



**自然语言处理工程报告**

课程：自然语言处理

姓名：徐翊航

学号：2019217093

班级：物联网19-1班

日期：2022年1月16日

**目录**

**1、实验一语料库的收集与整理 3**

**1.1研究背景 3**

**1.2模型方法 4**

**1.3系统设计 5**

**1.4系统演示与分析 6**

**1.5对本门课的感想、意见和建议 8**

**2、实验二词汇知识库使用技术 9**

**2.1研究背景 9**

**2.2模型方法 10**

**2.3系统设计 11**

**2.4系统演示与分析 11**

**2.5对本门课的感想、意见和建议 15**

**3、实验三中文分词技术应用 15**

**3.1研究背景 15**

**3.2模型方法 15**

**3.3系统设计 16**

**3.4系统演示与分析 17**

**3.5对本门课的感想、意见和建议 20**

1、实验一语料库的收集与整理

1.1研究背景

语料库就是存放语言材料的数据库。那么，顾名思义，语料库语言学就是

基于语料库进行语言学研究的一门学问。具体一点讲，语料库语言学是研究自然语言机读文本（或称“电子文本”）的采集、存储、标注、检索、统计等方法的一门学问，其目的是通过对客观存在的大规模真实文本中的语言事实进行定量分析，为语言学研究或自然语言处理系统开发提供支持。语料库语言学研究的内容十分广泛，涉及语料库的建设和利用等多个方面，归纳起来，可以大致包括如下几方面的内容： ①语料库的建设与编纂；②语料库的加工和管理；③语料库的应用，包括在语言学研究（言语、词汇和语义研究等）中的应用和在自然语言处理中的应用。

语料库可以按语种划分可以分为单语种语料库和多语种语料库；按记载媒体不同可以分为单媒体语料库和多媒体语料库；按照地域区别可以分为国家语料库和国际语料库等。这里主要介绍以语料代表性和平衡性为主要区分依据的“平衡语料库与平行语料库”、以语料库用途为主要区分依据的“通用语料库与专用语料库”、以语料分布时间为主要区分依据的“共时语料库与历时语料库”和以语料库内容加工程度划分的“生语料与标注语料库”。具体如下：

（1）平衡语料库与平行语料库

平衡语料库着重考虑的是语料的代表性与平衡性。张普（2003）曾经提出语料采集的七项原则：语料的真实性、语料的可靠性、语料的科学性、语料的代表性、语料的权威性、语料的分布性和语料的流通性。其中，语料的分布性还要考虑语料的科学领域分布、地域分布、时间分布和语体分布等。

（2）通用语料库与专用语料库

所谓的通用语料库实际上与平衡语料库是从不同角度看问题的结果，或者说是与专用领域对举的结果。为了某种专门的目的，只采集某一特定领域、特定地区、特定时间、特定类型的语料构成的语料库就是专用语料库。例如，新闻语料库、科技语料库、中小学语料库、北京口语语料库等。

（3）共时语料库与历时语料库

所谓共时语料库是为了对语言进行共时研究而建立的语料库。按照索绪尔的观点，共时研究是指研究大树的横断面所见的细胞和细胞关系，即研究一个共时平面中的元素与元素的关系。无论所采集语料的时间段有多长，只要研究的是一个平面上的元素或元素的关系，就是共时研究，所建立的语料库就是共时语料库。

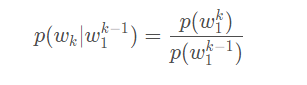
所谓的历时语料库是为了对语言进行历时研究而建立的语料库。根据历时语料库得到的统计结果就不像共时语料库的统计结果是一个频次点，而是依据时间轴的等距离抽样得到的若干频次变化形成的演变曲线，我们把这种曲线称为变化“走势图”。

（4）生语料与标注语料库

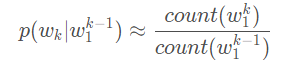
所谓生语料是指没有经过任何加工处理的原始语料数据。组织者只是简单地把语料收集起来，不加任何标注信息。标注语料库是指经过加工处理、标注了特定信息的语料库。根据加工程度不同，标注语料库又可以细分为分词语料库（主要指汉语）、分词与词性标注语料库、树库、命题库、篇章树库等。

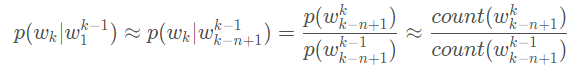
1.2模型方法

在此实验中主要采用N-gram模型对文本进行处理。n-gram作为一种基于统计语言，我们可以利用数学的语言对其进行描述。我们的目标是建立下面的模型：

模型中的基本元素是，只需要找到这个概率的表示方法即可，根据贝叶斯公式：

这里要我们要注意的是上的k并不是表示幂次，表示的是序列这也就解释了上面贝叶斯公式的分子为什么不是因为已经将包含进去，所以将这二者简写在一起。根据大数定理，上式可以表示为：

其中表示的是语料库中序列出现的频次。n-gram的基本思想，是它做了一个n-1阶的Markov假设，即认为一个词出现的概率只与它前面的n-1个词相关，公式如下：



根据这个概率可以对给定的序列预测下一个词最有可能是什么。

1.3系统设计

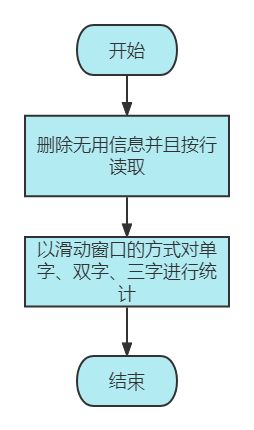
在实验一中我利用java完成了实验的要求，在这个实验中我主要完成了两个部分的内容，第一个部分是我利用n-gram模型对宋词预料进行了分词，我先去除文本当中一些无用的东西，然后按照行读取，每一行对应一首词，之后对单字、双字、和三字都进行统计，并且记录它们出现的次数，统计结果将在1.4节进行展示,统计的结果将用于实验二当中宋词的生成。宋词处理流程图如图1所示：

图1 宋词处理流程图

第二个部分是我同样也是用java进行的，这个部分是依据老师给出的1998-01-2003版带音.txt文件并且根据n-gram模型进行的词频的统计并且生成词典，在这个部分我首先对文件中一些无用的词例如日期等信息进行删除，然后依据词和词之间的间隔符进行统计生成词典，除此之外，我还对所给文件的词的词性、词之间转换频率也做了统计。

