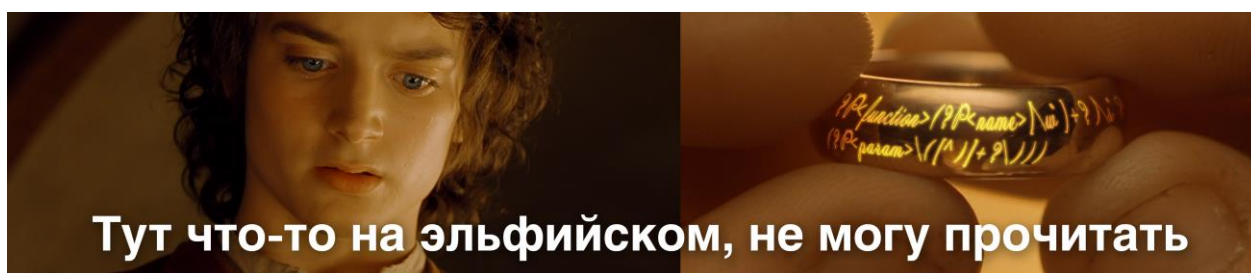


Как перестать бояться и полюбить регулярные выражения

- [Блог компании Миландр](#),
- [Регулярные выражения](#)

В работе и повседневной жизни мы буквально окружены текстами: книги и статьи, письма и чаты, веб-страницы и программный код. Да что там — сама биологическая жизнь основана на ДНК. И для эффективной работы по поиску или модификации текста уже давно был придуман мощный инструмент — регулярные выражения, или *regex*. Однако по ряду причин далеко не все их используют, или имеют дело с ними только ~~когда заставили~~ при крайней необходимости. В этой статье я попытаюсь исправить эту печальную ситуацию.



Чтобы функции отыскать, воедино созвать и единою чёрною волей сковать.

На Хабре уже есть ряд хороших статей, которые помогут разобраться в основах:

- [Регулярные выражения в Python от простого к сложному](#)
- [Регулярные выражения, пособие для новичков](#)

Мы не будем углубляться в работу регулярных выражений или пытаться в полном объёме покрыть все возможности, а рассмотрим на самых элементарных примерах, насколько просто начать пользоваться этим замечательным инструментом.

Иногда случается так, что регулярные выражения приходят в нашу жизнь совершенно неожиданно. Например, вы разбираетесь с валидацией пароля и сталкиваетесь со следующим:

```
^(?=.*[a-z])(?=.*[A-Z])(?=.*[0-9])(?=.*[!@#$%^()])[a-zA-Za-яA-Я0-9!@#$%^()*
```

```
] {1,}$
```

после чего настроение неизменно падает. Даже если и удалось продраться через нагромождение символов, то желание возвращаться к регулярным выражениям

наверняка сильно уменьшится. Человек же со стороны вообще может принять это за инопланетный язык и начать сторониться аки чумы: «Зачем мне ещё один язык программирования, да к тому же такой сложный?».



Гэндальф ещё Серый, а потому боится использовать regex.

Поиск

Для экспериментов нам понадобится старый добрый [Властелин Колец](#) и любой текстовый редактор с поддержкой регулярных выражений. В данной статье будет использоваться [Notepad++](#) (простой, удобный и расширяемый – очень рекомендую).

Копируем текст без оглавления в приложение, Ctrl+f, отмечаем «Regular expression» и вперёд!

Пример окна поиска в Notepad++.

