

唐诗宋词分析

17042201002 李劼翔

17042201003 黄芷玲

17042201004 樊甜甜

1. 小组成员及分工

17042201002 李劼翔 进行数据分析和建立模型

17042201003 黄芷玲 对数据进行爬取和整理、撰写文稿

17042201004 樊甜甜 对结果进行整理和绘图、撰写文稿

1. 作品背景和意义

通过对唐诗宋词的对比，了解诗词文化的精妙之处，展示我们对传统文化的传承与热爱。并通过机器学习，能够巧妙地对出诗词的下一句，让我们从中感受到诗词文化的优雅趣味。

1. 作品使用了哪些第三方库，简单介绍该第三方库

requests 网络爬虫

bs4 在网络爬虫的过程中，生成文档树以及抓取标签，可以代替正则表达式加快爬取进程。

wx, 设计GUI界面

genism, 建立词向量，对词和词的关系进行分析

jieba, 中文分词

wordcloud, 词云，对文本的整体提取主要词汇并以图片显示

numpy, 对矩阵进行操作，本程序中主要配合matplotlib

matplotlib， 绘图

1. 作品功能及运行过程（有截图最好）

程序目录：

1.**爬虫.py :**

从古诗词网（https://www.gushiwen.org）上爬取全唐诗和全宋词，并保存到./doc文件夹中，分别为全唐诗.txt 和全宋词.txt

2. **CleanDoc.py**

将文本中包含的作者名字、标题和词牌名等可能会影响分析结果的东西剔除。将结果保存到./doc文件夹中。分别为 tang.txt, song.txt

3.**单字频统计.py**

将诗中所有的单字进行统计对比。包括四个季节中出现的次数以及高频字。结果均为图片保存到./img文件夹中

4. **WordDivideClean.py**

将文本利用jieba进行分词，进行清理标点符号以及非汉字字符以后，保留两个字及以上的词以备后续工作。将结果文本保存到./doc 文件夹，分别是JieTang.txt 和 JieSong.txt。

