宋词生成系统工程报告

一、实验背景

宋词生成系统，基于词法句法，考虑其平仄，诗歌的质量主要体现在句法和语义上，一方面,诗歌作为自然语言的一种文学形式的表达,有着严格的句法要求。另一方面,诗歌的语义则包括了主题与词句的连贯、风格的统一、情感与意境的传达等等。语义层次最关键的问题是如何使产生的诗句看起来更有意义,使句与句之间更有连贯性,而不是毫无关联的词汇或句子的堆砌。

经过网上资料的查询，发现了绝大多数之前的方法是采用了模板的方式，根据一系列限制，如押韵，重音，词频等，结合基于语料库和字典式资源，来生成诗歌。例如，2008 年 Tosa 等人和 2009 年 Wu 等人的俳句生成器模型就是根据从语料库和额外的词汇资源提取出的规则来扩大用户的查询需求。2009 年的 Netzer 等人通过已经建立的联想词汇库生成俳句等等。

生成诗歌研究的另一种方法是采用遗传算法，例如 2003 年 Manurung 和 2010 年 Zhou 等人的研究。在 2012 年 Manurung 等人的研究中发现，所有机器产生的诗歌必须满足语法性，语义性，诗性的限制。他们的模型会产生几个候选诗歌，然后使用随机搜索的方法找出满足上述性质的诗歌。

第三种方法从统计机器翻译和相关的文本生成应用中得到启发。2010 年 Greene 等人从一个他们后来使用的诗歌文本语料库以及加权有限状态转换器中推断出了诗的韵律。2008 年 Jiang 和 Zhou 通过基于 SMT 方法的短语模型生成了两句诗。2012 年的 He 等人将 Jiang 和 Zhou 和方法延伸到了四行诗。

本次实验要求将采取最简单的方式，统计 ci.txt 文件中宋词的单字词，双字词，三字词出现的频率，统计几个词牌名下词的模板，根据模板随机挑选统计结果中出现频率较高的几个字词生成新的宋词。这是一种最简单的宋词生成方式，在本次实验中不考虑宋词的语义平仄韵律，只求生成一篇读起来像宋词的宋词生成系统。

二、系统设计

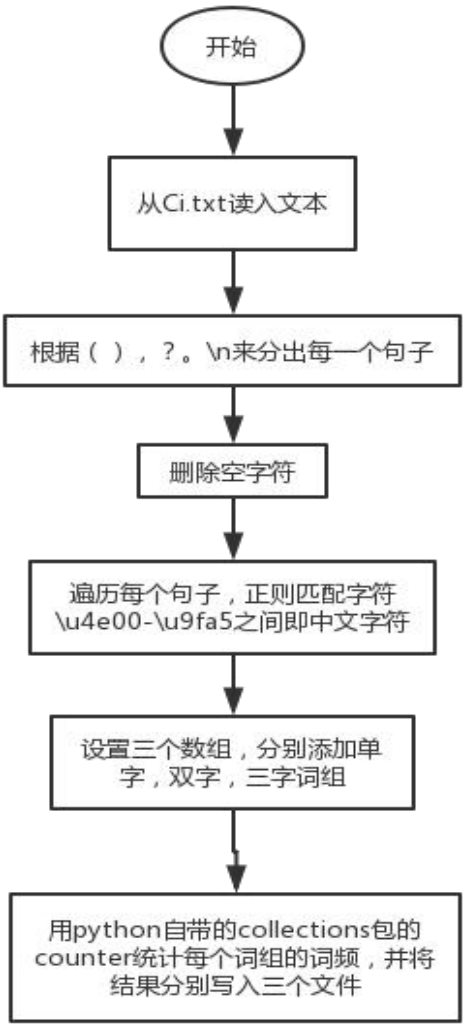
开始
从Ci.txt读入文本
根据（ ），？。\n来分出每一个句子
删除空字符
遍历每个句子，正则匹配字符\u4e00-\u9fa5之间即中文字符
设置三个数组，分别添加单字，双字，三字词组
用python自带的collections包的counter统计每个词组的词频，并将结果分别写入三个文件

