Национальный исследовательский университет
Высшая Школа Экономики
Факультет социальных наук
ОП Политология

"Искатель 3000": руководство пользователя

Выполнил:
Медведев Виктор

1 О программе

1.1 Введение

Искатель 3000 – это узкоспециализированный софт, предназначенный прежде всего для специалистов Social Science и политических учёных, применение которого возможно самым широком кругом пользователей. Функционально Искатель 3000 позволяет упростить работу по сбору основных качественных и количественных данных по федеральным, региональным должностным лицам и депутатам. При помощи Искателя 3000 пользователь может очень быстро найти интересующее его должностное лицо, получив в выдаче антикоррупционные (предвыборные) декларации за весь период его работы в государственных органах, а также сводку последних упоминаний в рейтингах и публикациях крупнейших политических телеграм каналов России. Программа обращается к двум источникам: declarator.org, крупнейшей и самой подробной базе должностных лиц и их деклараций, а также крупнейшему агрегатору российских политических национальных и региональных новостей - телеграм каналу «16 негритят», получая таким образом из первого данные о зарегистрированных доходах и расходах должностного лица, а из другого – наиболее важную и актуальную информацию, связанную с интересующим пользователя должностным лицом или депутатом. Искатель 3000 не имеет соответствующих аналогов, и поэтому де-факто занимает свободную рыночную нишу.

1.2 Общие особенности применения

Компилированный характер данных, получаемый в результате применения Искателя 3000, открывает широкие возможности для упрощения процесса кросс-валидации данных, актуального для специалистов и людей, увлеченных реальной политикой: оправданы ли суммы доходов, зарегистрированные в декларации, и если нет – то откуда они могли появиться в декларации? Это работает и наоборот: могут ли коррупционные скандалы, отраженные в СМИ, быть связаны с неоднозначными данными в антикоррупционных декларациях? При обращении к этим вопросам релевантно применение Искателя 3000, способного предложить удобное и комплексное представление данных для решения волнующих исследовательских проблем.

Если в результате работы с Искателем 3000 вы и ваша жизнь стали объектом интереса со стороны уполномоченных органов, разработчик проекта не несёт за это ответственности.

2 Подготовка к работе с программой

2.1 Установка Selenium

Для начала работы необходимо установить пакет Selenium по ссылке. Обратите внимание, что на Вашем устройстве должен быть установлен браузер Google Chrome версии не ниже 85. После загрузки распакуйте файл и скопируйте путь до него, чтобы не потерять - например:

C:/Users/user/Desktop/chromedriver.exe

2.2 Установка пакетов

После этого откройте в Jupyter Notebook ipynb-файл с нашей программой и запустите установку дополнительных пакетов из первой ячейки: для этого необходимо нажать на ячейку с установкой пакетов (Рисунок 1) и запустить ее через Run.

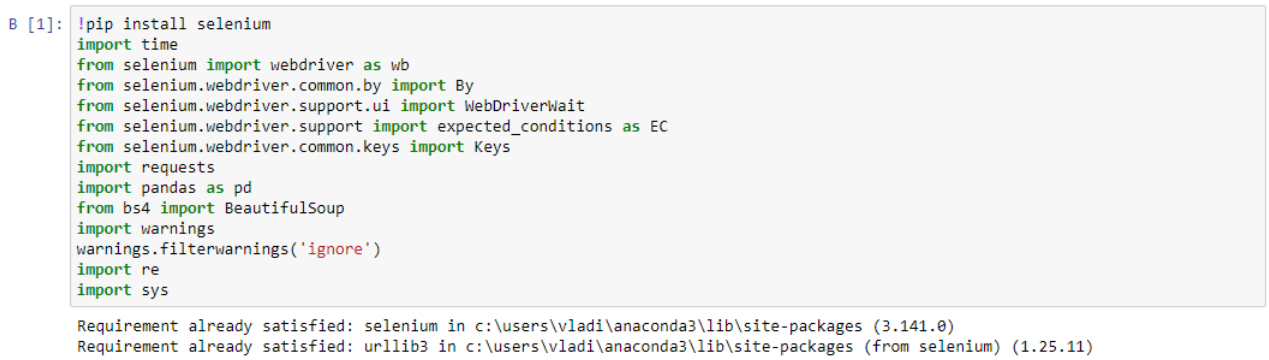


Рис. 1: Ячейка с дополнительными пакетами

После того, как все пакеты загрузились (это можно понять по изменению цвета кружка в правом верхнем углу с черного на белый), переходите к следующему шагу.

