В четвертой главе описан процесс разработки приложения «News Analyzer» и результаты его работы.

В приложении приведен листинг модуля реализующий EM-алгоритм.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** рассмотрено современное состояние проблемы информационного переизбытка в сети Интернет, а также дается обоснование актуальности темы диссертационной работы.

В первой главе рассматриваются проблемы обработки текстов на естественном языке, кластеризации таких текстов, а также особенности структуры интернет-новостей. Производиться описание существующих подходов к машинному обучению и дается их краткое описание.

Во второй главе выделяются отличия подхода кластеризации от подхода классификации. Рассматриваются наиболее распространенные подходы к тематическому моделированию, описание их достоинств и недостатков. Проводится анализ поставленной задачи и на основе проведенного исследования производится выбор подходящих алгоритмов.

В третьей главе представлены платформы и сервисы, позволяющие реализовать выбранные подходы. Приводится краткое описание моделей облачных сервисов. Описаны особенности IBM Bluemix и его когнитивных сервисов. Производится выбор языка программирования.

В четвертой главе описан процесс разработки приложения «News Analyzer». Описана реализация алгоритмов тематического моделирования и приведена сравнительная оценка, на основе которой делается заключение о том, что наиболее подходящим методом является робастный вероятностный латентный семантический анализ.

ЗАКЛЮЧЕНИЕ

В работе была затронута задача обработки больших объёмов данных, являющаяся нетривиальной для исследователя. Были проанализированы основные проблемы, с которыми приходится сталкиваться, обрабатывая тексты на естественном языке: нормализация текстов, выделение словаря, построение лингвистических моделей. Были проанализированы основные подходы предоставления облачных технологий.

В рамках исследования была спроектирована и реализована специализированная система News Analyzer, которая позволяет осуществлять кластеризацию новостных статей и поиск по ним. В работе были рассмотрены алгоритмы машинного обучения, позволяющие строить лингвистические модели без участия пользователя. Были выделены различия между алгоритмами, их слабые и сильные стороны и применимость к отдельным классам задач. Для нахождения алгоритма, наиболее подходящего для решения поставленной задачи, были изучены особенности структуры интернет-новостей.

Была выполнена выработка алгоритмов обработки текста и кластерного анализа, алгоритмов и методик интерпретации результатов. Для этого были разработаны распределённые алгоритмы предварительной обработки и интерпретации результатов, а также применены средства когнитивного анализа IBM Watson Natural Language Understanding. До непосредственного старта разработки приложения был произведен качественный анализ предметной области. Была продумана и описана архитектура будущего приложения. Приложение было реализовано согласно разработанной архитектуре. Также были проведены различные виды тестирования приложения, что помогло выявить ошибки реализации приложения и их исправить. Разработанная система является масштабируемой. Полученные результаты можно применять для маркетинговых исследований с целью продвижения продукции или услуг, написание статей и книг, отражающих наиболее актуальные темы и многие другие.

В качестве продолжения исследования можно реализовать следующие идеи:

1. Использование метаданных статей, которые позволят ещё сильнее ранжировать результаты.
2. Попробовать реализовать и применить другие алгоритмы кластеризации, или усилить уже реализованный.
3. Применение других видов графической визуализации.
4. Применение полученных результатов для других видов текстов на естественном языке, с целью расширения применения приложения.

5. Написание собственной библиотеки алгоритмов классификации и кластеризации позволяющей использовать разные подходы в зависимости от задач.

Область знаний, подвергнутая исследованию, на сегодняшний день является очень актуальной благодаря росту вычислительной мощности современных компьютеров и большим возможностям для хранения данных, а также желанию корпораций и организаций проводить анализ своей деятельности, выбирать правильную стратегию и, таким образом, увеличивать прибыль.

Основные научные результаты диссертации

1. Предложен подход к кластеризации новостного потока. В рамках подхода используется облачная платформа IBM Bluemix, которая предоставляет сервисы для развертывания приложения. Для предварительной нормализации текста интернет новостей используются когнитивные сервисы IBM Watson.
2. Проведена апробация предложенного подхода в рамках реализации приложения для поиска по новостному потоку. В качестве дополнительной задачи был реализован сентимент-анализ текста интернет новостей.

Рекомендации по практическому применению результатов

1. Полученные результаты формируют практическую и теоретическую базу для разработки программного обеспечения с использованием тематического моделирования и облачных вычислений. Они могут быть использованы для разработки новых и улучшения существующих систем.
2. Полученные результаты могут служить наглядным примером и теоретической базой для студентов и магистрантов исследующие данные области знаний.
3. Доработка написанных позволит визуализировать полученные модели для изучения особенностей обрабатываемых данных.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

1. Чугаинов К. В., Пилецкий И. И. Методы тематической кластеризации новостных статей / К. В. Чугаинов, И. И. Пилецкий. // Научно-практические исследования №2 (ISSN 2541-9528) – Омск: Дельта, – 2017.

2. Чугаинов К. В., Шумовая и фоновая составляющая в тематическом моделировании / К. В. Чугаинов, И. И. Пилецкий. // Научно-практические исследования №3 (ISSN 2541-9528) – Омск: Дельта, – 2017.