Кто такие специалисты «Data Scientist»?
Data Scientist — это специалист по работе с данными для решения задач бизнеса. Он работает на стыке программирования, машинного обучения и математики.

Что входит в обязанности специалиста Data Scientist»?
В основные обязанности дата-сайентиста входит сбор и анализ данных, построение моделей, их обучение и тестирование.

Датасаентист обрабатывает массивы данных, находит в них новые связи и закономерности, используя алгоритмы машинного обучения, и строит модели. Модель — это алгоритм, который можно использовать для решения бизнес-задач.
В банках модели помогают точнее принимать решения о выдаче кредита, в страховых компаниях — оценивают вероятность наступления страхового случая, в онлайн-коммерции — увеличивают конверсию маркетинговых предложений.
Анализ данных — это часть работы датасаентиста.

Задачи специалиста по Data Science
Задачи различаются от компании к компании. В крупных корпорациях датасаентист работает с несколькими направлениями. Например, для банка он может решать задачу кредитной оценки и заниматься процессами распознавания речи.
Этапы работы над задачей у датасаентистов из разных сфер похожи:
— выяснение требований заказчика;
— решение принципиального вопроса «Целесообразно ли решать задачу методами машинного обучения?»;
— подготовка данных, их разметка;
— принятие метрик оценки эффективности модели;
— разработка и тренировка модели машинного обучения;
— защита экономического эффекта от внедрения модели;
— внедрение модели в производственные процессы и продукты;
— сопровождение модели.
Каждая новая итерация позволяет лучше понять проблемы бизнес—а, уточнить решение. Поэтому каждый этап повторяется снова и снова для развития модели и обновления данных.

Выделить отличительные аспекты их деятельности, общее мнение о профессии

Отличия специалиста Data Science от других профессий

. Анализ данных — это одна из функций сайентиста, чей основной результат работы заключается в создании моделей и кода, основанных на анализе данных.

Основная задача Data Science специалист — это извлечение полезной информации для бизнеса из больших объемов данных, выявление закономерностей, создание и проверка гипотез путем моделирования и разработки нового программного обеспечения.

Такие специалисты используют ряд инструментов для достижения своей цели: -пакеты статистического моделирования, -технологии больших данных.

Data Science охватывает области знаний, такие как математика (математический анализ, матстатистика и матлогика), информатика (разработка программного обеспечения, баз данных, моделей и алгоритмов машинного обучения, Data Mining) и системный анализ (методы анализа предметной области, Business Intelligence). Data Science является одной из самых востребованных и высокооплачиваемых ИТ-профессий в настоящее время.

Преимущества концепции Data Science

Они позволяют:

1. Прогнозировать текущий доход и эффективность бизнеса, а также понимать, в каком направлении движется компания, благодаря анализу больших объемов данных.
2. Моделировать новые тактики и стратегии.
3. Автоматизировать любые процессы, уменьшить затраты и повысить эффективность бизнеса, используя методы Data Science.
4. Предоставлять клиентам решения, разработанные на базе искусственного интеллекта, что способствует повышению качества продуктов и услуг.

Что такое библиотеки?

Библиотеки - это наборы функций и инструментов, написанные другими людьми ранее, которые помогают в написании кода расширяют возможности языка программирования Python.

Также существует платформа Hugging Face, которая содержит в себе коллекцию готовых современных предварительно обученных Deep Learning моделей (это разновидность машинного обучения, которая позволяет компьютерам решать более сложные задачи. Модели глубокого обучения также могут самостоятельно создавать новые функции). Например модель NLP.

Существует множество различных библиотек. Вот некоторые из них:

1. Библиотеки от Hugging Face:
    - Transformers

Это библиотека, которая предоставляет инструменты и интерфейсы для простой загрузки и использования моделей из Hadding Face. Это позволяет вам экономить время и ресурсы, необходимые для обучения моделей с нуля. Модели Transformers поддерживают три фреймворках: PyTorch, TensorFlow и JAX. Transformers не является набором модулей, из которых составляется нейронная сеть, как например PyTorch. Вместо этого Transformers

предоставляет несколько высокоуровневых абстракций, которые позволяют работать с моделями в несколько строк кода.