1. Точное совпадение с фразой

Попробуем найти Исилдура (того, что у Саурона кольцо подрезал) на страницах «Властелина колец»

```
Find result - 62 hits
Search "Исилдур" (62 hits in 1 file)
D:\!\ WORK\Хабр\Властелин колец.txt (62 hits)
Line 1672: сече; и сын Элэндила Исилдур отсек Саурону палец вместе с Кольцом и взял
Line 1675: А Кольцо пропало. Оно кануло в воды Великой Реки Андуина. Ибо Исилдур
Line 8032: Средиземье, о могучем короле Элэндилье Высоком, о его сыновьях Исилдуре и
Line 8058: но и сам Саурон был повержен Исилдуrom, который завладел Кольцом Всевластья.
Line 8061: эпоха. А им, оказывается, завладел Исилдур?
Line 8064: родилось. Но Исилдур оставил Кольцо у себя. Видели это лишь мы с Сэрданом -
Line 8065: все остальные воины погибли, - и мы попытались предостеречь Исилдура, но он
```

Поздравляю! Вы только что использовали регулярное выражение «**Исилдур**». Тут всё просто – совпадение один к одному. Есть еще магия регистров букв (отмечаем галочку «Match Case»), но по умолчанию она обычно не включена. «**Исилдур**», «**исилдур**», «**ИсИлДУР**» дадут один и тот же результат. Если же мы знаем регистры букв искомого слова, то можем указать их и отбросить лишние результаты. Однако это может сыграть злую шутку, когда при последующем поиске может потеряться часть результатов. Убирайте галочку, если регистры букв могут быть разными.

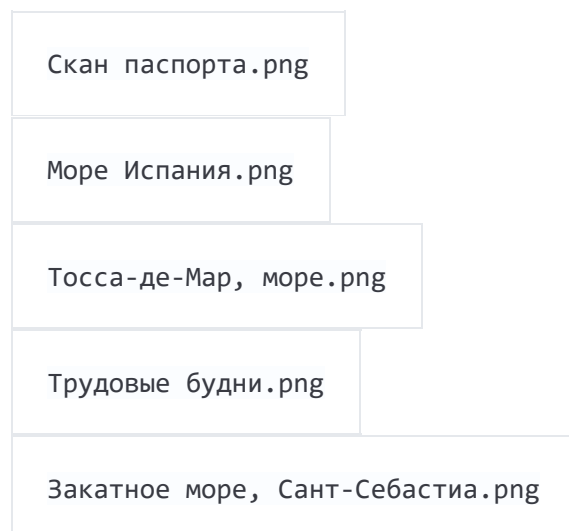
Веб браузеры, текстовые редакторы, мессенджеры – все имеют такую функцию. Однако не все об этом знают. Жаль, что у бумажных книг нет такой возможности.

Первый шаг сделан, и ничего не мешает двигаться дальше.

2. Символы подстановки

Еще одна распространённая возможность — это **символ подстановки** (wildcard) звёздочка «*». Она заменяет любое количество любых допустимых символов. Самый распространённый пример – поиск файлов.

Скажем, мы хотим найти фотографии отдыха на море. Для этого в строке поиска введём «***.png**» – и получим список всех фотографий в папке:



Уже хорошо. А что будет, если мы добавим слово «море» до или после звездочки?
море*.png

Море Испания.png

***море.png**

Тосса-де-Мар, море.png

Следующий вариант вернёт то, что нужно:

***море*.png**

Море Испания.png

Тосса-де-Мар, море.png

Закатное море, Сант-Себастья.png

«***море*.png**» — наглядный пример *шаблона поиска*, или *маски* — мы что-то знаем о структуре искомой строки (в данном случае расширение «**.png**» в конце и где-то слово «**море**»), в остальных местах предполагаем любые символы.

Регулярные выражения, как развитие подстановочных символов, имеют в своём арсенале огромный набор разнообразных средств, с помощью которых мы можем более тонко составлять шаблон поиска. Но пока познакомимся с аналогом звёздочки — точка и звёздочка «**.***».

3. Точка со звёздочкой «**.***»

Забудем о значении звёздочки в подстановочных символах – в регулярных выражениях она обозначает «любое количество, от 0 и до бесконечности». А точка обозначает «любой строковый символ». Вместе получаем «любое количество любых строковых символов».

