文计第二次作业

吴明轩 1300015852

1-2. 使用作业中所提供的代码读入txt文件

3. 删除第一行的数据，用split函数根据逗号将后面的部分拆分，注意删去最后的换行符号，并将出现次数、频率和百分比等数据组合成tuple类型

4. 使用filter函数就可以筛选长度大于等于2的条目

5-6. 遍历整个字典进行统计，需要注意的是第六题还需要根据频数进行统计

7. 题目中似乎只要求指出每个二元组出现在哪些词语中而不要求统计，因此我们考虑构建一个dict，其key为所有二元组，而value为所有包含该二元组的词语组成的字符串。做法和第5-6题类型，只是此时的value不是加上对应的频数而是加上包含该字符串的原始词语。

8. 结果参见frequency.xls，手工删除后的结果已经合并到该xls文件中。从中我们可以看到，常见词语包括两种类型，一是主要角色的名字，比如“宝玉”，“少平”等，二是一些常用的词语，比如“一个”，“我们”等。两本书的区别在于红楼梦中主角色彩很浓厚，前10中有四个是角色的名字，宝玉凤姐贾母和黛玉；而平凡的世界中前10只有两个是角色的名字。此外，可以明显感觉到两本书的用语随时代的变化也很不一样：红楼梦中的“如今”在平凡的世界中已经变成了“现在”，红楼梦中“说道”在平凡的世界中没有对应词语，应该是因为现在的小说通常只会用“说”一个单字来开始对话，而这种单字没有被程序识别到。因此，我们能感觉到白话文与古代（哪怕是清代）在用词上还是很不一样的。

需要注意的是我在筛选100词时标准比较松，基本上除非是“了这”这种明显不是二元组的词我都选择了保留。我没有读过平凡的世界，因此在有些词是否是人名地名的判断上可能有错误，不过我想这些小细节在统计上来看应该没有造成太大的误差。