（由于文本量巨大，分离和清洗至少需要半个小时，为节省时间特意保留成文本。读取文本的时间远小于分词清理的时间）

5. **词频统计.py**

对已经分离好的文本，进行分析。得出单独出现的高频词，和唐诗宋词里共同出现的高频词。

6. cloud.py

利用 wordcloud 对唐诗和宋词绘制图片。

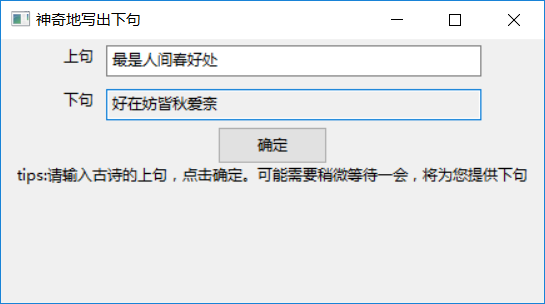
7. **古诗接下句.py**

利用word2vec 尝试再给出古诗上句的情况下给出下句。另外，还用wxpython设计了图形界面。

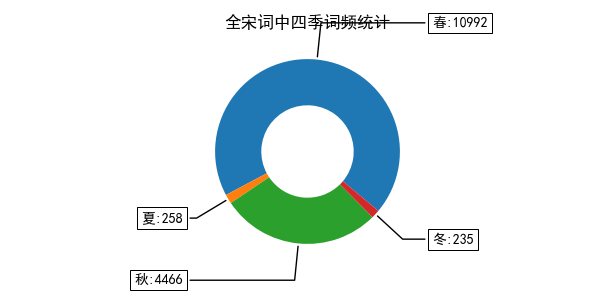
结果及过程：

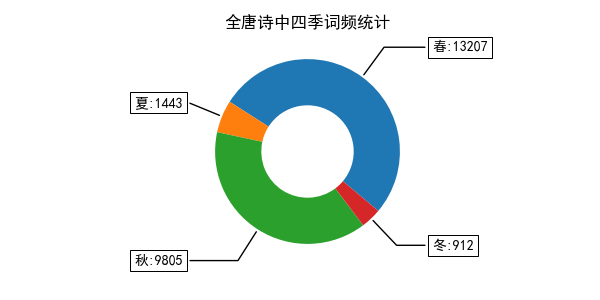
