Регулярные выражения

В дальнейшем мы будем часто использовать текстовые данные. Для обработки таких данных используются регулярные выражения.

Регулярные выражения – это небольшой язык встроенный в Python, который позволяет производить поиск подстроки в строке, разбивать строку на подстроки или заменять часть строки.

Например, мы хотим заменить все буквы «ё» в строке, на букву «е». Конечно, мы можем это сделать с помощью for:

string = 'ёжик нёс яблоки под ёлку'  
print(string)  
string = list(string)  
  
for i in range(len(string)):  
 if string[i] == 'ё':  
 string [i] = 'е'  
  
string = ''.join(string)  
print(string)

Вывод на консоль:

ёжик нёс яблоки под ёлку

ежик нес яблоки под елку

А вот как это выглядит с использованием регулярных выражений:

import re  
  
string = 'ёжик нёс яблоки под ёлку'  
print(string)  
string = re.sub('ё', 'е', string)  
print(string)

Итак, для того чтобы использовать регулярные выражения, необходимо подключить модуль re.

import re

re – это библиотека для работы с регулярными выражениями в Python. sub(pattern, replace, string) – метод, предоставляемый библиотекой, который принимает шаблон, который необходимо найти, подстроку, на которую будет заменять шаблон и саму строку, в которой будет производиться замена, и возвращает новую строку, в которой произведена замена.

Наиболее часто используемые методы из библиотеки re:

* re.search;
* re.findall;
* re.sub;
* re.split;

re.search(pattern, string) – метод, позволяющий найти первое вхождения шаблона в строку:

import re  
  
string = 'мама мыла раму'  
result = re.search('а', string)  
print(result)

Вывод на консоль:

<re.Match object; span=(1, 2), match='а'>

Метод возвращает объект, в котором находится кортеж из двух элементов – с какого по какой индекс был найден шаблон и найденная подстрока.