1.4系统演示与分析

实验一中是对于词频和词性的转换进行了统计，因此这里直接展示生成的结果如图2~8所示，具体如下。

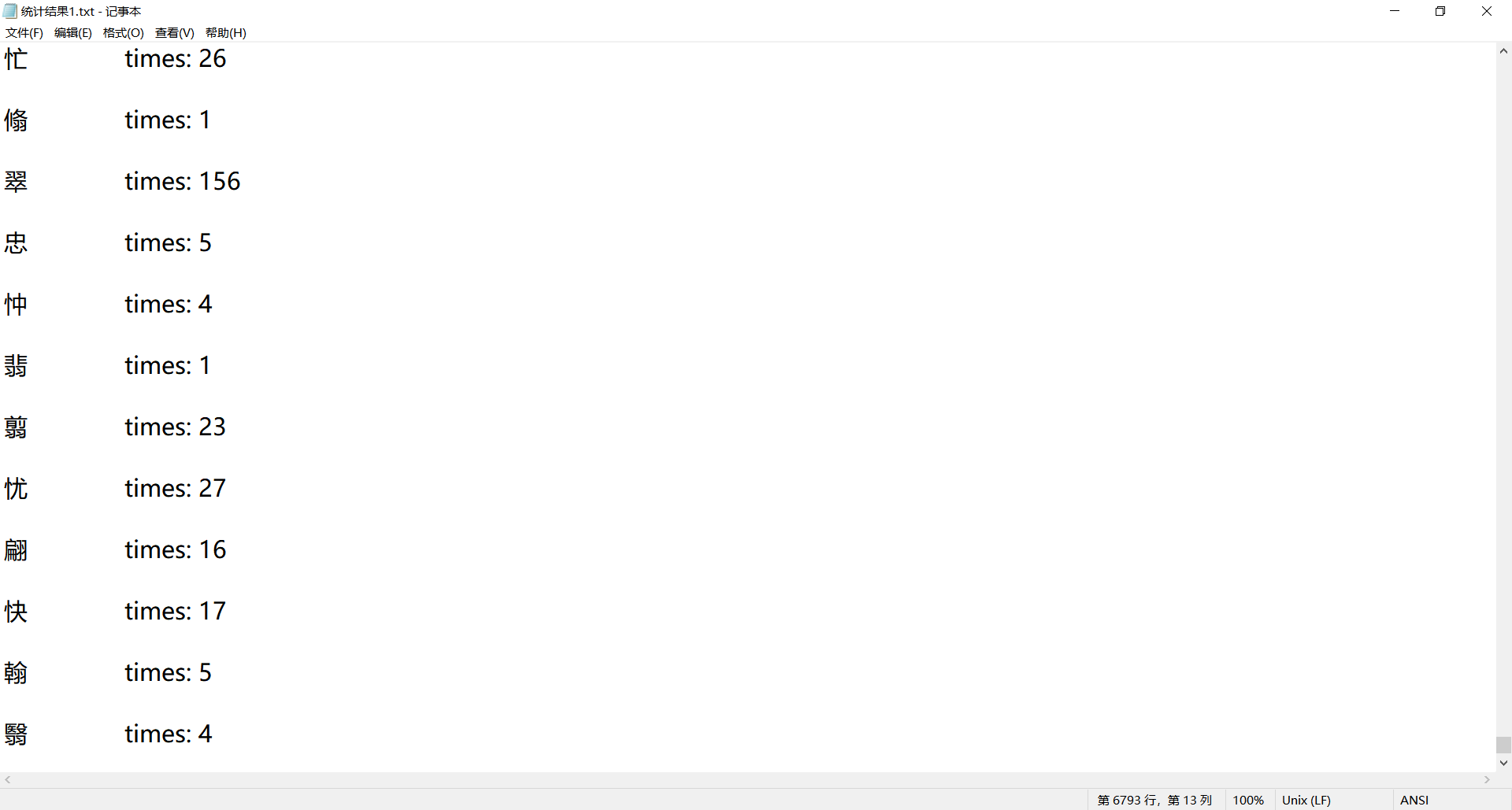


图2 宋词单字统计结果

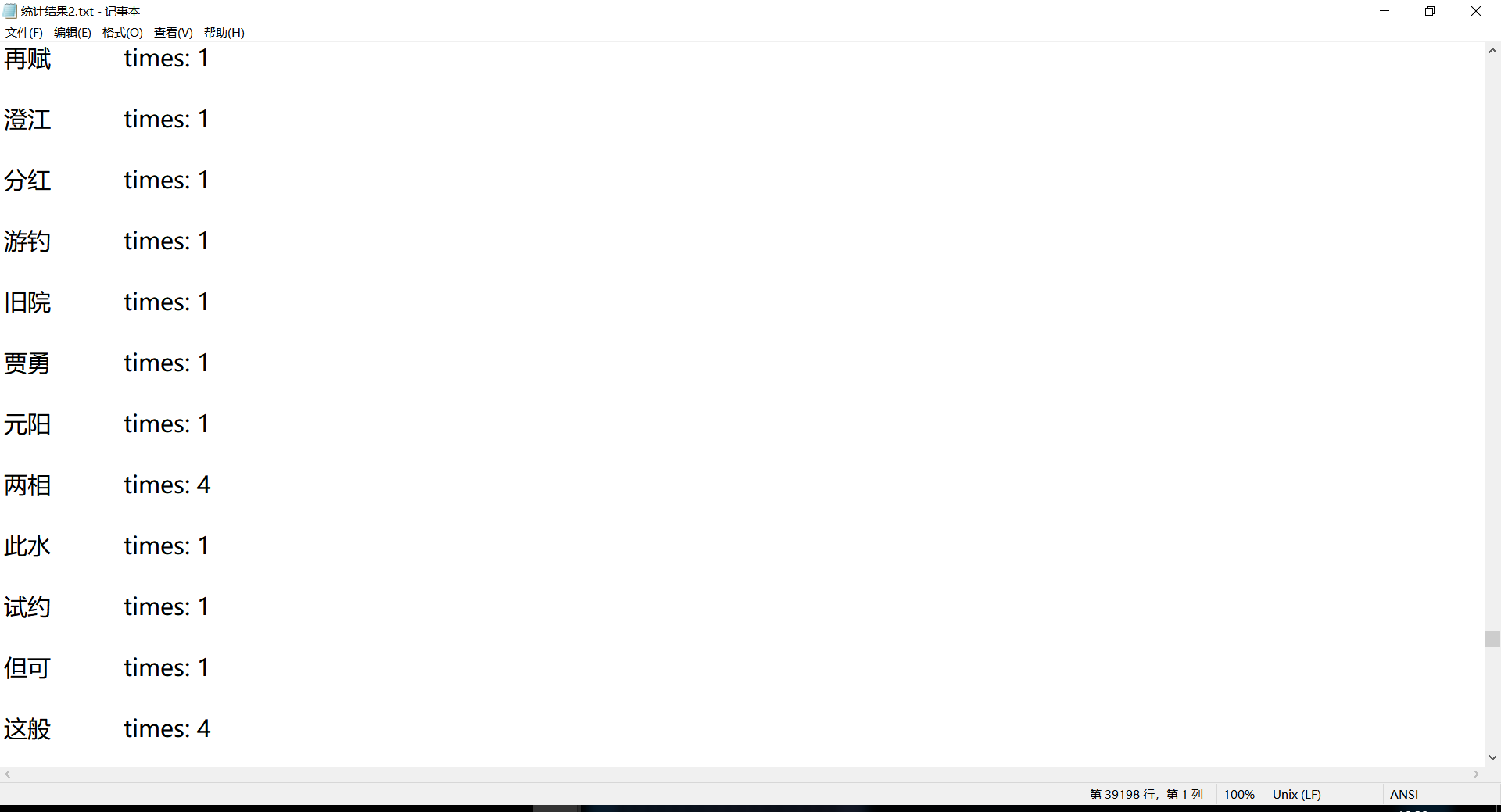


图3 宋词双字统计结果

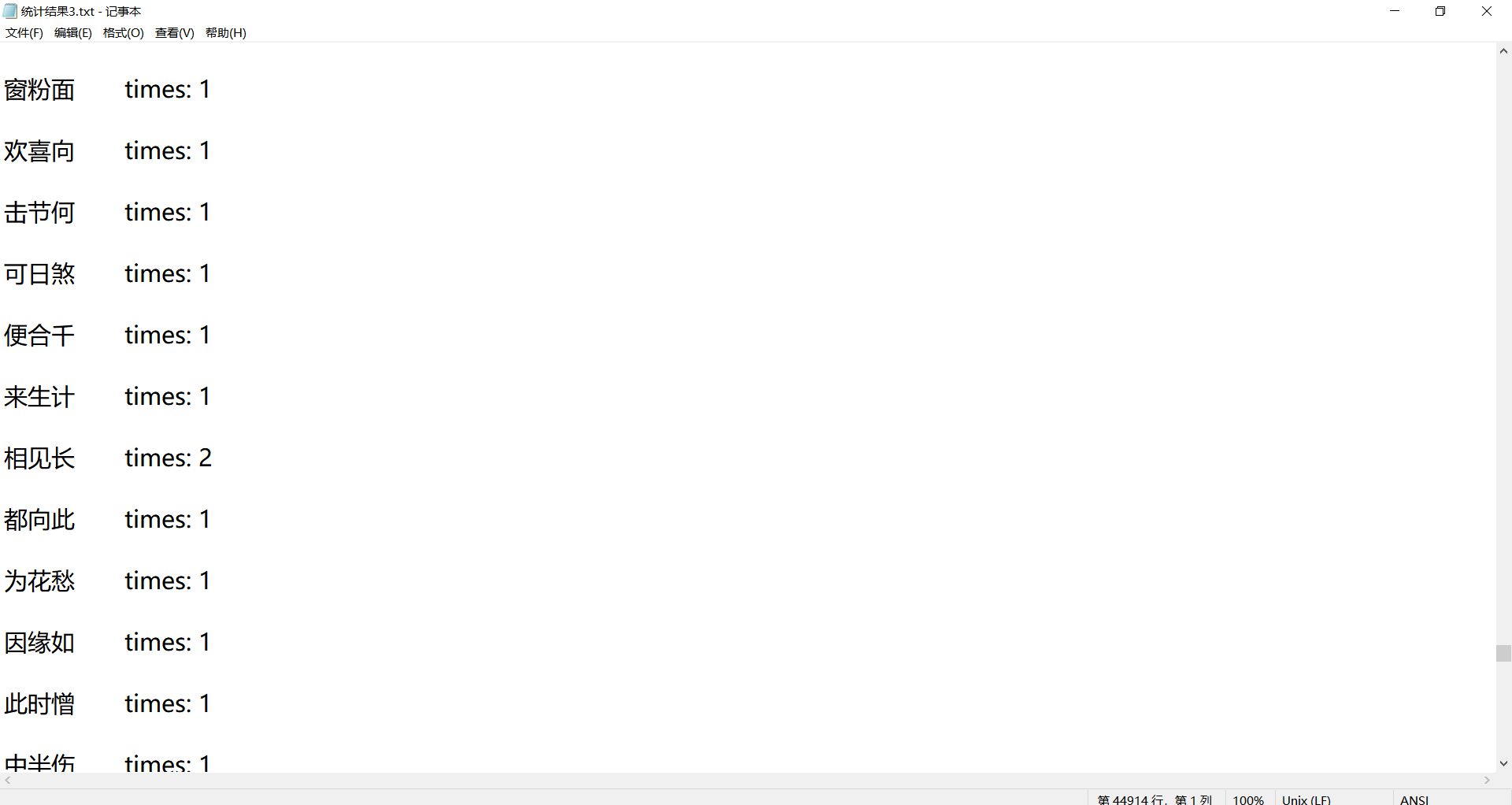


图4 宋词三字统计结果

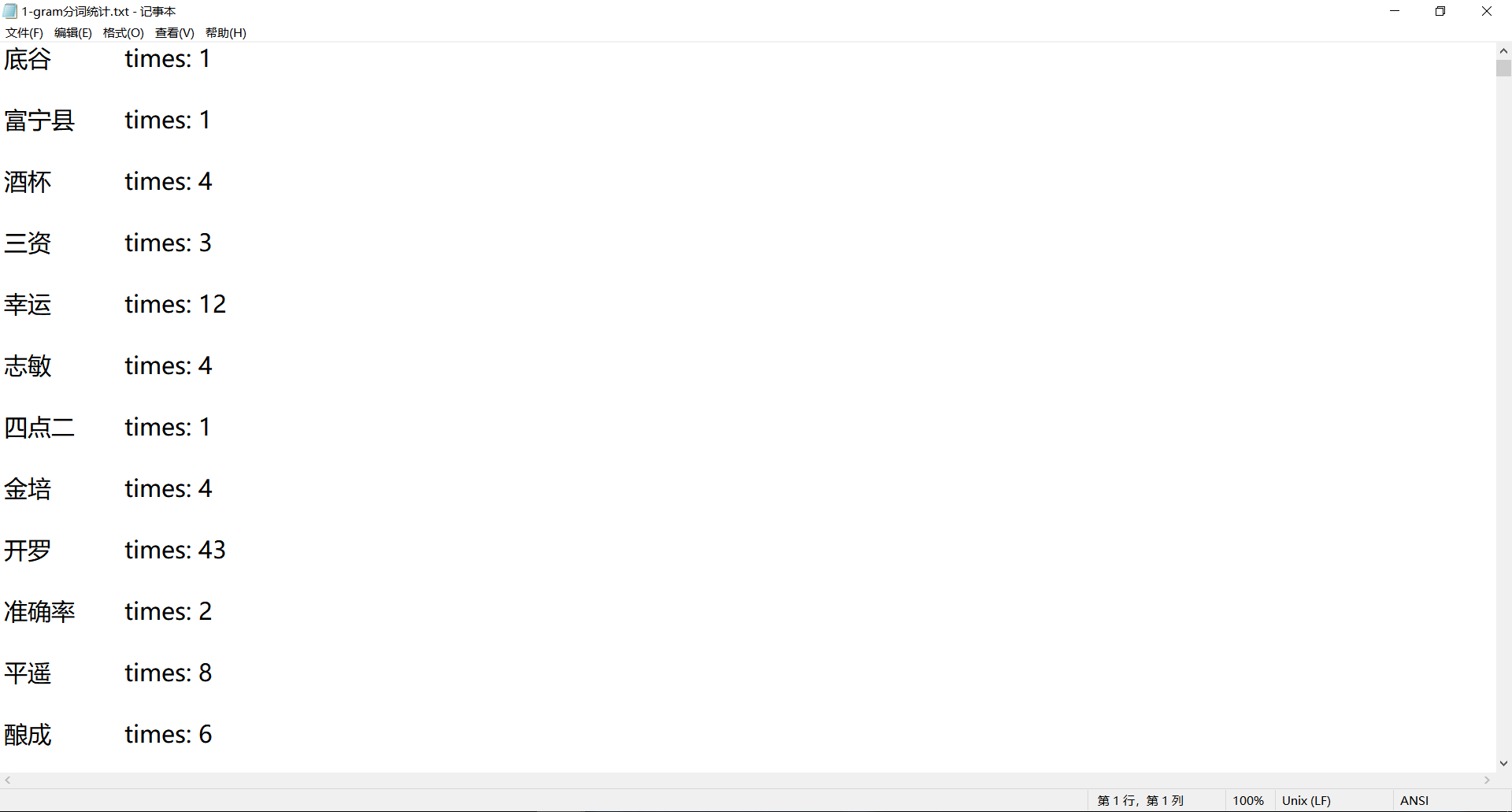


图5 1-gram分词统计

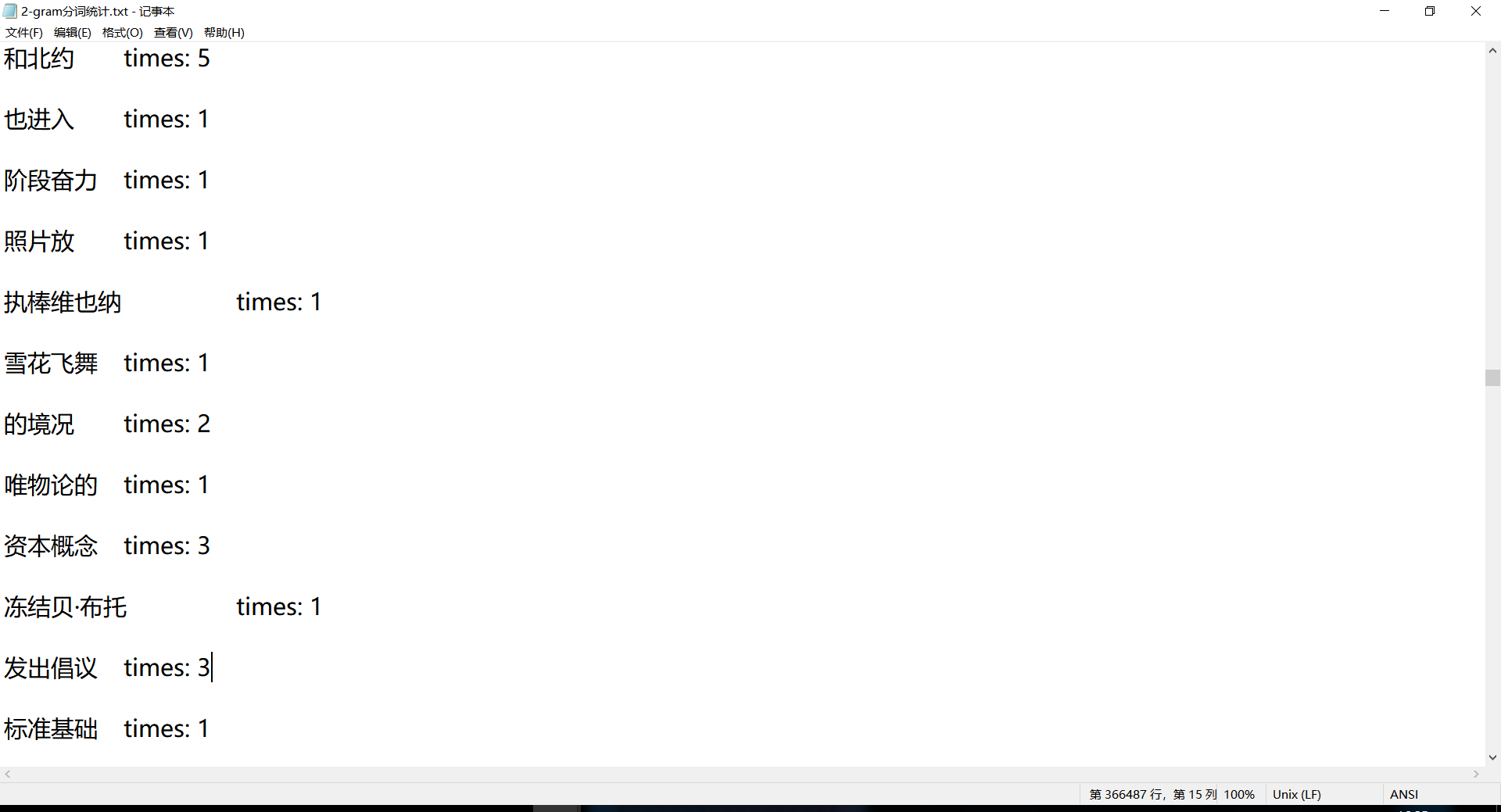


