Например нам нужно извлекать из проектов, размещенных на github. Например возьмем список из проектов на github, вообще это зависит от Вас, сколько проектов Вы хотите включить, будете извлекать. Какие данные мы можем извлекать из проектов

PyGithub/ PyGithub

Назв.организации/Название проекта.

Мы можем извлечь полное имя проекта, колличество pull requests на вытягивание из проекта, 61 pen 842 Closed. Можем получить либо открытые либо все запросы, а также учесть полеый запрос.

Например сейчас извлечем общее число запросов. Также аналогично промотры watches, star, fork

# Подключим библиотеку pandas для обработки файлов, манипулирования, анализа файлов, можем читать csv, json файлы, можем создавать свои собственные образы данных и сохранять их в формате json, csv. Мы используем pandas, чтоб сохранить извлеченные данные в .csv

import pandas as pd

project\_list = ['apache/any23', 'apache/groovy', 'YuriyGamiy/dj\_start']

# create function  
def extract\_project\_info():  
 # создадим пустой фрейм данных df\_project - это как пустой лист excel  
 df\_project = pd.DataFrame()  
 print(type(df\_project))  
  
 for project in project\_list:  
 # создадим объект класса github, которому передадим token  
 g = Github(access\_token)  
 # при появлении исключений, превышения лимитов, далее мы можем рассмотреть обработку исключений github  
 # извлекаем информацию о каждом проекте  
 repo = g.get\_repo(project)  
 # для суммарного получения запросов создадим переменную, состояние all  
 pl\_rq = repo.get\_pulls(state='all')  
  
 # сохраним наши данные во фрейме данных, добавляем словарь и вторым параметром флаг True в ignore\_index  
 # в словаре указываем, что мы хотим извлечь  
 df\_project = df\_project.append({  
 'Project\_ID': repo.id,  
 'Name': repo.name,  
 'Full\_name': repo.full\_name,  
 'Language': repo.language,  
 'Forks': repo.forks\_count,  
 'Stars': repo.stargazers\_count,  
 'Watchers': repo.subscribers\_count,  
 'Pul\_req.count': pl\_rq.totalCount,  
  
 }, ignore\_index=True)  
  
 # Данные будут извлечены и сохранены в этом файле .csv  
 # df\_project.to\_csv('../\_2\_Dataset/project\_dataset.csv', sep=',', encoding='utf-8', index=True)  
 df\_project.to\_csv('project\_dataset.json', sep=',', encoding='utf-8', index=True)  
 # index=True - каждая строка в файле будет иметь числовой индекс  
 # в этом примере мы получили представление о том как извлекать данные из конкретных проектов  
  
# вызов function  
extract\_project\_info()