**Аннотация**

*Урок посвящен технологии хранения однотипных записей в файлах, форматам с фиксированной и с произвольной длиной записи (DSV, TSV, CSV). Особое внимание уделено форматам с произвольной длиной записи, методам работы с ними при помощи строковых функций, а также специализированной библиотеки csv и PyQT.*

**Хранение записей в файлах**

Любой файл может содержать последовательность байт, которая интерпретируется человеком и прикладными программами различным образом. Например, байт с десятичным значением 65 может означать и число 65, и код латинской буквы A в зависимости от договоренности по формату. При разработке приложений очень быстро возникла необходимость хранения в файлах больших массивов однородных объектов (записей), имеющих некоторый набор полей или характеристик. По сути такой массив представляет собой таблицу со строками и столбцами. Каждый столбец имеет свой, как правило, примитивный тип: целое или вещественное число, строка и т. д.

Очень важно, что формат при этом может быть совершенно различным. Главное, чтобы все программы, которые с ним работают, были написаны в соответствии со спецификацией этого формата, то есть **понимали** его.

В процессе эволюции из всех форматов для хранения массивов записей прижились два: с фиксированной и произвольной длиной записи. А подлинного искусства обработка и хранение таблиц достигли в [реляционных базах данных](https://ru.wikipedia.org/wiki/%D0%A0%D0%B5%D0%BB%D1%8F%D1%86%D0%B8%D0%BE%D0%BD%D0%BD%D0%B0%D1%8F_%D0%B1%D0%B0%D0%B7%D0%B0_%D0%B4%D0%B0%D0%BD%D0%BD%D1%8B%D1%85), о которых мы поговорим уже очень скоро.

**Форматы с фиксированной длиной записи**

Такие форматы практически всегда имеют байтовую, а не текстовую структуру.

Размер каждого поля фиксируется таким образом, чтобы иметь минимальную, но достаточную длину в байтах для представления всего диапазона значений. Как правило, на этом этапе возникают проблемы с выбором длины строковых данных.

Для примера рассмотрим расшифровку записи для хранения информации о школьнике:

* Имя: строка(50)
* Фамилия: строка(50)
* Возраст: однобайтовое число
* Пол: строка(1)
* Адрес: строка(100)

Как видно из примера, в мире нашей программы существует ряд ограничений, например, нет адресов, которые содержат более 100 символов.

Преимущества бинарных форматов с фиксированной длиной записи аналогичны преимуществам по работе с массивами: мы за константное время можем взять любую по счету запись, просто вычислив, по какому смещению в файле она располагается относительно его начала.

Примером файлов такого рода является, например, [BMP-файл](https://upload.wikimedia.org/wikipedia/commons/c/c4/BMPfileFormat.png).

Его структура такова: в первых 54 байтах хранится так называемый **заголовок**. *Иногда его еще называют мета-данными, то есть данными о данных.* Он информирует о размере изображения, сжатии и других характеристиках всего файла. После заголовка подряд записывается информация о каждом пикселе изображения, на каждый пиксель отводится ровно 3 байта, по одному байту на каналы R, G, B. То есть информация об изображении представляет собой массив (список) кортежей из трех чисел — значений red, green и blue составляющих цвета пикселя.

Фиксированный формат имеет и недостатки. Один из самых больших в том, что, как только появляется запись, одно из полей которой **не влезает** в предопределенную длину, мы не сможем использовать данный формат или вынуждены будем его видоизменить.

Причем, расширяя размер поля для данной записи, мы расширяем ее и для всех остальных, расходуя понапрасну ресурсы памяти.

Одна из самых широко известных ситуаций, связанных с фиксированным форматом, — [проблема 2000 года](https://ru.wikipedia.org/wiki/%D0%9F%D1%80%D0%BE%D0%B1%D0%BB%D0%B5%D0%BC%D0%B0_2000_%D0%B3%D0%BE%D0%B4%D0%B0).

**Форматы с произвольной длиной записи**

Если длина полей в одной записи не фиксирована (например, мы говорим, что адрес клиента может быть любой длины), нам нужно решить две проблемы:

1. Как мы будем понимать, где **границы полей**.
2. Как мы будем понимать, где **границы записей**.

Решать их можно по-разному: например, перед каждым полем делать служебное поле из 4 байт, которое показывает размер в байтах следующего поля. Таким образом, в нашем случае поле не сможет содержать более 4 Гб данных.

Можно пойти другим путем и ввести специальные разделители. Поговорим про этот случай. Воспользуемся модельным примером и рассмотрим [прайс-лист](https://yastatic.net/s3/lyceum/content/resources/ikea.txt) товаров в магазине «Икея».

Вот небольшой фрагмент этого файла:

> keywords price product\_name

> МОРУМ, Ковёр, безворсовый 6999 МОРУМ

> МОРУМ, Ковёр, безворсовый 6999 МОРУМ

> ИДБИ, Придверный коврик 649 ИДБИ

> ХОДДЕ, Ковёр, безворсовый 1399 ХОДДЕ

> ОПЛЕВ, Придверный коврик 599 ОПЛЕВ

> ОПЛЕВ, Придверный коврик 599 ОПЛЕВ

Разделителем записей был выбран символ **новой строки**, а разделителем полей — **табуляция** (хотя на самом деле увидеть символы табуляции можно с трудом).

