

Регулярные выражения

Поиск в огромном массиве текста

```
<script>espn_ui.Helpers.translate.init();</script>
<script type="text/javascript">
var data = {"omniture":{"columnist":"lowe_zach", "league":"nba", "countryRegion":"en-us", "hier1":"nba:story", "section":"nba", "source":"espn.com", "pageName":"nba:story", "storyInfo":"22258759+zach-low-blake-griffin-trade-future-la-clippers-detroit-pistons", "sections":"nba:story", "site":"espn", "premium":"premium-no", "convrSport":"basketball", "pageURL":"www.espn.com/nba/story/_/id/22258759/zach-low-blake-griffin-trade-future-la-clippers-detroit-pistons", "lang":"en_us", "prop35":"2018-01-30", "contentType":"story", "sport":"basketball", "account":"wdgespcom", "siteType":"full", "prop58":"isIndex=false"}, "chartbeat":{"loadPubJS":false, "path":"/nba/story/_/id/22258759/zach-low-blake-griffin-trade-future-la-clippers-detroit-pistons", "zone":"www.espn.com.us.nba", "domain":"www.espn.com", "loadViaJS":true, "title":"Zach Lowe on the Blake Griffin trade and future for LA Clippers, Pistons", "sections":"nba", "authors":"story"}, "q": {"cid":"P07264C85-15CD-4A80-8E56-B5BFA6D93296", "vc":"b01"}, "general":{"ci":"600140", "assetid":"N/A", "segB":"N/A", "sfcode":"N/A"}, "pnza":{"apid":"P07264C85-15CD-4A80-8E56-B5BFA6D93296", "vc":"b01"}, "espn":{"apid":"P07264C85-15CD-4A80-8E56-B5BFA6D93296", "vc":"b01"}, "espnin":{"apid":"P07264C85-15CD-4A80-8E56-B5BFA6D93296", "vc":"b01"}, "fantasy":{"apid":"P302B69D5-F1DD-4E7A-BF8D-3E60F0EB5E5A", "vc":"c07"}, "watchespn":{"apid":"P07264C85-15CD-4A80-8E56-B5BFA6D93296", "vc":"b01"}, "espnportes":
```

Как найти тут все
ссылки??!

Поиск в огромном массиве текста

Т.е. я ищу:

www.текст.com/что-то там/...

шаблон

Где понадобится?

- парсинг результатов вычислений
- парсинг сайта
- поиск и редактирование информации в человеческой речи
- работа с путями и файлами в ОС
- ...

Во время изучения
чего-то нового,
я самозабвенно
выдумываю
невероятные
ситуации, в которых
это умение поможет
мне спасти мир

О нет! Убийца должно
быть последовал
за ней в отпуск!



Но чтобы узнать где он, нам нужно
прочитать 200 Мб писем в поисках
чего-то похожего по формату с адресом!



— Это безнадежно!

Всем расступиться

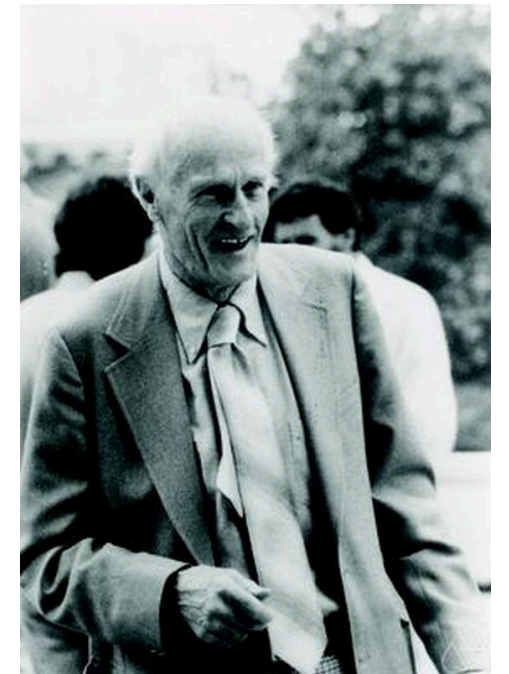
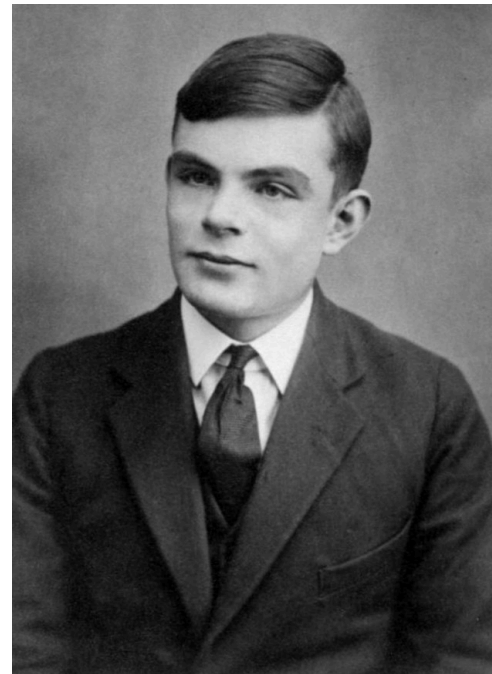


