HW4

Missing-Semester of Your CS Education

6 June 2021

Problem

1. 学习一下这篇简短的交互式正则表达式教程.

2. 统计 words 文件 **/usr/share/dict/words** 中所有**有且仅有 1 个**元音字母 (a, e, i, o, u) 的, 且**长度不小于 5** 的单词, 并统计这些单词中含有每一个元音字母的单词个数. 含有哪个元音字母的单词出现的最多? sed 的 **y** 命令, 或者 **tr** 程序也许可以帮你解决大小写的问题.

例如单词表:

```
git
supply
shrink
glycyphyllin
rhythm
psychology
```

中, 符合条件的单词有 **supply**, **shrink**, **glycyphyllin** 三个, 这些单词中有两个 (**shrink**, **glycyphyllin**) 含有 i, 一个 (**supply**) 含有 u, 含有 i 的单词出现的最多.

3. 进行原地替换听上去很有诱惑力, 例如: **sed s/REGEX/SUBSTITUTION/ input.txt > input.txt**. 但是这并不是一个明智的做法, 为什么呢? 还是说只有 **sed** 是这样的? 查看 **man sed** 来完成这个问题.

4. 选择一个你喜欢的网站, 使用 **curl** 命令获取其 HTML 代码, 统计其每一种 HTML 标签的使用次数 (**使用正则表达式**), 这些标签使用次数的平均数、中位数和最常用的标签. 例如: `<div>` 和 `<div class="foo" id="bar">` 都是一种标签 `div`.

*5. 在网上找一个类似这个或者这个的数据集。或者从这里找一些. 使用 **curl** 获取数据集并提取其中两列数据, 如果您想要获取的是 HTML 数据, 那么 pup 可能会更有帮助; 对于 JSON 类型的数据, 可以试试 jq. 请使用一条指令来找出其中一列的最大值和最小值, 用另外一条指令计算两列之间差的总和.

补充说明:

Add your answer here!

十分重要!

1. 我们建议您使用 $LaTeX$ 来完成作业。但是如果您对此不熟悉，那么您也可以使用 $Markdown$ 来编写 (这里是一个非常小和高效的软件，称为Typora)。实际上，我们不太推荐 $Word$，但是用 $Word$ 完成作业是可以的，但是记住提交作业的格式是 **"PDF"**，而不是 ".md" 或 ".docx"

2. 我们知道在网站上有一个官方答案，但我们希望那只是Reference，而不是您的最终答案。您不仅应该编写代码并将终端窗口粘贴到您的计算机中，还应该尝试解释这些命令工作的原因。我们希望你的作业包含更多的文字来解释原因，而不仅仅是复制粘贴终端窗口。

3. 每个人都应当独立完成作业。这并不是一个团队合作项目!!!

4. 这个作业的截止日期是8 月 1 日，但是我们建议您在6 月 14 日之前完成并提交。

5. **The Chinese Homework Webpage**

6. 本次作业初始满分为 8 分，每道题目 2 分 (第五题可以选择性完成，获取 HTML 格式的数据占 1 分，获取 JSON 格式的数据占 1 分)