2.3 Настройка дирекций и выгрузки данных

Перед тем, как запускать ячейки с парсерами, необходимо предварительно совершить действия в двух частях программы:

1. В первой же строчке первого парсера изменить путь, по которому располагается драйвер Selenium, на тот, по которому Вы его распаковали (см. первый пункт). Строчка выделена на рисунке ниже.

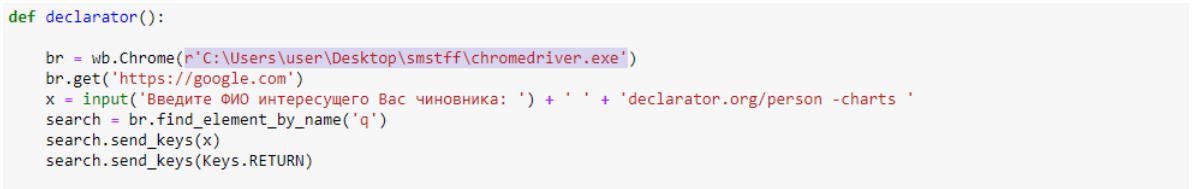


Рис. 2: Строка с путем драйвера Selenium-1

Необходимо проделать те же действия с первой строкой второй ячейки с парсером телеграма.

```
br = wb.Chrome(r'C:\Users\vladi\Documents\chromedriver_win32\chromedriver.exe')
time.sleep(1)

br.get('https://web.telegram.org/#/im?p=@Gubery')
time.sleep(3)

br.find_element_by_css_selector('body > div.page_wrap > div > div.login_page > 
time.sleep(2)
number = input('Введите номер телефона слитно и без +7: ')
br.find_element_by_css_selector('body > div.page_wrap > div > div.login_page > 
time.sleep(2)

br.find_element_by_css_selector('body > div.page_wrap > div > div.login_page > 
```

Рис. 3: Строка с путем драйвера Selenium-2

2. *(Если вы не собираетесь выгружать полученные данные в excel, то можете пропустить этот пункт.)* В одних из последних строчек парсеров программы необходимо изменить название будущего датафрейма - по умолчанию они будут загружаться в ту же папку, где находится файл программы, поэтому необходимо, чтобы название файла не дублировалось с уже существующими там файлами такого же формата. Строчки так же выделены на рисунках ниже.

```
df2.style.set_properties(**{'text-align': 'left'})
f = input('Хотите ли вы скачать данные в excel-формате? Ответьте да или нет: ')
if f == 'да':
    df2.to_excel("df_dec4.xlsx")
    return df2
```

Рис. 4: Строка с будущим названием файла excel-1

```
df['Пост'] = l_news
df['Дата поста'] = l_date

f_n = input('Хотите ли вы скачать данные в excel-формате? Ответьте да или нет: ')
if f_n == 'да':
    df.to_excel("df_news4.xlsx")
    df
else:
    df
```

Рис. 5: Строка с будущим названием файла excel-2

3 Выбор цели исследования

Для запуска программы повторите действия из пункта 2: нажмите на ячейку, после этого запустите её через Run. В отдельной ячейке Вам будет предложено ввести ФИО интересующего Вас чиновника. Нажмите на появившееся поле и введите ФИО (можете ввести только фамилию и имя, но в таком случае есть риск, что программа предоставит неверные данные). После ввода перепроверьте ФИО на наличие опечаток и нажмите Enter.

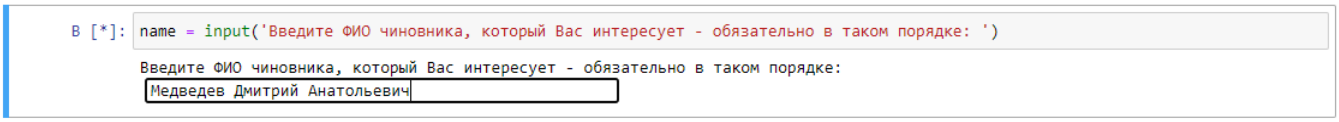


Рис. 6: Ввод данных о чиновнике

4 Парсинг сайта Декларатор

4.1 Работа программы

При запуске программы у Вас откроется новое окно браузера Chrome. **Рекомендуем развернуть его (перевести в полноэкранный режим), затем свернуть и забыть**.

