本次文计大作业为利用 python 对《红楼梦》的 120 回内容进行基本的统计分析并展示分析结果，以下为具体内容。

一、按要求完成以下内容：

（1）利用 python 程序 将 红楼梦.txt 文件分割为三个新的 txt 文件。

红楼梦.txt 共包含 120 回 的内容，按照（1-40, 41-80, 81-120）的切分方式， 使得新的每个 txt 文件中包含 40 回目的内容，分别命名为 *part1.txt*、*part2.txt*、 *part3.txt*。

（在完成此小题后，后面三道小题和第二道大题的统计结果应使用 python 写入到一个 result.txt 的文件中。）

（2）统计以下五个副词在三部分文档中的各自的出现次数。

副词：越发、难道、可巧、不曾、原是

对这五个副词在三个文档中的出现次数都要进行记录，最后应该得到 15 个记录。

（3）统计以下虚词的在三部分文档中的出现次数分别占三部分文档所有文字的百分比。

虚词：或、亦、方、即、皆、仍、故、尚、呀、吗、咧、罢、么、呢、让、向、往、就、但、越、再、更、很、偏。

对每一个文档，都统计这些虚词的总共出现次数，计算它们的总次数占所有字数的百分比，即频率。每个文档得出一个频率，最后应得到三个频率。

（4）统计这三部分文档的平均段落长度和平均句子长度。

二、学习使用 python 工具包。

统计三部分文档中的出现次数最多的前 100 个二元组。二元组是指分词后的句子中，连续出现的两个词语。如这句话 “出现次数最多的二元组” ，分词后应为“出现 次数 最多 的 二元组”，它的二元组有“出现-次数”、“次数-最多”、“最多-的”和“的-二元组”。

在统计二元组出现次数之前，首先需对三个文档内容进行分词，而后才可统计出现次数最高的 100 个二元组。分词工具推荐使用北京大学中文分词工具包 pkuseg-python。使用方式见：

https://github.com/lancopku/pkuseg-python

对三部分文档，均统计出现次数最多的前 100 个二元组，并将统计结果写入 result.txt 文件。

三、自由发挥，参考第一道大题，选择两条可以体现文本特征的统计指标，尽量不要与第一、二大题的统计内容相似。

四、将前两道大题的统计结果制作一份 PPT，具体要求如下：

（1）PPT 不少于 10 页；

（2）自定义母版并在 PPT 中使用；

（3）使用 PPT 中的项目符号及编号列表；

（4）对某项统计结果应用图表中的柱状图进行可视化表示；
（5）对某项统计结果应用表格进行展示；
（6）整体美观漂亮。

最终大作业提交内容应包含四部分：
1. 处理文本的 python 文件（请将所有代码内容写在一个.py 文件中）；
2. 程序运行所需要的数据文件，如 *红楼梦.txt 文件* ；
3. 记录统计结果的 result.txt 文件；
4. 展示统计结果的 PPT。

请将这四部分存于一个文件夹并将文件夹命名为 学号_姓名_文计大作业 。将此文件夹压缩为 .rar 文件，提交此压缩文件。