**https://dc.cloud.alipay.com/index#/home**

**2018蚂蚁金服自然语言处理比赛心得（ATEC）**

2018年07月11日 22:08:45 阅读数：422 标签： [NLP](http://so.csdn.net/so/search/s.do?q=NLP&t=blog) [相似度](http://so.csdn.net/so/search/s.do?q=相似度&t=blog) [自然语言处理](http://so.csdn.net/so/search/s.do?q=自然语言处理&t=blog) 更多

个人分类： [NLP](https://blog.csdn.net/mengchunyanjx/article/category/7795715)

碰巧在微信公众号看到蚂蚁金服举办的金融NLP，自己想着去试试，最后遗憾没进前100，复赛资格都没有~忧伤。

1、官方分2次给数据集，第一次是3.6万条，这些数据作为A榜训练集；

2、第二次是6.4万条，同时开启B榜的排名，A榜关闭；

赛题分析：

该赛题是处理客服机器人的对话系统，判断自己的相似度，同义为1，不同义为0.

1、数据是格式为：行号\t句子1\t句子2\t标签；

2、全部是中文句子，该领域属于垂直领域，有许多的自定义词语；

3、分词采用结巴，先用哈工大的停用词，先不够造个人专属词典。

4、采用简单的朴素贝叶斯多项式，将问题看做文本二分类；

初步测试结果：得分0.68，还算可以，奋斗

5、想继续提高，就需要大量精力来做数据预处理，官方给的标签准确率为96%，错别字不少，词语前后顺序不一致的也不少，对话句子属于短文本，前后顺序先不考虑影响。

6、坑爹的问题在后面，，正负样本差距太大。3.6万条预料，竟然只有7000千多的正样本。尴尬。这就头大了，我的理解是需要构造正样本，确保样本均匀，不知道各位有何看法？？

7、没时间去构造了，直接喂数据，这次加了用户词典，通用词也用数据处理后的，效果提高了一丢丢。。大哭。

工作项目忙得不行~~再来看，B榜开始了。。

8、A榜的模型在B榜只跑出0.23，，WHAT??

9、又多了6.4万条预料 ，这数据处理一个人看起来都怕，，有点没信心了。。

10、下班回来继续整数据处理，NLP工程师70%的时间都在干数据处理吧。。。

11、这次采用RNN-LSTM嵌词方法。先分词好，用gensim的word2vec训练词向量，维度是300维；

12、将数据分词8:2，8作为训练集，2作为测试机，训练中设置交叉验证；

13、句长度取15，训练之后，预测，看样子有那么一点点提高，预测分数0.32;

14、感觉数据还是太不干净，又没有太多人力去继续整理了，加attention试试，

15、训练完成，跑分，0.41，，算了，不整了，也没时间了，自己技术还是太渣。。

各位，如果看到，还麻烦留言，我期待你们的好方法，那几个0.6+的，太厉害了，

请教几个问题：

1、如何有效快速的做数据预处理？

2、有朋友用CNN都能做到0.6+，可否讲讲？

谢谢~2018.7.11