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

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之间有任意字符，

[ ]（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）

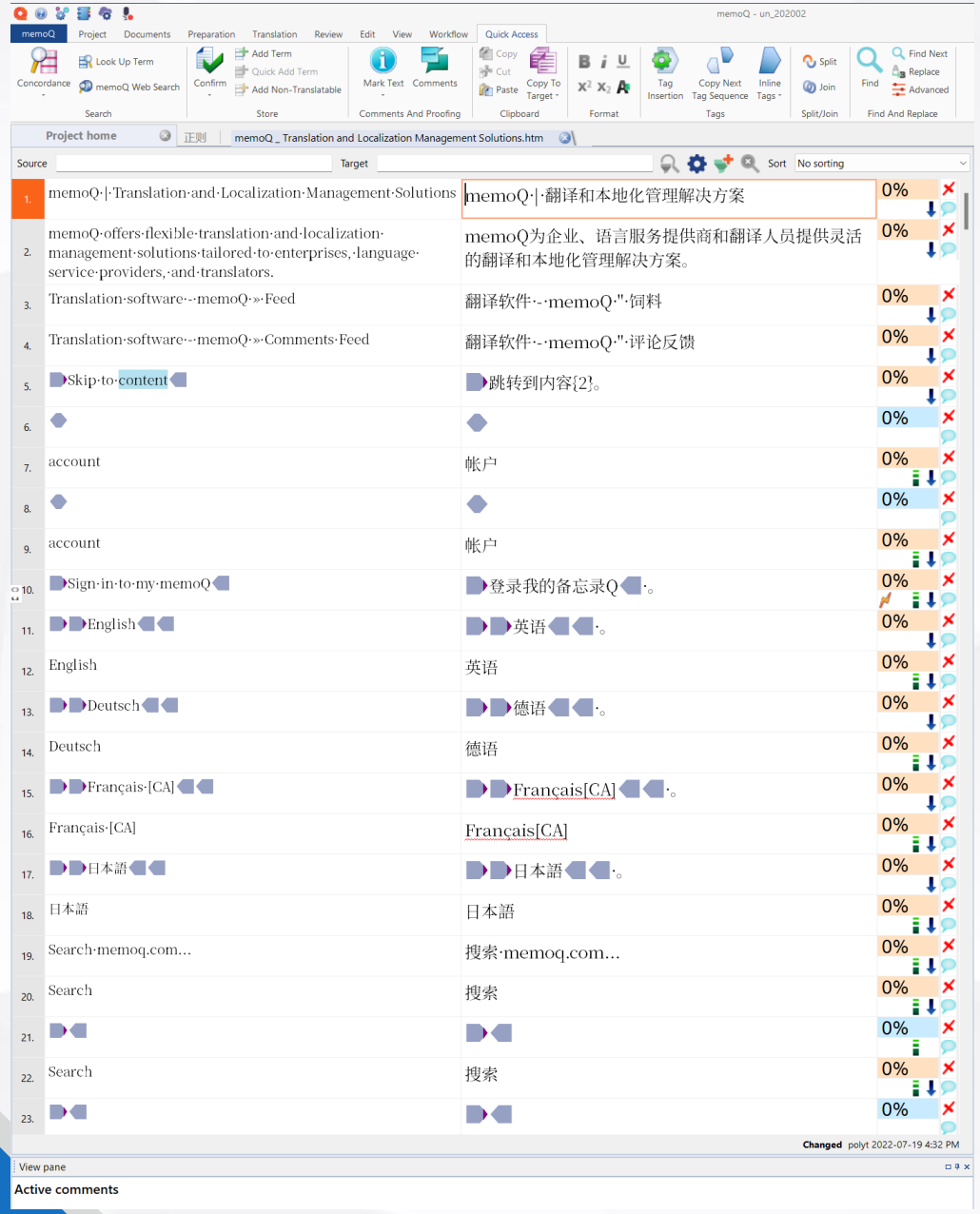
03如何学习正则表达式

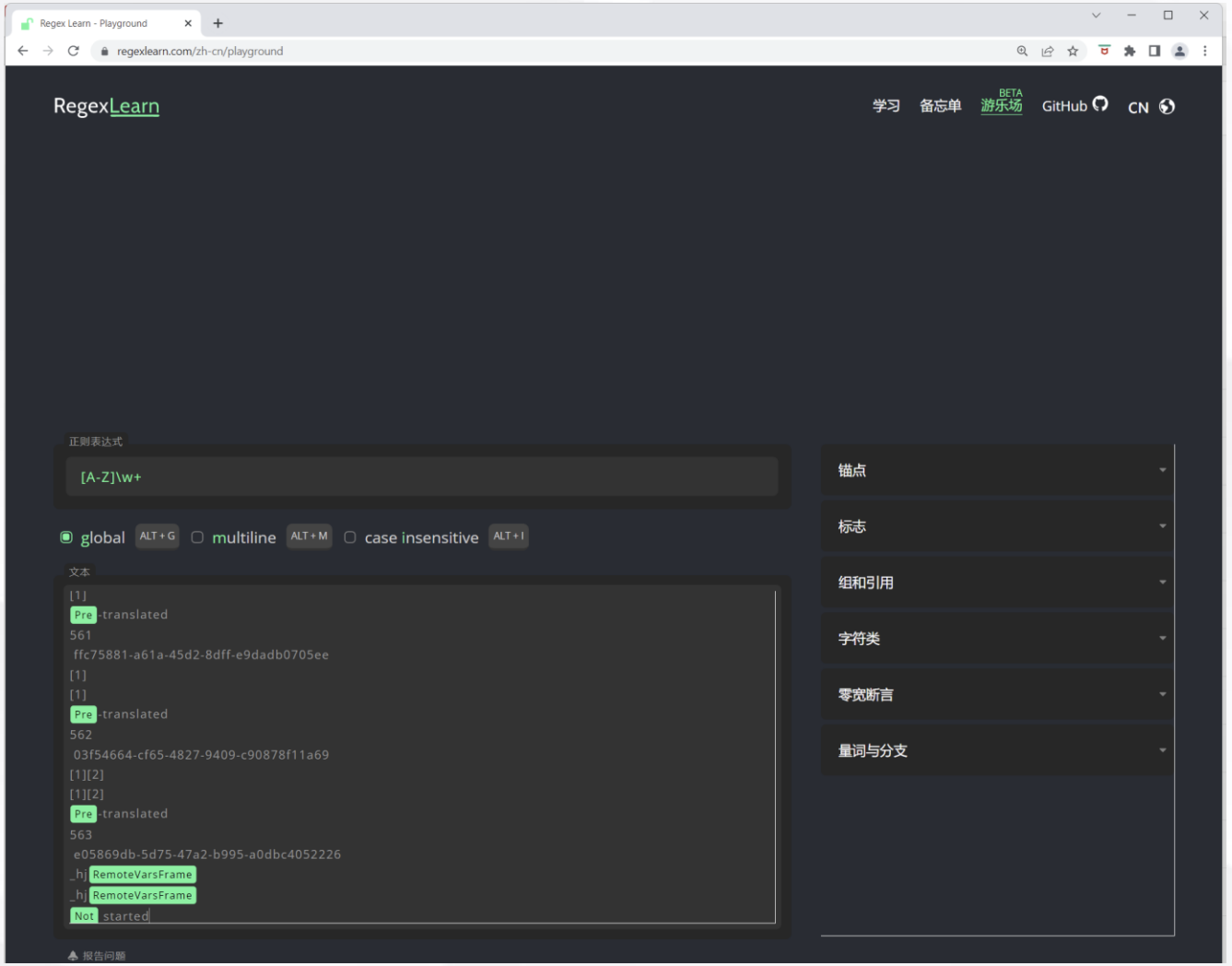
任何技能的最佳学习方式，就是在使用中学习，开课！

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





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	• [一-龥] • [a-zA-Z]	字符范围查“字符映射表”
2	3, 4	标点符号讹化：» vs "	• [»"“”]	
3	5, 48	标签讹化：	• [\{\}\[\]]	
4	5, 10	句尾标点符号问题 	• [^\.]$ • 。$	
5	10，153	非译字符串误译： 	• 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	• (?i)translator • (?i)(翻译人员|翻译机|translator|译者|译员)?	
7	315, 318	筛选带特殊符号的非译单元：gamevil_logo	• [^]*_[^]*	
8	6,8,21 …	筛选没有任何有意义文字的句段	• ^[\W]*$	
9	378, 517	查找和删除中文标点符号前后的空格	• \s(?=[，。？！；：]) • (?<=[，。？！；：])\s • \s(?=[，。？！；：])|(?<=[，。？！；：])\s	
10	364-370	查找过分断句（像字幕那种）	• ^[a-z]	

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

[2]4.5 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.	[2]4.5 星（满分 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 VillageHealth (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

achieve a 4.5 out of 5 star rating by Medicare 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