20201103 作业

说明：

以下各题依赖于文件“50 万人名.txt”，其编码格式为 utf-8.

1、 统计每个姓出现的次数，不考虑复姓问题(复姓视为单姓，第一个字为姓)，按降序排列。第一列为姓的汉字，第二列为对应数量，列与列之间用英文逗号分隔；

2、 统计名(姓之外的字)中每个字出现次数，相当于统计名称常用字，不考虑复姓问题(视为单姓)，按降序排列；第一列为汉字，第二列为对应数量，列与列之间用\t 分隔，仅输出前 200 个汉字；

3、 统计名(第一个字为姓)中有两字重复出现的次数，相当于”丽丽”出现的次数，”萌萌”出现的次数。如果是四个字，如：欧阳阳春，阳阳算重复；欧阳丽丽，丽丽算；花花美，花花不算；欧阳春阳，阳阳二字不算；第一列为重复的两个字，第二列为对应的次数，列与列之间用\t 间隔。按降序排列，如数量超过 200，按 200 对计算；

4、 统计 4 字姓名中前两字出现频次，并输出两字排名前 80%的前两字（即前两字累计出现次数达到 80%，假定所有 4 字姓名前两字的综合为 1000 频次，欧阳 100 次，西门 40 次，则欧阳站 10%，西门占 4%，累计 14%，一直累计到 80%为止），以及对应次数。输出时，第一列为前两字，第二列为出现次数，列与列之间用“→”间隔。

提交内容包括：每个问题的程序源代码(每个问题对应 1 个程序，程序命名为:学号_编号_20201103.py)；产生的数据(命名规则如程序，扩展名为 txt)，程序源码注释丰富，变量命名规范。将上述内容压缩到“学号_20201103.zip”中；

计分规则：

- 1 个正确，得 40 分；
- 2 个正确，得 70 分
- 3 个正确，得 90 分
- 4 个正确，得 100 分