Государственное бюджетное общеобразовательное учреждение

средняя общеобразовательная школа №138

имени Святого благоверного князя Александра Невского

Калининского района Санкт-Петербурга

Курсовая работа

«Алгоритмы машинного обучения в решении проблем учебных заведений»

Выполнил:

Ученик 9а класса

Фиолетов Эдуард

Руководитель:

О.О.О

Санкт-Петербург

2020

Содержание

1. Вступление

2. Основная часть

2.1 Теоретическая часть

2.1.1 Что такое наука о данных?

2.1.2 Как работает Аналитик данных

2.1.3 Составляющие Data Science

2.2 Практическая часть

3. Заключение

4. Список литературы

4.1 Ссылки

4.2 Литература

1. Вступление

*Отправной точкой в этом проекте стала 1 неделя карантина, когда мне попалась реклама "Яндекс Практикум". Они обещали профессиональную переподготовку за несколько месяцев, на должность "Data Scientist", из которых первые две недели бесплатно. К концу бесплатного периода мне так сильно понравилась эта область, что ко мне стало приходить понимание – это будет кормить не только меня, но и окружающих. Все лето я изучал дисциплины этой отрасли, и понял, что это задача на всю жизнь.*

Спорим не угадали автора этого вступления? Да всё правильно, его писал не я, а одна из самых продвинутых нейросетей по обработке естественного языка (NLP) – «GPT-3Large» под разработкой международной компании «OpenAI», с дообучением нашего «Сбера».

Да, к сожалению, мне не выдали разрешение на использования API от компании основателя, на авторы выложили исходный код в общий доступ, чем и пользуются различные компании и отдельные разработчики, модернизируя изначальную модель.

Вся моя работа заключалась в том, чтобы указать заголовок, внутренние теги, опорные точки текста (имена, даты, места), размер нужного материала, и дополнительные настройки (использование ненормативной лексики, длинна чисел с плавающей точкой etc.)

Даже если вы раскусили меня, и нашли, чем машина себя выдала, есть ещё множество областей, где человек уже бессилен. Например, разработчики подразделения компании «NVIDIA» - «NVlabs», создали генеративно-состязательную нейросеть, которая создаёт изображения людей, которых никогда не существовало. В большинстве случаев, вероятность угадывания работы нейросети среди реальных фотографий, равняется случайной. То-есть при отсутствии видимых артефактов, вы не сможете сказать наверняка, где реальный человек. Также существуют нейросети, генерирующие текста и музыку для музыкантов, картины в стиле современного искусства, самих деятелей культуры и ещё много чего ещё.

К чему я это всё? Да к тому, что по-факту, машина прошла «Тест Тьюринга»!

А что это значит? А значит это то, что будущее уже стучится в дверь, и просит, чтобы ты уже начал изучать науку о данных - Data Science

А конкретно в этой работе я буду рассматривать места применения машинного обучения в школах, вузах и других образовательных учерждениях.

2. Основная часть

2.1 Теоритическая часть

2.1.1 Что такое наука о данных?

“[Наука о данных](http://bit.ly/quaesita_datasci) — это дисциплина, позволяющая сделать данные полезными” - это мое любимое определение науки о данных.

Ещё есть цитата википедии:

“Раздел информатики, изучающий проблемы анализа, обработки и представления данных в цифровой форме. Объединяет методы по обработке данных в условиях больших объёмов и высокого уровня параллелизма, статистические методы, методы интеллектуального анализа данных и приложения искусственного интеллекта для работы с данными, а также методы проектирования и разработки баз данных.”