图6 2-gram分词统计

词性统计：

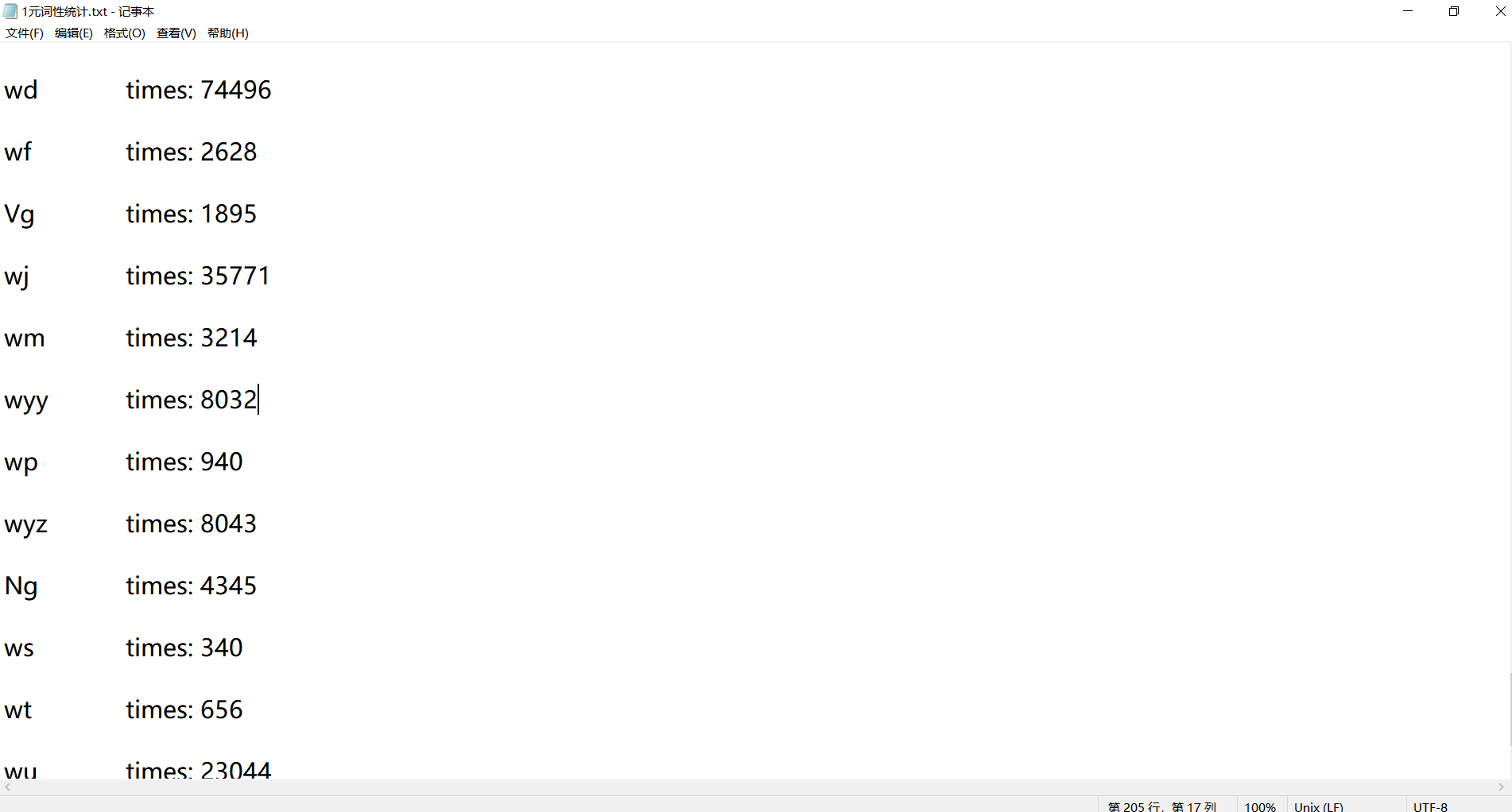


图7 一元词性统计

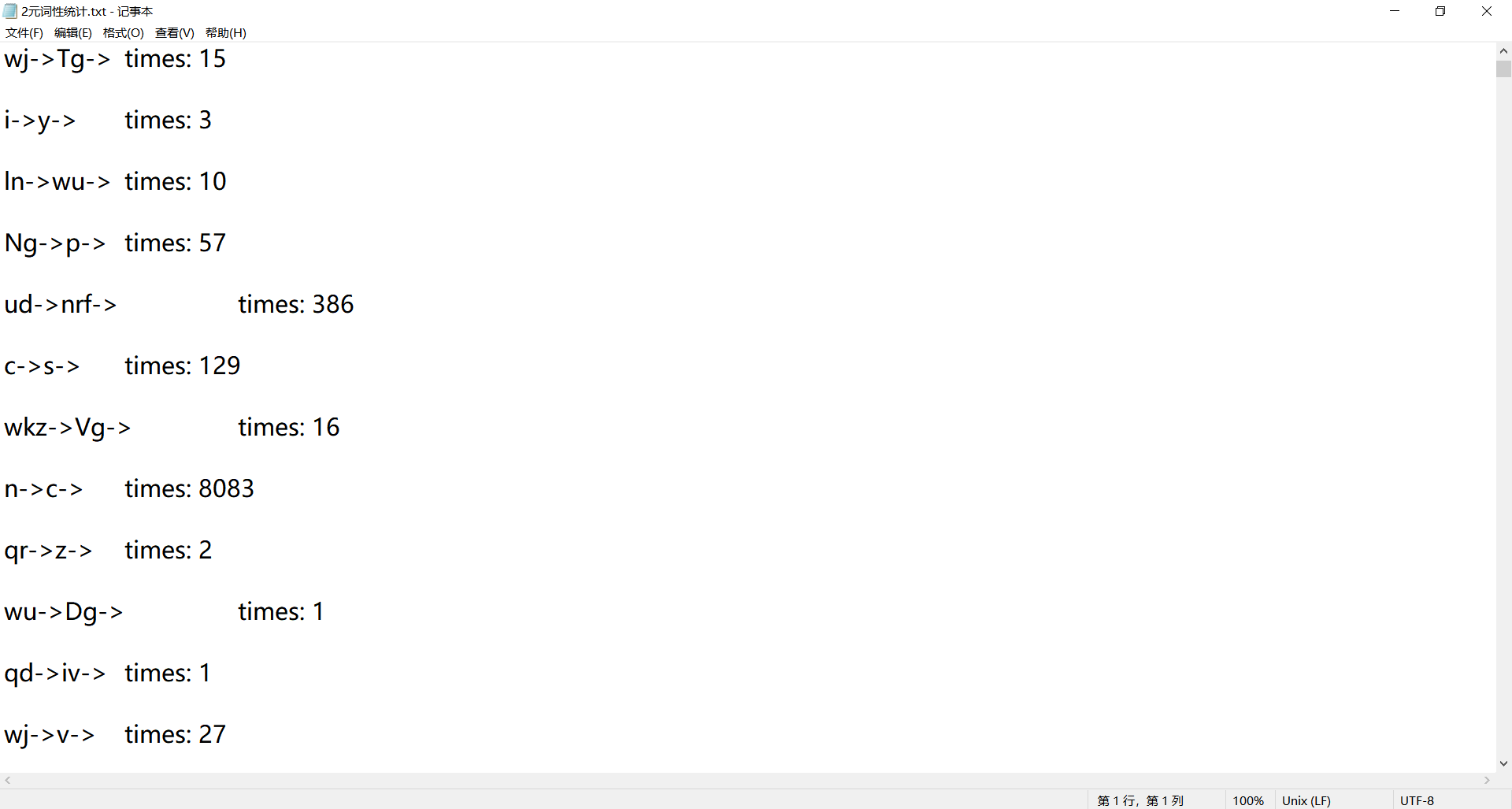


图8 二元词性统计

经过观察以及简单验证，以上结果均正确，最终将结果输出到了文件当中，成功有效的完成了实验一。

1.5对本门课的感想、意见和建议

实验一是较为简单的，词频的统计是自然语言处理当中很多事情的基础，我在做这个实验之前重新查看了教学群里发送的PPT更加了解后开始本次实验。而这一部分的代码实现也是比较简单的，代码的工作量并不是很大，但是让我更加了解了n-gram模型，通过这个模型可以较好的解决词频的统计问题，也提高了我的代码编写水平，激发了我对自然语言的兴趣。

2、实验二词汇知识库使用技术

2.1研究背景

词汇知识库是根据词义（而不是词形）组织的词汇信息，从某种意义上讲，他是一部语义词典。例如当前的WordNet词汇知识库就是按找语义关系进行组织的。语义关系看作是同义词 集合之间的一些指针，语义关系是双向的。如果词 义{x1 , x2 , …} 和{y1 , y2 , …}之间有一种语义关系R，则在{y1 , y2 , …}和{x1 , x2 , …} 之间也有语义关系R，属于这两个同义词集合的单词之间的关系也是R。四种语义关系分别为：