- Evaluate

  Библиотека для простой оценки моделей машинного обучения и наборов данных. С помощью одной строки кода вы получаете доступ к десяткам методов оценки для различных областей. Является продуктом Hadding Face.

- Accelerate

  Библиотека, которая позволяет запускать один и тот же код PyTorch в любой распределенной конфигурации, добавляя всего четыре строки кода. Является продуктом Hadding Face.

2. Stanza

   Содержит инструменты, которые можно использовать для преобразования строки и анализа текста, содержащей текст на любом языке в списки предложений и слов с целью создания базовых форм этих слов, для анализа зависимости синтаксической конструкции и распознавания именованных сущностей

3. Natasha

   Решает основные задачи NLP для русского языка: токенизация, сегментация предложений, встраивание слов, разметка морфологии, извлечение фактов, нормализация фраз и разбор синтаксиса. Библиотека поддерживает Python 3.5+ и PyPy3, не требует GPU, зависит только от NumPy.

   Библиотеки от проекта Natasha:

   - Ipymarkup

     Примитивная библиотека от проекта Natasha, нужна для подсветки подстрок в тексте, визуализации NER. Инструкция по установке, пример использования в репозитории Ipymarkup. Библиотека похожа на displaCy и displaCy ENT, бесценна при отладке грамматик для Yargy-парсера.

   - Razdel

     Библиотека Razdel — часть проекта Natasha, делит русскоязычный текст на токены и предложения. Он решает задачи разбора морфологии и синтаксиса, которые имеют смысл только для отдельных слов внутри одного предложения.

4. Pandas

Предназначена для анализа уже структурированных данных, размещенных в таблице.

5. Seqeval

   Разработана специально для оценки качества классификации в задаче распознавания именованных сущностей. Оценивает точность классификации для каждой сущности по отдельности.

6. PyTorch

   Это фреймворк, который применяет метод обучения за счёт применения решений множества сходных задач для языка программирования Python с открытым исходным кодом. Используется для решения различных задач: компьютерное зрение, обработка естественного языка(язык на котором общаются люди). Разрабатывается преимущественно группой искусственного интеллекта Facebook.

7. Plotly

   Бесплатная графическая библиотека с открытым исходным кодом, которую вы можете использовать в коммерческих целях, в последние годы набирающая популярность в Data Science среде. Plotly работает offline и позволяет строить интерактивные визуализации, т. е. изучать какие-то данные «на лету» (не перестраивая график, изменяя масштаб, включая/выключая какие-то данные), и строить полноценный интерактивный отчёт. Ключевое преимущество перед аналогами — удобство построения сложных интерактивных визуализаций — полноценных мини-приложений, которые делают результат работы аналитика более доступным для конечного пользователя.

8. Spark NLP

   Библиотека обработки текста с открытым исходным кодом для расширенной обработки естественного языка для языки программирования Python, Java и Scala. Библиотека предлагает предварительно обученные модели нейронных сетей, конвейеры и встраивания, а также поддержку для обучения пользовательских моделей. Также для этой него есть вспомогательная библиотека Spark NLP Display, которая нужна для визуализации аннотаций, созданных с помощью Spark NLP.

9. Pymorphy2

   Морфологический анализатор для русского языка, написанный на языке Python и использующий словари из OpenCorpora. Он может приводить слово к нормальной форме (например, “люди -> человек”, или “гулял -> гулять”), ставить слово в нужную форму (например, ставить слово во множественное число, менять падеж слова и т.д.), возвращать грамматическую информацию о слове (число, род, падеж, часть речи и т.д.)

10. Mystem3

Разработан Ильёй Сегаловичем в компании «Яндекс». Программа работает на основе словаря и способна формировать морфологические гипотезы о незнакомых словах. То есть программа MyStem производит морфологический анализ текста на русском языке.}

Что такое NLP? Возможности NLP.

