Apache Spark и PySpark

Инструкция для того, что начать использовать pyspark для учебных целей.

# Docker

Есть два популярных инструмента для анализа данных. Устанавливайте один из них, который вам удобнее.

Предварительно необходимо установить [docker desktop](https://www.docker.com), чтобы использовать готовые образы. По [ссылке](https://www.docker.com/get-started) нужно скачать и установить Docker Desktop (при установке могут потребоваться установка дополнительных компонент, следуйте инструкции при установке).

## Zeppelin

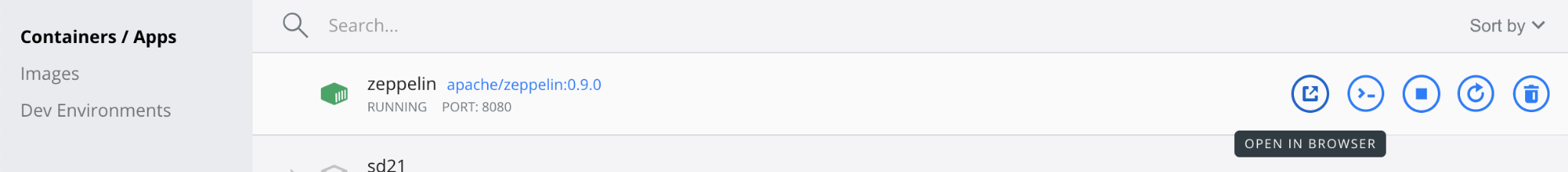
Установка:

* В командной строке (terminal, либо powershell) выполняем следующую операцию:

docker run -p 8080:8080 --rm -v $PWD/notebook:/notebook -e ZEPPELIN\_NOTEBOOK\_DIR='/notebook' --name zeppelin apache/zeppelin:0.9.0

, где /notebook - директория, которой будут лежать ваши ноутбуки;

* дождитесь скачивания и запуска контейнера;
* в docker появится новый запущенный контейнер;

