**程序说明**

环境：python: 3.6.5，jieba: 0.39，gensim: 3.6.0

1. 首先需要下载搜狗实验室的新闻语料，因为文件比较大，我在提交的代码中删除了，下载地址：https://www.sogou.com/labs/resource/ca.php，其中的1.02G的完整版。然后将里面的txt文件解压到sougouca文件夹，并运行sougouca文件夹中的get\_sougou\_file.py，这时会在上级目录生成一个大概2.65G的sougou.txt文件。
2. 运行get\_jieba\_distinguting.py，得到DL散度排序结果。
3. 再运行get\_jieba\_result.py，得到train\_jieba\_result\_not\_stop\_word.txt，用来求解tf-idf的结果。
4. 运行best/jieba\_tf\_idf.py的得到tf-idf结果，结果为submit\_ubuntu.csv文件，此文件为tf-idf结果，内容为20000个词。
5. 运行get\_jieba\_result\_for\_entropy.py得到entropy\_result.csv文件。
6. 下面一个模型是我队友做的，在Textrank\_code文件夹里，里面有程序说明文档，根据文档，将最终的结果Article4\_sent3\_2.5.2\_3.5.2\_key3.csv复制到best文件下即可。
7. 然后运行best/get\_mixed\_of\_three\_models.py融合三个模型的结果，就可以得到三个模型的交集结果submit\_merge.csv，这是含有二千多个词的好结果。
8. 然后为了得到种子词，我们将上面生成的submit\_merge.csv与收集的一些词典dic\_main.txt求交集，即运行get\_seed\_words.py得到seeds.csv。
9. 然后我们从刚才三个模型的结果中抽取20000个词，作为候选词集，即运行get\_candidate\_word.py，生成含有20000个词的candidate\_words.csv。
10. 最后使用种子词集seeds.csv和candidate\_words.csv通过相似度方法求解候选词集中的好词，即运行./word2vec\_get\_sort.py得到submit\_seed\_to\_word2vec\_result.
11. 最后运行best/get\_finnal\_result.py融合submit\_merge.csv和submit\_seed\_to\_word2vec\_result得到最终结果。

时间比较紧，代码可能比较乱，敬请谅解。可能有一些超参和之后设置有所不同而没有改过来，所以如有问题请随时联系我。

姓名：刘辉

电话：18801220652

邮箱：liuhui@iie.ac.cn