В результате мы получили текстовый файл, который легко читается и человеком, и компьютерной программой.

Назовем такой формат **TSV**.

**Формат TSV**

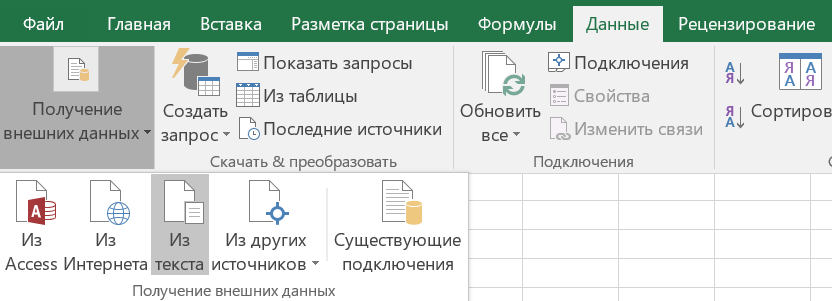
**TSV**

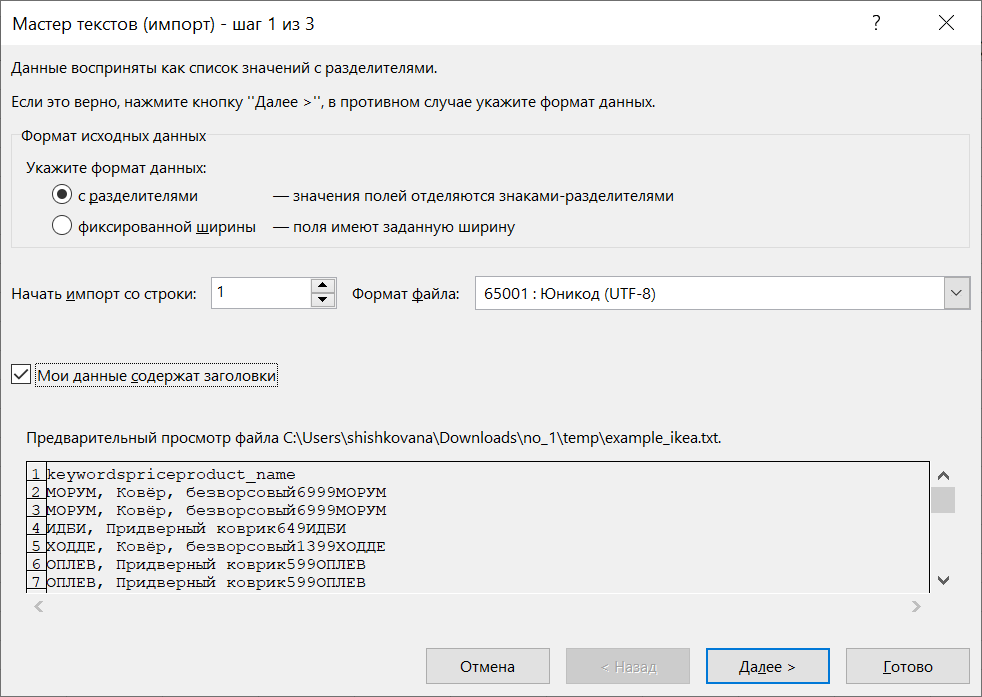
TSV (англ. tab separated values — «значения, разделенные табуляцией») — текстовый формат для представления таблиц баз данных. Каждая запись в таблице — строка текстового файла. Каждое поле записи отделяется от других символом табуляции, а точнее — горизонтальной табуляции.

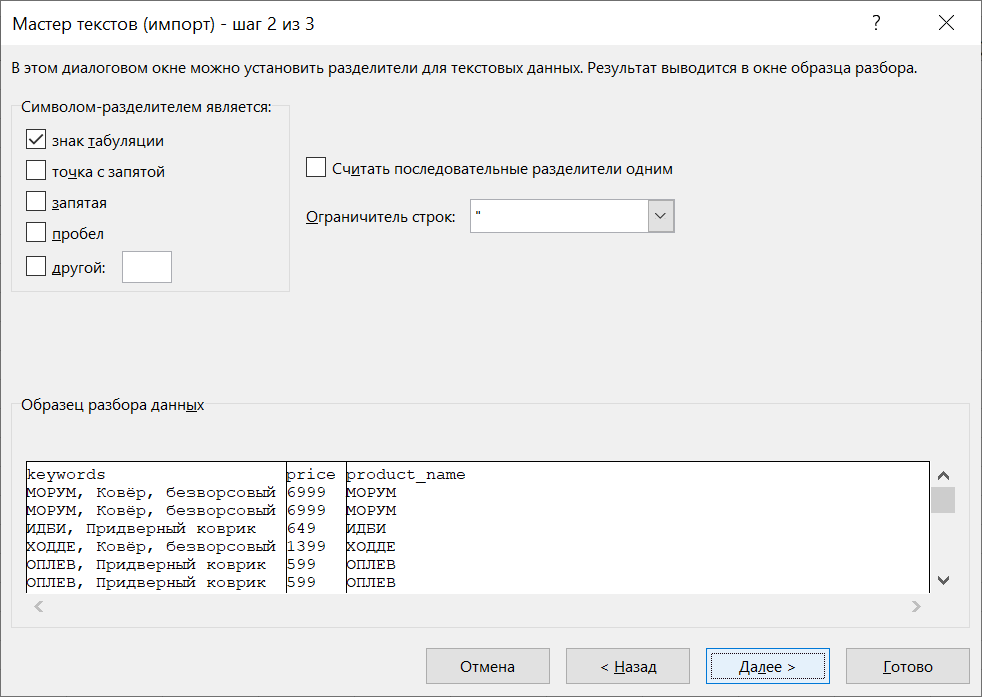
**TSV** — форма более общего формата **DSV** («значения, разграниченные разделителем», от англ. delimiter separated values).