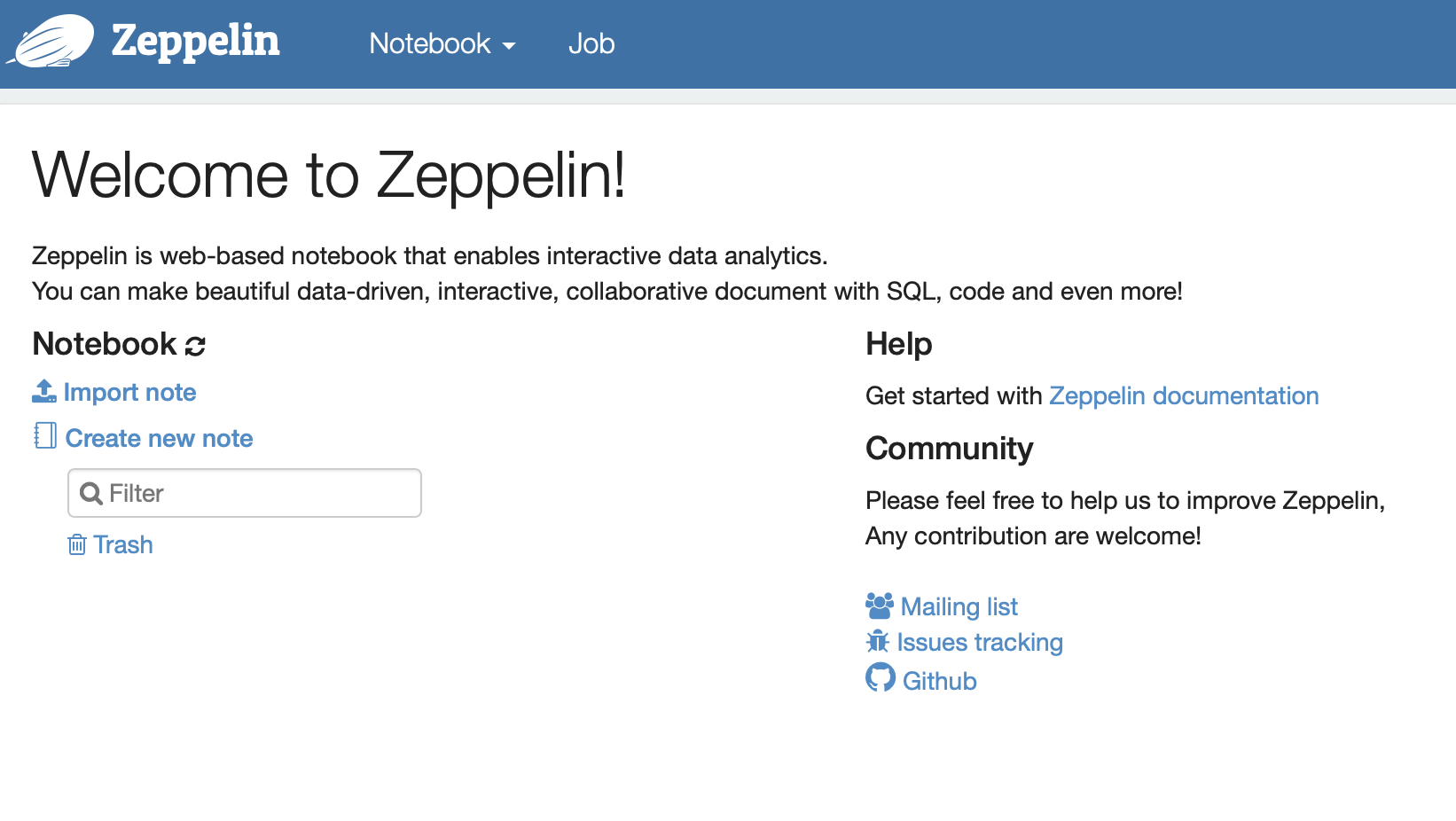


* нажимаем “Открыть в браузере”, либо сразу переходим по ссылке localhost:8080;
* откроется zeppelin.

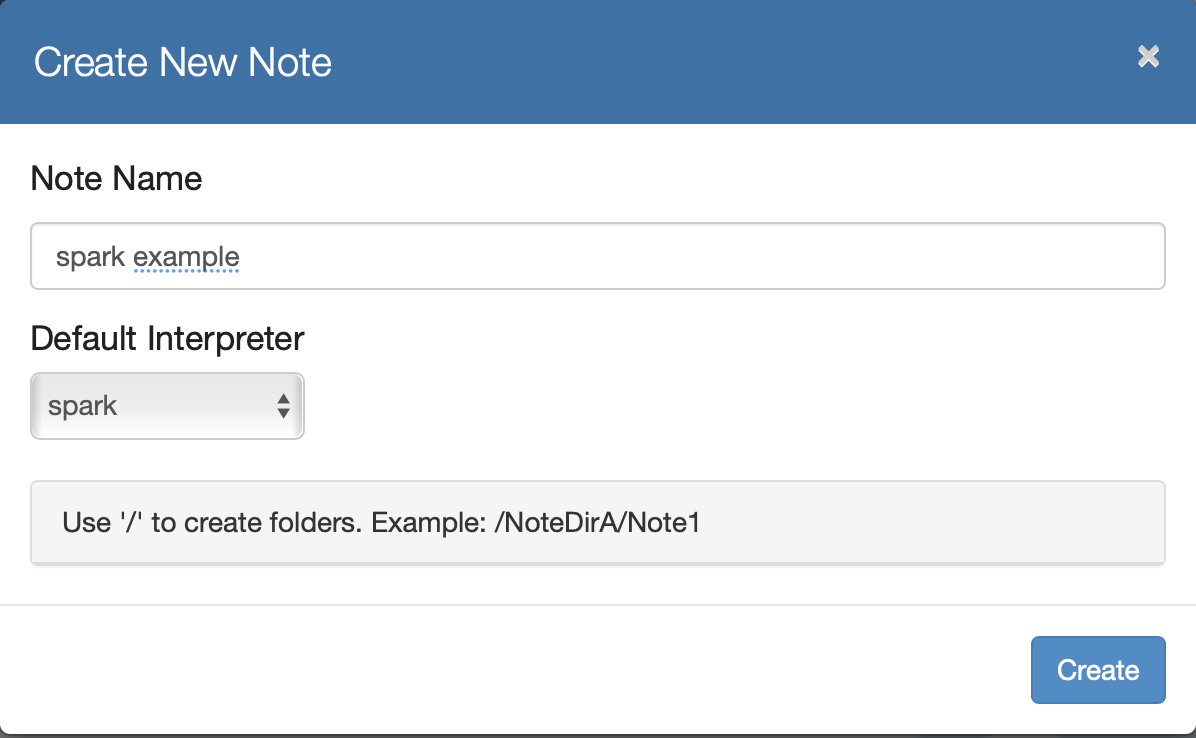
[Подробнее про установку](https://zeppelin.apache.org/download.html).