Точка не включает в себя символ новой строки, поэтому поиск будет производится до конца строки, или, другими словами, абзаца. Посмотрим, как часто Фродо и Сэм оказывались вместе: для этого поищем по шаблонам «**Сэм.*Фродо**» и «**Фродо.*Сэм**»:

```
Find result - 103 hits
Search "Сэм.*Фродо" (103 hits in 1 file)
Search "Фродо.*Сэм" (112 hits in 1 file)
D:\! WORK\Хабр\Властелин колец.txt (112 hits)
Line 508: (наряду с выдуманной): она явно составлена по примечаниям фродо или Сэммиума
Line 2095: - Господин фродо, хоть вы-то! - в ужасе выкрикнул Сэм. - Вы-то хоть не
Line 2299: фродо захлопнул и запер круглую дверь, а ключ отдал Сэму.
Line 2504: - То есть как? - сурово спросил фродо, метнув на Сэма изумленный
Line 2841: фродо уселся возле Сэма и принялся за еду.
Line 2889: фродо изумленно поглядел на Сэма, будто и в нем ждал странных перемен.
Line 2981: фродо привалился спиной к стволу и закрыл глаза. Сэм с Пинотом, сидя
Line 3051: фродо и Сэм замерли, а Пин сделал еще несколько шагов. Ворота
Line 3481: - На кого уж теперь надеяться, - проворчал фродо, и Сэм опустил
Line 3493: - Так-то так, - сказал фродо, глядя на ухмыляющегося Сэма. - Только
Line 3914: В надежде на помощь и в страхе перед новой опасностью фродо и Сэм оба
Line 3939: - Помогите! - кинулись навстречу ему фродо и Сэм.
Line 4498: тесной цепочкой. фродо во главе, за ним Сэм, Пин и Мерри. Тропе, казалось,
Line 5268: фродо, Пин и Сэм, один за другим, вошли в свою темную комнату. Мерри не
```

215 совпадений. Кто бы сомневался...



Можно это сделать проще, одним запросом. Для этого используем специальный знак ИЛИ «|»: альтернативные варианты разделяются вертикальной чертой и окружаются скобками:

(Фродо.*Сэм|Сэм.*Фродо)

```
Find result - 213 hits
Search "(фродо.*Сэм|Сэм.*фродо)" (213 hits in 1 file)
D:\! WORK\Хабр\Властелин колец.txt (213 hits)
Line 508: (наряду с выдуманной): она явно составлена по примечаниям фродо или Сэммиума
Line 2095: - Господин фродо, хоть вы-то! - в ужасе выкрикнул Сэм. - Вы-то хоть не
Line 2116: - Ничего не поделаешь, Сэм, - грустно сказал фродо. Он вдруг понял, что
Line 2198: Балке, чуть подальше Зайгорода. Всем, кроме Сэма, фродо говорил, что туда он
Line 2294: Пин сидел и спал, откинувшись на свой мешок. Сэма не было. фродо
Line 2299: фродо захлопнул и запер круглую дверь, а ключ отдал Сэму.
Line 2449: За ним выпростались и Сэм с фродо - и тоже прилегли к обочине, мешки
Line 2504: - То есть как? - сурово спросил фродо, метнув на Сэма изумленный
Line 2700: его под руку и не давал упасть. А Сэм шагнул рядом с фродо и шел как во сне -
Line 2750: уже видел седьмой сон, а Сэм все сидел у ног фродо: крепился, крепился - и
Line 2841: фродо уселся возле Сэма и принялся за еду.
Line 2889: фродо изумленно поглядел на Сэма, будто и в нем ждал странных перемен.
Line 2981: фродо привалился спиной к стволу и закрыл глаза. Сэм с Пинотом, сидя
Line 3051: фродо и Сэм замерли, а Пин сделал еще несколько шагов. Ворота
Line 3474: - Сэм! - только и мог воскликнуть фродо. Он даже не знал, смеяться,
```

Можно заметить, что такой поиск выдал меньше совпадений — 213. Так получается, потому что, например, строчку «Рядом с Фродо шел всхлипывающий Сэм, да и сам Фродо безмолвно плакал.» отдельные запросы найдут два раза.