NLP (Natural Language Processing) или **обработка естественного языка** представляет собой область исследований в области искусственного интеллекта, занимающуюся разработкой методов и технологий для взаимодействия между компьютерами и естественным языком человека.

Основная цель NLP: создание моделей и алгоритмов, способных понимать, интерпретировать и генерировать человеческий язык с высоким уровнем точности.

Генерация текста: Модели NLP способны создавать текст на основе предоставленных данных. Это может быть использовано для генерации контента, написания статей и других задач.

Распознавание речи: Такая функция необходима для множества приложений, которые используют голосовые команды или взаимодействуют с людьми в чатах.

Ответы на вопросы: Модели могут формировать ответы на вопросы, анализируя предоставленный контекст и предоставляя информативные ответы.

Машинный перевод: NLP модели могут переводить текст с одного языка на другой, сохраняя смысл и структуру предложений.

Чат-боты: Создание чат-ботов, способных вести диалоги с пользователями и выполнять различные задачи.

Анализ тональности: Модели NLP могут определять эмоциональный окрас текста, такой как положительный, отрицательный или нейтральный.

Обработка команд: Модели могут интерпретировать текстовые команды и выполнять соответствующие задачи.

Обобщение текста: Способность выделения ключевых аспектов текста и создания кратких обобщений.

Обучение с подкреплением: Модели могут улучшать свои навыки на основе обратной связи и опыта.

Как работает NLP

Работа NLP включает несколько основных этапов:

1. **Токенизация**: зачастую первым шагом в NLP является разделение текста на токены или слова. Токены могут быть помещены в массив для дальнейшей обработки.

2. **Лемматизация и стемминг**: для облегчения анализа, слова могут и быть приведены к базовой форме (лемме) или усечены до основы (стеммер). Это позволяет объединить разные формы одного слова для достижения более точной обработки.

3. **Частеречная разметка**: процесс определения части речи каждого слова в тексте. Это важно для понимания грамматики и смысла предложения.

4. **Синтаксический анализ**: выявление связей между словами в предложении и построение дерева синтаксической структуры. Это позволяет понять зависимости между словами и их роли в предложении.

5. **Семантический анализ**: определение смысла текста и выявление его значения. Семантический анализ используется для выделения в тексте семантических единиц, т.н. сущностей. Сущности подразделяются на именованные и неименованные.

6. **Разрешение кореферентности**: идентификация ссылок на предметы и местоимения в тексте и соотнесение их с соответствующими сущностями или объектами.

7. **Анализ ошибок**: оценка и исправление ошибок, связанных с пониманием и обработкой текста. Это может включать в себя коррекцию

грамматических ошибок, определение настроения или тональности текста и т. д.

8. **Генерация текста**: создание текста с помощью компьютера на основе входных данных и предшествующего анализа. Это может включать в себя формирование ответов на фиксированные вопросы, создание текстовых резюме и т.д.

Где и как использует NLP

Маркетинг

Обработку естественного языка используют для анализа отзывов клиентов, чтобы понять, как улучшить продукт или услугу. С помощью анализа можно собрать информацию о том, что говорят пользователи в социальных сетях. А затем провести семантический анализ, чтобы определить, насколько положительно отзываются о компании клиенты и какие проблемы есть у клиентов.

Робототехника

Роботы, которые взаимодействуют с человеком, должны правильно воспринимать и выполнять его команды. Здесь не обойтись без NLP — речь необходимо сначала перевести в текстовый формат, а затем в понятные для машины инструкции.

Человекоподобный робот София использует NLP, чтобы воспринимать речь и эмоциональное состояние говорящего, а также генерировать собственные ответы. Но до универсального интеллекта ей далеко. Обычно журналисты заранее передают разработчику Софии список вопросов, которые собираются обсудить с роботом.

