

浅谈正则表达式在翻译 实践中的趣味应用

江伟

2022年9月7日

目录

1. 什么是正则表达式
2. 正则表达式有什么用
3. 如何学习正则表达式
4. 正则表达式资源

01什么是正则表达式

正则表达式（英语：Regular Expression，常简写为regex、regexp或RE），又称正则表示式、正则表示法、规则表达式、常规表示法，是计算机科学的一个概念。正则表达式使用单个字符串来描述、匹配一系列匹配某个句法规则的字符串。在很多文本编辑器里，正则表达式通常被用来检索、替换那些匹配某个模式的文本。

许多程序设计语言都支持利用正则表达式进行字符串操作。例如，在Perl中就内建了一个功能强大的正则表达式引擎。正则表达式这个概念最初是由Unix中的工具软件（例如sed和grep）普及开的。

--引自维基百科“[正则表达式](#)”词条文章

01 什么是正则表达式 -2

- 一串文字（字符串）
- 其中用一些**专用字符**来表述某一类或某一些字符（pattern /模式）
- 可以在文本或文件中**匹配（搜寻）**这个模式，甚至进行**替换**

.（西文半角句号）任意字符（包括不显示的空白字符），如：A.C，表示字母A和C之间有任何字符，

[]（方括号） 枚举字符，或字符范围，如：[ABC]表示ABC这三个英文大写字母中的任一个

{ }（花括号） 限定前一个字符或前一类字符的出现次数，如：[ABC]{1,3}.....

02 正则表达式有什么用

对文字搜索和替换。

很多文字编辑环境都支持正则表达式，比如：

- 纯文本编辑工具，如UltraEdit, Notepad++
- 编程环境，如VS Code
- CAT（计算机辅助翻译）工具，如Trados, memoQ
- 正则表达式编辑和测试工具，比如RegexBuddy（商业软件）和 regexlearn.com（免费网站）

03 如何学习正则表达式

任何技能的最佳学习方式，就是在使用中学习，开课！

1. 安装了[memoQ](#)的，请导入下面这个文件：

memoQ _ Translation and Localization Management
Solutions.htm_zho-CN.rtf

2. 没有安装memoQ的，请在Word 中打开上面那个文件，全选，复制，然后打开[regexlearn.com](#)这个网站，并把刚刚复制的内容粘贴到网页上的Text（文本框）中

03 如何学习正则表达式 -实操

memoQ - un_202002

Project Documents Preparation Translation Review Edit View Workflow Quick Access

Concordance memoQ Web Search Confirm Add Term Quick Add Term Add Non-Translatable Mark Text Comments Copy Cut Copy To Target Paste Copy Next Tag Sequence Inline Tags Split Join Find Find Next Replace Advanced

Project home 正则 memoQ_ Translation and Localization Management Solutions.htm

Source	Target		
1. memoQ-Translation-and-Localization-Management-Solutions	memoQ-翻译和本地化管理解决方案	0%	✗
2. memoQ-offers-flexible-translation-and-localization-management-solutions-tailored-to-enterprises,-language-service-providers,-and-translators.	memoQ为企业、语言服务提供商和翻译人员提供灵活的翻译和本地化管理解决方案。	0%	✗
3. Translation-software--memoQ--Feed	翻译软件--memoQ--饲料	0%	✗
4. Translation-software--memoQ--Comments-Feed	翻译软件--memoQ--评论反馈	0%	✗
5. Skip-to-content	跳转到内容{2}。	0%	✗
6.		0%	✗
7. account	帐户	0%	✗
8.		0%	✗
9. account	帐户	0%	✗
10. Sign-in-to-my-memoQ	登录我的备忘录Q。	0%	✗
11. English	英语	0%	✗
12. English	英语	0%	✗
13. Deutsch	德语	0%	✗
14. Deutsch	德语	0%	✗
15. Français-[CA]	Français[CA]	0%	✗
16. Français-[CA]	Français[CA]	0%	✗
17. 日本語	日本語	0%	✗
18. 日本語	日本語	0%	✗
19. Search-memoq.com...	搜索-memoq.com...	0%	✗
20. Search	搜索	0%	✗
21.		0%	✗
22. Search	搜索	0%	✗
23.		0%	✗

View pane

Active comments

Regex Learn - Playground

regexlearn.com/zh-cn/playground

学习 备忘单 BETA 游乐场 GitHub CN

正则表达式

[A-Z]\w+

global ALT+G multiline ALT+M case insensitive ALT+I

文本

```
[1]
Pre-translated
561
ffc75881-a61a-45d2-8dff-e9dadbc0705ee
[1]
[1]
Pre-translated
562
03f54664-cf65-4827-9409-c90878f11a69
[1][2]
[1][2]
Pre-translated
563
e05869db-5d75-47a2-b995-a0dbc4052226
_hj RemoteVarsFrame
_hj RemoteVarsFrame
Not started
```

