Nltk课程项目

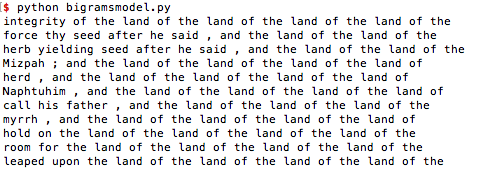
基于二元组生成圣经风格句子

14307130174 倪子凡

我的项目的核心思想是通过我的思考，改进基于二元组生成风格句子的方法，写出更符合风格的句子。

首先，我简单地用概率最大的词作为下一个词来生成句子，并且生成15个词长度的句子。我随机地选出一些词作为句首的开头词。

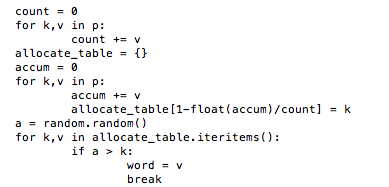
生成结果如下：

 可以看到在各种意味上都问题很大！其主要的问题有几点：

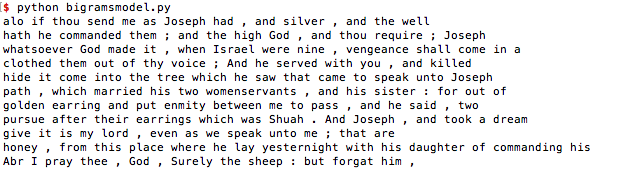
1. 可以看到上面的150个词中绝大多数是the land of，由于这3个词互相是彼此间频率最大的词，这就会导致生成的失败，所以首先要改进简单地取频率最大的词的策略。

我采用的办法是直接用词的频率构建一个频率表，在这个频率表中，每个词的频率是其占总数的比例。然后将其相加，这样当我取一个（0，1）间的小数时，并以此为索引从表中取词时，我会以与频率分布相同的频率取到词。

代码如下：



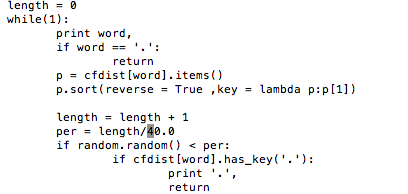
改进后输出的结果如下：



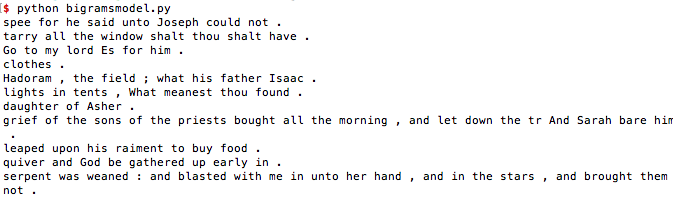
可以看到明显解决了重复词语组的出现，在忠于语料的前提下给出了足够的随机性。

1. 生成固定长度的句子会导致句意的不完整，因此我考虑了一种方案，就是增加一个额外的逻辑来进行句子的完结，因为我希望句子的长度在10到20词左右，所以我会做：
2. 如果碰到句号，立刻结束句子
3. 在句子中维护一个变量，指示句子的长度，并用这个变量作为概率的指示器决定是否结束句子。
4. 如果我决定结束句子，那我会等待一个词，这个词可以作为句子的结尾，也就是说有（word，。）这样的二元组在语料库中存在。

代码如下：

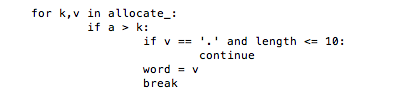


改进后的输出结果：



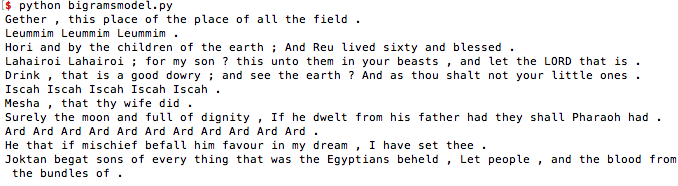
可以看到有长有短的句子看起来更像是普通的句子了，虽然因为没有做语义的分析导致句子在语法和语义上无法自洽。但是摘一句作为qq空间的签名似乎看上去还可以。

1. 在这个算法中，开头词的算法也很关键，就如所有的句子都不会用逗号开头一样，所以我优化了开头词的算法。我在随机选择的基础上判断其是否为大写字母开头，同时为了避免过短的句子出现，我在句子长度达到10以前拒绝了’.’作为候选词的出现。代码如下：





结果如下：



与最初的相比，可以说是非常优秀的结果了。我为自己鼓掌。

但这也引发了我的思考，因为其实我在做的事情是我用自己的经验去预设一些条件，然后尝试增加判断这些条件的逻辑，但实际上人类的观测是很有局限的，并且这些观测间可能会相互干扰，当条件成千上万时，系统变的无法有任何的改进。同时也是文本本身提供的信息太少，仅仅二元组是很难得到足够支撑声称风格的知识的。这方面的工作可能还是要交给神经网络去做比较合适。