Я знаю регулярные
выражения



Математика

- 1970е – У. Маккалок и У. Питтс (USA) и А.Тьюринг (Eng).
Теория конечных автоматов и теория формальных языков.
- 1951 – С.Кинли (USA)
Регулярные (распознаваемые)
множества и языки
- 1962 – первые реализации
в программировании (SNOBOL)



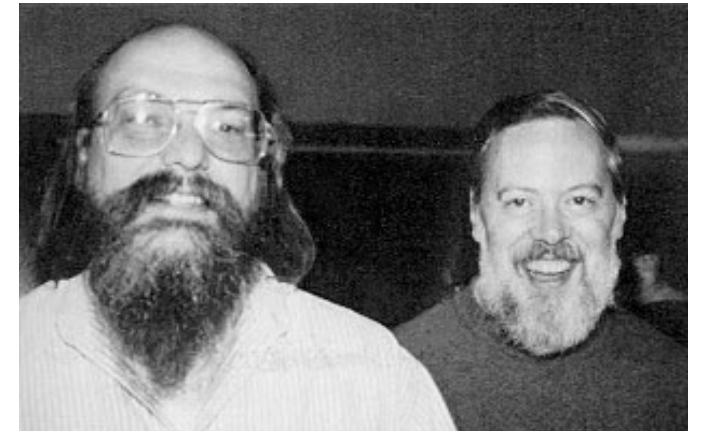
IT

- 1970е - Кен Томпсон и Деннис Ритчи внедряют Regex в UNIX vi, ed и др. редакторы

- 1980е – PostgreSQL

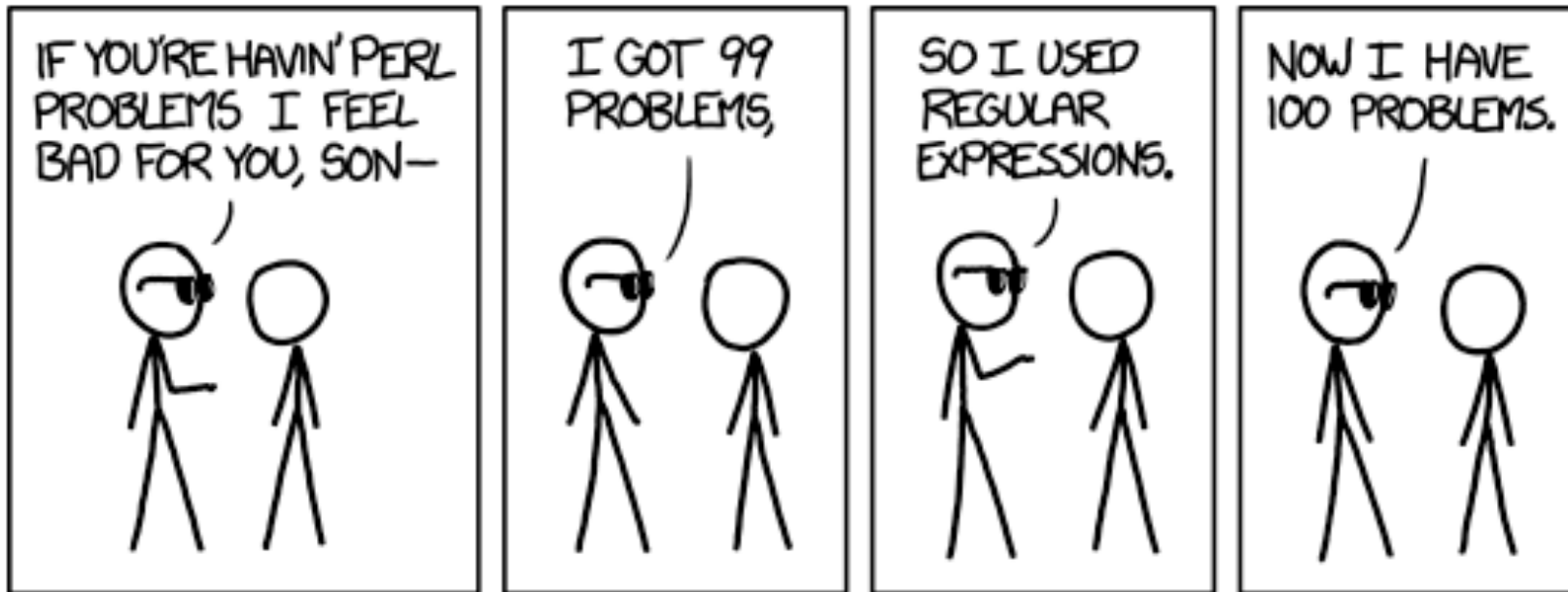
- и понеслась

```
$ ed fstab
Newline appended
116
%1
/dev/hda2 / ext2 defaults 1 1\r$
/dev/hdb1 /home ext2 defaults 1 2\r$
/dev/hda1 swap swap pri=40 0 0$
3s/40/42/
w fstab
116
-
```