1. 同义关系；
2. 反义关系；
3. 上下位关系或称从属/上属关系：如{枫叶}是{数}的下位，{树}是{植物}的下位；
4. 部分关系或称部分/整体关系；

在众多词汇知识库中，知网在其中占有非常重要的作用，它具有4个基本观点分别是：

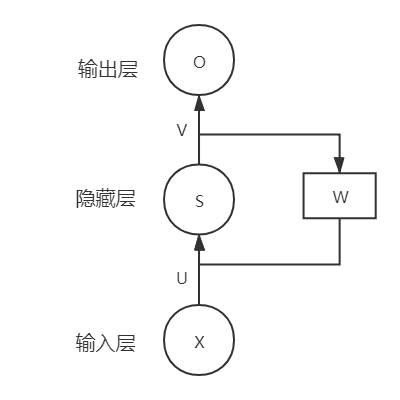
1. NLP系统最终需要更强大的知识库的支持；
2. 知识是一个系统，是一个包含着各种概念与概念之间的关系，以及概念的属性与属性之间的关系的系统。一个人比另外一个人有更多的知识说到底是他不仅掌握了更多的概念，尤其重要的是他掌握了更多的概念之间的关系以及概念的属性与属性之间的关系。
3. 关于知识库建设，他提出应首先建立一种可以被称为知识系统的常识性知识库。它以通用的概念为描述对象，建立并描述这些概念之间的关系。
4. 首先应由知识工程师来设计知识库的框架，并建立常识性知识库的原型。在此基础上再向专业性知识库延伸和发展。专业性知识库或称百科性知识库主要靠专业人员来完成。这里很类似于通用的词典由语言工作者编纂，百科全书则是由各专业的专家编写。

而其特色就是知网作为一个知识系统，名副其实是一个网而不是树。它所着力要反映的是概念的共性和个性，例如：对于“医生”和“患者”，“人”是它们的共性。同时知网还着力要反映概念之间和概念的属性之间的各种关系。

2.2模型方法

在本实验中我使用了两种模型来实现，首先，我是用了N-gram模型完成宋词的输出，因为在实验一中我已经统计了词中的单字、双字、和三字的频率，我引入了一个随机文本生成办法，它利用我们在实验一中生成的单字、双字、三字的字典文本来进行构造，由于对于每一个词牌每个位置拜访的字是类似的，因此我们可以利用这个方法来进行生成，最后将文字拼接起来组成生成的宋词。

第二个模型是采用的RNN，因为我考虑到当采用第一种方法时生成的词可能不是特别具有文学性，因此我采用了RNN模型，这也是我具体介绍的。RNN是一种特殊的神经网络结构, 它是根据"人的认知是基于过往的经验和记忆"这一观点提出的。 它与DNN,CNN不同的是: 它不仅考虑前一时刻的输入,而且赋予了网络对前面的内容的一种'记忆'功能。RNN之所以称为循环神经网路，即一个序列当前的输出与前面的输出也有关。具体的表现形式为网络会对前面的信息进行记忆并应用于当前输出的计算中，即隐藏层之间的节点不再无连接而是有连接的，并且隐藏层的输入不仅包括输入层的输出还包括上一时刻隐藏层的输出。也正是考虑到这个，因此他更加适合与生成诗歌文本。

 图9 RNN模型

上图中其中U是输入层到隐藏层的权重矩阵，O也是一个向量，它表示输出层的值；V是隐藏层到输出层的权重矩阵。循环神经网络的隐藏层的值s不仅仅取决于当前这次的输入x，还取决于上一次隐藏层的值S。权重矩阵 W就是隐藏层上一次的值作为这一次的输入的权重。

我计划采用BPTT算法来训练RNN。其实本质还是BP算法，只不过RNN处理时间序列数据，所以要基于时间反向传播，故叫随时间反向传播。BPTT的中心思想和BP算法相同，沿着需要优化的参数的负梯度方向不断寻找更优的点直至收敛。综上所述，BPTT算法本质还是BP算法，BP算法本质还是梯度下降法，那么求各个参数的梯度便成了此算法的核心。

在此处要防止模型出现过拟合和欠拟合的现象。如果发生过度拟合则会把噪声数据的特征也学习到了，导致模型泛化能力下降，在训练集上表现很好，但是在测试集上表现很差。这时我需要通过增大训练集、增加正则化等方法对模型进行优化。而欠拟合会导致模型不能够很好地拟合数据，我们就应该通过进一步添加特征项或减小正则化参数等方法来实现，最后实现的效果由于第一种模型。

2.3系统设计

对于N-gram模型，通过用户选择需要生成的词牌，这里我设置了“卜算子”、“满江红”、“醉花阴”、“江城子”四种，选择完成后点击生成便可以得到需要的宋词。

对于RNN模型，首先我先训练样本，我通过网上搜集40000多首的唐诗，然后我们需要进行汉字的embedding，embedding的研究已经取得了很大的进展，在这里我们只是简单地进行处理，简单来说我统计所有汉字的词频，然后按照词频从高到低进行排序，这样我就获得了每个汉字和一个列表序号的映射关系。通过了embedding我们就可以将每一首诗会转化为一个多维向量，维度的个数代表汉字的个数。我们利用RNN神经网络对每一首诗进行训练，RNN的神经网络的搭建现在都比较固定了。 搭建好神经网络之后我们就可以进行训练了，我们采用分批训练，每64首训练一次。 训练结束后保存模型，我们下次直接使用这个模型，采用随机开始，这样每次都生成不同的诗。当然这里涉及到了停止的问题，我会在每一首诗的后面加一个截断符，这样网络就会学习到这样的特征。

2.4系统演示与分析

N-gram模型生成结果图如图11、12所示。

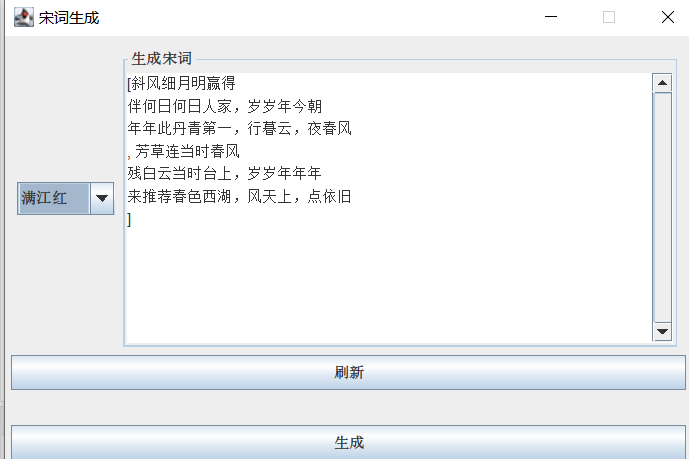


图11 N-gram模型生成结果（1）

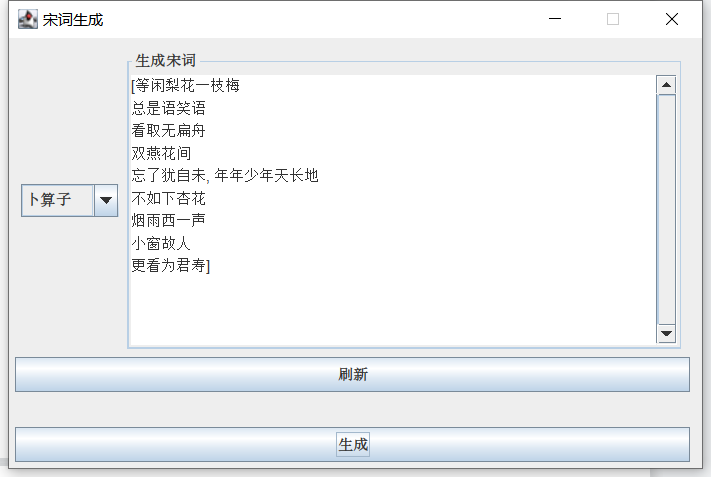


图12 N-gram模型生成结果（2）

RNN模型生成结果代码：

1. def gen\_poetry():
2. def to\_word(weights):
3. t = np.cumsum(weights)
4. s = np.sum(weights)
5. sample = int(np.searchsorted(t, np.random.rand(1)\*s))
6. return words[sample]
8. \_, last\_state, probs, cell, initial\_state = neural\_network()
9. result = ""
11. with tf.Session() as sess:
12. sess.run(tf.global\_variables\_initializer())
14. saver = tf.train.Saver(tf.all\_variables())
16. module\_file = tf.train.latest\_checkpoint('./train\_dir')
17. print(module\_file)
18. saver.restore(sess, module\_file)
20. state\_ = sess.run(cell.zero\_state(1, tf.float32))
22. x = np.array([list(map(word\_num\_map.get, '['))])
23. [probs\_, state\_] = sess.run([probs, last\_state], feed\_dict={input\_data: x, initial\_state: state\_})
24. word = to\_word(probs\_)
25. *#word = words[np.argmax(probs\_)]*
26. poem = ''
27. while word != ']':
28. poem += word
29. x = np.zeros((1, 1))
30. x[0, 0] = word\_num\_map[word]
31. [probs\_, state\_] = sess.run([probs, last\_state], feed\_dict={input\_data: x, initial\_state: state\_})
32. word = to\_word(probs\_)
33. *#word = words[np.argmax(probs\_)]*
34. result = poem
35. return result