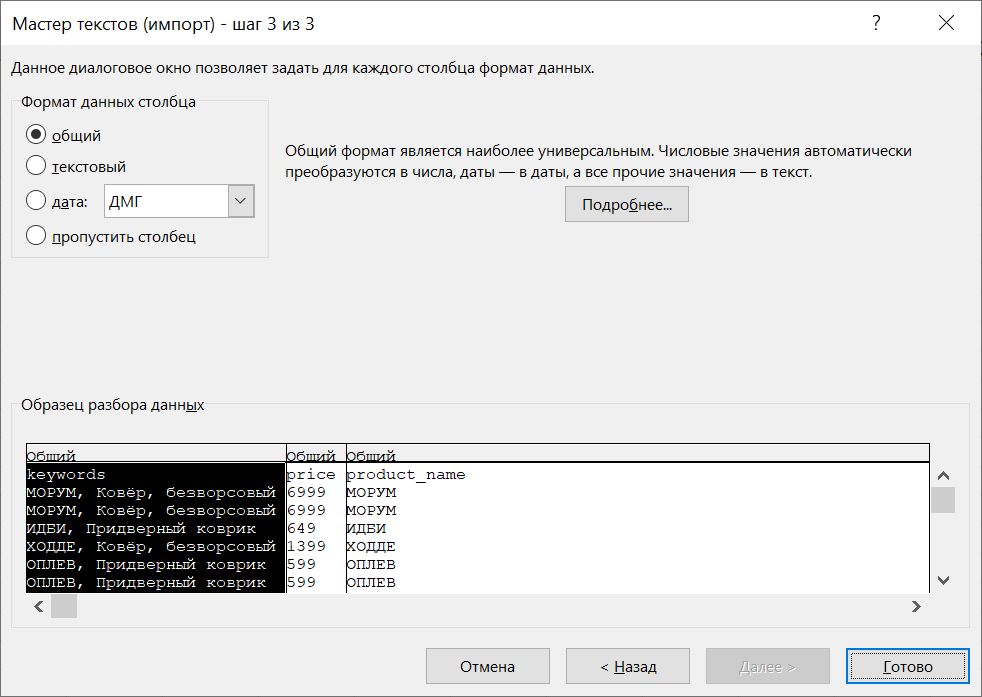
Программы для работы с электронными таблицами (MS Excel, LibreOffice Calc) поддерживают импорт из такого формата.

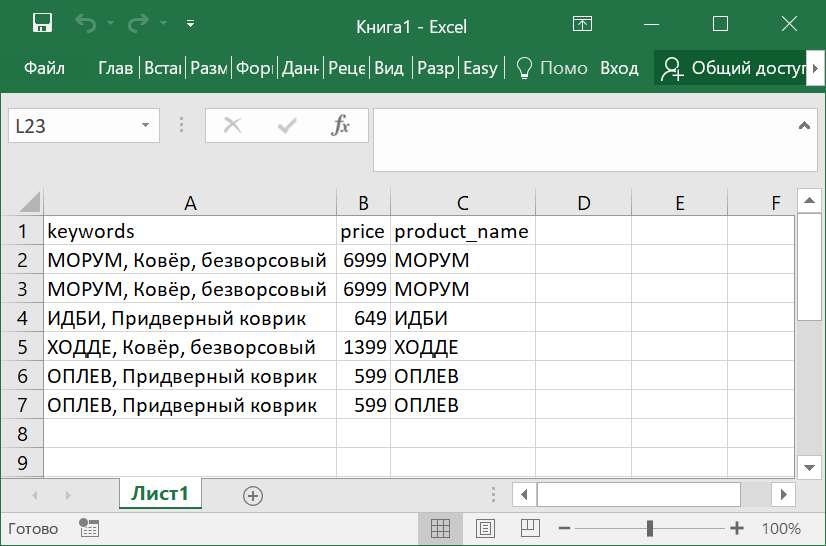
Импорт из текстового файла в Excel











Посмотрим, как работать с таким форматом в Python. Тут нет ничего сложного, поскольку у нас есть мощная функциональность по работе со строками, в частности, метод split.

data = open('files/ikea.txt', encoding='utf-8').read()

for row in data.split('\n')[:10]:

print(row.split('\t'))

['keywords', 'price', 'product\_name']

['МОРУМ, Ковёр, безворсовый', '6999', 'МОРУМ']