Также, чтобы было понятней чем же занимается специалист по анализу данных, можно привести несколько примеров задач, которые ставят перед аналитиком:

1. Обнаружение аномалий, например, ненормальное поведение клиента, мошенничества

2. Персонализированный маркетинг — электронные рассылки, ретаргетинг, системы рекомендаций

3. Количественные прогнозы — показатели эффективности, качество рекламных кампаний и других мероприятий

3. Скоринговые системы — обработка больших объёмов данных, помощь в принятии решений, например, о предоставлении кредита

2.1.2 Как работает Аналитик данных

Уже из определения вы могли понять, что тут очень много ***различной*** работы. Поэтому следует прояснить основные чести работы data scintist’а:

1. Формулировка вороса

Оценка масштабов, времени, сложности

2. Предобработка данных

Очистка сырых данных, анализ достоверности данных, research

3. Анализ чистых данных

Самая интересная часть работы. Разрабатываем модели.

4. Визуализация

Представление итоговой информации в красивых графиках

5. Реакция

Принятие риешения на основе итогов анализа

В реальном мире очень мало специалистов, которые могут эффективно работать во всех областях, поэтому для каждой дисциплины существуют более узкие специальности:

* Проблема - Аналитик BI
* Задача - Data Analyst
* Сбор данных - Data Engineer
* Аналитика - Data Science
* Разработка - ML Engineer

2.1.3 Составляющие Data Science

В данном пункте стоит разграничить состовляющие обучения, и состовляющие самой науки:

Основные термины:

1. Искусственный интеллект (Artificial Intelligence) — область, посвящённая созданию интеллектуальных систем, работающих и действующих как люди.

2. Машинное обучение (Machine learning) — создание инструмента для извлечения знаний из данных. Модели ML обучаются на данных самостоятельно или поэтапно

# 3. Глубокое обучение (Deep learning) — создание многослойных нейронных сетей в областях, где требуется более продвинутый или быстрый анализ и традиционное машинное обучение не справляется.

4. Аналитика данных (Data Mining) — переводится как добыча данных, извлечение информации, раскопка данных, интеллектуальный анализ данных, средства поиска закономерностей, извлечение знаний, анализ шаблонов, "извлечение зерен знаний из гор данных", раскопка знаний базах данных, информационная проходка данных, "промывание" данных.

Остановки на пути обучения:

1. Математика

В Data Science из царицы наук берётся 4 раздела: Статистика, Линейная алгебра, Теория вероятностей, Мат. Анализ. Эти фундаментальные математические знания действительно важны, чтобы анализировать результаты применения алгоритмов обработки данных. Сильные инженеры в машинном обучении без такого образования есть, но это скорее исключение.

2. Программирование

Де-факто в этой сфере абсолютную победу одержал язык программирования “Python”. Почему? Ответ прост: Потому-что “питон” позволяет забыть о скучных частях программирования, и предлагает среду, где мы можем быстро набросать пришедшие в голову и воплотить их в жизнь.

3. Data Mining

Важный исследовательский процесс. Он включает анализ скрытых моделей данных в соответствии с различными вариантами перевода в полезную информацию, которая собирается и формируется в хранилищах данных для облегчения принятия деловых решений, призванных сократить расходы и увеличить доход

4. Список литературы

Ссылки

https://openai.com/ - *сайт создателей нейросети “GPT-3Large”*

https://thispersondoesnotexist.com/ - *проект нейросети от “NVlabs” по созданию изображений людей, которых никогда не существовало*

https://ru.wikipedia.org/wiki/Наука\_о\_данных

https://netology.ru/blog/01-2020-gid-po-data-science

Литература

Работа с данными в любой сфере: Как выйти на новый уровень, исполь-

зуя аналитику / Кирилл Еременко ; Пер. с англ. — М. : Альпина Паблишер,

2019 — 303 с. - ISBN 978-5-9614-2582-6

Python и машинное обучение: машинное и глубокое обучение с ис­ пользованием Python, scikit-learn и TensorFlow 2, / Sebastian Raschka Vahid Mirjalili / 3-е изд.: Пер. с англ. - СПб. : ООО "Диалектика", 2020. - 848 с.: ISBN 978-5-907203-57-0