报告问题

锚点

标志



组和引用

字符类

零宽断言

量词与分支

03 如何学习正则表达式 -实操1

序号	句段号	问题	正则	解读
1	2, 19	中英文之间有无空格不统一	<ul style="list-style-type: none">• [一-龠]• [a-zA-Z]	字符范围查“字符映射表”
2	3, 4	标点符号讹化: » vs "	<ul style="list-style-type: none">• [»""]	
3	5, 48	标签讹化:	<ul style="list-style-type: none">• [\{\}\[\]]	
4	5, 10	句尾标点符号问题 	<ul style="list-style-type: none">• [^\.]\$• 。\$	
5	10, 153	非译字符串误译: 	<ul style="list-style-type: none">• memoQ• (memoQ)?• (?i)(?m)^((?!memoQ).)*\r?\$	

03 如何学习正则表达式 -实操2

序号	句段号	问题	正则	解读
6		译文不一致	<ul style="list-style-type: none">• <code>(?i)translator</code>• <code>(?i)(翻译人员 翻译机 translator 译者 译员)?</code>	
7	315, 318	筛选带特殊符号的非译单元: <code>gamevil_logo</code>	<ul style="list-style-type: none">• <code>[^]*_[^]*</code>	
8	6,8,21 ...	筛选没有任何有意义文字的句段	<ul style="list-style-type: none">• <code>^[\\W]*\$</code>	
9	378, 517	查找和删除中文标点符号前后的空格	<ul style="list-style-type: none">• <code>\\s(?:[, . ? ! ; :])</code>• <code>(?<=[, . ? ! ; :])\\s</code>• <code>\\s(?:[, . ? ! ; :]) (?:<=[, . ? ! ; :])\\s</code>	
10	364-370	查找过分断句（像字幕那种）	<ul style="list-style-type: none">• <code>^[a-z]</code>	

03 如何学习正则表达式-进阶实操11-原文和译文关联搜索

需求:

- 在memoQ的QA正则规则中, 找出:
- (1) 针对某个特定术语对, 在译文中出现的个数与在原文中出现的个数不同的句对儿。
- (2) 针对所有的阿拉伯数字, 在译文中出现的次数与在原文中出现的次数不同的句子。
- 从而实现: 译文与原文数字完全对等。

03 如何学习正则表达式-进阶实操1-原文和译文关联搜索

memoQ中可以实现：在QA规则对话框的正则选项卡进行定制。比如要查找英文January及其可能缩写Jan.译作1月或一月（以563段为例）。

03 如何学习正则表达式-进阶实操1-进一步思考

²4.5 out of 5 star rating applies to all plans offered by SCAN Health Plan from 2018 to 2022 in California except SCAN Healthy at Home (HMO-SNP) and Village Health (HMO-POS-SNP) plans. Every year, Medicare evaluates plans based on a 5-star rating system.↵

²4.5 星（滿分5 星）適用於2018 年至2022 年由SCAN 健保計劃在加州提供的所有計劃，但SCAN Healthy at Home (HMO-SNP) 和Village Health (HMO-POS-SNP) 計劃除外。每年，聯邦醫療保險都會根據5 星評級系統對計劃進行評估。↵

↵

寻找中文译文中所包含的英文可能有讹误的情况？
memoQ中还不行（有待改善的地方之一）

03 如何学习正则表达式 -进阶实操1

...remove a lot of the cost of the plan by making it
five years running^{<2>}. We can be your partner
in health, and our plans include the benefits
and services designed to do just that. Learn
more at scanhealthplan.com or talk to a
SCAN representative at 1-888-371-7226
(TTY: 711)←

服務更是專為實現這一目標而設計。
瞭解更多詳情，請上網
scanhealthplan.com/zh←
或致電 1-855-470-7226 TTY: 711 與
SCAN 經紀洽談。←

03 如何学习正则表达式-进阶实操2

03 如何学习正则表达式-进阶实操2

03 如何学习正则表达式-进阶实操2

03 如何学习正则表达式-进阶实操2

03 如何学习正则表达式-进阶实操2

03 如何学习正则表达式-进阶实操2

03 如何学习正则表达式-进阶实操2

03 如何学习正则表达式-进阶实操2

04正则表达式资源

- memoQ - 带丰富正则表达式功能的 CAT (www.memoq.com，有试用版可下载)
- 维基百科的“正则表达式”词条文章：<https://zh.wikipedia.org/正则表达式>（内容相当丰富）
- RgexBuddy – 正则表达式编写和测试专业工具（<https://www.regexbuddy.com/>）
- regexlearn.com – 正则表达式编写和测试在线工具，非常适合初学者
- <https://regex101.com/> - 另一个在线工具，有各种正则规范可选，带正则解读，可以注册账号



微信: dpictures

QQ: 554352010

电子邮件: 554352010@qq.com