

# Регулярные выражения

Семинар 12

# Методы анализа текстов

- Регулярные выражения
- Алгоритмы поиска данных

# Регулярные выражения

**Регулярные выражения** (англ. regular expressions) — формальный язык поиска и осуществления манипуляций с подстроками в тексте, основанный на использовании метасимволов.

# Постановка задачи

Вам необходимо найти все строки, отвечающие некоторому шаблону.

Каким образом можно записать этот шаблон?

Что должно входить в этот шаблон?

Шаблон последовательно сравнивается с символами входной строки. Результатом является совпадение или несовпадение строки с шаблоном.

# Регулярные выражения

A

B - конкретные символы

C

AB – символ A за которым идет символ B

ABC – последовательность ABC

# Регулярные выражения

$A \mid B$  – символ  $A$  или символ  $B$  (только один)

$AB \mid AC$  – строка  $AB$  или строка  $AC$  (одна из двух)

# Регулярные выражения

В квадратных символах записывается набор символов, в котором проводится поиск текущего символа.

[A B C D E F 0 1 2 3 4 5 6 7 8 9] – набор символов

A–Z – интервал символов (только в [] )

[A–F 0–9] – точно такой же набор символов

# Регулярные выражения

Круглые скобки разделяют группы символов  
(AB | AC) [0-9] – AB или AC после которых  
идет цифра



# Регулярные выражения

$A?$  — символ  $A$ , которого может не быть

$CAT(T|H)?(Y|IE)$

$CAT[TH]?(Y|IE)$

# Регулярные выражения

$\wedge$  — начало строки

$\$$  — конец строки

$\cdot$  (точка) — любой символ

# Регулярные выражения (особенности)

[ ^ 1 – 6 ] – символы КРОМЕ 1-6 (здесь ^ используется как исключение набора символов)

[ – / . ] – символы -, / и . (не любой символ)

\ . – символ точки (вне квадратных скобок)

\ [ – символ открывающей квадратной скобки

\ ( – символ открывающей круглой скобки

\\ – обратный слеш

и так далее

# Регулярные выражения (специальные символы)

`\n` символ перевода строки

`\r` символ возврата каретки

`\t` символ табуляции

`\xhh` вставка символа с кодом 0xhh

`\d` цифра (0-9)

`\D` не цифра

`\s` пустой символ

`\S` непустой символ

`\w` обычно все буквы, цифры и подчеркивание

`\W` все, кроме `\w`

# Регулярные выражения (повторы)

- \* – полная итерация (повторяется 0 или больше раз)
- + – положительная итерация (повторяется 1 или больше раз)
- { 1 , 3 } – повторяется от 1 до 3 раз
- { , 3 } – повторяется от 0 до 3 раз
- { 3 , } – повторяется от 3 раз
- { 3 } – повторяется ровно 3 раза

# Регулярные выражения (примеры)

$[1-9][0-9]^*|0$  – целое неотрицательное  
число

$-?[1-9][0-9]^*|0$  – целое  
положительное/отрицательное число или  
ноль

# Классы символов

Класс знаков	Описание	Шаблон	Число соответствий
[ character_group ]	Соответствует любому одиночному символу, входящему в character_group. По умолчанию при сопоставлении учитывается регистр.	[ae]	"a" в "gray" "a", "e" в "lane"
[^ character_group ]	Отрицание: соответствует любому одиночному символу, не входящему в character_group. По умолчанию символы в character_group чувствительны к регистру.	[^aei]	"r", "g", "n" в "reign"
[ first - last ]	Диапазон символов: соответствует одному символу в диапазоне от первого до последнего.	[A-Z]	"A", "B" в "AB123"
.	Подстановочный знак: соответствует любому одиночному символу, кроме \n.  Для сопоставления символа точки ( . или \u002E) перед ней нужно поставить дополнительную обратную косую черту (\.).	a.e	"ave" в "nave" "ate" в "water"
\w	Соответствует любому алфавитно-цифровому знаку.	\w	"I", "D", "A", "1", "3" в "ID A1.3"
\W	Соответствует любому символу, который не является буквенно-цифровым знаком.	\W	" ", ".", " в "ID A1.3"
\d	Соответствует любой десятичной цифре.	\d	"4" в "4 = IV"
\D	Соответствует любому символу, не являющемуся десятичной цифрой.	\D	" ", "=", " ", "I", "V" в "4 = IV"

# Квантификаторы

Квантор указывает количество вхождений предшествующего элемента (знака, группы или класса знаков), которое должно присутствовать во входной строке, чтобы было зафиксировано соответствие. Кванторы состоят из языковых элементов, приведенных в следующей таблице.