1. 根据上句写下句

利用wxpython 设计出框架，利用全唐诗建模，再根据输入的每个字进行匹配得出下句



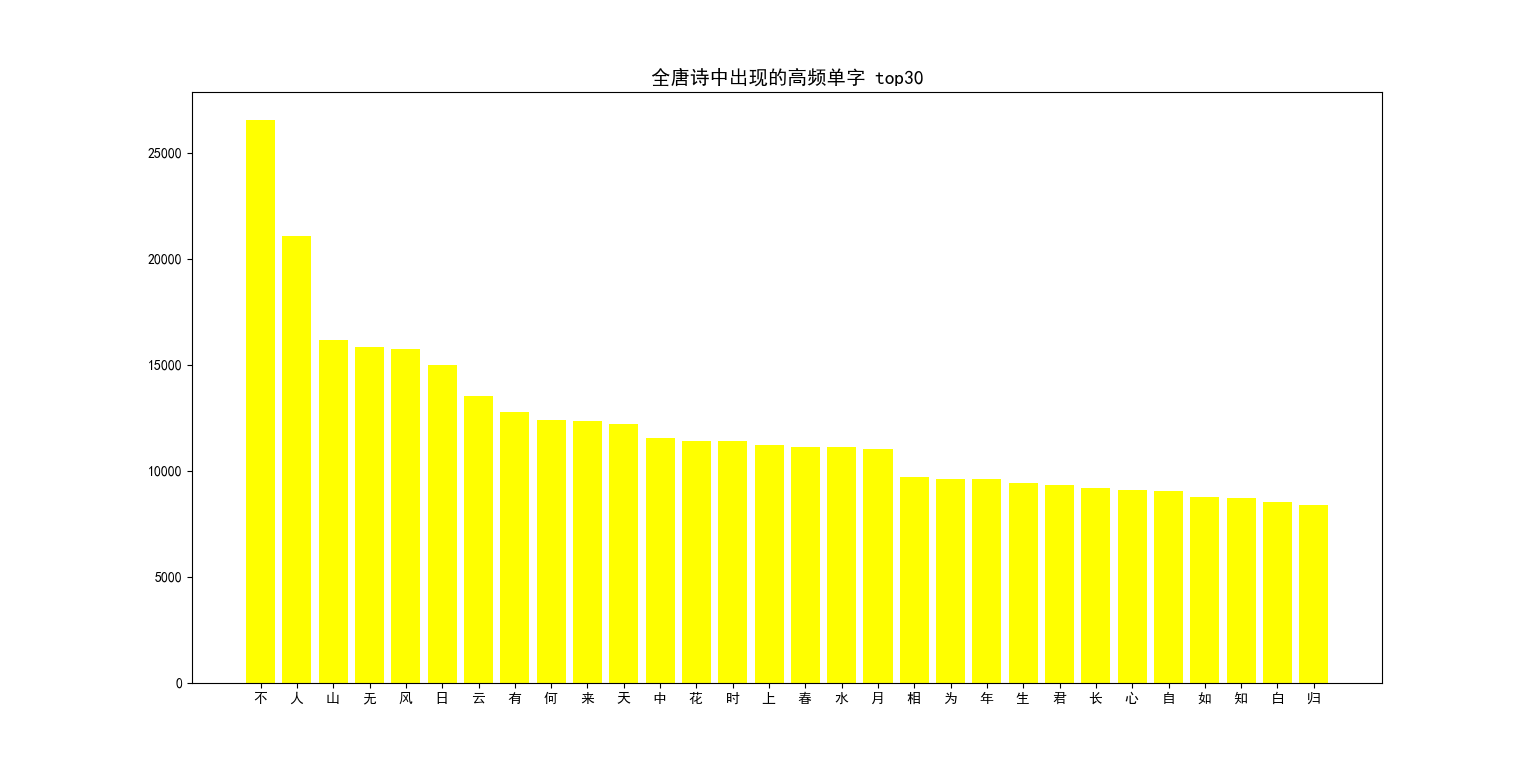
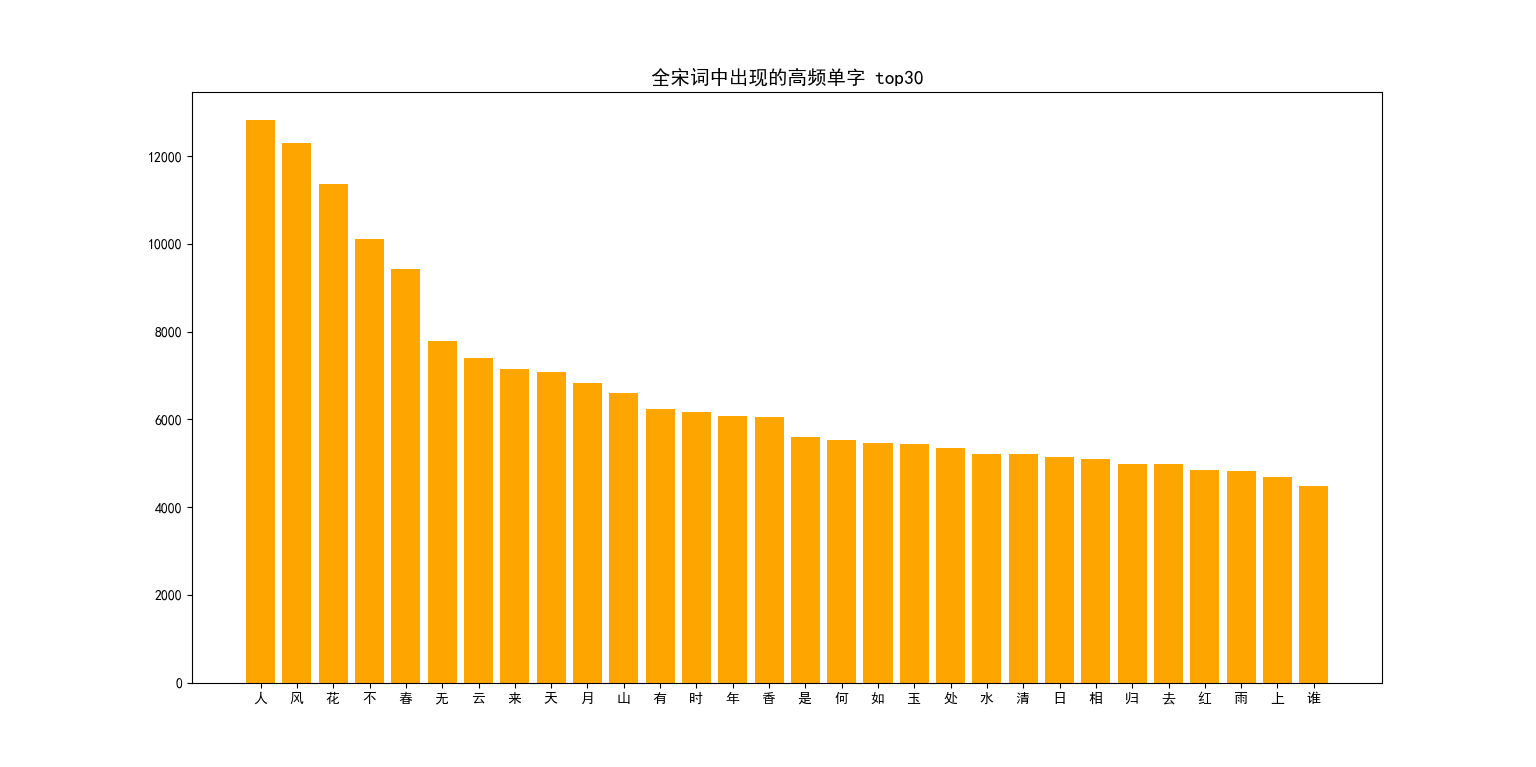
2. 唐诗宋词中出现的四季对比