['МОРУМ, Ковёр, безворсовый', '6999', 'МОРУМ']

['ИДБИ, Придверный коврик', '649', 'ИДБИ']

['ХОДДЕ, Ковёр, безворсовый', '1399', 'ХОДДЕ']

['ОПЛЕВ, Придверный коврик', '599', 'ОПЛЕВ']

['ОПЛЕВ, Придверный коврик', '599', 'ОПЛЕВ']

['ЮНКЭН, Брикеты', '89', 'ЮНКЭН']

['БУНСЁ, Детское садовое кресло', '1199', 'БУНСЁ']

['ИКЕА ПС ВОГЭ, Садовое лёгкое кресло', '1999', 'ИКЕА ПС ВОГЭ']

Вспомним списочные выражения и сразу сделаем «двумерный массив», а потом обратимся к цене пятого по счету товара:

table = [r.split('\t') for r in data.split('\n')]

print(table[5][1])

599

Мы можем также отсортировать элементы по цене и напечатать 10 самых дешевых товаров:

table = table[1:]

table.sort(key=lambda x: int(x[1]))

for r in table[:10]:

print(r)

['СМОРИСКА, Стопка', '6', 'СМОРИСКА']

['СМОРИСКА, Стакан', '12', 'СМОРИСКА']

['ДИСТАНС, Контейнер', '12', 'ДИСТАНС']

['ДИСТАНС, Контейнер', '12', 'ДИСТАНС']

['ОППЕН, Миска', '25', 'ОППЕН']

['ДАРРОКА, Стакан', '25', 'ДАРРОКА']

['АНТАГЕН, Щетка для мытья посуды', '25', 'АНТАГЕН']

['ВАНКИВА, Рама', '25', 'ВАНКИВА']

['ХОППЛЁС, Доска разделочная', '27', 'ХОППЛЁС']

['ДАРРОКА, Стакан д/виски', '29', 'ДАРРОКА']

**Формат CSV**

Одним из самых распространенных форматов **DSV** стал **CSV** — формат с разделителем полей — запятой (англ. comma separated values). Наш файл будет выглядеть в нем [вот так](https://yastatic.net/s3/lyceum/content/resources/ikea.csv):

> keywords,price,product\_name

>"МОРУМ, Ковёр, безворсовый",6999,МОРУМ

>"МОРУМ, Ковёр, безворсовый",6999,МОРУМ

>"ИДБИ, Придверный коврик",649,ИДБИ

>"ХОДДЕ, Ковёр, безворсовый",1399,ХОДДЕ

>"ОПЛЕВ, Придверный коврик",599,ОПЛЕВ

>"ОПЛЕВ, Придверный коврик",599,ОПЛЕВ

>"ЮНКЭН, Брикеты",89,ЮНКЭН

>"БУНСЁ, Детское садовое кресло",1199,БУНСЁ

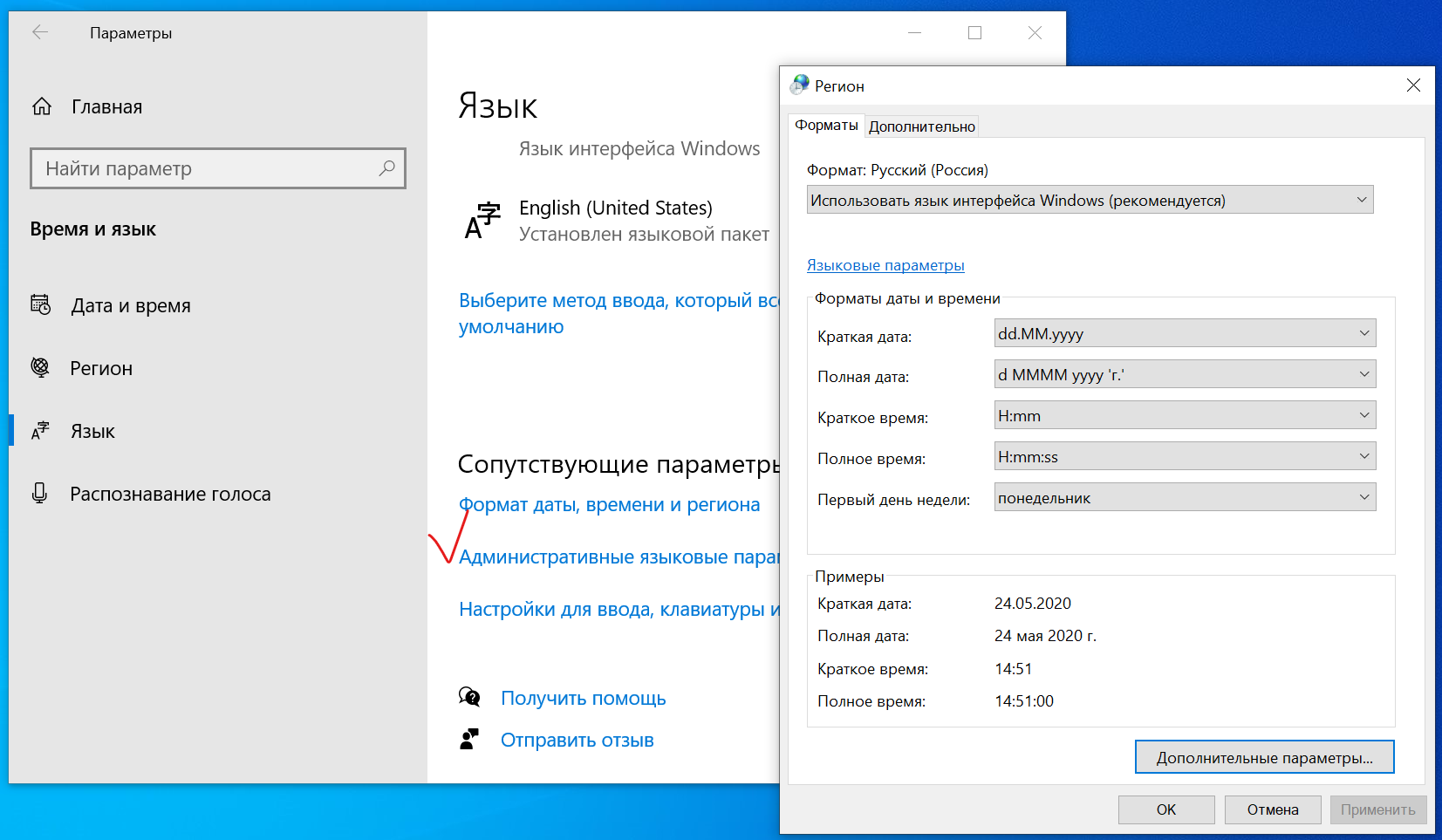
>"ИКЕА ПС ВОГЭ, Садовое лёгкое кресло",1999,ИКЕА ПС ВОГЭ

