Тема «Работа с файлами»
Самостоятельная работа
Чтение данных из файла

Выполните указанные ниже задания в виде отдельного файла Jupyter notebook, сохраните его и прикрепите к данному заданию в системе (в имени файла укажите вашу фамилию и инициалы, а также номер модуля и название данной темы).

В ходе этой самостоятельной работы будем работать файлом data.txt (и его аналогом в Excel data.xlsx), в котором собрана аналитика по рекламной кампании.

Для удобства использования в различных задачах исходный файл разбит на небольшие производные файлы с ограниченным объемом данных.

Эти файлы находятся в архиве, который нужно распаковать в ту же самую папку, в которой находится Ваш Jupyter notebook.

Сначала рассмотрим случай файла, в котором содержится только один столбец файла data.txt с ID пользователей (столбец user_id). Файл user_ids.txt

Так выглядит столбец с ID пользователей:
1010
1036
1041
1041
1042
...

Выполните чтение этого файла. Выведите содержимое файла на экран. Какое значение будет выведено последним на экран?

Чтение содержимого файла в список

Выполните чтение файла 'user_ids.txt' в список под названием «user_ids». Выведите на экран первые 5 элементов списка user_ids. Какой ID будет в этом наборе последним?

Сколько всего элементов в этом листе? (длину списка можно определить с помощью функции len).

Какое количество уникальных ID в рассматриваемом файле?

Чтение файлов с заголовками

Какая строчка появится первой при чтении файла user_ids_headers.txt обычным способом?

Выполните чтение данного файла таким образом, чтобы заголовок не учитывался.

Чтение данных из файлов с несколькими столбцами

Выполните чтение файла 'data_3_columns.txt' таким образом, чтобы учитывалась структура столбцов.

Перевод значений в численный вид

При чтении файла 'data_3_columns.txt' появляется проблема - третий столбец должен быть числом (стоимость заказа). Но он выводится в кавычках, еще и через запятую (следствие копирования данных из Excel).

При этом разделителем в числе всегда является точка (в большинстве языков программирования).

Python позволяет привести к единому виду большие объемы данных. Чтобы исправить запятые на точки можем использовать метод replace. Первым аргументом ставим запятую (элемент, который ищем и заменяем в строке). Вторым - точку (элемент, на который заменяем).

Заодно заменим элементы столбца понятными названиями переменных (medium, source, amount_paid), которые отражают, какие данные у нас в каком столбце.

```
with open( 'data_3_columns.txt', 'r' ) as f:
    for line in f:
        line = line.strip().split('\t')
        medium = line[0]
        source = line[1]
        amount_paid = line[2].replace(',', '.')
        print( line )
        print( source, medium, amount_paid )
```

```
['seo', 'google', '20,20']
google seo 20.2
['sem', 'yandex', '15,60']
yandex sem 15.6
['email', 'promo', '13,20']
promo email 13.2
```

```
['sem', 'yandex', '9,80']
yandex sem 9.8
['sem', 'google', '14,80']
google sem 14.8
...
```

Таким образом, можно заменить запятые на точки, но переменная amount_paid все еще типа string:

```
type( amount_paid )
str
```

Это не даст нам производить с ней вычисления, как с числом. Например, невозможно подсчитать сумму заказов или найти самый дорогой.

Переведем стоимость заказа из строк в тип float с помощью одноименной функции float:

```
with open( 'data_3_columns.txt', 'r' ) as f:
    for line in f:
        line = line.strip().split('\t')

        medium = line[0]
        source = line[1]
        amount_paid = float( line[2].replace(',', '.') )

        print( source, medium, amount_paid )
```

Какое значение amount_paid будет в последней строчке файла? Не забывайте, что разделителем должна быть точка.

Подсчет суммы в файлах

Ранее третий столбец был преобразован в числовой вид. Посчитаем сумму этого столбца в переменную total_sum.

При работе используем все тот же файл data_3_columns.txt

Для понимания процесса на каждом шаге будем выводить на экран сумму total_sum накопленным итогом. Берем код из прошлого шага и добавляем необходимые строки:

```
total_sum = 0
with open( 'data_3_columns.txt', 'r' ) as f:
    for line in f:
```

```
        line = line.strip().split('\t')
        amount_paid = float( line[2].replace(',', '.') )
        total_sum += amount_paid
        print( 'Текущая сумма расходов: {:.2f}'.format( total_sum ) )
```

```
Текущая сумма расходов: 20.20
Текущая сумма расходов: 35.80
Текущая сумма расходов: 49.00
Текущая сумма расходов: 58.80
Текущая сумма расходов: 73.60
...
```

Какое значение суммы получаем в последнем примере?

Посчитаем сумму не для всех строк, а при следующем условии: нам необходимо брать только те строки, у которых источник source равен google. Для этого в цикл надо добавить всего одну строчку:

```
total_sum = 0
with open( 'data_3_columns.txt', 'r' ) as f:
    for line in f:
        line = line.strip().split('\t')

        medium = line[0]
        source = line[1]
        amount_paid = float( line[2].replace(',', '.') )

        if source == 'google':
            total_sum += amount_paid
            print( 'Текущая сумма расходов google: {:.2f}'.format( total_sum ) )
```

```
Текущая сумма расходов google: 20.20
Текущая сумма расходов google: 35.00
Текущая сумма расходов google: 49.40
Текущая сумма расходов google: 63.40
Текущая сумма расходов google: 86.00
...
Текущая сумма расходов google: 1318.80
```

Какова сумма amount_paid для строк, у которых source == 'yandex' и medium == 'seo'? Напоминание - одновременные условия в операторе if перечисляются с помощью and.

Библиотека openpyxl

Выполните чтение файла data.xlsx. Какое значение у ячейки E1?

Чтение содержимого Excel-файла в цикле

Давайте возьмем столбцы с E по I (файл data.xlsx) и посчитаем сумму заказов в этом файле (т.е. сумму значений в столбце I). Выведем сначала первые 5 строк:

```
for line in sheet[ 'E2:I5' ]:
    medium = line[0].value
    source = line[1].value
    amount_paid = line[4].value
    print( source, medium, amount_paid )
```

```
google seo 20.2
yandex sem 15.6
promo email 13.2
yandex sem 9.8
```

При чтении Excel-файла с помощью библиотеки openpyxl мы похоже получили столбец I сразу в нужном числовом типе.

Проверьте, какой тип у переменной amount_paid? Напоминание: для проверки типа переменной можно воспользоваться функцией type

Посчитаем сумму столбца I в нашем файле. Для этого пройдемся со второй по последнюю строки столбцов E-I. Сумму значений будем накапливать в переменной total_amount_paid.

```
total_amount_paid = 0
for line in sheet[ 'E2:I295' ]:
    amount_paid = line[4].value
```

Допишите алгоритм подсчета суммы в столбце I. Какое значение суммы total_amount_paid должно получиться?