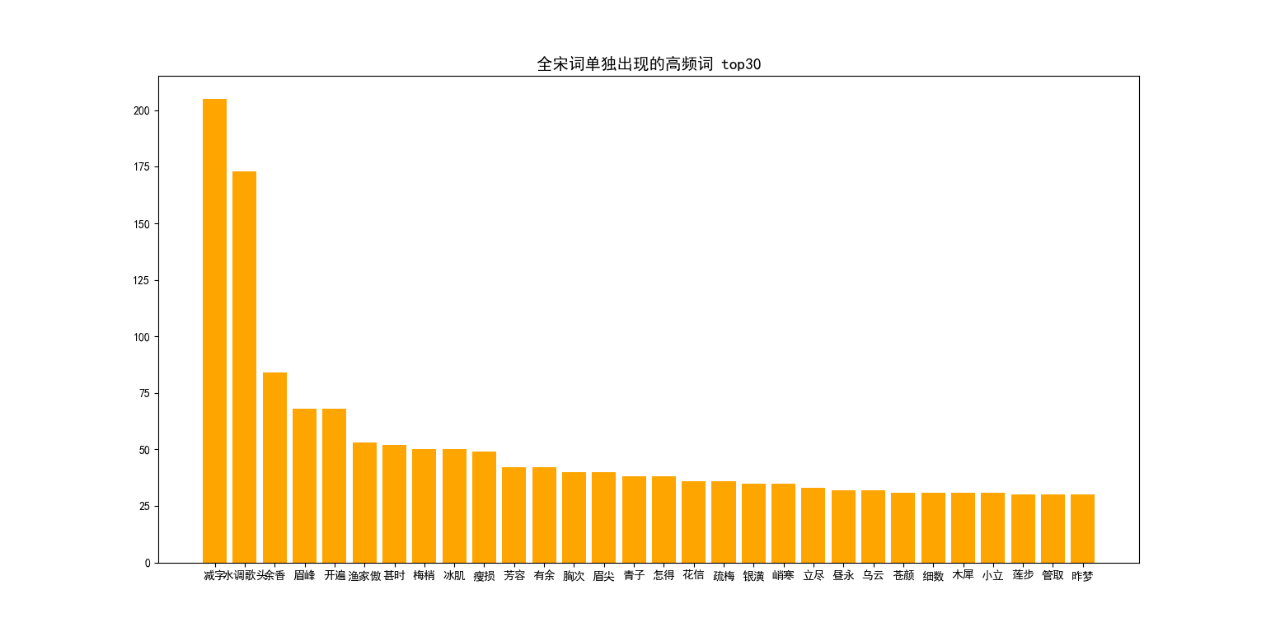
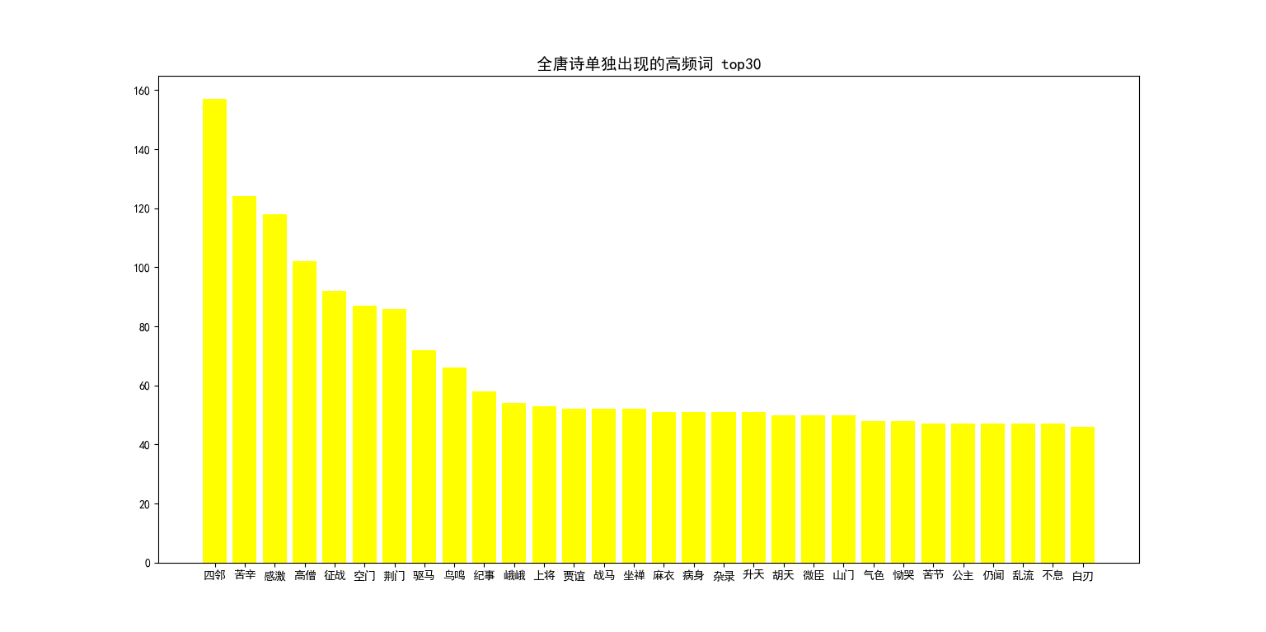