Для всех форматов DSV проблемой является символ-разделитель полей в данных. В этом случае вводят так называемый **разделитель текста**, в качестве которого выступают двойные кавычки, а если в поле встречается сам разделитель текста, то его удваивают. Например, **ООО "Светлана"** при записи в файл превращается в **"ООО ""Светлана"""**.

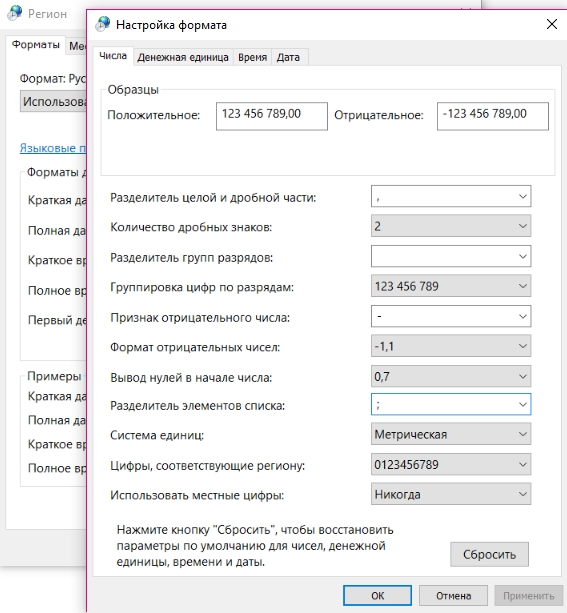
Вот почему в приведенном фрагменте CSV-файла первое поле в кавычках — внутри него есть запятые.

Нужно отметить, что в CSV могут быть другие разделители, например, точка с запятой. Очень часто это регулируется настройками операционной системы (параметр «Разделитель элементов списка» в ОС Windows):

Параметры — Время и язык — Язык:



Дополнительные параметры:



**Библиотека CSV**

Несмотря на то, что DSV-форматы просты, отсутствие четких стандартов в выборе разделителей и экранировании символов привели к тому, что с ними лучше работать при помощи специализированных библиотек, а не в стиле «использования функции» split().

Для работы с такими форматами в Python есть модуль [csv](https://docs.python.org/3/library/csv.html" \t "_blank).

В модуле есть два основных объекта: reader и writer, созданные, чтобы читать и создавать csv-файлы соответственно.

Приведем пример использования **читателя** с почти полным набором значений, указав:

* Кодировку файла
* Символ-разделитель
* Разделитель текста

Объект reader дает доступ к построчному итератору полностью аналогично работе с файлом или списком.

Общности ради в следующем примере мы покажем, что разделителем может быть любой символ.

import csv

with open('files/ikea.csv', encoding="utf8") as csvfile:

reader = csv.reader(csvfile, delimiter=';', quotechar='"')

for index, row in enumerate(reader):

if index > 10:

break

print(row)

['keywords', 'price', 'product\_name']

['МОРУМ, Ковёр, безворсовый', '6999', 'МОРУМ']

['МОРУМ, Ковёр, безворсовый', '6999', 'МОРУМ']

['ИДБИ, Придверный коврик', '649', 'ИДБИ']

['ХОДДЕ, Ковёр, безворсовый', '1399', 'ХОДДЕ']

['ОПЛЕВ, Придверный коврик', '599', 'ОПЛЕВ']

['ОПЛЕВ, Придверный коврик', '599', 'ОПЛЕВ']

['ЮНКЭН, Брикеты', '89', 'ЮНКЭН']

['БУНСЁ, Детское садовое кресло', '1199', 'БУНСЁ']

['ИКЕА ПС ВОГЭ, Садовое лёгкое кресло', '1999', 'ИКЕА ПС ВОГЭ']

['КУНГСХОЛЬМЕН, Садовый табурет', '5500', 'КУНГСХОЛЬМЕН']

Давайте разберем построчно, что происходит в этом коде.