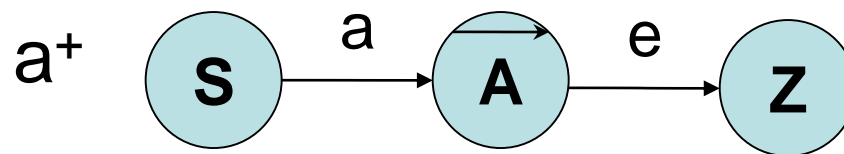
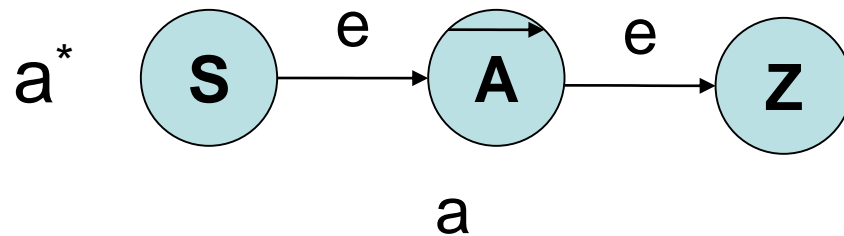
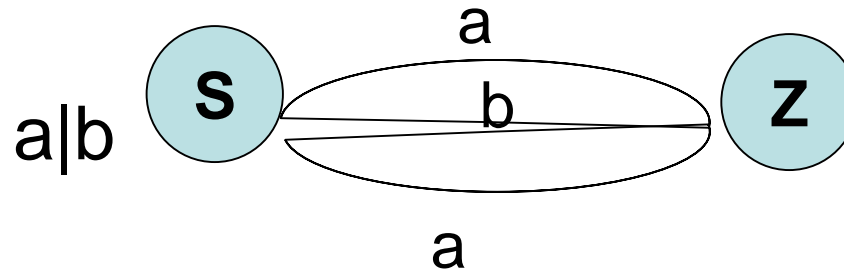
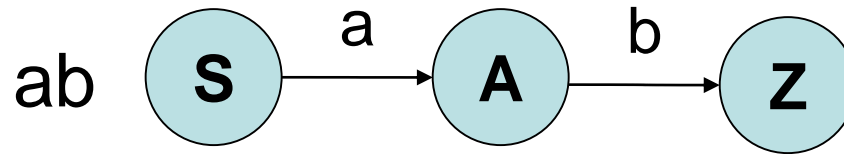
Квантификатор	Описание	Шаблон	Число соответствий
*	Соответствует предыдущему элементу ноль или более раз.	\d*\.\d	".0", "19.9", "219.9"
+	Соответствует предыдущему элементу один или более раз.	"be+"	"bee" в "been", "be" в "bent"
?	Соответствует предыдущему элементу ноль или один раз.	"rai?n"	"ran", "rain"
{ n }	Предыдущий элемент повторяется ровно n раз.	",\d{3}"	",043" в "1,043.6", ",876", ",543" и ",210" в "9,876,543,210"
{ n , }	Предыдущий элемент повторяется как минимум n раз.	"\d{2,}"	"166", "29", "1930"
{ n , m }	Предыдущий элемент повторяется как минимум n раз, но не более чем m раз.	"\d{3,5}"	"166", "17668"  "19302" в "193024"
*?	Предыдущий элемент не повторяется вообще или повторяется, но как можно меньшее число раз.	\d*?\.\d	".0", "19.9", "219.9"jkbhdbv



