本次作业使用《笑傲江湖》作为研究对象。其文件名为笑傲江湖.txt。本次使用的库有：

Jieba：“结巴”中文分词

Numpy：Numerical Python

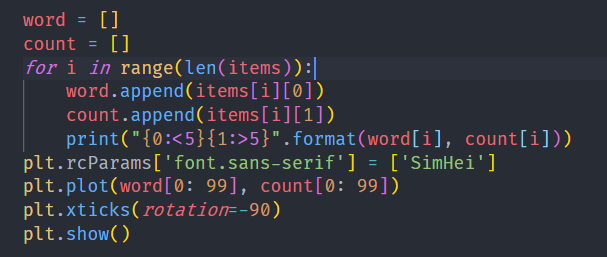
Matplotlib：a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms

Re：正则表达式

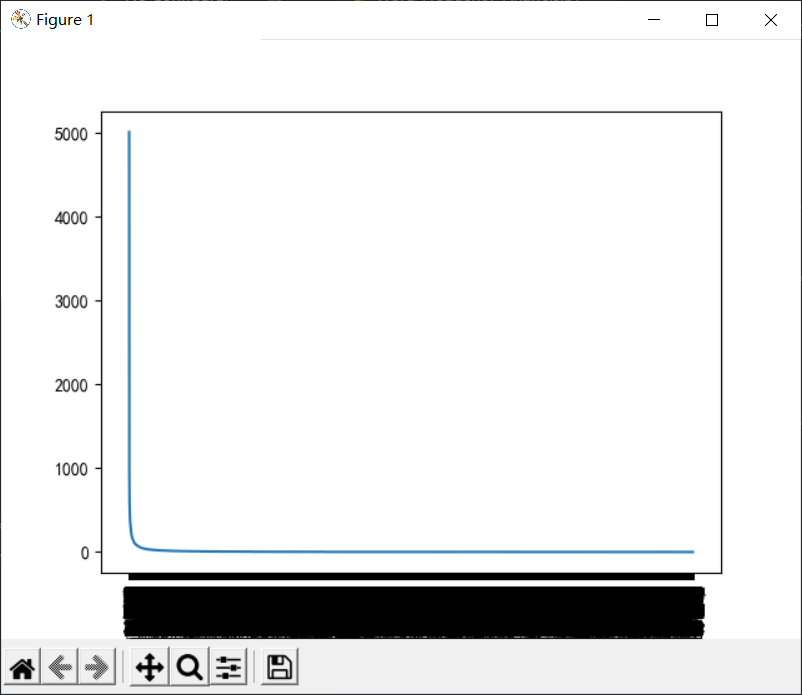
首先读取文件笑傲江湖.txt



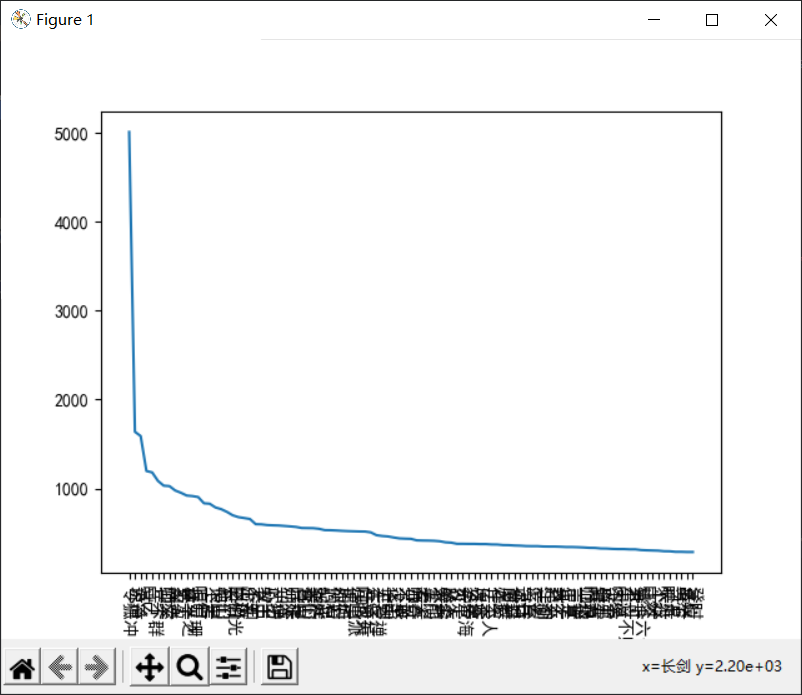
使用jieba的精确模式分词，并排除掉单个的连接词和各种单个的标点符号，将每个词以及它的出现次数用字典counts存起来，顺便按出现次数从大到小排序