Мы пользуемся менеджером контекста with, чтобы просто открыть наш файл с кодировкой UTF-8, а потом создаем объект reader, говоря ему про символы-разделители полей и строк.

Объект reader может служить итератором (и использоваться в цикле for) по строкам, каждая из которых представляет собой список. При создании reader мы указываем, что символ-разделитель записей delimiter в нашем файле — точка с запятой, а символ кавычек quotechar — двойные кавычки. Кроме того, мы используем enumerate, чтобы посчитать строки.

Отметим, что исходный файл содержит подписи полей в первой строке, что будет снова нам мешать (например, при сортировке строк). Для корректной работы мы должны были бы исключить первую строку из обработки.

Но в модуле csv есть специальный объект **DictReader**, который поддерживает создание объекта-словаря на основе подписей к полям.

**DictReader** не просто словарь, а словарь, который отслеживает порядок ключей после их добавления, — **OrderDict**, что будет дальше видно в примере. Дополнительно почитать про **OrderDict** можно [тут](https://docs.python.org/3.6/library/collections.html).

Теперь мы можем обращаться к полям не по индексу, а по **названию**, что делает программу еще более понятной.

Найдем топ-10 самых дорогих товаров (как вы думаете, какая запись более понятна: int(x['price']) или int(x[1])?):

with open('files/ikea.csv', encoding="utf8") as csvfile:

reader = csv.DictReader(csvfile, delimiter=';', quotechar='"')

expensive = sorted(reader, key=lambda x: int(x['price']), reverse=True)

for record in expensive[:10]:

print(record)

{'keywords': 'МОРУМ, Ковёр, безворсовый', 'price': '6999', 'product\_name': 'МОРУМ'}

{'keywords': 'КУНГСХОЛЬМЕН, Садовый табурет', 'price': '5500', 'product\_name': 'КУНГСХОЛЬМЕН'}

{'keywords': 'ИКЕА ПС ВОГЭ, Садовое лёгкое кресло', 'price': '1999', 'product\_name': 'ИКЕА ПС ВОГЭ'}

{'keywords': 'ХОДДЕ, Ковёр, безворсовый', 'price': '1399', 'product\_name': 'ХОДДЕ'}

{'keywords': 'БУНСЁ, Детское садовое кресло', 'price': '1199', 'product\_name': 'БУНСЁ'}

{'keywords': 'МОРУМ, Ковёр, ворсовый', 'price': '999', 'product\_name': 'МОРУМ'}

{'keywords': 'ИДБИ, Придверный коврик', 'price': '649', 'product\_name': 'ИДБИ'}

{'keywords': 'ОПЛЕВ, Придверный коврик', 'price': '599', 'product\_name': 'ОПЛЕВ'}

{'keywords': 'ОПЛЕВ, дверный коврик', 'price': '99', 'product\_name': 'ОПЛЕВ'}

{'keywords': 'ЮНКЭН, Брикеты', 'price': '89', 'product\_name': 'ЮНКЭН'}

Мы привели цены к типу int, потому что строки сравниваются в лексикографическом порядке (по алфавиту). Например:

print('11' > '100')

True

print(11 > 100)

False

Использование объекта для записи (writer) аналогично «читателю» (reader):

with open('files/квадраты.csv', 'w', newline='') as csvfile:

writer = csv.writer(

csvfile, delimiter=';', quotechar='"', quoting=csv.QUOTE\_MINIMAL)

for i in range(10):

writer.writerow([i, i \*\* 2, "Квадрат числа %d равен %d" % (i, i \*\* 2)])

В этом случае использовался опциональный параметр функции open() newline. Он отвечает за переводы строк при чтении или записи в текстовый файл. По умолчанию имеет значение None, в этом случае все разделители строк преобразуются в \n. Если в файле оказывается лишний перевод строки, то следует использовать этот параметр в режиме newline='', тогда \n будет преобразован в пустую строку.

Выполните этот код и посмотрите, что получилось.

Записывать в csv-файл можно и с помощью DictWriter, аналогичного DictReader. Но нужно ему указать, какие заголовки должны быть в файле и какие значения им соответствуют у каждой записи. Для этого сначала подготовим список словарей:

import csv

data = [{

'lastname': 'Иванов',

'firstname': 'Пётр',

'class\_number': 9,

'class\_letter': 'А'

}, {

'lastname': 'Кузнецов',

'firstname': 'Алексей',

'class\_number': 9,

'class\_letter': 'В'

}, {

'lastname': 'Меньшова',

'firstname': 'Алиса',

'class\_number': 9,

'class\_letter': 'А'

}, {

'lastname': 'Иванова',

'firstname': 'Татьяна',

'class\_number': 9,

'class\_letter': 'Б'