IT

- 1980e - Ларри Уолл и его Perl



IT

сейчас поддержка regex есть во всех адекватных редакторах текста:
Sublime, nano, atom, notepad++ ... Свои реализации есть и в Word.

```
26 https://sci-hub.se/https://doi.org/10.1016/j.ijheatmasstransfer.2009.01.024
27 https://sci-hub.se/https://doi.org/10.1142/S0218348X20500498
28 https://doi.org/10.1142/S0218348X20500553
29 https://doi.org/10.1142/S0218348X20500206
30 https://doi.org/10.1142/S0218348X2050022X
31 https://doi.org/10.1142/S0218348X20500048
32 https://doi.org/10.1142/S0218348X20500097
33 https://doi.org/10.1142/S0218348X20500103
34 https://doi.org/10.1142/S0218348X20500139
35 Y. Zhao, S. Gong, C. Zhang, Z. Zhang and Y. Jiang, Fractal characteristics of crack propaga-tion in coal under impact
loading, Fractals 26(2) (2018) 1840014
36 Mandal, T. Roychowdhury, K. Chirom, A. Bhat-tacharya and R. B. Singh, Complex multifractal nature in mycobacterium tuberculosis
genome, Sci. Rep. 7 (2017) 46395
37 S. Dutta, Decoding the morphological differences between himalayan glacial and fluvial landscapes using multifractal
analysis, Sci. Rep. 7 (2017) 11022
```

.* Aa “ ” ↻ ☐ .*\\.org.*

Find

Find Prev

Find All

Да что же это такое?

- Регулярные выражения (regex – regular expressions) – это **последовательность** [спец] символов, позволяющая находить в тексте **совпадение** с одним или несколькими **шаблонами**
- *.txt – регулярное выражение, с помощью которого я буду искать все файлы формата “.txt”
- regex - это математический концепт, который может быть по-своему реализован в разных языках программирования

Оператор	Описание
.	Один любой символ, кроме новой строки \n.
?	0 или 1 вхождение шаблона слева
+	1 и более вхождений шаблона слева
*	0 и более вхождений шаблона слева
\w	Любая цифра или буква (\W — все, кроме буквы или цифры)
\d	Любая цифра [0-9] (\D — все, кроме цифры)
\s	Любой пробельный символ (\S — любой непробельный символ)
\b	Граница слова
[..]	Один из символов в скобках ([^..] — любой символ, кроме тех, что в скобках)
\	Экранирование специальных символов (\. означает точку или \+ — знак «плюс»)
^ и \$	Начало и конец строки соответственно
{n,m}	От n до m вхождений ({,m} — от 0 до m)
a b	Соответствует a или b
()	Группирует выражение и возвращает найденный текст
\t, \n, \r	Символ табуляции, новой строки и возврата каретки соответственно

полный список тут: <https://habr.com/ru/post/349860/>

**When you do a regex expression correctly
first try without using google for help**



Как можно

Регулярка	Её смысл
simple text	В точности текст «simple text»
\d{5}	Последовательности из 5 цифр \d означает любую цифру {5} — ровно 5 раз
\d\d/\d\d/\d{4}	Даты в формате ДД/ММ/ГГГГ (и прочие куски, на них похожие, например, 98/76/5432)