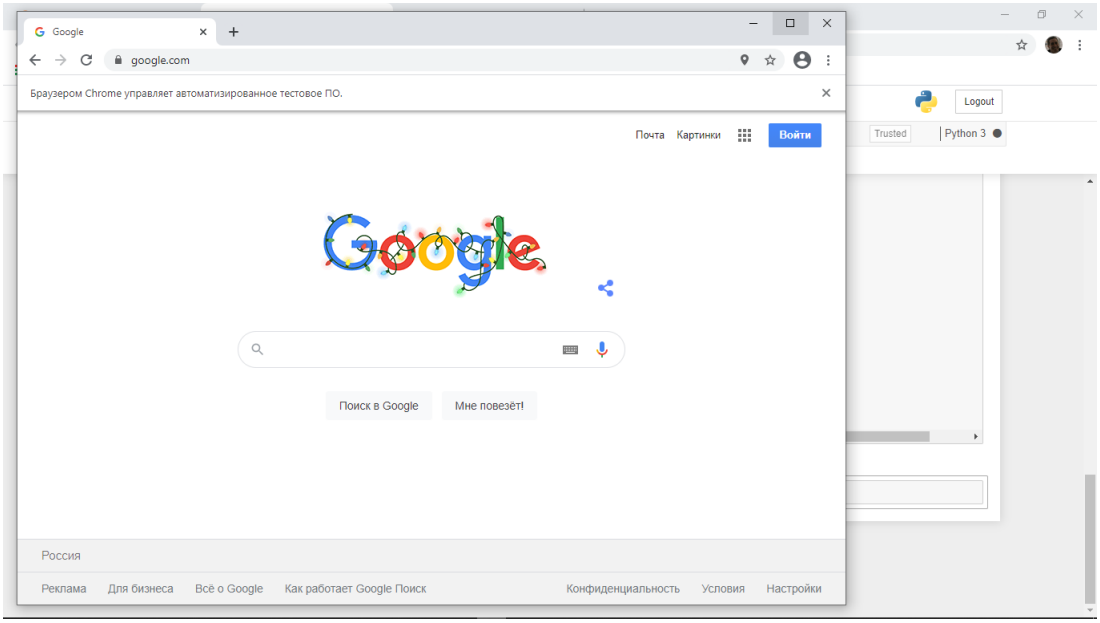


Рис. 7: Окно браузера, открытое программой

Примерно через 20 секунд Вам будет предложено скачать данные в формате excel. Если вам необходимо выгрузить данные в excel формате, введите "да"(строго маленькими буквами), если Вам достаточно данных в блокноте - введите "нет". После ввода нажмите Enter.

После этого Вы увидите дата-фрейм со значениями дохода по всем имеющимся декларациям, площадью недвижимости, а также количеством транспорта. Последний

```
    else:
        return df2


declarator()
```

```
Хотите ли вы скачать данные в excel-формате? Ответьте да или нет:
```

Рис. 8: Выгрузка данных в excel

столбец "Дополнительно"предоставляет более подробную информацию о доходах: доходы супруга или супруги, детей, марки машин и так далее. Чтобы подробнее с ней ознакомиться, откройте файл excel - он появится в той же папке, что и наша программа.

```
Хотите ли вы скачать данные в excel-формате? Ответьте да или нет: да
```

Out[23]:

	Декларация, статус, учреждение	Доход, руб.	Недвижимость, м2	Транспорт, шт.	Дополнительно
0	Антикоррупционная декларация 2019 Заместитель...	11051195	400	2	Доход 11 051 195,00 руб. Недвижимое имущество ...
1	Антикоррупционная декларация 2018 Председател...	9917511	400	3	Доход 9 917 510,76 руб. Недвижимое имущество З...
2	Антикоррупционная декларация 2017 Председател...	8565296	400	3	Доход 8 565 296,33 руб. Недвижимое имущество З...
3	Антикоррупционная декларация 2016 Председател...	8586975	400	3	Доход 8 586 974,69 руб. Недвижимое имущество З...
4	Антикоррупционная декларация 2015 Председател...	8767883	400	3	Доход 8 767 882,96 руб. Недвижимое имущество З...
5	Антикоррупционная декларация 2014 Председател...	8051574	400	3	
6	Антикоррупционная декларация 2013 Председател...	4259525	400	3	Доход 4 259 525,23 руб. Недвижимое имущество З...
7	Антикоррупционная декларация 2012 Председател...	5814351	400	3	Доход 5 814 351,09 руб. Недвижимое имущество З...
8	Антикоррупционная декларация 2011 Президент Р...	3371353	400	3	Доход 3 371 353,27 руб. Недвижимое имущество З...
9	Антикоррупционная декларация 2010 Президент Р...	3378674	216	2	Доход 3 378 673,63 руб. Недвижимое имущество З...
10	Предвыборная декларация 2010 Государственная ...	3378674	400	2	Доход 3 378 673,63 руб. (Администрация Президе...
11	Антикоррупционная декларация 2009 Президент Р...	3335281	216	2	Доход 3 335 281,39 руб. Недвижимое имущество З...
12	Антикоррупционная декларация 2008 Президент Р...	4139726	216	1	Доход 4 139 726,00 руб. Недвижимое имущество З...
13	Предвыборная декларация 2006 кандидат на пост...	7000748	400	1	Доход 7 000 748,00 руб. (зарплата, Аппарат Пра...

