婴儿姓名 Python 练习

社会保障中心为我们提供了以下数据，数据中记录了美国每年新生婴儿最受欢迎的名字。

请在名为“babynames”的文件夹下找到相关文件。观察这些 html 文件所包含信息。请思考如何能够从这些 html 文件中导出文件中的数据。

第一部分

打开“babynames.py”文件，编写你的 extract_names(filename) 函数。这个函数的输入应该为诸如 baby1990.html 这样的文件名，返回一个列表（list）。列表第一个元素应该为年份，随后为名字及其对应的排序。名字按英文字母顺序排列。例如 [‘2006’，‘Aaliyah 91’，‘Abagail 895’，‘Aaron 57’，...]。修改你的 main（）使得其调用你写的 extract_names(filename) 函数并打印返回结果，（main（）中已经包含了对应变量处理语句，你可以直接在 command line 使用它）。文末有用于抓取年份和名字的正则表达式的提示，你如果遇到困难可以参考。需要注意的是，对于抓取普通网页，regular expression 不总能很好的解决问题。但是我们案例中提供的网页有统一的格式，因此可以很好的完成工作。

我们在这应该将男女名字同时处理。在有些年份，有的名字会出现不止一次，你可以选择其中的一个来作为排名依据。可选挑战：当你遇到这个情况的时候，试着选出较小的那个数。

把整个程序拆分成不同的部分，在每一个部分完成后可以通过运行并打印结果来确保代码的正确。有经验的程序员经常采用这个技巧来逐步检查，逐步推进，而不是一口气将整个程序写完。

把每一步结束之后的阶段性结果打印出来能帮助你思考如何在下一步处理这些结果。Python 非常适用于这种分步编程方式。例如，首先尝试让你的程序提取并打印出年份，然后调用 sys.exit(0)。下列是我们推荐的将程序拆分开的方案:

- 从文件提取所有的文本信息并且打印
- 定位并提取年份，将其打印
- 提取名字和排名并打印
- 将名字和排名信息存入字典（dict）中
- 创建[年份，名字 排名，...]的列表（list）并打印
- 让 main()用上你前步获得的列表

之前，我们编写的函数直接将结果输入到了默认的输出设备。如果我们能让函数返回所提取的数据，那么函数的可用性将会提高。（你在开发过程中任然可以在函数内逐步打印你的阶段性结果）

让你的 main()调用你编写的 extract_names()。当你在 command line 运行你的 python 程序时，对应每一个 command line 变量都应该输出一个如下格式的文本总结。通过这个聪明的使用 join 的方法，你可以将之前要求的 list 变为我们希望的文本总结的格式：`text = '\n'.join(mylist) + '\n'`。

每一个文本总结应有以下格式

```
2006
Aaliyah 91
Aaron 57
Abagail 895
Abbey 695
Abbie 650
...
```

Part B

下一步，我们希望在 command line 运行该程序的时候，如果命令中有 --summaryfile 这样的 flag，程序对每一个“foo.html”会将对应的文本总结写入到新建的“foo.html.summary”文件中。

当上述的--summaryfile 功能实现后。通过在 command line 运行程序，并在程序中运用"./babynames.py --summaryfile baby*.html"代码，可以一次性生成所有的总结文本文件。（shell 会寻找所有符合"baby*.html"格式的文件名，并将所有文件名传入 sys.argv 中）

当所有数据整理到对应的总结文本文件中后，你可以通过以下语句来观察这些名字排名随时间变化的趋势。

```
$ grep 'Trinity ' *.summary

$ grep 'Nick ' *.summary

$ grep 'Miguel ' *.summary

$ grep 'Emily ' *.summary
```

正则表达式提示-- 年份: r'Popularity\sin\s(\d\d\d\d)' 名字: r'<td>(\d+)</td><td>(\w+)</td>\<td>(\w+)</td>'