Чат-боты

На основе NLP работают многочисленные инструменты для генерации текстов, например GigaChat от «Сбера» и YandexGPT от «Яндекса». Они отвечают на вопросы пользователей, генерируют тексты на разные темы и в разных форматах, составляют отчёты и так далее.

Медицина

Технологии NLP используются для озвучки текста в программах и устройствах для людей с нарушениями речи.

Синтезировать речь умели и 15 лет назад, но тогда для этого комбинировали предзаписанные MP3-файлы и она звучала неестественно. С помощью NLP можно превращать текст в речь в реальном времени. А ещё для каждого

пользователя можно сгенерировать оригинальный и уникальный голос на основе его собственного.С помощью синтезатора речи общался с миром известный учёный Стивен Хокинг.

Финансовая аналитика

Финансовые институты в России могут использовать NLP для анализа и обобщения текстовых данных из новостей, финансовых отчетов и социальных медиа. Это может помочь предсказывать тенденции на рынке, оценивать риски и принимать более информированные решения в сфере инвестиций.

Обработка договоров

Юридические компании могут использовать технологии NLP для автоматизированной обработки и анализа текстов договоров. Это ускоряет процессы составления и проверки юридических документов, а также помогает выявлять ключевые аспекты и риски.

Основными IT-инструментами специалистов Data Science являются: SQL, PYTHON, R и Machine Learning

SQL — это язык программирования, предназначенный для работы с наборами фактов, отношениями между ними и извлечения данных из базы данных. В программах управления реляционными базами данных, таких как Microsoft Office Access, язык SQL используется для работы с данными. В отличие от многих языков программирования, SQL удобочитаем и понятен даже новичкам. Как и многие языки программирования, SQL является международным стандартом, признанным такими комитетами по стандартизации, как ISO и ANSI.

На языке SQL описываются наборы данных, помогающие получать ответы на вопросы. При использовании SQL необходимо применять правильный синтаксис. Синтаксис — это набор правил, позволяющих правильно сочетать элементы языка. Синтаксис SQL основан на синтаксисе английского языка и имеет много общих элементов с синтаксисом языка Visual Basic для приложений. Понимание принципов работы SQL помогает создавать более точные запросы и упрощает исправление запросов, которые возвращают неправильные результаты.

Python — это активно развивающийся скриптовый язык, который используют для решения большого объема самых разноплановых проблем и задач. Он поможет построить модели, оценить гипотезу и построить связи между данными. Python пригодится в создании компьютерных и мобильных приложений, его применяют в работе с большим объемом информации, при разработке web-сайтов и других разнообразных проектов, используют в машинном обучении.

Язык R — один из самых распространённых в научной среде. Им пользуются учёные, которым нужно проводить статистические исследования и строить модели. Вне научной среды язык программирования R очень востребован. Статистические исследования важны для двух коммерческих специальностей — аналитиков данных и специалистов по Data Science. Им регулярно нужно проводить математические расчёты на основе выборок данных, и язык R для этого подходит идеально. В сфере анализа данных и машинного обучения без R — никуда.

Машинное обучение (Machine Learning) — это направление искусственного интеллекта, сосредоточенное на создании систем, которые обучаются и развиваются на основе получаемых ими данных.

В результате изучения профессии Data Scientist мы познакомились не только с общей информацией о специальности, её отличительными свойствами, но и с инструментами, которыми пользуются специалисты из этой области на постоянной основе, такие как: SQL,Python,R, Machine learning , модель NLP и “ библиотеки “

Каждый член команды получил бесценный и позитивный опыт не только в сфере программирования, но и анализа информации, выделения главных аспектов из большого массива данных. Эти навыки пригодятся нам не только при сдаче экзаменов, но и во взрослой жизни. Мы также узнали, что эта профессия только развивается в нашей стране, что дает огромный простор в деятельности в этой среде

Работа над проектом оставила у нас лишь позитивные эмоции и опыт, который мы никогда не забудем и будем всегда благодарны за возможность исследования новой, развивающейся профессии