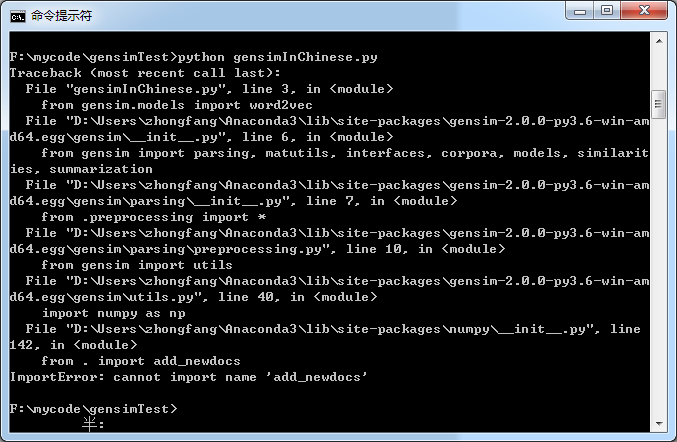
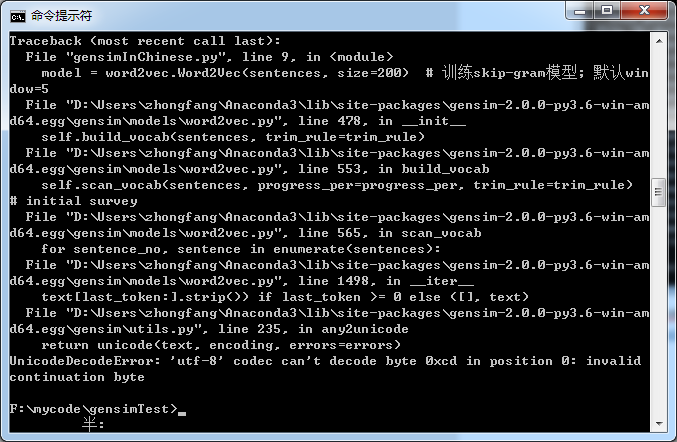
# Genism 中文分词与训练

## python cannot import name 'add\_newdocs'



根据stackoverflow <http://stackoverflow.com/questions/22465066/importerror-cannot-import-name-add-newdocs> ，如果import numpy都不成功的话，就要重装numpy

Genism utf8 编码问题



试了把python转成utf8 以及utf8+dom，都没成功

根据stackoverflow，<http://stackoverflow.com/questions/34427678/utf-8-decode-error-when-loading-a-word2vec-module>

model = gensim.models.Word2Vec.load\_word2vec\_format('/data5/momo-projects/user\_interest\_classification/code/word2vec/vectors\_groups\_1105.bin', binary=True, unicode\_errors='ignore')

sentences = word2vec.Text8Corpus(u"huxike\_seg.txt")

model = word2vec.Word2Vec(sentences, size=200, unicode\_errors='ignore')

改为



在gensim调用加入两个参数

看了一下，估计是要把分词结果存为utf-8

试过了，还是不行

<http://stackoverflow.com/questions/27659985/error-utf8-codec-cant-decode-byte-0x80-in-position-0-invalid-start-byte> 说要用load函数

gensim.models.Word2Vec.load(' xxxx’)

而且两句话合并成一句话了

额不太对，上面的是load已经训练好的模型

转UTF8 要点击“转为UTF”，而不是“以UTF编码”！！！

继续参考

<http://blog.csdn.net/churximi/article/details/51472300>

<http://blog.csdn.net/mebiuw/article/details/52303622>

<http://blog.csdn.net/lixintong1992/article/details/50387007>

<http://zake7749.github.io/2016/08/28/word2vec-with-gensim/>