HW2

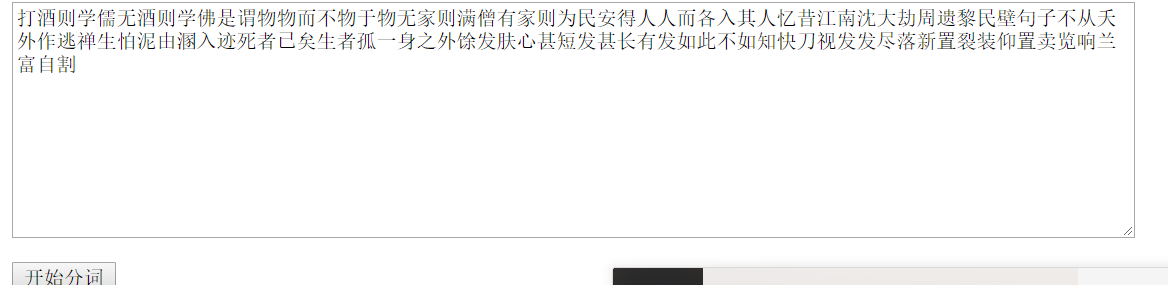
PB16080377聂雷海

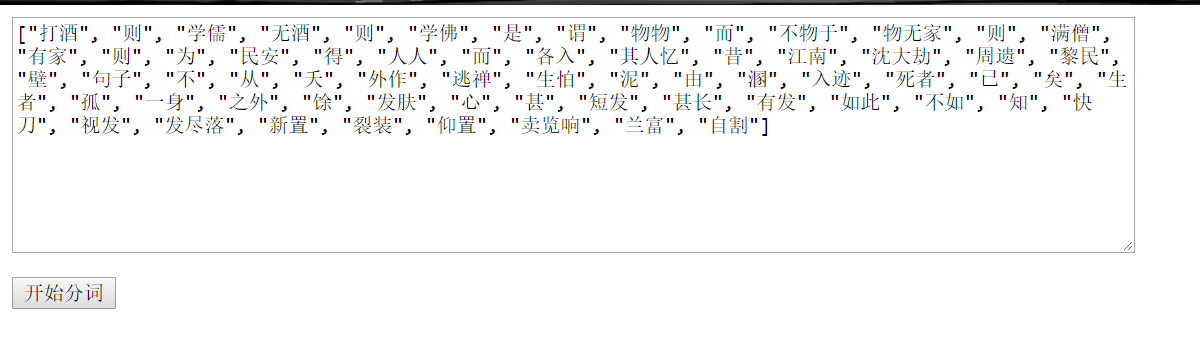
1. 下载或者在线测试一到两个开源的中文分词工具，测试一下其对古汉语和古诗词的分词表现，给出测试结果和截图。然后讨论一下古汉语和古诗词分词与现代汉语分词相比的难点主要是什么？

Answer:

作业采用课上提到的中文分词工具 结巴 分词的 c++ 版本。古汉语采用补勤诗存的一句话。

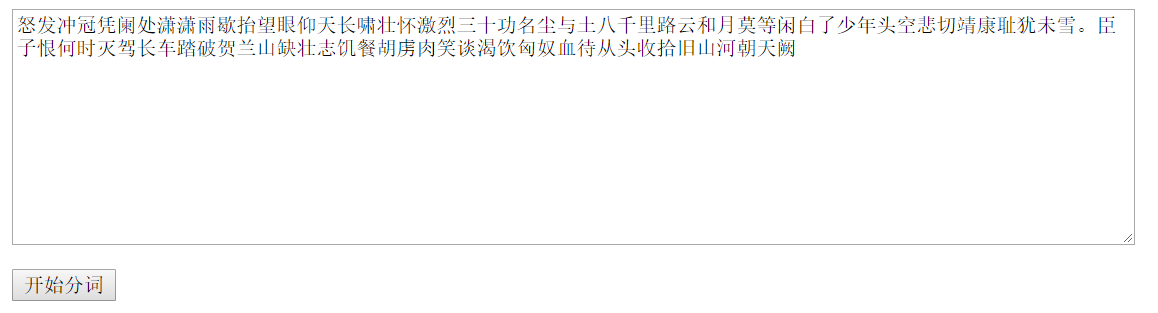
(url: <https://ctext.org/wiki.pl?if=en&chapter=857349&remap=gb>)

