代码运行帮助文档

ID:824858863

团队成员：张元晨（济南大学 2014级 材料科学与工程学院 电子材料系）

联系方式：<Tel:17862903881>

QQ:824858863

学号：20140121534

邮箱：[824858863@qq.com](mailto:824858863@qq.com)

指导老师：无

如果有严重问题希望及时联系我，代码纯手打，保证原创

文件结构：

-baiduthink:

-baiduread:

-code:

main\_tijiao.py

-DATA:

-jiaba2.txt(自定义词典，来源：http://www.jianshu.com/p/206caa232ded)

-n3605.txt(自定义词典，来源：http://www.jianshu.com/p/206caa232ded)

-weizhi1.txt(自定义词典，来源：http://www.jianshu.com/p/206caa232ded)

-out:

- outdata\_20w\_main\_tijiao\_8\_29.txt

我的运行环境：

Windows 10

Python 3.5.2(32位版)

部分代码如下所示：

\_\_author\_\_ = 'think\_zhangyuanchen'

#-\*- coding: UTF-8 -\*-

#原始核心文档

import jieba

import jieba.posseg as pseg

from math import sqrt

road\_2='E:/OneDrive/workspace/baiduthink'

注意事项：

1. 请注意Python版本
2. 根据文件夹所在位置修改代码运行路径，只修改以上代码中的红色背景部分即可
3. 请导入程序运行所依赖的包(结巴分词下载： <https://github.com/fxsjy/jieba/tree/jieba3k>)：

import jieba

import string

import sys

import re

import jieba.posseg as pseg

from math import sqrt

1. 代码运行时如果出现不支持中文错误请将代码中的中文注释删去

主要思路：

1. 首先清洗训练数据以及测试数据集，将Python不支持的字符除去
2. 模型训练：

将核心实体附近的词记录到字典中，并记录出现频率，获得与实体相关的特征

1. 模型输出

将测试集中的每一句话进行分词，并计算每一词可能是核心实体的概率，输出最可能的核心实体