Рис. 9: Итоговый дата-фрейм

5 Парсинг телеграм канала 16 негритят

5.1 Работа программы

Для запуска парсера телеграма Вам понадобится профиль в мессенджере. Пожалуйста, зарегистрируйтесь, чтобы Вы смогли работать с дальнейшим кодом. Запускаем также, как и во втором пункте, ячейку с кодом. Снова откроется окно браузера Chrome. Рекомендуем развернуть его (перевести в полноэкранный режим).

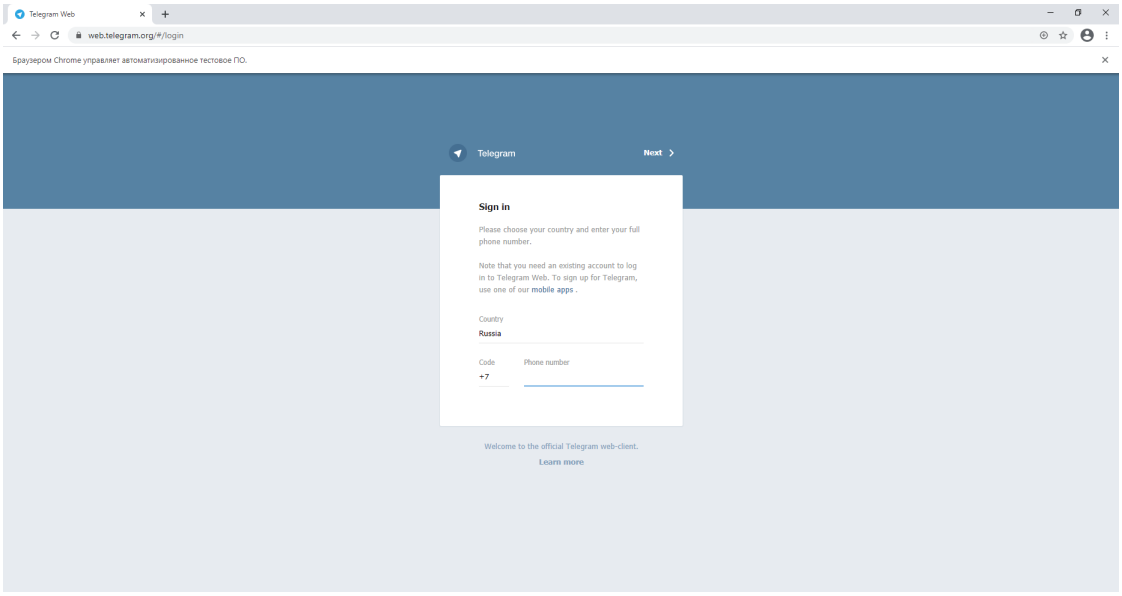


Рис. 10: Окно браузера, открытое программой

Использование кода требует предварительной авторизации в телеграме. Вам будет предложено ввести номер телефона, к которому привязан профиль мессенджера. Ввод должен обязательно осуществляться слитно без первой цифры номера, зарегистрированного в Российской Федерации.

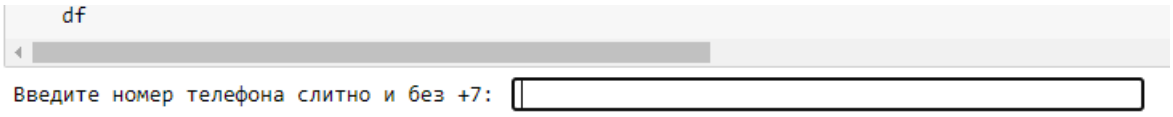


Рис. 11: Окно ввода номера телефона

Вам будет отправлен Telegram пятизначный код подтверждения входа в аккаунт мессенджера. После ввода нажмите на Enter. На введение кода у Вас есть целый час, поэтому, если ничего не приходит в течение полуминуты, просто подождите.

