



Библиотеки обработки данных для языка Python

ИУ-5

Краткий план лекции

- NumPy
- Pandas
- PandaSQL
- DASK

Рекомендуемая книга

- Библиотеки NumPy и Pandas хорошо описаны в книге Дж. Вандер Пласа.



Библиотека NumPy

- Библиотека [NumPy](#) предназначена для научных вычислений. Основным типом данных является многомерная матрица.
- Сайт библиотеки - <http://www.numpy.org/>
- Введение на русском языке - <https://habr.com/ru/post/352678/>
- Введение для начинающих (Kaggle) - <https://www.kaggle.com/abdullahsahin/numpy-tutorial-for-beginner>
- Введение - <https://docs.scipy.org/doc/numpy/user/quickstart.html>
- Документация - <https://docs.scipy.org/doc/numpy/reference/index.html#reference>
- Другие тьюториалы по NumPy:
 - Часть 1 - <https://www.machinelearningplus.com/python/numpy-tutorial-part1-array-python-examples/>
 - Часть 2 - <https://www.machinelearningplus.com/python/numpy-tutorial-python-part2/>
 - 101 упражнение по NumPy - <https://www.machinelearningplus.com/python/101-numpy-exercises-python/>
 - 100 упражнение по NumPy - <http://www.labri.fr/perso/nrougier/teaching/numpy.100/>
 - Сопряжение размерностей матриц (broadcasting) - <https://docs.scipy.org/doc/numpy/user/basics.broadcasting.html>

Библиотека Pandas

- [Pandas](#) - это библиотека, предназначенная для чтения, хранения, записи и обработки наборов данных (датасетов).
- Основная структура данных в Pandas это денормализованная таблица данных. Pandas одновременно обладает некоторыми характеристиками электронной таблицы (Excel) и реляционной СУБД.
- В чем Pandas похож на электронную таблицу:
 - В Pandas основной структурой данных является одна таблица (а не схема связанных таблиц как в реляционной СУБД).
 - Таблица должна содержать полный набор данных, предназначенных для дальнейшего построения моделей машинного обучения.
 - Таблица данных, как правило, является денормализованной.
 - В данных могут быть пропущенные значения.
- В чем Pandas похож на реляционную СУБД:
 - Таблица данных (как и реляционная таблица) состоит из типизированных столбцов (атрибутов).
 - Строка таблицы соответствует записи в реляционной БД.
 - Над таблицами возможно выполнение операций реляционной алгебры – соединение (join), группировка и другие операции.

Библиотека Pandas

- Официальная документация - <http://pandas.pydata.org/pandas-docs/stable/>
- Русскоязычное руководство - <https://khashtamov.com/ru/pandas-introduction/>
- Введение для начинающих (Kaggle) - <https://www.kaggle.com/abdullahsahin/step-by-step-pandas-tutorial-for-beginner>
- Интерактивное руководство - <https://www.datacamp.com/community/tutorials/pandas-tutorial-dataframe-python>
- 101 упражнение по Pandas - <https://www.machinelearningplus.com/python/101-pandas-exercises-python/>
- 12 наиболее полезных техник Pandas - <https://www.analyticsvidhya.com/blog/2016/01/12-pandas-techniques-python-data-manipulation/>

Библиотека *PandaSQL*

- Предназначена для выполнения SQL-запросов над наборами данных Pandas.
- Официальный сайт - <https://github.com/yhat/pandasql>
- Библиотека использует синтаксис SQLite - <https://www.sqlite.org/lang.html>
- Примеры использования:
 - <https://habr.com/ru/post/279213/>
 - https://github.com/miptgirl/udacity_engagement_analysis/blob/master/pandasql_example.ipynb

Библиотека *Dask*

- Библиотека Dask предназначена для распараллеливания вычислений при обработке данных.
- Преимущества библиотеки - <http://docs.dask.org/en/latest/why.html>
- Сравнения с библиотеками для обработки больших данных:
 - <http://docs.dask.org/en/latest/spark.html>
 - <https://matthewrocklin.com/blog/work/2018/08/28/dataframe-performance-high-level>