

Home assignment 4

Maksimov Andrey

Task 1

С помощью библиотеки `requests` обращаемся по данному URL к странице и выкачиваем данные в формате `.zip`. Далее с помощью `os.getcwd()` получаем текущую директорию и распаковываем файлы. Затем через `data.iloc[:,1].head()` выводим ответ.

Task 2

Используем `data.head()` и возможность напрямую обращаться и проводить операции с колонками `DataFrame`.

Task3

Парсим `xml` через `BeautifulSoup`. Указываем класс `wikitable` и выделяем таблицы. Затем я делал `prettify` и смотрел на само дерево, что дало понимание необходимых тегов. После этого идет банальный парсинг. (Комментарии в самом коде добавят ясности)

Task4

Используем `value.counts()`

Task5

`respond_language = data['HaveWorkedLanguage'].dropna()` – получаем исходную колонку, затем идем по ней и добавляем в `list()` новые языки.

Task6

`language_series = pd.Series(0, index = languages)` – создадим `Series`, в который будем в последствии класть наши языки. Брать языки будем из колонки `data['HaveWorkedLanguage'].dropna()`

Task7

Создадим копии колонок `data['HaveWorkedLanguage']` и `data['Country']`

Затем через `for country, country_language in zip(data_result, respond_language)`: будем параллельно идти по странам и словам и складывать в `tmp_dict` для каждой страны языки, которые указали респонденты.

Затем применим `Counter` для каждой из стран и определим самые популярные языки