RNN模型生成结果如图13、图14所示。

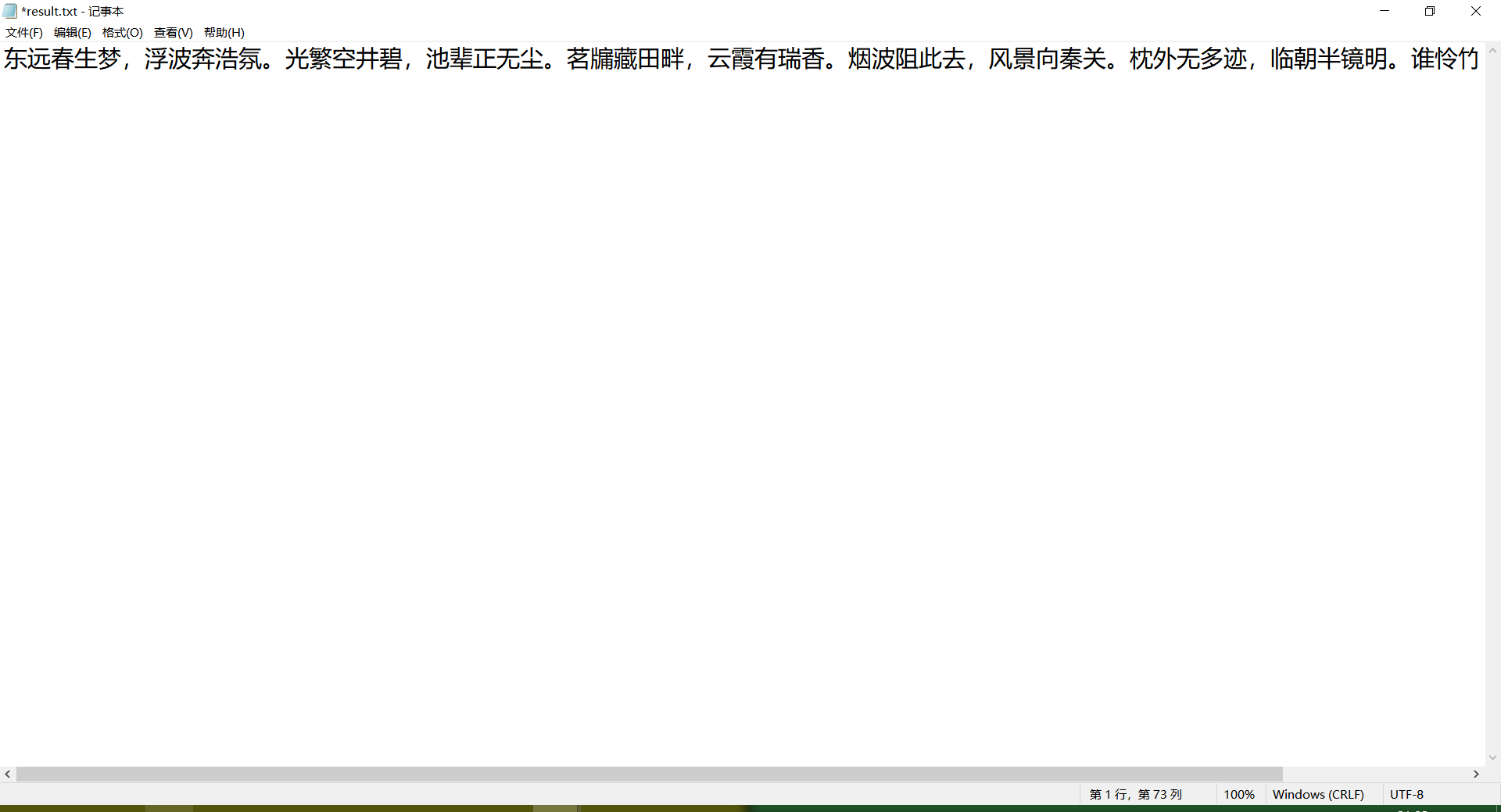


图13 RNN模型生成结果（1）

诗内容为：

东远春生梦，浮波奔浩氛。光繁空井碧，池辈正无尘。茗牖藏田畔，云霞有瑞香。烟波阻此去，风景向秦关。枕外无多迹，临朝半镜明。谁怜竹洞里，终可遣忘衡。

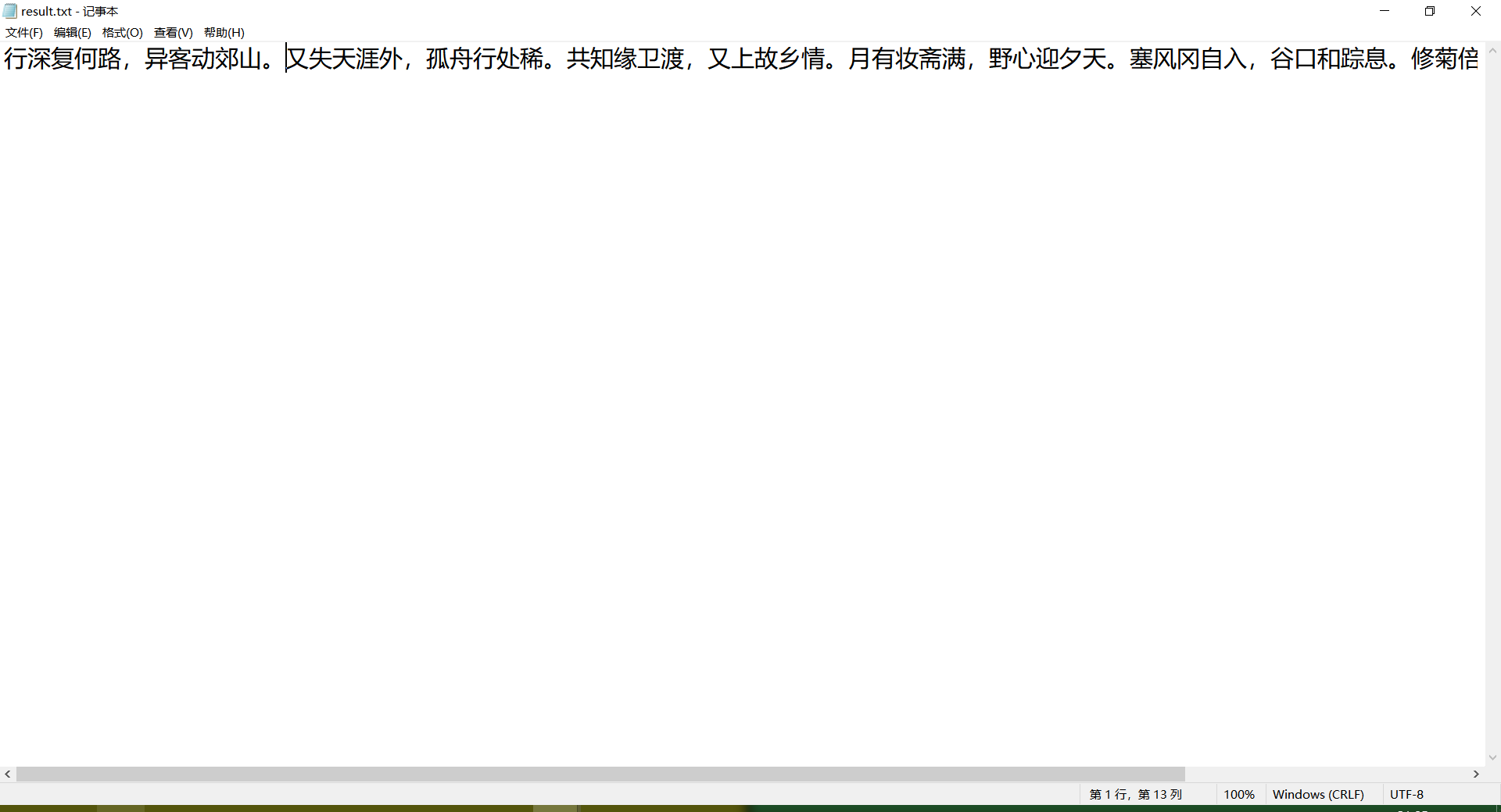


图14 RNN模型生成结果（2）

诗内容为：

行深复何路，异客动郊山。又失天涯外，孤舟行处稀。共知缘卫渡，又上故乡情。月有妆斋满，野心迎夕天。塞风冈自入，谷口和踪息。修菊倍傍人，结人难相慰，还是若云栖。

由生成结果可知此模型能够比较好的生成诗词，RNN模型具有比较好的效果相对于N-gram模型。

2.5对本门课的感想、意见和建议

在这次实验中，我利用了两种模型来生成诗词，让我对这两个模型有了比较深刻的印象，但是实验的时间还是比较有限，在假期中我也会继续完善这一部分的内容，争取搭建一个更好的模型。