Чтобы узнать, как часто упоминается Гэндальф в полной форме, используем шаблон:

Гэндальф (Серый|Белый)

```
Find result - 11 hits
Search "Гэндальф (Белый|Серый)" (11 hits in 1 file)
D:\! WORK\Хабр\Властелин колец.txt (11 hits)
Line 355: Гэндальф Серый, а с ним тринадцать гномов: царь-изгнанник Торин Дубошит и
Line 8594: "Ты просишь помощи, Гэндальф Серый? - с издевкой в голосе переспросил
Line 9479: Впереди отряда шел Гэндальф Серый; по правую руку от него - Арагорн,
Line 12115: отправился Гэндальф Серый. Мне давно хотелось повидать его вновь, но границ
Line 14652: которым я пока не вправе говорить. Вел Отряд Гэндальф Серый.
Line 17013: статья, и всеобщая гибель. Я - Гэндальф, Гэндальф Белый, но черные силы
Line 18528: неведомой подмогой. Ты - великий волшебник, о Гэндальф Белый!
Line 19933: - Чудеса, да и только! - осклабился Саруман. - Вот он, Гэндальф Серый -
Line 19966: он, и голос его властно загремел. - Гляди, я уж не тот Гэндальф Серый, кого
Line 22990: - Гэндальф! - воскликнул Фродо. - Так я и подумал. Гэндальф Серый, наш
Line 27294: врасплох. У нас побывал Гэндальф Серый, и мы изготовились ехать на восток,
```

Всё. Имея на руках рассмотренное выше, можно начинать исследовать текстовые моря или болота логов. Начните с малого, исследуйте на возможность поиска те приложения, которыми вы часто пользуетесь. Задавайте поиску простые вопросы, и со временем вы сами почувствуете желание сделать поиск более сложным.

А если остались силы, то можно копнуть чуть глубже и взять на вооружение ещё два базовых понятия, изучив которые вы станете истинными начинающими regex-волшебниками.



Саруман понимает, какая мощь в его руках, и не боится её. Поэтому он всегда в курсе всех дел.

4. Количество повторений или квантификаторы

С одним **квантификатором** мы уже познакомились – звёздочка. Есть ещё несколько способов задать количество:

{4}	точное количество
{1,5}	диапазон
{2,}	от числа до бесконечности
{,7}	от 0 до числа

Некоторые часто используемые диапазоны имеют собственные символы:

* или {0,}	любое количество
? или {0,1}	ноль или один
+ или {1,}	есть как минимум один

Спросим, как часто упоминается одно и то же слово в одном абзаце:

```
Find result - 1 hit
Search "(Хоббит.*){3,}" (1 hit in 1 file)
D:\! WORK\Хабр\Властелин колец.txt (1 hit)
Line 4804: - Хоббиты! Целых четыре хоббита! Судя по выговору, из самой Хоббитании,
Search "(Бильбо.*){3,}" (0 hits in 0 files)
Search "(Бильбо.*){2}" (7 hits in 1 file)
D:\! WORK\Хабр\Властелин колец.txt (7 hits)
Line 11: Бильбо на восток и возвращении домой. Как раз по милости Бильбо хоббиты и
Line 635: Когда старик с помощью Бильбо и гномов разгрузил повозку, Бильбо роздал
Line 1113: более, что заложил его еще отец Бильбо. Дурри тут же простил Бильбо, что
Line 2069: Бильбо. Будет годовщина - тут-то и в путь, по примеру Бильбо. Без этого
Line 2361: А на Бильбо и правда очень похоже, особенно на Бильбо в последние годы,
Line 3364: при Бильбо, но раз ему угодно было секретничать, так я и не болтал. Бильбо я
Line 8260: Бильбо очень нужна Властелину Мордора? А Бильбо, как известно, нашел Кольцо.
```

Бильбо 2 раза — «(Бильбо.){2}» — упоминается в 7 случаях, а большее количество раз в абзаце — «(Бильбо.){3,}» — ни разу. Слово «Хоббит» встречается аж 3 раза в строке, как часть слов «Хоббиты» «хоббита» и «Хоббитании».