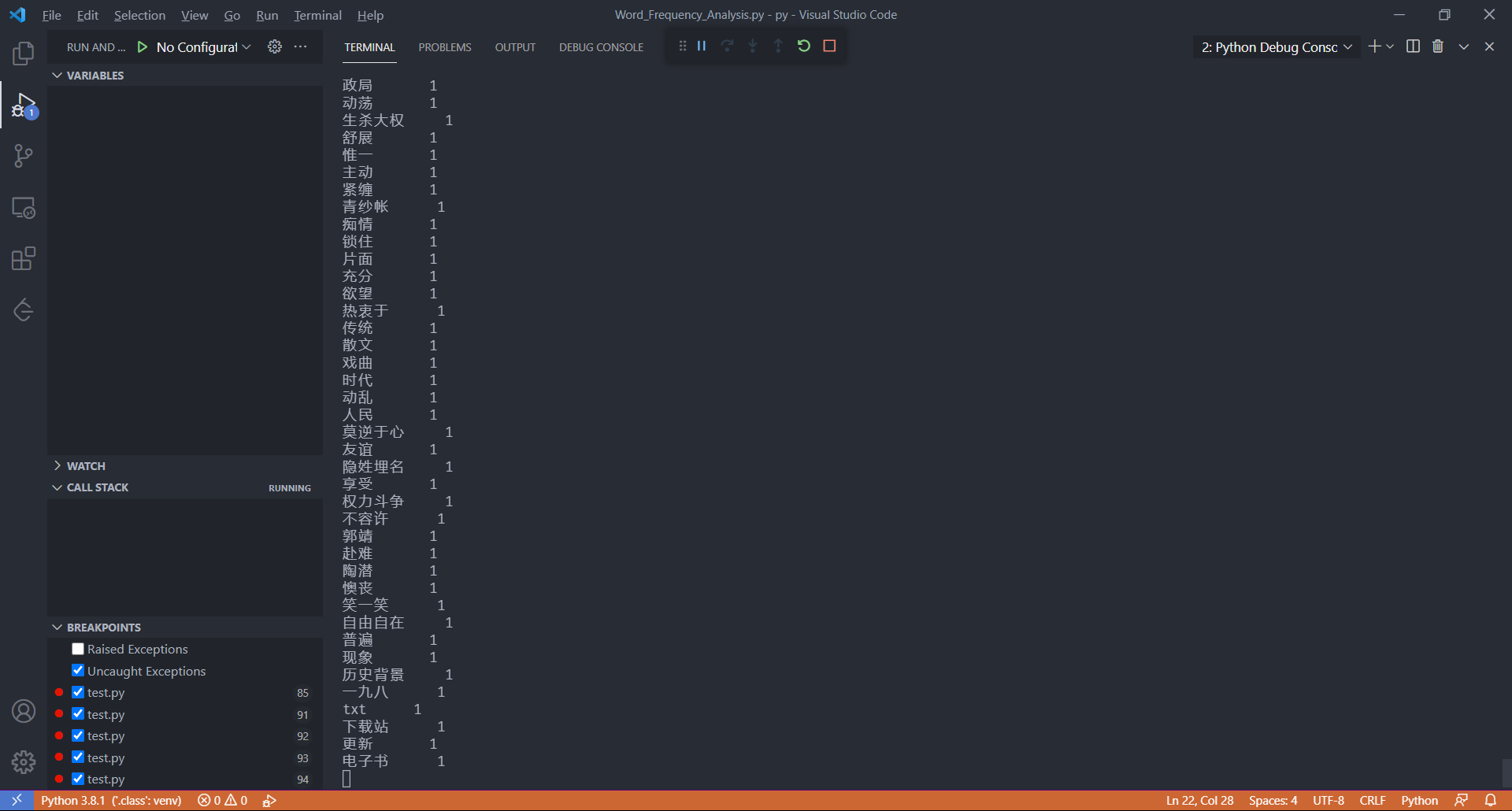
然后分别以字典的key和value作为横纵坐标画图。这里由于低频词太多导致曲线过于陡峭（见下图）所以画图只使用前100个key。