3、中文分词技术技术

3.1研究背景

分词将短文本和长文本处理为最小单位粒度是词或词语的过程。方法包括基于字符串匹配的分词方法、基于理解的分词方法、基于统计的分词方法和基于规则的分词方法。但是和大部分西方语言不同，书面汉语的词语之间没有明显的空格标记，句子是以字串的形式出现。因此对中文进行处理的第一步就是进行自动分词，即将字串转变成词串。比如“中国建筑业呈现新格局”分词后的词串是 中国／建筑业／呈现／新／格局。

为什么中文分词如此重要呢，是因为它是处理中文的语义分析，文本分类，信息检索，机器翻译，机器问答等问题的基础。如果分词效果不好，很有可能会严重影响到后续的研究。因为中文存在交集歧义，组合歧义，无法在句子中解决的歧义，具有未登录词等等特征，使得中文分词很难。

3.2模型方法

在这个实验中我采用了正向最大匹配算法、逆向最大匹配算法和双向最大匹配算法三种。下面将对这三种进行介绍。

**正向最大匹配算法**：从左到右将待分词文本中的几个连续字符与词表匹配，如果匹配上，则切分出一个词。但这里有一个问题：要做到最大匹配，并不是第一次匹配到就可以切分的。我们来举个例子：待分词文本：content[]={"中"，"华"，"民"，"族"，"从"，"此"，"站"，"起"，"来"，"了"，"。"}词表：dict[]={"中华"， "中华民族" ， "从此"，"站起来"}，那么其具体过程为：

（1）从content[1]开始，当扫描到content[2]的时候，发现"中华"已经在词表dict[]中了。但还不能切分出来，因为我们不知道后面的词语能不能组成更长的词(最大匹配)。

（2）继续扫描content[3]，发现"中华民"并不是dict[]中的词。但是我们还不能确定是否前面找到的"中华"已经是最大的词了。因为"中华民"是dict[2]的前缀。

（3）扫描content[4]，发现"中华民族"是dict[]中的词。继续扫描下去：

（4）当扫描content[5]的时候，发现"中华民族从"并不是词表中的词，也不是词的前缀。因此可以切分出前面最大的词——"中华民族"。

由此可见，最大匹配出的词必须保证下一个扫描不是词表中的词或词的前缀才可以结束。

**逆向最大匹配算法**：算法的过程为：

（1）输入最大词长maxWordLength，字典wordDict，待分句子。

（2）从待分句子的末尾开始向前截取长度为maxWordLength的子句，进行分词。

（3）对一个子句的分词过程为，首先判断子句是否在字典中，若在，则保存这个子句，并从原句中删除这个子句，转到2。若不在，则判断子句长度是否为1，若为1，则将单字保存，从原句中删除单字，转到2。若不为1，则将子句中最右边的一个字删除，形成新的子句，转到3。

**双向最大匹配法：**该算法是将正向最大匹配法得到的分词结果和逆向最大匹配法的到的结果进行比较，从而决定正确的分词方法。据SunM.S. 和 Benjamin K.T.（1995）的研究表明，中文中90.0％左右的句子，正向最大匹配法和逆向最大匹配法完全重合且正确，只有大概9.0％的句子两种切分方法得到的结果不一样，但其中必有一个是正确的（歧义检测成功），只有不到1.0％的句子，或者正向最大匹配法和逆向最大匹配法的切分虽重合却是错的，或者正向最大匹配法和逆向最大匹配法切分不同但两个都不对（歧义检测失败）。这正是双向最大匹配法在实用中文信息处理系统中得以广泛使用的原因所在。

3.3系统设计

在做这个实验时，我首先需要读取一个已经分词完成的词典，然后我利用python做了一个图形化界面，在这个界面中可以输入你需要分词的文本，随后就能给出利用三种匹配法各自生成的分词以后的文本。运行程序后截图如图15所示。



图15 运行界面截图

通过这个系统也可以发现双向最大匹配算法要优于前面两种算法，因为我们在选择标准时，是通过观察分词的数量，分词的数量越少则越好，而当分词数量相同的情况下，则比较单个词的数量，越少则越好。