5. Представление символов

Точка (любой символ) — один из примеров *представления символов*. Другими примерами являются цифра «\d», цифра или буква «\w», набор символов «[abc]», «[1-9]», «[,.;:]».

Например, мы хотим узнать, как часто числа используются в книге: «\d+» или «\d{1,}» – как минимум одна цифра подряд:

```
Find result - 19 hits
Search "\d+" (19 hits in 1 file)
new 70 (19 hits)
Line 4: 1. О хоббитах
Line 137: пережив поветрие Черной Смерти в 37г. (Л.Х.), они плодились и множились,
Line 139: погибли голодной смертью, но уже и Дни Нужды (1158-1160) ко времени нашего
Line 139: погибли голодной смертью, но уже и Дни Нужды (1158-1160) ко времени нашего
Line 157: Зеленополье в 1147г.(Л.Х.), когда Бандобрас Крол наголову разгромил
Line 238: 2. О трубочном зелье
Line 254: зелье в своем огороде во дни Изенгрима Второго, год примерно 1070-й согласно
Line 286: 3. О благоустройстве Хоббитании
Line 293: прирезаны к Хоббитании в 1164 году (Л. Х.).
Line 353: 4. О том, как нашлось Кольцо
Line 356: двенадцать его сотоварищей. Апрельским утром 1341 года от заселения
Line 455: Он вернулся домой на пятьдесят втором году жизни, 22 июня 1342 года
Line 455: Он вернулся домой на пятьдесят втором году жизни, 22 июня 1342 года
Line 457: не собрался праздновать свое стоодиннадцатилетие (год 1401). Тут и начало
```

Находим номера глав и даты. Если не брать совсем уж легендарные времена (37г.), то рассказ ведётся о втором тысячелетии по летоисчислению хоббитов (Л. Х.) Оценим масштаб саги, для этого найдём 4 цифры подряд — \d{4}

```
Find result - 16 hits
Search "\d{4}" (16 hits in 1 file)
D:\! WORK\Хабр\Властелин колец.txt (16 hits)
Line 139: погибли голодной смертью, но уже и Дни Нужды (1158-1160) ко времени нашего
Line 139: погибли голодной смертью, но уже и Дни Нужды (1158-1160) ко времени нашего
Line 157: Зеленополье в 1147г. (Л.Х.), когда Бандобрас Крол наголову разгромил
Line 254: зелье в своем огороде во дни Изенгрима Второго, год примерно 1070-й согласно
Line 293: прирезаны к Хоббитании в 1164 году (Л. Х.).
Line 356: двенадцать его сотоварищей. Апрельским утром 1341 года от заселения
Line 455: Он вернулся домой на пятьдесят втором году жизни, 22 июня 1342 года
Line 457: не собрался праздновать свое стоодиннадцатилетие (год 1401). Тут и начало
Line 5446: Хоббитании 1418-й.
Line 9743: зимы 1311 года, когда на засыпанную снегом Хоббитанию напали полчища белых
Line 34816: надгробный камень и насадили рощу. Так завершилась битва у Приречья в 1419
Line 34817: году, последняя битва в Хоббитании; предыдущая была в 1147-м, в Северном
Line 35098: Был 1420 год сказочно погожий: ласково светило солнце, мягко, вовремя и
Line 35148: Так и порешили. Весною 1420 года Сэм Скрэмби женился на Розе Кроттон (в
Line 35183: Время шло; настал 1421 год. В марте Фродо опять было плохо, но он
```

