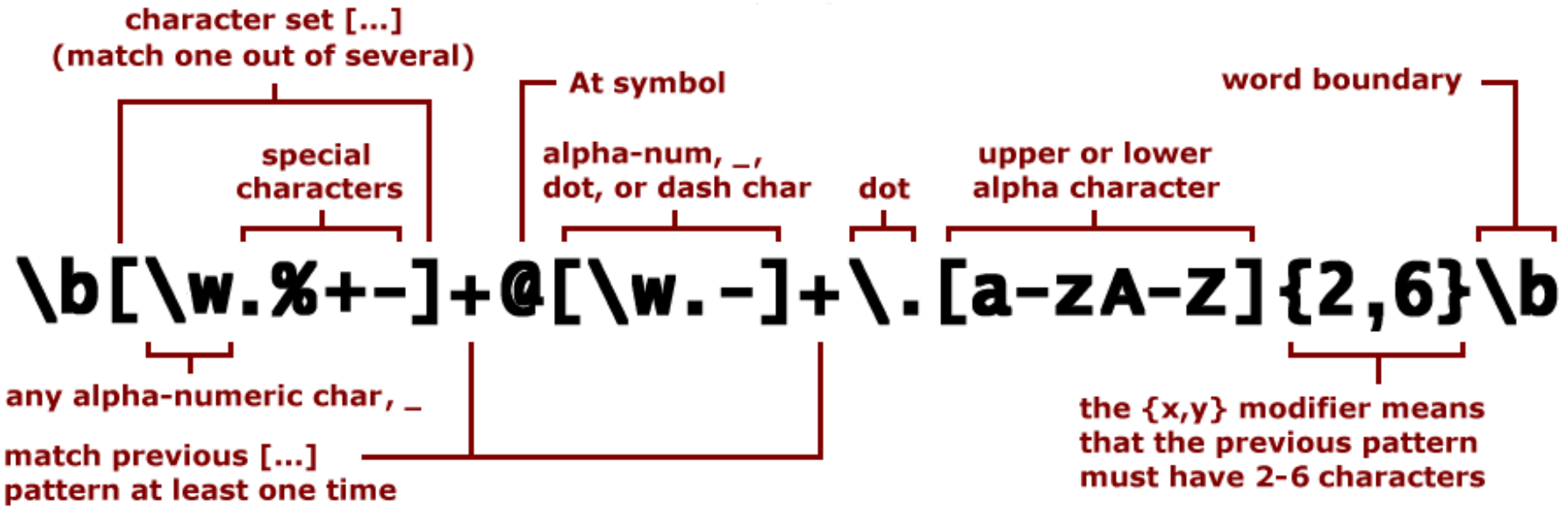
写正则表达式的正确姿势

丁海星 2016.06.11

讲真，这是什么鬼？


character set [...]
(match one out of several)
special characters
At symbol
alpha-num, _, dot, or dash char
dot
upper or lower alpha character
word boundary
\b[\w.%+-]+@[\w.-]+\.[a-zA-Z]{2,6}\b
any alpha-numeric char, _
match previous [...] pattern at least one time
the {x,y} modifier means that the previous pattern must have 2-6 characters

误解 & 理解

- 正则语法的反人类（关键字缩写，没有缩进、空格）容易让我们把注意力集中到语法本身上。
- 正则是一门语言，我们不仅要学习它的语法、词汇，更需要用它写出了漂亮的文章。
- 在自然语言处理中，正则只是不太友好的强大工具，重要的是对自然语言的思考和理解。

举个栗子

找新三板企业的负面新闻

必需品

- 新闻数据
  - 数据集尽量保证平均采样（时间，空间，目的）
  - 不要停止获取更多好数据的尝试。
- 目的明确
  - 最好有标注语聊，这样目的非常明确。
  - 大部分情况没有标注语料，沟通远比结果重要。（召回率和准确率要求不一样。有时候 60% 准确率就满意，有时候 95% 准确率也会不满意。）

争取获取

- 领域知识
  - 企业相关的实体。
  - 逻辑关系。
- 成本预算
  - 问题的本质是成本和效果之间采取平衡。
  - 预估问题的复杂度。
  - 计算机器的工作量所占比重。

自我修养

自我修养

- 自然语言理解能力
  - 比如：句法，变形，同义词 ..
- 复杂性理解和处理能力

工具

- 编辑器
- 快捷键
- 正则可视化

代码库

- 统计排序
- 相似度比较
- 常用词典
- 常用词表
- 句法分析工具
- 符号级数据清理
- 编码处理工具
- 聚类算法

调优

- 用一个曲线混一口饭吃。
- 必须想办法对分布排序，否则没有这一口饭吃。
- 只处理 Head ，正确 / 错误都要做。

关键

长期、复杂、艰苦、需要互相协助的旅程：

程序和正则分离！
知识和规则分离！

其他

正则能解决什么问题？

语用确定的特定领域内，上下文相关的字形级别语义标注问题。

注意正则没有推理。

正则效率问题

- 正则的效率非常高，但是大坑是隐秘的。
- （写个死循环一样的当然就很慢了~）

- 数据量大、复杂度在 O(n*n) 以上的问题，要采取一些手段。
- 比如： AC 自动机、双太树、 ...

- 实际问题中，思路开阔，正则上场时机把握好。

误区

- 不要想利用已有的词典解决问题。
- 不要想建立自己用的通用词表。
- 现有 NLP 工具要在适当场合，恰当使用。
- 比如：词性，实体识别，句法分析
- One more thing … 规则离不开统计，机器学习是正则表达式的基友。

谢谢大家