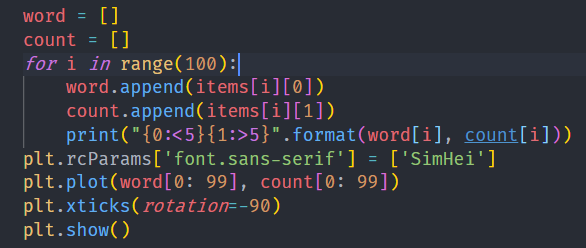
下图为只选用前100个key的曲线

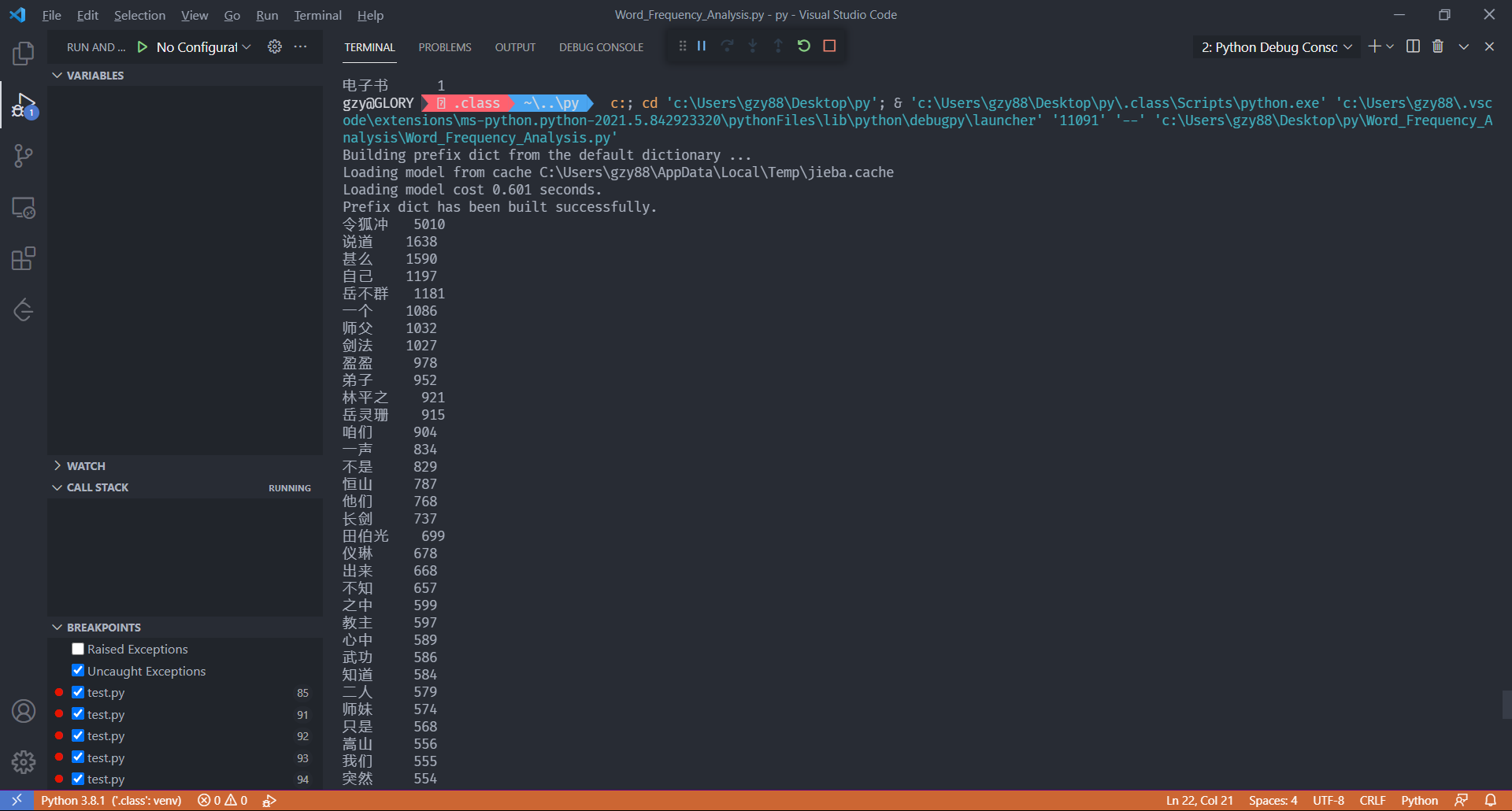


下面输出文字版结果（由于我的terminal设置原因只能显示最后200个的结果）

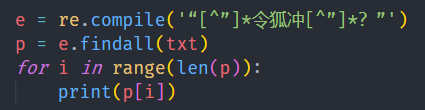


下面把代码稍稍改一下使其只输出前100个key的结果

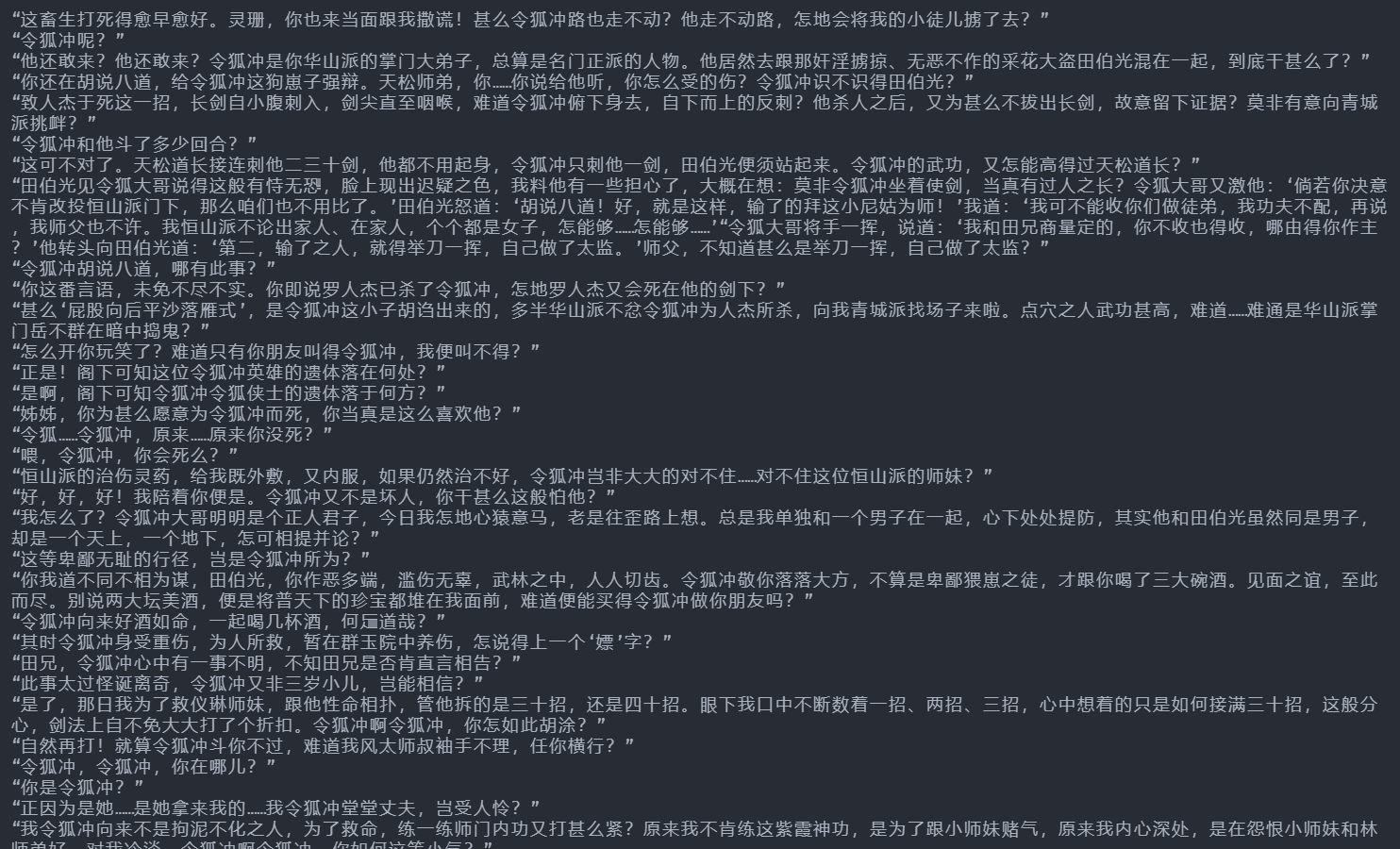




最后，我想知道有关令狐冲的一些问题，因此使用正则表达式，找出了文本中的所有含有令狐冲的对话中的问句



只展示部分结果



本次实验用到的结巴分词库，在GitHub上有十分详细的中英文文档，方便我们理解其算法以及使用方式。读了他我深受启发，同时也不得不感叹他人算法的高明。我们现在小手一动，一行代码就能解决的事可能是他人在背后付出了无数时间的开发以及维护的结果。因此我们在使用这些库的时候应该不忘去给予一些力所能及的反馈，比如在GitHub上点个star或者看看有没有自己能够解决的issue。多跟其他人讨论、交流才能更好地开拓视野，提升技能。