

Анализ данных и машинное обучение в гидрологии

Неделя 3. Практикум

План

1. `git pull`;
2. jupyter notebook;
3. ссылки на полезные ресурсы.

Обновление локального репозитория курса

1. Переходим в нашу локальную папку, в которой хранятся материалы занятий:

```
$ cd Documents/DA_and_ML_in_hydrology
```

2. Обновляем репозиторий:

```
$ git fetch
```

Ура, вы готовы к сегодняшнему практикуму!



Запуск среды разработки

```
$ jupyter notebook
```

Навигация

Week3 -> 03_pandas.ipynb

jupyter 03_pandas Last Checkpoint: 12 minutes ago (autosaved)

File Edit View Insert Cell Kernel Widgets Help [Python [conda root]]

Pandas

[Pandas](#) - стандарт в области анализа данных на Python.

pandas

$$y_{it} = \beta'x_{it} + \mu_i + \epsilon_{it}$$

Основные функции библиотеки pandas:

- быстрый и эффективный процесс чтения/записи файлов различных форматов (csv -> SQL);
- быстрый и эффективное манипулирование данными (добавление, репликация и т.д.);
- эффективная работа с рядами данных, с пропусками в данных;
- удобная работа со срезами, индексами, изменением формы, подвыборками;
- широкий функционал группировки и сводных таблиц;
- отличная работа с временными рядами;
- высокая производительность (основной код на C и Cython);
- бесплатное программное обеспечение.

Learning by doing

Давайте скачаем данные проекта [MOPEX](#) и посмотрим, что можно с ними сделать.

Репозиторий данных проекта MOPEX - ftp://hydrology.nws.noaa.gov/pub/gcjp/mopex/US_Data/:

- /Basin_Characteristics/usgs431.txt - список водосборов;
- /Us_438_Daily/ - директория с основными данными (осадки, испарение, температура);

In [1]:

```
# Загрузим файл со списком водосборов
# Прим.: именно этой командой мы можем загрузить весь интернет (вряд ли стоит это делать)
!wget ftp://hydrology.nws.noaa.gov/pub/gcjp/mopex/US_Data/Basin_Characteristics/usgs431.txt
```

In [2]:

```
# Проверим загрузку, имя файла
!ls
02156500.dly  03_pandas.ipynb  usgs431.txt
```

In [3]:

```
# импортируем pandas и numpy
import pandas as pd
import numpy as np
```

In [4]:

```
# Открываем файл, считываем только нужные нам столбцы (идентификатор, координаты, площадь)
f = open("usgs431.txt", 'r')

basin_id = []
basin_long = []
basin_lat = []
basin_area = []
for line in f:
    s = line.split()
    basin_id.append(s[0])
    basin_long.append(s[1])
    basin_lat.append(s[2])
    basin_area.append(s[3])

# Создаем рабочий DataFrame
basin_list = pd.DataFrame({'ID': basin_id,
                           'long': basin_long,
                           'lat': basin_lat,
                           'area': basin_area})
```

Чем занять себя две недели?

1. Всё для изучения Python: 181 бесплатный материал;
2. earthpy.org - python в науках о Земле;
3. Python Digest (Rus);
4. Убедиться в глобальном потеплении.