1. 数据来源：

唐诗宋词，该数据中只存在诗的正文，去掉了标题。

数据集：总共有34646首诗,1721655个字(6110个去重后的字)

1. 数据处理：

目标：将数据转换成二维数组，因为每首诗的长度不同

1. 判断标点符号，获取诗词内容
2. 根据唐诗宋词的规律，进行异常剔除
3. 添加B+content+E 获取诗词完整内容
4. 取出所有诗中所有的字构成一维数组
5. 以字为key，该字出现的次数为value形成字典,按value从大到小排列{'不'：6000,'的'：5800,}
6. 对计数结果进行由大到小排序，返回的结果是一个数组，数组元素为(某个字，该字的次数),[('不'：6000),('的'：5800),]
7. 按照出现次数由大到小的顺序取出所有的字放到一个小括号中，字与字之间用逗号隔开,('不','的',)
8. 末尾加一个空格,('不','的', , ,' ')
9. 为每个字打上位置标签，从0开始，形成字典,高频次的字在前面{'不'：0,'的':1, , ,' ':6110}
10. 每首诗中的字都能在word\_int\_map中找到位置标签，poems\_vector是二维矩阵，行数为诗的个数，列代表每首诗的字的位置标签