нетология

Регулярные выражения

Елена Никитина

Елена Никитина

Руководитель проектов ООО “Аналитические программные решения”

План занятия

1. Что такое регулярные выражения и чем они отличаются от простого поиска
2. Синтаксис регулярных выражений
3. Модуль re в Python
4. Примеры использования регулярных выражений

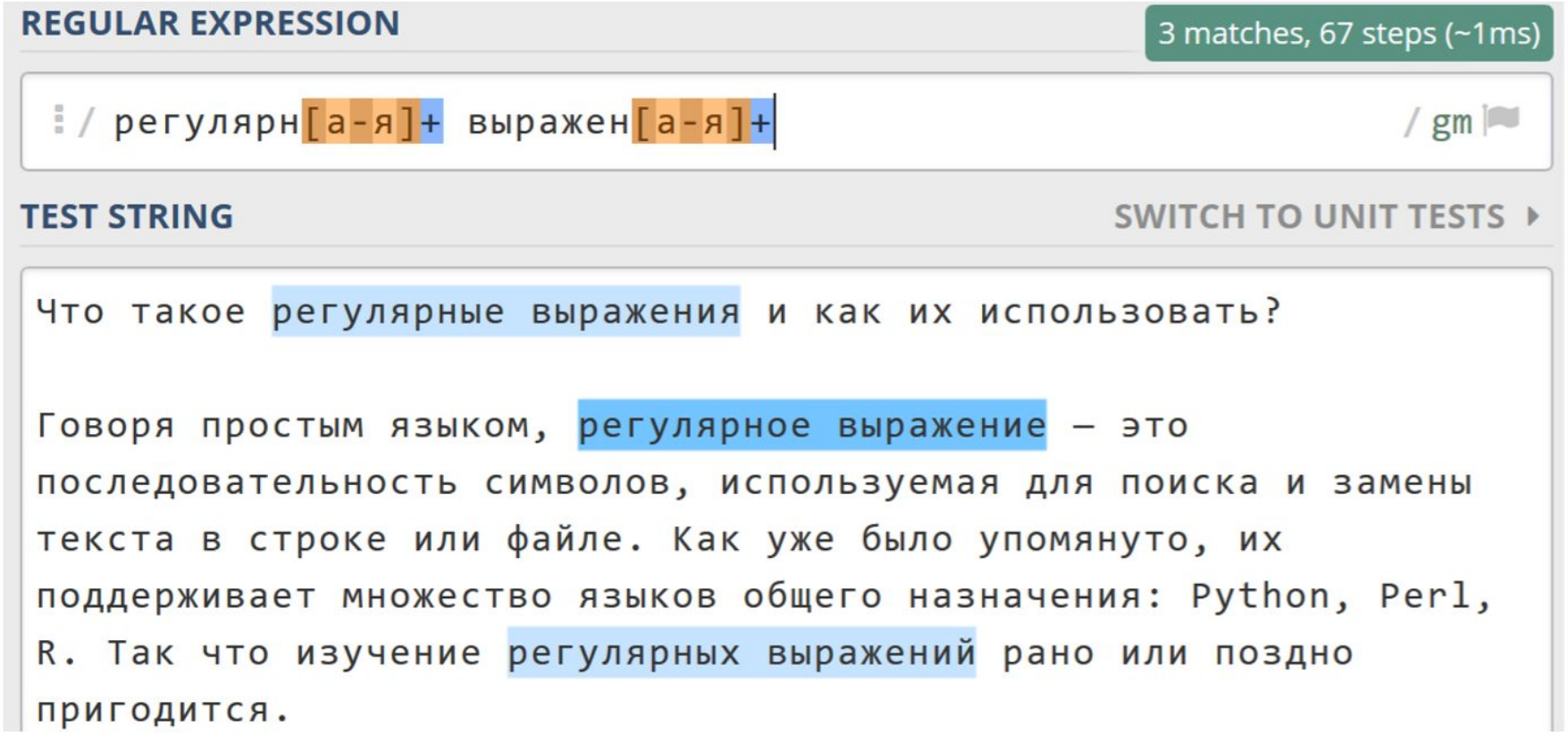
Регулярные выражения

Регулярные выражения (Regular Expressions, regex) – это простой язык поисковых запросов для поиска шаблонов в тексте.

