Федеральное государственное образовательное бюджетное учреждение
высшего профессионального образования
«Финансовый университет при Правительстве Российской Федерации»

Департамент анализа данных, принятия решения и финансовых технологий

Курсовая работа

по дисциплине **"Технологии анализа данных и машинное обучение"** на тему:

"Классификация текстов методами машинного обучения"

Вид исследуемых данных: Корпус новостей с сайта Lenta.Ru

Выполнила:
студентка группы ПМ17-1
Баданина Н. Д.
Научный руководитель:
доктор техн. наук, профессор
Судаков Владимир Анатольевич

Москва 2020

Содержание

Введение

На сегодняшний день технологии развиваются с экспоненциальной скоростью. В научном мире появилась целая новая область знаний, которая требует изучения - это Искусственный Интеллект (ИИ). В ИИ, как подраздел можно включить машинное обучение и его алгоритмы. Одним из примеров алгоритмов машинного обучения являются нейронные сети. Пик развития машинного обучения начался ориентировочно с 2015 года, когда началась активная цифровизация, внедрение современных цифровых технологий, бизнесов, уход в онлайн. Это подтолкнуло компании к вложению средств в изучение области ИИ.

В данной работе речь пойдет об алгоритмах обработки текста на естественном языке. Голосовые помощники, чат-боты, умные устройства для дома позволяют компаниям привлечь дополнительную прибыль. Технологии, основанные на распознавании естественных языков создают новый интерфейс для взаимодействия с пользователем. Таким образом, создается эффект геймификации, что увеличивает возврат клиентов (retention) и уменьшает отток (churn).

На Российском рынке умные колонки с голосовыми помощниками стоят в малом количестве домохозяйств. Этот рынок развит в США, но имеет большой потенциал и в странах СНГ. На данный момент чат-боты используются не столько для увеличения продаж, сколько для уменьшения операционный затрат. К примеру, в банковском мобильном приложении можно задать вопрос в чат и ответит не оператор, а бот. При этом, компания экономит на затратах на операторов.

Примером внедрения анализа естественного языка може служить поддержка "тегов рекомендаций", реализованная, к примеру, Netflix, YouTube. Тег - метоинформация о фрагменте контента важная для поиска и рекомендации. Теги определяют свойства описываемого ими контента и могут использоваться для группировки схожих элементов и предложения описательных названий для таких групп.

В речевом анализе аудиоданные преобразуются в текст, к которому можно применить алгоритмы NLP.

0.1 Сложности при обработке текстов на естественном языке

Основной сложностью при обработке текстов на естественном языке средствами языков программирования является понимание алгоритмом контекста, в рамках которого идет обработка отдельного слова. Зачастую в тексте используются слова в переносном значении или в значении, которое установили собеседники между собой по договоренности. При существовании множества смыслов язык должен быть изботочен. Избыточность является серьбезной проблемой при построении алгоритмов NLP так как разработчики не могут и не будут указывать буквальный смысл каждого ассоциативного слова. Единицей анализа текста является лексема - строка кодированных байтов, представляющая собой текст. Лексема "батарея" изменила свой смысл с течением времени. Так, в текстах 19 века и позже можно увидеть это слово для обозначения артиллерийского подразделения из нескольких орудий. В современных публикациях лексема используется для обозначения хранилища, преобразующего химическую энергию в электронную.

0.2 Перспективы развития технологии

Развитие технологии классификации текстов началось с введения спам-фильтров для почты. Приложения, основанные на использовании естественного языка только начинают распространяться, но в будущем могут взять на себя задачи, которые сейчас решаются стандартными формами и интерфейсами.

1 Теоретическая часть

1.1 Задача классификации текстов

Целью машинного обучения является подгонка существующих данных под некоторую модель, создание представления реального мира, помогающего принимать решения или генерировать прогнозы на основе новых данных, путем поиска закономерностей в них. На практике для этого выбирается семейство моделей, определяющих связи между целевыми и входными данными, задается форма, включающая параметры и особенности, а затем с помощью некоторой оптимизации минимизируется ошибка модели на обучающих данных. Затем обученной модели можно передавать новые данные, на сонове которых она будет строить прогноз и возвращать метки, вероятности, признаки принадлежности или значения. Задача состоит в том, чтобы найти баланс между свособностью с высокой точность. находить закономерности в известных данных и способностью обобщени для анализа данных, которые модель не видела прежде. Многие приложения для анализа естественного языка включают не одну, а целое множество моделей машиноого обучения, взаимодействующий между собой и влияющих друг на друга. Модели могут повторно обучаться на новых данных, нацеливаться на новые пространства решений.

1.2 Модели классификации

2 Практическая часть

2.1 Работа с данными

2.2 Инструменты разработки

Заключение

Список использованных источников

[1] Applied Text Analysis with Python by Benjamin Bengfort, Rebecca Bilbro, and Tony Ojeda (O'Reilly). 9781491963043

Приложение

Исходный код можно найти по ссылке.