3.4系统演示与分析

系统演示1：图16所示

输入分词文本：为人民服务。

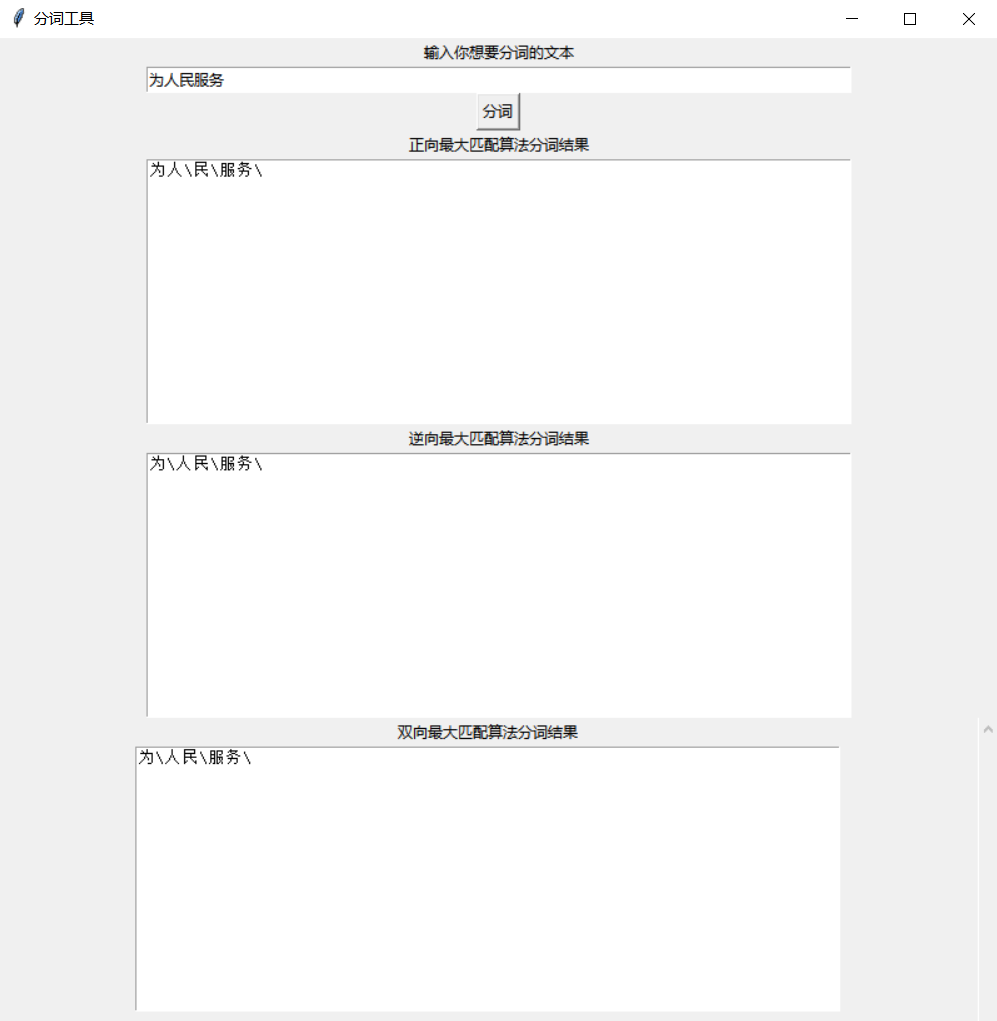


图16 分词演示结果1

由上图可知，采用逆向最大匹配算法分词的结果和双向最大匹配算法的分词结果一致二者要优于正向最大匹配算法的结果，二者为理想结果。

系统演示1：图17所示

输入分词文本：联合国邮政管理局14日宣布，为庆祝北京2022年冬奥会的召开，联合国将发行主题为“体育促进和平”的邮票。

这是联合国首次为冬奥会发行邮票。此次发行的邮票分别展现了冰球运动员、单板滑雪运动员、花样滑冰运动员、雪车运动员、冰壶运动员以及高山滑雪运动员形象。

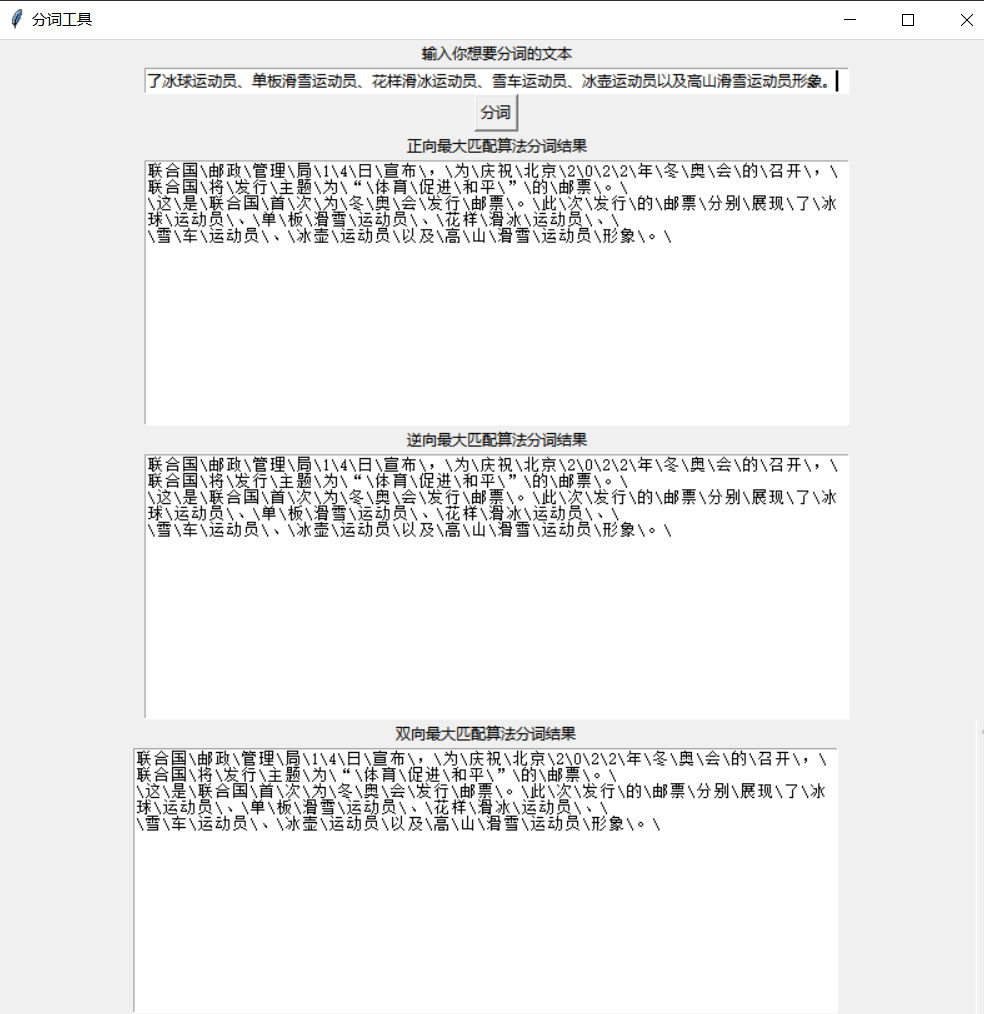


图17 分词演示结果2

最后也实现了将输出结果保存到文本中如图18所示：

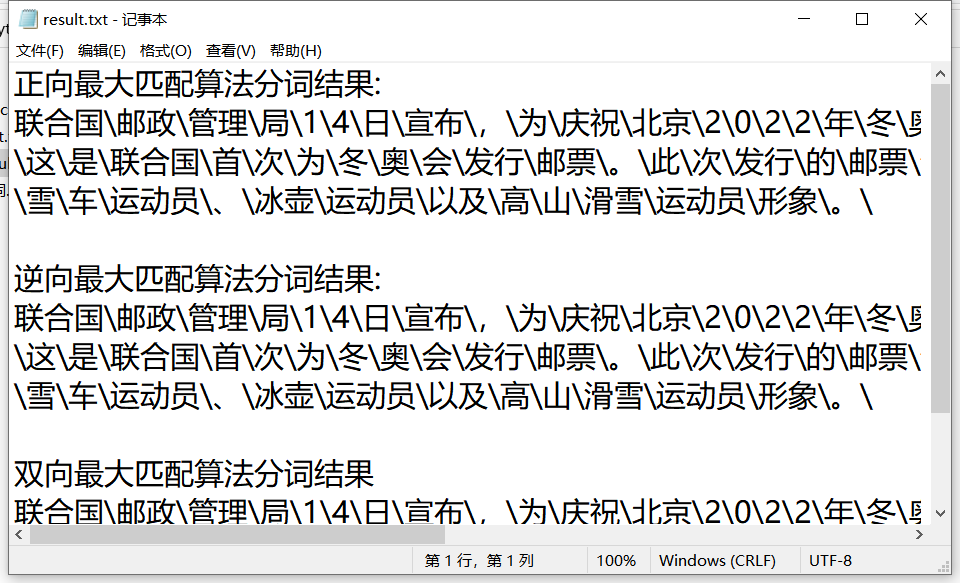


图18 生成文本结果

3.5对本门课的感想、意见和建议

通过这次分词实验让我更加了解了FMM、BMM、RMM的相关算法，达到了实验目的，在后续的假期里我会继续优化算法希望能和实验二一样通过建立相关神经网络模型来实现分词。在一次线上腾讯会议中，我们听取了几名老师所做的报告，其中一篇名为图网络表征的研究使我影响深刻，当今时代人工智能在语言识别、人脸识别、图像识别方面都有着巨大的突破，很多方面都已经达到甚至超过人类的水平。但是当前深度学习与标准机器学习的区别在于自动构造特征。对于广泛应用场景中的图结构数据，结构化的图数据的特性，会导致挖掘困难。这主要存在与以下四个方面：

1. 非欧特性：节点数量分布不均匀；
2. 置换不变性：节点间没有固定的排序顺序；
3. 链接关系：存在同质、异构等多种关系；
4. 不完备性：节点属性和网络结构不完整；

这会导致典型的深度学习模型例如DNN、RNN、CNN难以直接适配，因此我们要通过网络表示学习将网络中节点嵌入到低维的向量表示空间中，在低维度表示空间中进行网络重构和网络推断任务，这样也会使得算法更加高效、并行。网络应用的任务在当今十分热门，网络分析任务可以用来处理节点分类、链接预测、社区检测等，并且在多个领域应用包括社交网络、分子网络、知识图谱、智慧教育。

对于当今的一些网络表征相关工作，比DeepWalk它就基于word2vec，假设在网络中具有相同结构的节点，那么它们的网络向量表征也比较相似，也正是通过这一点来通过随机游走来构建句子序列，均匀从相邻节点中采样生成训练样本序列。而对于Node2vec,它基于DeepWalk，设置了偏置的截断式随机游走探索邻居结构多样性，这种偏置搜索是一种平滑了的深度优先搜索和广度优先搜索，在二者之间找到一个平衡。但是这些研究都还面临一些挑战在表征语义方面，网络节点的语义多面性与向量表征空间的单一性存在冲突，其次在表征空间中网络结构的非欧特性与欧式表征空间的局限性。因此我们为了解决这个问题需要进行多面语义属性网络表征来解决属性网络中节点多套向量表征学习问题从而表示出节点在多面语义上不同的相似关系，这是因为属性网络中节点存在语义多面性包括以下两点：

1. 节点链接形成有多种不同因素，从而存在不同的语义关系；
2. 节点在不同语义上相似性不同；

对于多面语义属性网络表征现有的方法包括单个表征学习方法和多面表征学习方法，这具有一定的局限性有以下几个方面：

1. 使用单个向量表征空间难以表示节点间的多面相似关系；
2. 需要重复多次模型生成表征，导致大量训练参数冗余；
3. 忽略语义间的关联性，需要大量监督训练数据来学习多套向量表征。

因此提出了多面语义属性网络向量表征-MCNE，它首次提出在单个向量空间中学习节点多面语义相似性的问题，设计了二值掩码层将单个向量表征分割成多个条件子空间，利用图卷积网络建模多面语义交互关系，并基于多任务学习框架联合学习。它是基于图神经网络建模高阶多语义交互，利用多任务联合学习，基于BPR损失函数学习单语义相似性，基于多任务框架学习多语义相似性，在经过测试之后此方法也能够表现出较好的结果。但是也存在一些挑战：

1. 缺乏知识先验：当前网络表征技术依赖大规模样本数据并难以有效利用先验知识，导致算法结果与人的先验知识或专家知识相冲突。
2. 缺乏因果关系：基于数据驱动的分析技术只能挖掘数据间的关联关系，难以揭示数据变量之间本质的因果关系，导致模型鲁棒性不足、泛化能力弱等问题。
3. 缺乏可解释性：网络表征技术的黑盒属性导致模型可解释性差，存在模型决策不可信问题

可见在未来还有很多东西需要我们不断的努力，通过这次实验以及听以上报告内容，我对自然语言处理有了浓厚的兴趣，在以后的学习道路上我也会在不断奋斗。