Например, можно:

- найти и заменить любое количество пробелов на один
- найти все e-mail или телефоны в тексте, даже если они записаны по-разному
- находить все слова и фразы независимо от окончаний, числа, рода, падежа

Как работают регулярные выражения



https://regex101.com/

Синтаксис. Спецсимволы

.	Любой символ
^	1) начало строки, 2) инвертирование («всё, кроме»)
$	Конец строки
*****	Любое количество вхождений, от 0 до бесконечности
+	Количество вхождений от 1 до бесконечности
?	0 или 1 вхождение {n} точное количество вхождений – n раз
{n, m}	Количество вхождений не менее n и не более m раз
****	Символ экранирования. Например, символ точки: \.
|	Символ “или”. Найдет любой из шаблонов (например, 500|100 - будет находить и 500, и 100)

Синтаксис. Спецсимволы

[]	Набор символов, любой из которых может встретиться в тексте. *Например, [а-яёА-ЯЁ] — любая буква русского алфавита в любом регистре*
\d	Любая цифра. Аналогично [0-9]
\D	Все, кроме цифры. Аналогично [^0-9]
\w	(для unicode) Любые буква, цифра и символ подчёркивания. Для ASCII то же самое, но работает только для латинских букв
\W	Все, кроме букв, цифр и символа подчёркивания
\s	Любой пробельный символ, включая сам пробел: [\t\n\r\f\v]
\S	Все, кроме пробельных символов
(...)	Группировка символов (вместо **...** — текст или регулярное выражение)

Основные функции модуля re

- **re.match** (pattern, string, flags=0)

  Ищет по заданному шаблону строго с начала текста
- **re.search** (pattern, string, flags=0)

  Ищет во всём тексте, возвращает первое совпадение
- **re.findall** (pattern, string)

  Ищет во всём тексте, возвращает список всех найденных совпадений
- **re.compile** (pattern, flags=0)

  Компилирует регулярное выражение в объект, который можно вызывать без “re”.

  Для многократного использования. Работает быстрее, чем “re”.
- **re.split** (pattern, string, maxsplit=0, flags=0)

  Разделяет строку по заданному шаблону
- **re.sub** (pattern, repl, string)

  Ищет шаблон в строке и заменяет его на указанную подстроку

re.match, re.search

```
 7   pattern = re.compile("регулярн[а-я]+ выражен[а-я]+")
 8   print("=== re.match ===")
 9   print(re.match(pattern, text))
10   result = re.match("Что", text)
11   print(result)
12   print(result.group(0))
13   print("First: {}, last: {}".format(result.start(), result.end()))
14   print("=== re.search ===")
15   print(re.search(pattern, text))
```

```
=== re.match ===
None
<_sre.SRE_Match object; span=(0, 3), match='Что'>
Что
First: 0, last: 3
=== re.search ===
<_sre.SRE_Match object; span=(10, 30), match='регулярные выражения'>
```

re.findall + re.compile

```
1   import re
2
3   pattern = re.compile("регулярн[а-я]+ выражен[а-я]+")
4   text = "Что такое регулярные выражения и как их использовать?\
5   Говоря простым языком, регулярное выражение – это последовательность символов,
    используемая для поиска и замены текста в строке или файле. Как уже было упомянуто, их
    поддерживает множество языков общего назначения: Python, Perl, R. Так что изучение
    регулярных выражений рано или поздно пригодится."
6
7   result = pattern.findall(text)
8   print(result)
```

```
Python 3.6.1 (default, Dec 2015, 13:05:11)
[GCC 4.8.2] on linux
['регулярные выражения', 'регулярное выражение', 'регулярных выраже
ний']
```