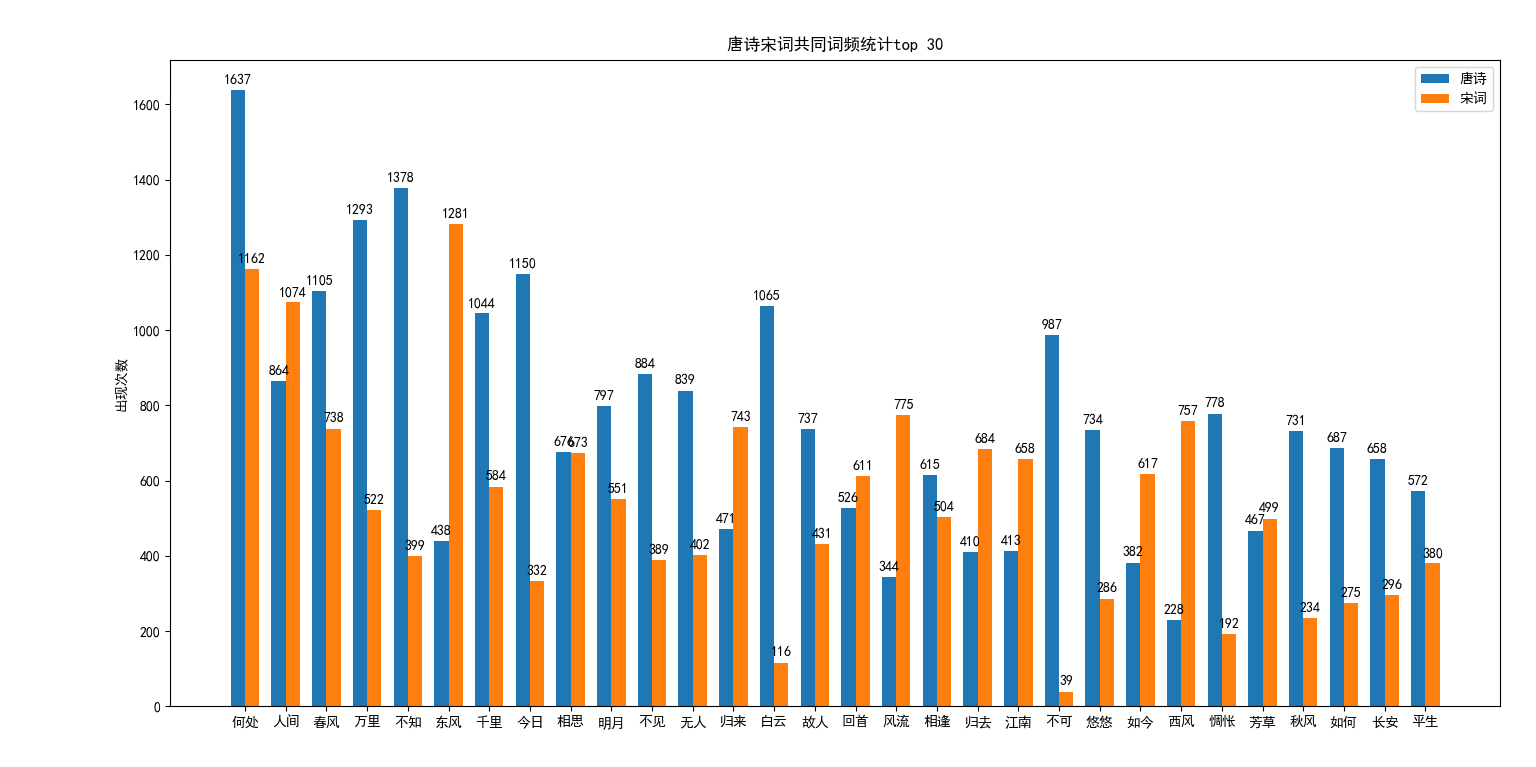
春秋的比例都相当的高，但是显然宋代词人更加偏爱春天

3. 在唐诗宋词中分别出现频率最高的单词字



4. 词频分析





5. 词云

宋词：



唐诗：



1. 作品的优点，特点

通过可视化清楚展示唐诗宋词中的高频词汇，对诗人的偏爱与个性也能窥见一斑，也能帮助我们更加全面概括地统筹了解一段历史甚至一个地区的风土人情、一个朝代的风起云涌，通过分析学习，能够实现简单的诗句匹配，对仗工整，语句优美。

1. 心得（在作品完成过程中，遇到了哪些困难，如何克服的，有什么收获）

尝试用LSTM训练诗词写藏头诗，但由于操作不当、训练时间过长（超过24小时），导致失败；后改用词向量进行简单的上下句对仗，最终能够工整地对出下一句诗。

爬虫爬取数据时得到的文本有大量数据需要清洗，才能进行后续操作，通过更加熟练地应用正则表达式，得到清洗干净的数据。

但是总的来说收获还是很大，在team中跟大家一起独立完整地完成一份设计，从前期的确定方向，到后面的搜集资料，再到正式的开始设计与编程，每一步都是自己踏踏实实走过来，感觉成就感很大，这种经历让我们收获颇丰。

1. 对本课程的意见和建议

老师上课过程中讲解认真，能够让同学们主动动手实践完成课堂任务，并能够适当拓展知识面，让同学们不拘泥于课堂知识，使大家在这堂课中受益匪浅。

课程中知识点较为分散，希望能够对某一内容更加深入，使大家能够更加了解与应用，在课程中让大家更加充分地进行实践。

老师的上课风格我们也很喜欢，自由、开放，让我发自内心的喜欢这门课，这门课虽然是选修，但是我对这门课的态度却比专业课还要认真，因为它不仅实用专业，还很生动有趣，刘老师上课的时候说过，计算机专业的学习主要靠课外的练习，仅仅靠着课上的东西是没有办法满足自己的，所以对python的学习我们会在以后的生活中继续下去的，python真的很好玩！