Получение и предобработка данных. Первичная работа с объектом DataFrame

Андрей Куртасов
Системный аналитик

Проверка связи

Отправьте «✚», если меня видно и слышно

Если у вас нет звука или изображения:

- перезагрузите страницу
- попробуйте зайти заново
- откройте трансляцию в другом браузере (используйте Google Chrome или Microsoft Edge)
- с осторожностью используйте VPN, при подключении через VPN видеопотоки могут тормозить

План семинара

1. Рекомендуемая литература и ресурсы
2. Разбор заданий из LMS
3. Понятие конвейера обработки данных
4. Упражнения на получение и предобработку данных

Рекомендуемая литература и ресурсы

МФТИ

Вопрос

Можете порекомендовать коллегам какую-либо литературу или ресурсы по тематике курса?

Пожалуйста, поделитесь своими идеями в чате или в документе (открыт для комментариев).

ИИ с точки зрения бизнеса и управления продуктами

1. Домингос, Педро. Верховный алгоритм: как машинное обучение изменит наш мир / П. Домингос; пер. с англ. В. Горохова. — М. : Манн, Иванов и Фербер, 2016. — 336 с.
2. Марр, Б. Искусственный интеллект на практике: 50 кейсов успешных компаний / Бернард Марр, Мэтт Уорд; пер. с англ. Е. Петровой. – М.: Манн, Иванов и Фербер, 2020. – 316 с.
3. Мин, Ц. Как Alibaba использует искусственный интеллект в бизнесе: Сетевое взаимодействие и анализ данных / Цзэн Мин; пер. с кит. — М.: Альпина Паблишер, 2022. — 360 с.
4. Амейзен Э. Создание приложений машинного обучения: от идеи к продукту. Пер. с англ.– СПб.: Питер, 2023. – 272 с.
5. Bratsis, I. The AI Product Manager's Handbook: Develop a product that takes advantage of machine learning to solve AI problems. – Packt Publishing Ltd, 2023 . – 250 p.

Язык Python

МФТИ

1. Курс «Программирование на Python»: stepik.org/course/67/
2. Лутц, Марк. Изучаем Python: авторитетный курс объектно-ориентированного программирования / Марк Лутц; перевод с английского Ю. Н. Артеменко. – 5-е изд. – Москва: Диалектика; Санкт-Петербург: Диалектика, 2020.
3. Reitz, K., Schlusser, T. The Hitchhiker's Guide to Python: Best Practices for Development. – O'Reilly, 2016. – URL: docs.python-guide.org
4. Игровой тренажер по Python: py.checkio.org

Машинное обучение

1. Элбон К. Машинное обучение с использованием Python. Сборник рецептов: Пер. с англ. – СПб.: БХВ-Петербург, 2020. – 384 с.
2. Рашка С. Python и машинное обучение: Пер. с англ. – М.: ДМК-Пресс, 2017. – 418 с.
3. Дайзенрот М., Фейзал А., Он Ч. Математика в машинном обучении. – СПб.: Питер, 2023. – 512 с.

МФТИ

Вопросы? Комментарии? Напишите в чат или поднимите руку

МФТИ

Разбор заданий из LMS

МФТИ

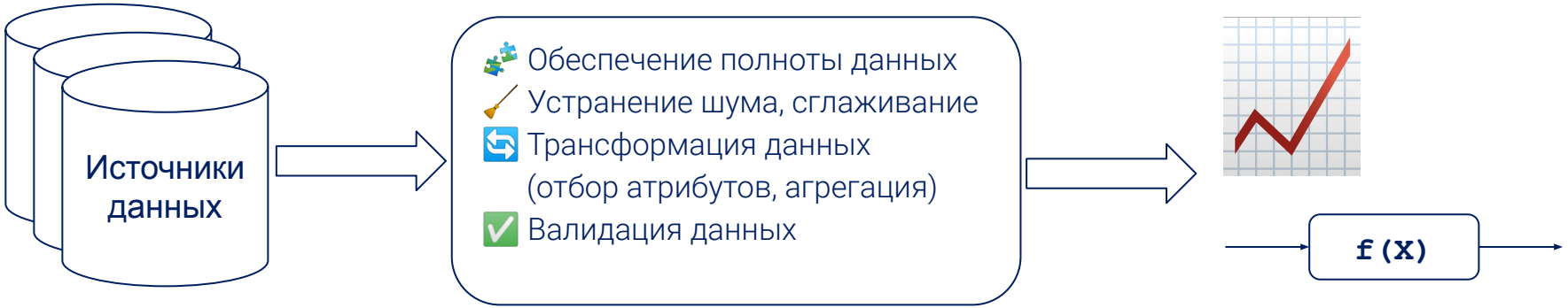
Проведем опрос

Как успехи с заданиями первой и второй недели?

МФТИ

Конвейер для обработки данных

Конвейер для обработки данных



CSV
TSV
Excel
SQL
JSON
XML
API

Получение:

```
csv
sqlite3
psycopg2
vertica_db_client
requests
```

Предобработка:

```
pandas
polars
numpy
scipy
```

Анализ и обучение:

```
pandas
polars
numpy
scipy
sklearn
pytorch
tensorflow
matplotlib
seaborn
```

Отчеты
Графики
Дашборды
Модели

pandas: основные возможности

МФТИ

1. Загрузка данных из разных источников (CSV, Excel, SQL).
2. Данные хранятся в объектах:
    a. Series – одномерная структура данных, хранящая данные одного типа. Напоминает словарь, ключи которого – индексы серии.
    b. DataFrame – двумерная структура данных, хранящая данные разных типов. Напоминает список словарей.
3. Срезы, сортировка, индексирование крупных наборов данных.
4. Объединение наборов данных (JOIN, UNION).
5. Построение сводных таблиц.
6. Обработка отсутствующих данных (показываются как NaN).

МФТИ

Вопросы? Комментарии? Напишите в чат или поднимите руку

Упражнения

Упражнение 1

1. Загрузить файл CSV с данными о координатах стран в программу на Python в виде DataFrame: github.com/google/dspl/blob/master/samples/google/canonical/countries.csv
2. Вывести список названий 10 стран, расположенных наиболее близко к экватору.

Упражнение 2

1. Загрузить в pandas датасет о нарушениях с портала открытых данных: data.mos.ru/opendata/7702051094-prinyatye-mery-administrativnogo-vozdeystviya-za-narusheniya-vyyavlennye-pri-osushchestvlenii-litsenzionnogo-kontrolya.
2. Изучить структуру датафрейма.
3. Задать для столбцов типы.
4. Удалить из датафрейма дубликаты (при наличии).
5. Вывести все возможные типы нарушений.
6. Вывести количество случаев по каждому типу нарушения.

МФТИ

Свободная дискуссия

Ваши вопросы? Пожелания?

До встречи!