Через некоторое время (в зависимости от мощности компьютера и популярности человека) Вам будет предложено скачать данные в формате xlsx. Если Вам необходимо выгрузить данные в excel, то также просто введите "да"(строго строчными буквами). Если Вам достаточно датафрейма, то введите "нет". После ввода нажмите на Enter.

При любой выбранной опции (скачивании xlsx или нет) Вы увидете дата-фрейм с новостями и датой, когда они были опубликованы.

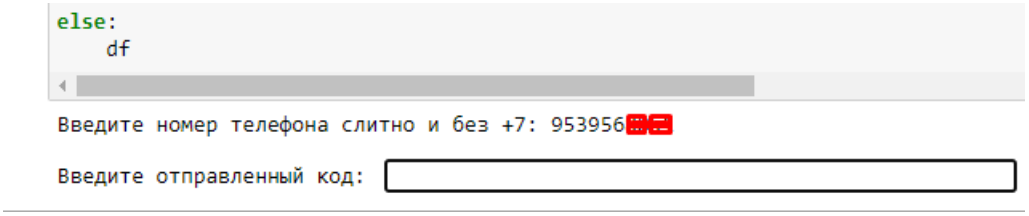


Рис. 12: Окно ввода отправленного кода

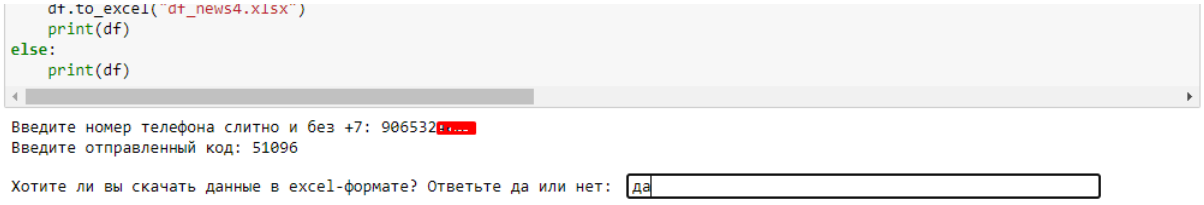
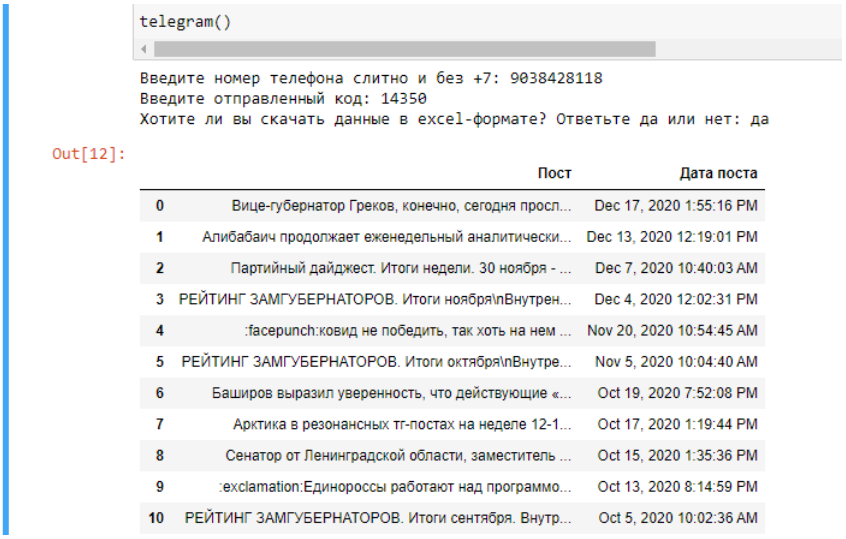


Рис. 13: Выгрузка данных в excel

После этого Вы увидите дата-фрейм со значениями дохода по всем имеющимся декларациям, площадью недвижимости, а также количеством транспорта. Последний столбец "Дополнительно"предоставляет более подробную информацию о доходах: доходы супруга или супруги, детей, марки машин и так далее. Чтобы подробнее с ней ознакомиться, откройте файл excel - он появится в той же папке, что и наша программа.



```
telegram()
```

```
Введите номер телефона слитно и без +7: 9038428118
Введите отправленный код: 14350
Хотите ли вы скачать данные в excel-формате? Ответьте да или нет: да
```