Использование:

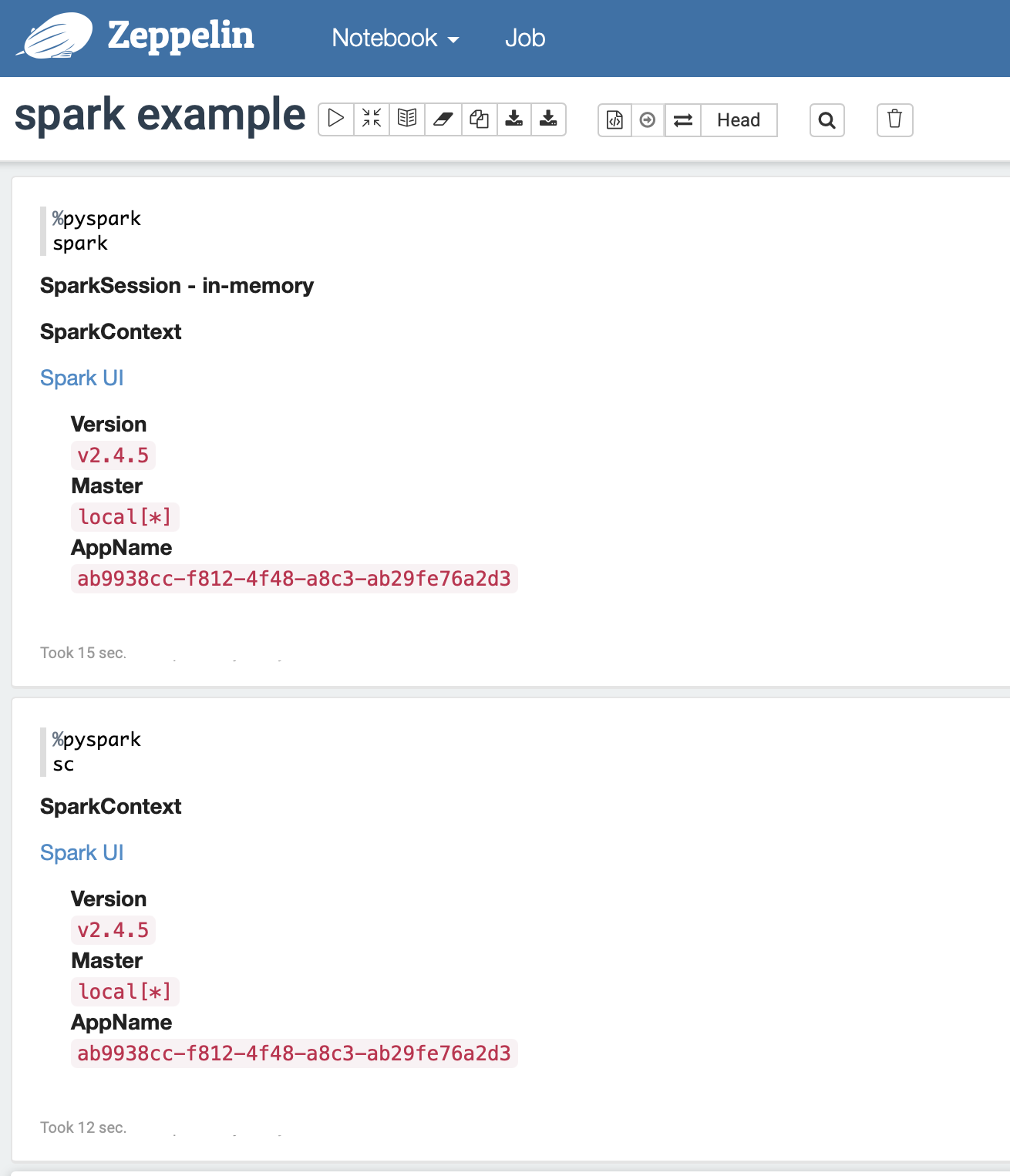
* По ссылке localhost:8080 после запуска откроется zeppelin, где можно создавать блокноты;



* zeppelin из коробки поддерживает spark, создаём блокнот (интерпретатор %spark);



* далее в ячейке прописываем интерпретатор %pyspark и можно писать код (блокнот уже содержит инициализированные SparkContext и SQLContext);



* подробнее [тут](http://zeppelin.apache.org/docs/0.9.0/index.html).

## Jupyter

Для jupyter так же есть готовый образ со spark.

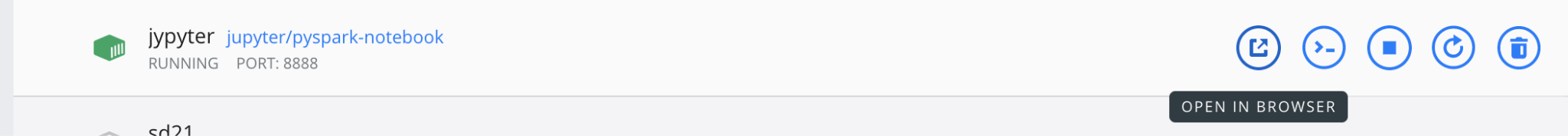
Установка:

* В командной строке (terminal, либо powershell) выполняем следующую операцию:

docker run -p 8888:8888 -v $PWD/notebook:/home/jovyan/work --name jypyter jupyter/pyspark-notebook start.sh jupyter notebook --NotebookApp.token=''