Как не стоит

```
(?:[a-z0-9!#$%&'*/+=?^_`{|}~-]+(?:\.(?:[a-z0-9!#$%&'*/+=?^_`{|}~-]+)*|"(?:[\x01-\x08\x0b\x0c\x0e-\x1f\x21\x23-\x5b\x5d-\x7f]|\\(?:[\x01-\x09\x0b\x0c\x0e-\x7f])*"))@(?:[a-z0-9](?:[a-z0-9-]*[a-z0-9])?\.)+[a-z0-9](?:[a-z0-9-]*[a-z0-9])?|\\(?:[?:(?:25[0-5]| 2[0-4][0-9]|[01]?[0-9][0-9]?)\.){3}(?:25[0-5]| 2[0-4][0-9]|[01]?[0-9][0-9]?)|[a-z0-9-]*[a-z0-9]:(?:[\x01-\x08\x0b\x0c\x0e-\x1f\x21-\x5a\x53-\x7f]|\\(?:[\x01-\x09\x0b\x0c\x0e-\x7f])+)\\))
```

* позволяет распознать в тексте email

Несколько заданий

- <https://regex101.com/>
- <https://tproger.ru/translations/regular-expression-python/>

1. Найти подстроку "cat" в строке
Проверить на: "CatcatCATCaT"

Решение: r"cat"

CatcatCATCaT

2. Найти все подстроки, в которых после символа "a" идёт 2 или 3 символа b.
Проверить на: "Ab", "Cgiabb_ab_abbbbb"

Решение: r"ab{2,3}"

3. Найти все подстроки, в которых две подстроки из букв нижнего регистра
отделены символом "_"

Проверить на:

"John_Smith_name_surname_Name_Surname"
"John_Smith_name_surname_Name_Surname"

Решение: r"[a-z]+_[a-z]+"

Предостережение

будьте аккуратнее, используя регулярные выражения в web-запросах (сайты/ssh/email-протоколы/...)

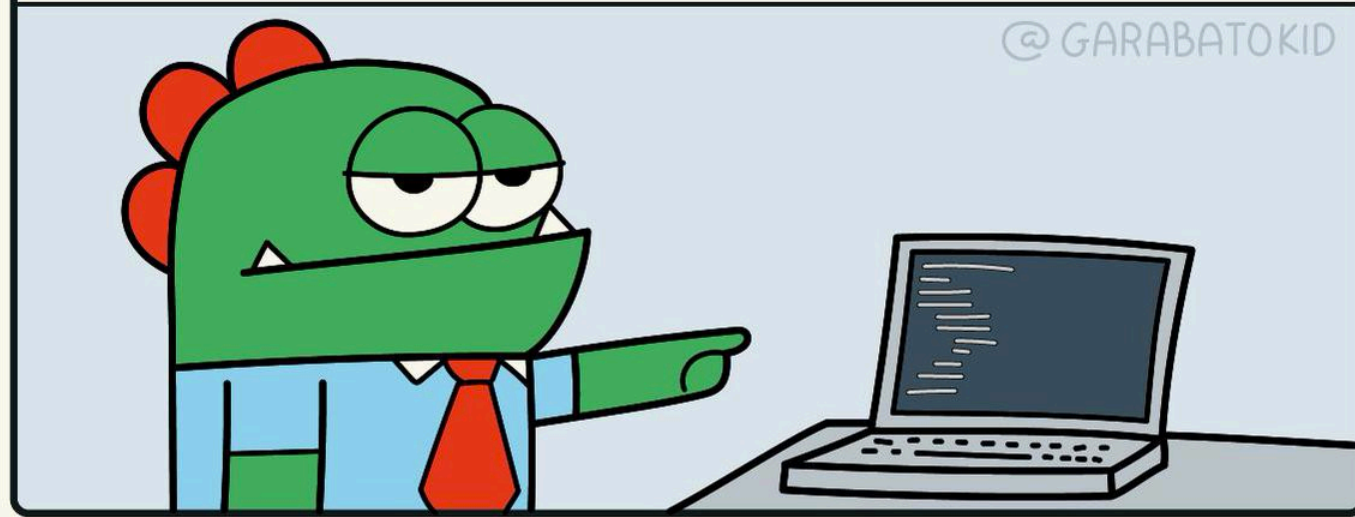
за такое запросто можно словить

используйте официальные API



HOW TO REGEX

STEP 1: OPEN YOUR FAVORITE EDITOR



STEP 2: LET YOUR CAT PLAY ON YOUR KEYBOARD



One more thing

G	H	P
Ссылка на репозиторий	Результаты теста 23/10	
https://github.com/AlexanderSwobodskii/Course_project.git	+	
https://github.com/Sergey1vstr/Semestral-Project	+	
https://github.com/Anastasiia-star/Kproject	+	
https://github.com/Anastasiia-star/Kproject	+	
	переписать	
	+	
	+	
https://github.com/Natalie-Palchikovskaya/my_project	+	
https://github.com/AfanasyevAA6/my-project	+	
https://github.com/anzh-bal/kursovaya	+	
	+	