Out[12]:

	Пост	Дата поста
0	Вице-губернатор Греков, конечно, сегодня просл...	Dec 17, 2020 1:55:16 PM
1	Алибабаич продолжает еженедельный аналитически...	Dec 13, 2020 12:19:01 PM
2	Партийный дайджест. Итоги недели. 30 ноября - ...	Dec 7, 2020 10:40:03 AM
3	РЕЙТИНГ ЗАМГУБЕРНАТОРОВ. Итоги ноября\nВнутрен...	Dec 4, 2020 12:02:31 PM
4	:facepunch:ковид не победить, так хоть на нем ...	Nov 20, 2020 10:54:45 AM
5	РЕЙТИНГ ЗАМГУБЕРНАТОРОВ. Итоги октября\nВнутре...	Nov 5, 2020 10:04:40 AM
6	Баширов выразил уверенность, что действующие «...	Oct 19, 2020 7:52:08 PM
7	Арктика в резонансных тг-постах на неделе 12-1...	Oct 17, 2020 1:19:44 PM
8	Сенатор от Ленинградской области, заместитель ...	Oct 15, 2020 1:35:36 PM
9	:exclamation:Единороссы работают над программо...	Oct 13, 2020 8:14:59 PM
10	РЕЙТИНГ ЗАМГУБЕРНАТОРОВ. Итоги сентября. Внутр...	Oct 5, 2020 10:02:36 AM

Рис. 14: Итоговый дата-фрейм

6 FAQ

6.1 Первый парсер не работает, что не так?

Мы включили в код некоторые индикаторы, которые помогают понять, где именно возникла ошибка. Вместо итогового дата-фрейма Вы можете увидеть перед собой два типа сообщений:

1. *'Пожалуйста, перезапустите программу.'*

   В данном случае программа не нашла ссылку на чиновника. Наиболее вероятная причина - изменение структуры кода страницы. Попробуйте перезапустить программу, переведя открывшееся окно в полноэкранный режим. Менее вероятно, но также возможна опечатка в ФИО чиновника: перепроверьте введенные данные и так же перезапустите программу.

2. *'К сожалению, мы не смогли найти деклараций о доходах данного чиновника. Попробуйте найти декларации другого чиновника.'*

   В данном случае структура кода точно не нарушена и точная причина - введенного Вами чиновника не существует в базе данных declarator.org. Перепроверьте введенные данные или введите ФИО другого чиновника.

6.2 Второй парсер не работает, что не так?

Обратите внимания на следующие пункты:

1. Проверьте верную последовательность и написания имени. Поиск по имени всегда должен начинаться с фамилии в правильном написании.
2. Программа повторяет действия пользователя и требует, поэтому таких же подгрузок и времени в зависимости от качества компьютерных характеристик и интернета. Если появляются ошибки типа NoSuchElementException, то попробуйте изменить время ожидания, увеличивая на 1 аргумент time.sleep.
3. Проверьте, что у Вас работа парсера происходит в полноэкранном режиме. В свёрнутом окошке html может структурироваться для мобильной версии, которая иначе работает. Из-за этого возможна некорректная работа кода. Если же не работает полноэкранная, то попытайтесь сделать обратные действия.
4. Популярные фамилии могут привести к попаданию однофамильцев в набор данных, поэтому требуется самостоятельный просмотр исследователем данных.
5. Данные могут выглядеть не очень приглядно из-за видео, фото, эмоджи, которые считываются в тексте, а также разделителей. Рекомендуем скачивать и смотреть xlsx файл, так как информация оказывается в более удобном виде.
6. Возможен пропуск некоторых значений. Чтобы избавиться от этого, попытайтесь ещё раз прогнать код с более большими значениями аргумента time.sleep().

6.3 Я хочу анализировать данные в таблице прямо в блокноте, как это сделать?

В данном случае после всех описанных шагов вам необходимо создать новую строку в блокноте и вписать туда код, показанный на рисунке ниже.

Рис. 15: Сохранение дата-фрейма в отдельную переменную

После этого просто повторите все шаги, которые были сделаны ранее. Отличие состоит в том, что теперь таблица будет сохранена в переменной df, с которой можно будет работать прямо в блокноте.