# Квантификаторы

Квантификатор	Описание	Шаблон	Число соответствий
<code>+</code>	Предыдущий элемент повторяется один или несколько раз, но как можно меньшее число раз.	<code>"be+?"</code>	<code>"be"</code> в <code>"been"</code> , <code>"be"</code> в <code>"bent"</code>
<code>??</code>	Предыдущий элемент не повторяется или повторяется один раз, но как можно меньшее число раз.	<code>"rai??n"</code>	<code>"ran"</code> , <code>"rain"</code>
<code>{ n }?</code>	Предыдущий элемент повторяется ровно n раз.	<code>"\d{3}?"</code>	<code>",043"</code> в <code>"1,043.6"</code> , <code>",876"</code> , <code>",543"</code> и <code>",210"</code> в <code>"9,876,543,210"</code>
<code>{ n , }?</code>	Предыдущий элемент повторяется как минимум n раз (как можно меньше).	<code>"\d{2,}?"</code>	<code>"166"</code> , <code>"29"</code> , <code>"1930"</code>
<code>{ n , m }?</code>	Предыдущий элемент повторяется не менее n и не более m раз (как можно меньше).	<code>"\d{3,5}?"</code>	<code>"166"</code> , <code>"17668"</code> <code>"193"</code> , <code>"024"</code> в <code>"193024"</code>

# Генерация конечного автомата по регулярному выражению





# Работа в Regex (примеры)

Regex101.com

← → ↻ 🏠 regex101.com ☆ ⚙️ A ⋮

regular expressions 101 @regex101 \$ donate 💖 sponsor 📧 contact 🚩 bug reports & feedback 📖 wiki 📄 whats new?

 SAVE & SHARE

 Save Regex **ctrl+s**

FLAVOR

⌘ PCRE2 (PHP >=7... ✓

⌘ PCRE (PHP <7.3)

⌘ ECMAScript (JavaScri...

⌘ Python 2.7

⌘ Golang

⌘ Java 8

FUNCTION

> Match ✓

✂ Substitution

📋 List

🧪 Unit Tests

TOOLS

📄 Code Generator

🔍 Regex Debugger

SPONSOR

**MOOVW**  
Jamstack at Scale

REGULAR EXPRESSION

6 matches, 12 steps (~1ms)

/ `[\\d]` / gm 📄

TEST STRING

123jhb445

EXPLANATION

▼ / `[\\d]` / gm

▼ Match a single character present in the list below

w

`[\\d]`

`[\\d]` matches a digit (equivalent to `[0-9]`)