}]

with open('dictwriter.csv', 'w', newline='') as f:

writer = csv.DictWriter(

f, fieldnames=list(data[0].keys()),

delimiter=';', quoting=csv.QUOTE\_NONNUMERIC)

writer.writeheader()

for d in data:

writer.writerow(d)

Результат:

"lastname";"firstname";"class\_number";"class\_letter"

"Иванов";"Пётр";9;"А"

"Кузнецов";"Алексей";9;"В"

"Меньшова";"Алиса";9;"А"

"Иванова";"Татьяна";9;"Б"

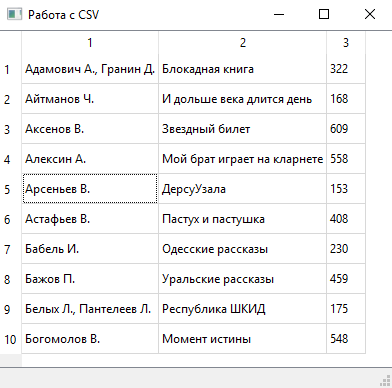
Скачать файл ikea.csv можно по [ссылке](https://yastatic.net/s3/lyceum/content/resources/ikea.csv).

**Использование графического интерфейса. Чтение из .csv файла**

Поскольку большинство программ используются людьми без опыта работы в командной строке, в пакетах, отвечающих за создание графического интерфейса, существуют специальные модули для работы с табличными данными. Для работы с таблицами в PyQT существует класс **Table Widget**. Давайте рассмотрим несколько примеров его использования.

Создадим простое приложение, которое будет отображать содержимое таблицы, которая хранится в формате .csv.

Для этого создадим в QtDesigner новую форму, добавим на нее **Table Widget**, а затем подключим интерфейс к нашей программе (в виде ui-файла или конвертируем полученный GUI в файл с расширением .py).



import csv

import sys

from PyQt5 import uic

from PyQt5.QtWidgets import QApplication

from PyQt5.QtWidgets import QMainWindow, QTableWidgetItem

class MyWidget(QMainWindow):

def \_\_init\_\_(self):

super().\_\_init\_\_()

uic.loadUi('UI1.ui', self)

self.loadTable('Data.csv')

def loadTable(self, table\_name):

with open(table\_name, encoding="utf8") as csvfile:

reader = csv.reader(csvfile,

delimiter=';', quotechar='"')

title = next(reader)

self.tableWidget.setColumnCount(len(title))

self.tableWidget.setHorizontalHeaderLabels(title)

self.tableWidget.setRowCount(0)

for i, row in enumerate(reader):

self.tableWidget.setRowCount(

self.tableWidget.rowCount() + 1)

for j, elem in enumerate(row):

self.tableWidget.setItem(

i, j, QTableWidgetItem(elem))

self.tableWidget.resizeColumnsToContents()

if \_\_name\_\_ == '\_\_main\_\_':

app = QApplication(sys.argv)

ex = MyWidget()

ex.show()

sys.exit(app.exec())

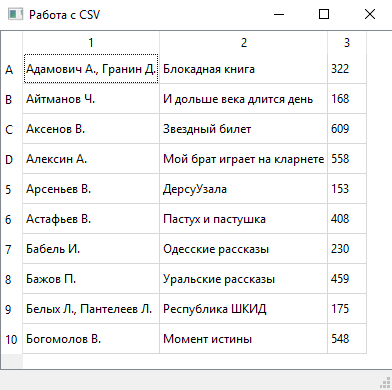
В методе loadTable() класса MyWidget происходит загрузка содержимого в виджет таблицы. Давайте посмотрим более детально.

Для отображения таблицы необходимо указать ее размеры: количество строк и столбцов. Если бы у нас была фиксированная матрица, мы бы легко узнали ее размеры, но в случае работы с .csv мы используем итератор, поэтому необходимо применить некоторые «лайфхаки».

Сначала, так же, как и в прошлых примерах, мы открываем файл для чтения. Поскольку reader является итератором, то, для того чтобы получить первую строку, необходимо воспользоваться функцией next().

Чтобы отобразить заголовки в виджете, воспользуемся методом setHorizontalHeaderLabels(). Кстати, существует также функция setVerticalHeaderLabels() — для задания вертикальных заголовков. По умолчанию, если программист не задает текст заголовков, это просто номера. Если в функцию передано недостаточное количество заголовков, после того как они «закончатся», следующие заголовки будут цифровыми. Например:

self.tableWidget.setVerticalHeaderLabels(['A', 'B', 'C', 'D'])



Зная количество заголовков, можно установить количество столбцов в нашей таблице, используя метод setColumnCount(). Но так как мы не знаем (если не преобразовали reader в список), сколько элементов в нашем итераторе, установим изначальное количество, равное 0, а затем будем на каждом шаге увеличивать его на единицу.

Основным методом при работе с QTableWidget является tableWidget.setItem(i, j, QTableWidgetItem(elem)). Он помещает в заданную с помощью координат ячейку соответствующее значение. Важно не забыть, что в таблице хранятся не текстовые, не числовые и не какие-либо другие данные, а данные типа QTableWidgetItem, поэтому не забывайте приводить значения к этому типу.

