《大数据计算机基础》2022 秋季学期作业 2

本次作业数据链接：

https://disk.pku.edu.cn:443/link/76AC064E68781B7C767EF3D985CD8A5D

数据有效期限截止：2022-10-23 23:59

本次作业要求如下：

1、本次作业数据规模较大，请处理时先用小规模数据调试程序，然后用大数据生成最终结果，sample.txt 是样例数据。观察样例数据，其每条数据的格式为 JSON，只处理 content 对应的内容；

2、最终利用 new2016zh.zip 中的数据，形成每个汉字后跟随高频字，最终形成类似联想输入法。即在程序中每输入完一个汉字后（不是从拼音推断可能的汉字，而是选择好某个汉字后），显示出其跟随的高频汉字，显示时按字频排序。由于没有学习 GUI 程序设计，可以多次多行显示输入即可；

3、利用 jieba 或其他分词扩展库（采用精准模式），形成每个词条（很多词语），并统计每个词语后跟随的高频汉字，即输入两个或多个汉字后，查找该汉字组合是否词语，然后显示该词语对应的高频汉字；如果一个组合既可以少字为词语，也可以多字为词语，按多字执行，如“繁荣昌盛”，“繁荣”和”昌盛”按“繁荣昌盛”一个词语执行；

4、利用上述数据，统计覆盖 80%文本的最少单字是哪些，并将这些汉字及其字频、累计字频按顺序输出到文件 HF_SingleHZ.txt 之中；

5、利用上述数据，统计词语出现频次，将排名最高的 10000 个词语输出到 HF_Word.txt 之中，包含词条、频次；

6、利用 N_Gram 和其他统计知识（不利用其他分词扩展库）形成词条，并与 jieba 进行比较，查找差异原因，并持续优化。

7、上述第 1-5 题为必须，第 6 题为加分选做【将挑选优秀作业展示】。形成数据的程序和利用数据输入显示的程序应分开，不宜合并在一起，以提升运行显示速度。程序命名自定，压缩后在教学网上提交。压缩包中需包含 readme.txt 文件，说明压缩包中每个文件的用途。