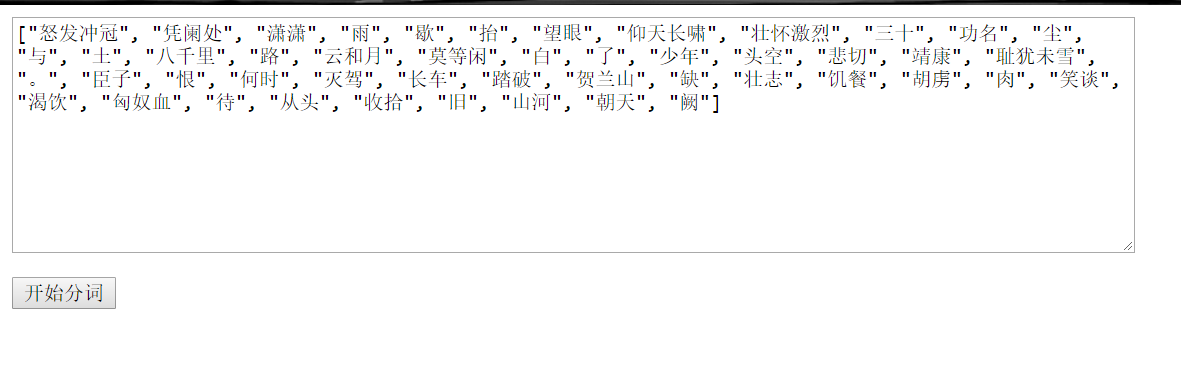




古诗词采用岳飞的《满江红》

(url: <https://baike.baidu.com/item/%E6%BB%A1%E6%B1%9F%E7%BA%A2%C2%B7%E6%80%92%E5%8F%91%E5%86%B2%E5%86%A0/8852069?fr=aladdin>)





从上面两个例子可以看出，判断古诗词总体尚可，但古汉语判断的确一般。

总的来说，分词对古文还是较弱。

在我看来，是现代汉语与古代汉语存在很大的不同。古代汉语一词多译（包括一字多译），对人来说，更是基于理解意思后方能正确断句。现代汉语句法功能确定，词类界限比较清晰，而古汉语则相反，并且语句多样，没有量词。最后，古汉语很多字、词是基于引申义的，倘若不结合语境，很难懂得分词。

2.假设词典中包括词｛王公，公子，研究，研究生，生命，起源｝以及所有单字集合，请分别给出句子“王公子在研究生命的起源”的FMM和BMM分词结果。

Answer:

FMM: 王公/子/在/研究生/命/的/起源

BMM: 王/公子/在/研究/生命/的/起源

3.下面的句子存在哪种类型的分词歧义？为什么？

a)吉林省长春药店

*1. 吉林 / 省长 / 春药 / 店*

*2. 吉林省长 / 春药 / 店*

*3. 吉林省 / 长春 / 药店*

answer:

1和2属于组合歧义，因为吉林，省长和吉林省长均为词典中的词。1和3属于交叉歧义，吉林，省长，春药，吉林省和长春均为词典中的词。2和3属于交叉歧义，吉林省长，吉林省，长春和春药均为词典中的词。

b)东北大学生联合会

*1. 东北 / 大学生 / 联合会*

*2. 东北大学 / 生 / 联合会*

*3. 东北 / 大学 / 生 / 联合会*

answer:

2和3是组合歧义，东北，大学，东北大学均为词典里的词。

1和3不是组合歧义和交叉歧义，大学，大学生均为词典里的词，但生 不是词典里的词。

1和2是交叉歧义，东北，大学生，东北大学，生都是词典里的词。

c)人大代表群体性事件

*1. 人大 / 代表 / 群体 / 性 / 事件*

*2. 人大代表 / 群体性事件*

*3. 人大代表 / 群体 / 性事 / 件*

*4. 人大代表 / 群体性 / 事件*

answer:

1和2属于组合歧义，人大，代表，人大代表，群体，性，事件，群体性事件都是词典里的词。

3和4属于交叉歧义，群体，性事，群体性，事件都是词典里的词。