re.split

```
3   text = "Что такое регулярные выражения и как их использовать?\
4   Говоря простым языком, регулярное выражение – это последовательность символов,
    используемая для поиска и замены текста в строке или файле. Как уже было упомянуто,
    их поддерживает множество языков общего назначения: Python, Perl, R. Так что
    изучение регулярных выражений рано или поздно пригодится."
5
6   pattern = re.compile("регулярн[а-я]+ выражен[а-я]+")
7   print("=== re.split ===")
8   result = re.split("[.?]", text)
9   print(result)
10  print("Всего предложений: {}".format(len(result)))
```

```
=== re.split ===
['Что такое регулярные выражения и как их использовать', 'Говоря просты
м языком, регулярное выражение – это последовательность символов, испол
ьзуемая для поиска и замены текста в строке или файле', ' Как уже было
упомянуто, их поддерживает множество языков общего назначения: Python,
Perl, R', ' Так что изучение регулярных выражений рано или поздно приго
дится', '']
Всего предложений: 5
```

re.sub

```
3   text = "Что такое регулярные выражения и как их использовать?\
4   Говоря простым языком, регулярное выражение – это последовательность символов, используемая для
    поиска и замены текста в строке или файле. Как уже было упомянуто, их поддерживает множество
    языков общего назначения: Python, Perl, R. Так что изучение регулярных выражений рано или поздно
    пригодится."
5
6   pattern = re.compile("регулярн[а-я]+ выражен[а-я]+")
7   print("=== re.sub ===")
8   result = pattern.sub("RegEx", text)
9   print(result)
```

```
=== re.sub ===
Что такое RegEx и как их использовать?Говоря простым языком, RegEx – это пос
ледовательность символов, используемая для поиска и замены текста в строке и
ли файле. Как уже было упомянуто, их поддерживает множество языков общего на
значения: Python, Perl, R. Так что изучение RegEx рано или поздно пригодится
```

Флаги

re.A, re.ASCII	ASCII-диапазон символов вместо Юникода
re.U, re.UNICODE	Использование диапазонов Юникода. Работает по умолчанию, можно не назначать
re.I, re.IGNORECASE	Игнорировать регистр символов
re.M, re.MULTILINE	Разбивать текст на строки при обработке. Нужен, в основном, для функций re.match и re.search
re.S, re.DOTALL	По умолчанию символ точки означает любой символ, кроме символа новой строки \n. Если назначить этот флаг, ограничение снимается

Модификаторы

```
3   text = "Что такое регулярные выражения и как их использовать?\
4   Говоря простым языком, регулярное выражение – это последовательность символов,
     используемая для поиска и замены текста в строке или файле. Как уже было
    упомянуто, их поддерживает множество языков общего назначения: Python, Perl,
    R. Так что изучение регулярных выражений рано или поздно пригодится."
5
6   pattern = "[\w]+"
7   result = re.findall(pattern, text, re.U)
8   print(result)
```

```
['Что', 'такое', 'регулярные', 'выражения', 'и', 'как', 'их', 'использовать', 'Говоря', 'простым
', 'языком', 'регулярное', 'выражение', 'это', 'последовательность', 'символов', 'используемая',
 'для', 'поиска', 'и', 'замены', 'текста', 'в', 'строке', 'или', 'файле', 'Как', 'уже', 'было',
'упомянуто', 'их', 'поддерживает', 'множество', 'языков', 'общего', 'назначения', 'Python', 'Per
l', 'R', 'Так', 'что', 'изучение', 'регулярных', 'выражений', 'рано', 'или', 'поздно', 'пригодит
ся']
```

Полезные ссылки

- Документация по регулярным выражениям
- Тестер регулярных выражений

Домашнее задание

Давайте посмотрим ваше домашнее задание.

- Вопросы по домашней работе задавайте в чате
- Задачи можно сдавать по частям
- Зачёт по домашней работе проставляется после того, как приняты **все задачи**

нетология

Задавайте вопросы и напишите отзыв о лекции

Елена Никитина