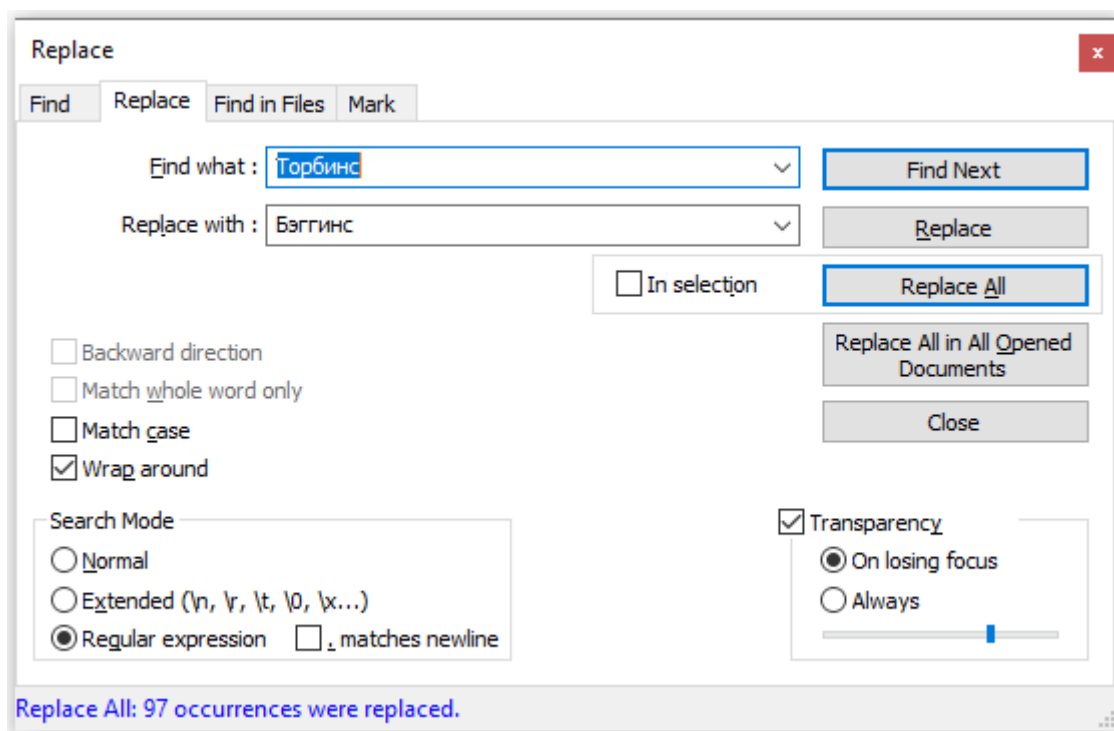
Первая дата говорит нам о начале табаководства у хоббитов в 1070 году, а Фродо уплывает от нас на Запад в 1421 — 351 год.

С остальными представлениями символов предлагаю познакомиться самостоятельно.

Замена



Как только фамилию Бильбо не переводили: Сумникс, Беббинс, в нашем случае — Торбинс, но для меня есть только один вариант — Бэггинс. Однако это не проблема: почти всегда рядом с возможностью поиска есть возможность замены. Найденное слово можно заменить по всему тексту в одно действие. Для этого воспользуемся вкладкой «Replace»:



Теперь можно читать книгу с привычным переводом. Но будьте осторожны: шаблон поиска может подойти не только желаемому слову, но и более длинному, содержащему в себе искомое. Тогда при замене «дача» на «деревня» мы неожиданно можем получить «удеревня» или «телепередеревня». Желательно предварительно провести поиск по такому же шаблону и проверить результаты. Усложняйте шаблон для более точного поиска. В случае «дача» мы можем добавить пробел до слова и пробел или знаки препинания в конце: « **дача**[.,!?:;] ».

Практика использования

Для создания, редактирования или исследования регулярных выражений советую использовать понравившийся вам онлайн-редактор. Для меня самым удобным оказался **REGEX101**:

Сегодня я уже не представляю свою работу без регулярных выражений, как и без слепого метода печати. А требовалось всего лишь начать использовать. Чего вам и желаю.