▼ Global pattern flags

`g` modifier: **g**lobal. All matches (don't return

MATCH INFORMATION

Match 1	0-1	1
Match 2	1-2	2
Match 3	2-3	3

QUICK REFERENCE

Search reference

🗑 All Tokens

★ Common Token... ✓

🕒 General Tokens

🔗 Anchors

A single charact... `[abc]`

A character exc... `[^abc]`

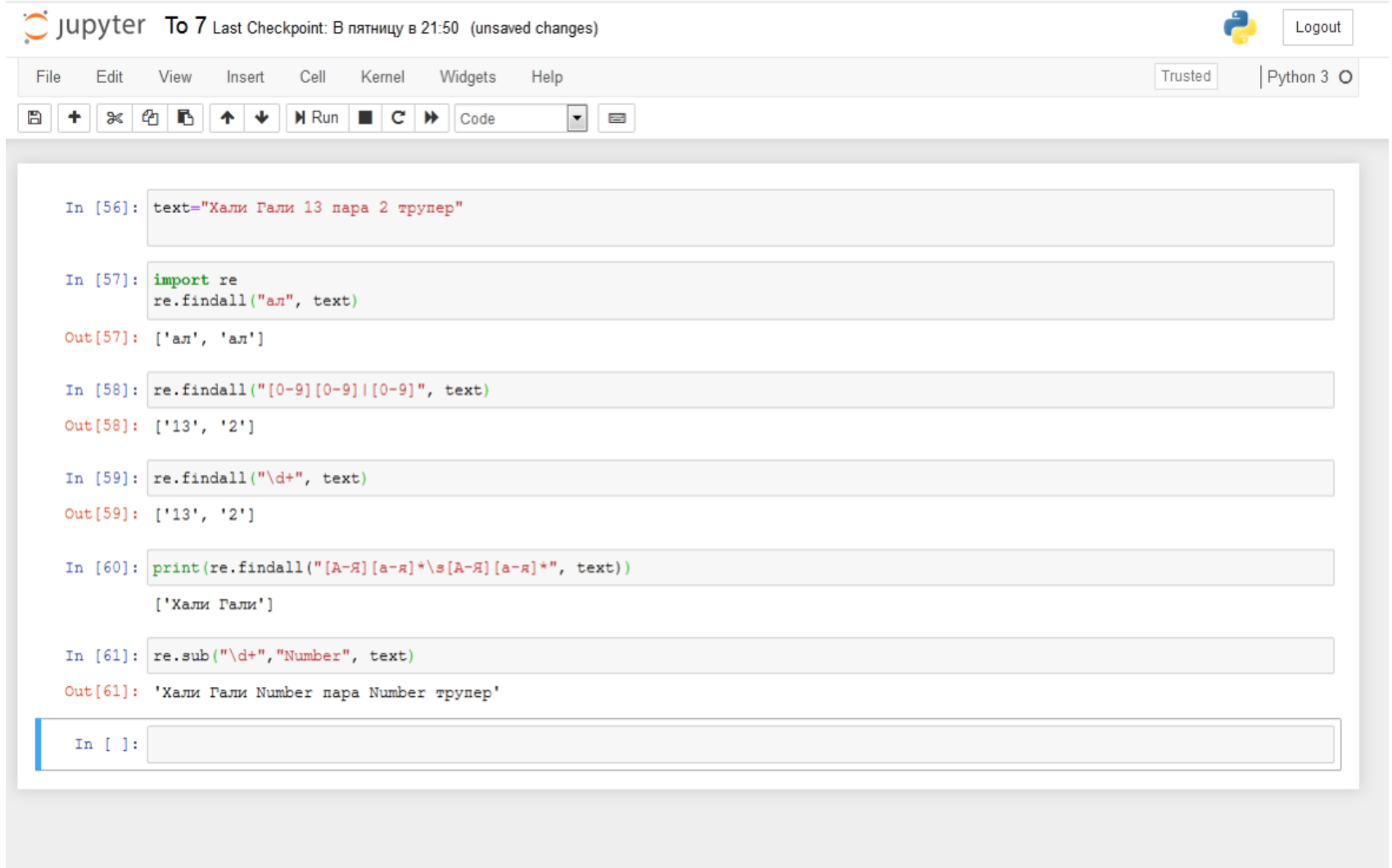
A character in th... `[a-z]`


A character not... `[^a-z]`

A character l... `[a-zA-Z]`

🪟 🔍 Введите здесь текст для поиска 🪟 📄 🗑 🔄 🛒 ⚙️ 📧 📁 📄 📄 📄 ENG 16:15 21.04.2021 🗨

# Работа в Python (примеры)



Jupyter To 7 Last Checkpoint: В пятницу в 21:50 (unsaved changes)  Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Code

```
In [56]: text="Хали Гали 13 пара 2 трупер"
```

```
In [57]: import re
re.findall("ал", text)
```

```
Out[57]: ['ал', 'ал']
```

```
In [58]: re.findall("[0-9][0-9]|[0-9]", text)
```

```
Out[58]: ['13', '2']
```

```
In [59]: re.findall("\d+", text)
```

```
Out[59]: ['13', '2']
```

```
In [60]: print(re.findall("[А-Я][а-я]*\s[А-Я][а-я]*", text))
['Хали Гали']
```

```
In [61]: re.sub("\d+", "Number", text)
```

```
Out[61]: 'Хали Гали Number пара Number трупер'
```

```
In [ ]:
```

# Пример

- [colab.research.google.com/drive/18tu9vgzoYzX7Xr93JuDAyBOBtzkBo-Kf#scrollTo=gjO\\_No88jBUf](https://colab.research.google.com/drive/18tu9vgzoYzX7Xr93JuDAyBOBtzkBo-Kf#scrollTo=gjO_No88jBUf)

# Задания

Написать регулярное выражение для:

9.30-10.50 – дата, идущая через слеш, дефис или точку

11.10-12.30 – домен сайта с двумя вложениями

13.00-14.20 – шестнадцатеричное число

14.40-16.00 – номер мобильного телефона в РФ