Очень часто непонятно заранее, какого размера будет содержимое той или иной ячейки в таблице. Для того чтобы размеры ячеек соответствовали длине текста, используем метод resizeColumnsToContents() после того, как заполним таблицу данными. Он растянет или сузит, при необходимости, размеры ячейки, чтобы пользователю был виден весь текст.

Согласитесь, что сейчас таблица выглядит блекло и неинтересно. Давайте раскрасим ее. Для этого нам понадобится импортировать модуль, отвечающий за работу с цветом: from PyQt5.QtGui import QColor. К сожалению, по умолчанию в PyQT нет функции, чтобы изменить цвет целого ряда таблицы, а есть только для одной ячейки. Так что придется написать такой метод самостоятельно.

def color\_row(self, row, color):

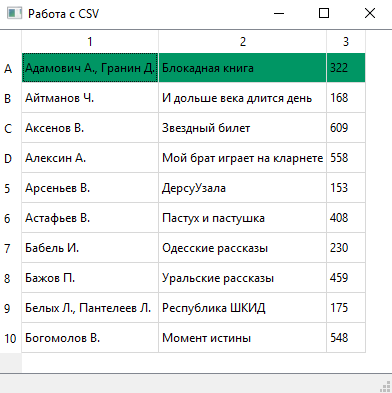
for i in range(self.tableWidget.columnCount()):

self.tableWidget.item(row, i).setBackground(color)

И теперь, если мы захотим выделить нулевой ряд цветом, нам достаточно лишь вызвать этот метод:

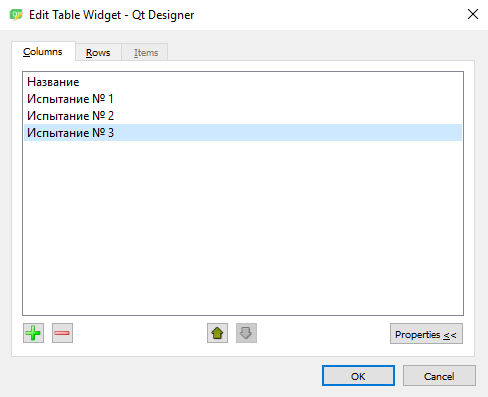
self.color\_row(0, QColor(0, 150, 100))

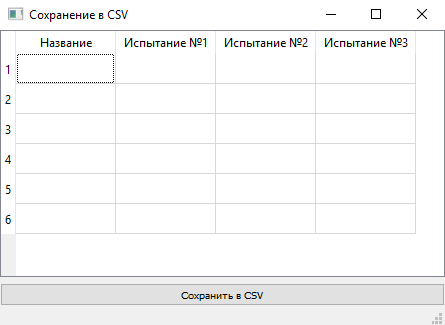
Получим такой результат:



**PyQT. Запись из таблицы в .csv файл**

Мы уже научились читать данные из файла и отображать их с помощью таблицы. Теперь попробуем обратный процесс. Для этого напишем программу, которая может использоваться при выставлении баллов команде за различные этапы на каком-нибудь соревновании. Известно, что и команд, и испытаний фиксированное число. Так что можно создать сетку для нашей таблицы, используя QtDesigner. Для этого необходимо навести курсор на виджет, нажать правую кнопку мыши, выбрать пункт меню **Edit Items...**, а затем указать данные.





Для сохранения данных напишем метод save\_2\_csv:

def save\_2\_csv(self):

with open('results.csv', 'w', newline='') as csvfile:

writer = csv.writer(

csvfile, delimiter=';', quotechar='"',

quoting=csv.QUOTE\_MINIMAL)

# Получение списка заголовков

writer.writerow(

[self.tableWidget.horizontalHeaderItem(i).text()

for i in range(self.tableWidget.columnCount())])

for i in range(self.tableWidget.rowCount()):

row = []

for j in range(self.tableWidget.columnCount()):

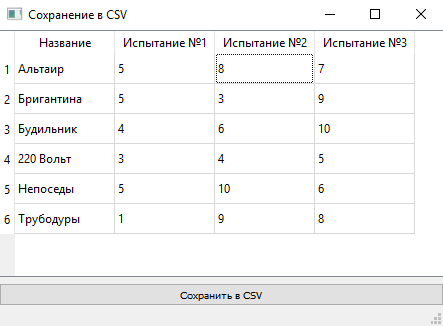
item = self.tableWidget.item(i, j)

if item is not None:

row.append(item.text())

writer.writerow(row)

К сожалению, в PyQT нет встроенной функции, которая возвращает список заголовков, так что пришлось использовать генератор и встроенный метод horizontalHeaderItem(i), который возвращает i-й заголовок. А затем каждая строка переносится в список, а этот список уже записывается в файл.



Результатом будет такой .csv-файлик:

|  |  |  |  |
| --- | --- | --- | --- |
| Название | Испытание №1 | Испытание №2 | Испытание №3 |
| Альтаир | 5 | 8 | 7 |
| Бригантина | 5 | 3 | 9 |
| Будильник | 4 | 6 | 10 |
| 220 Вольт | 3 | 4 | 5 |
| Непоседы | 5 | 10 | 6 |
| Трубодуры | 1 | 9 | 8 |