, где /notebook - директория, которой будут лежать ваши ноутбуки;

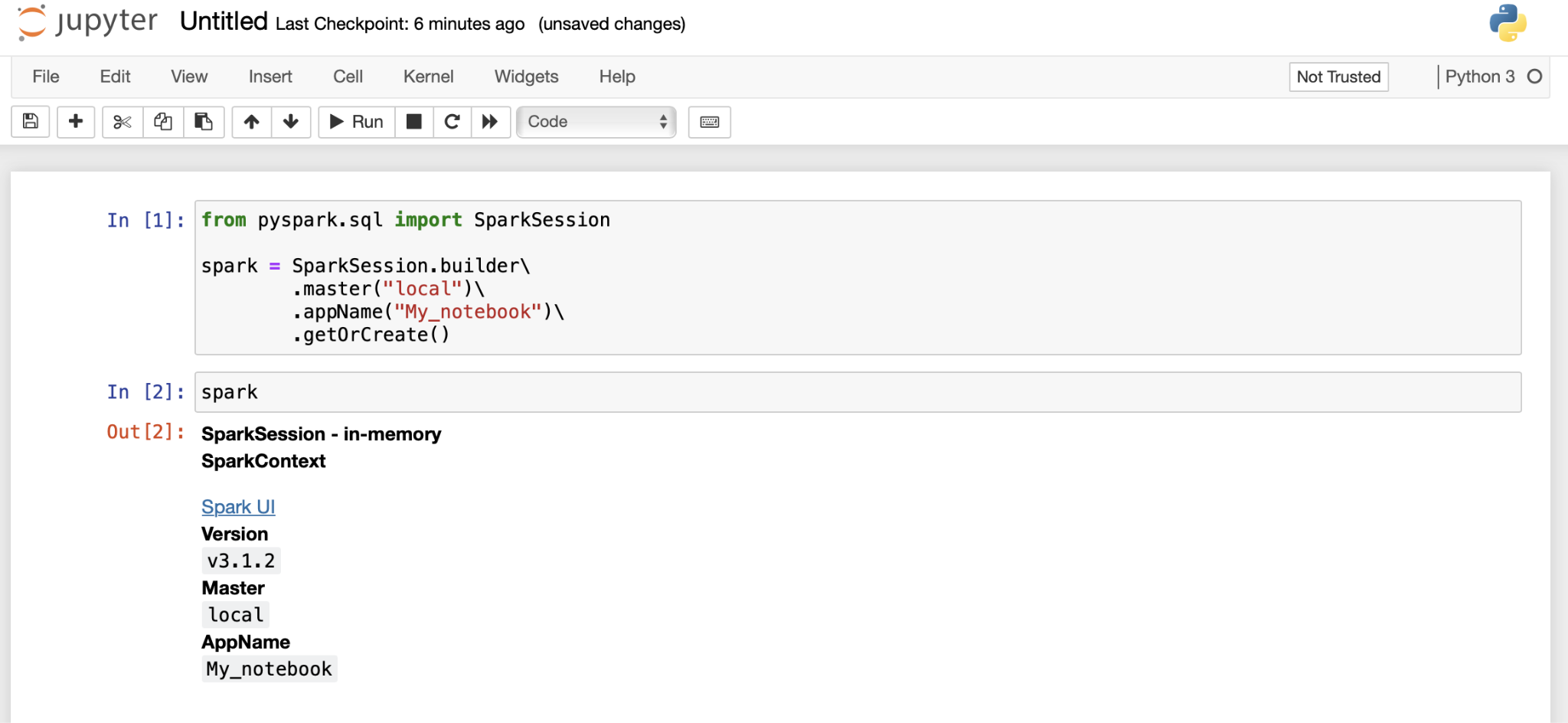
* дождитесь скачивания и запуска контейнера;
* в docker появится новый запущенный контейнер;



* нажимаем “Открыть в браузере”, либо сразу переходим по ссылке localhost:8888;
* откроется Jupyter.

Использование:

* Создаём новый блокнот;
* создаём SparkSession.



* готово, можно использовать pyspark. Подробнее про jupyter [тут](https://jupyter-notebook.readthedocs.io/en/stable/notebook.html).

# Colab - облачный jupyter от google

Пример блокнота:

<https://colab.research.google.com/drive/1cwztCPGUHtAwobs_dfZeLM6faqLgGIcA?usp=sharing>

# Локально без docker

1. Устанавливаем питон (например <https://docs.anaconda.com/anaconda/install/index.html> , куда уже включён jupyter)
2. Устанавливаем pyspark (pip install)
3. Возможно понадобится установить Java 8 (подробная инструкция [тут](https://spark.apache.org/docs/latest/api/python/getting_started/install.html))