图 1

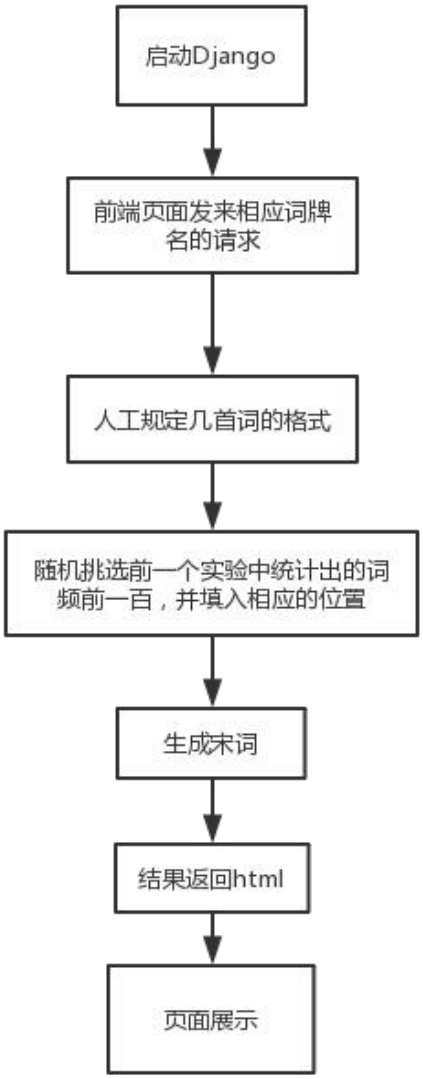

图 2

整个宋词生成系统分成了两个部分，统计词频（图 1）和随机生成宋词（图 2），最后前端 html 页面通过 django 显示最终的宋词生成结果。

三、系统演示与分析

词频统计结果如下：

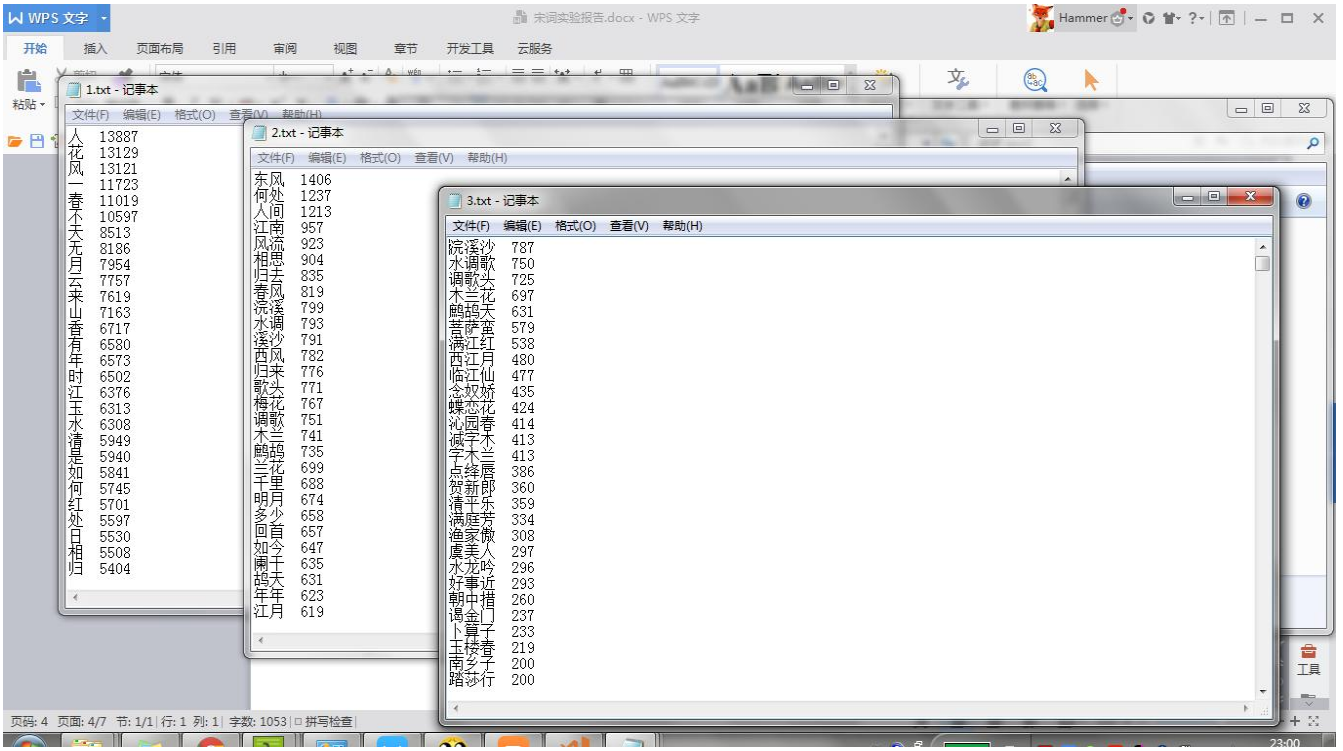

图 3

系统最后通过 html 页面显示，通过 django，将后台生成的相对应的词牌名的词的结果返回给前端页面，并进行展示，通过左边导航栏选择不同的词牌名随机生成不同的词。

图 4


127.0.0.1:8000/show#
宋词生成系统
请选择词牌名
酒泉子
苏幕遮
西江月
甘草子
生查子
酒泉子
长忆南国，
为相有来试新妆。
身世击楫门深闭。
何妨沈香亭。
江潮吴宫过重阳。
谁人杨花人间无。
成花杯盘金尊倒。
风玉早梅芳

图 5


127.0.0.1:8000/show#
宋词生成系统
请选择词牌名
酒泉子
苏幕遮
西江月
甘草子
生查子
酒泉子
长忆王孙，
旁人新样日相逢。
别时到长记当年。
为行奉劳歌。
桥南香疏宫仙子。
向此经纶倚危阑。
老来金泥玉玲珑。
试看说相思

图 6

图 7

图 3 是系统的首页，图 4，图 5 是酒泉子的词示例，图 6 显示了随机生成的苏幕遮。实验由于采用简单的随机生成，并没有特别多的技术难点，主要在 GUI 界面的展示上。

简单随机生成的方式的缺点也显而易见，生成的词语义上无法连贯，并且缺少了宋词独有的平仄韵律，如果要考虑语义的宋词生成应该需要更复杂一些的算法，例如 rnn 循环神经网络等算法进行模型的训练和宋词的生成。