# 工作日志-姬浩迪

2019年7月8日星期一

研究使用了下hanLP的分词算法，效果很不错，基本能够满足我们的需求，后续开发中可以直接使用来分析文本。

然后看了下之前组员在文本情感分析上提出的两种方案：

1. 建立情感词典用算法来分析。词典已经找好，问题在于大部分的词典都是按照褒义词、贬义词、中性词来分的，如果按照我们之前的设想分为喜怒哀惧4种等级的话，可能需要人工给词典添加等级权重，工作量是很大的。还有的难点在于算法的实现，包括词典的存取，对句子分词(这个可以做到了已经)，按照分词给每个句子计算感情值，加总文章的总权值。
2. 通过机器学习构建模型，自动分析文本感情。问题在于数据集难以满足要求，喜怒哀惧区分的数据集网上基本找不到，只有正面，负面的数据集。不过通过简单测试hanLP里自有的感情分析工具，能够学习简单的数据集做比较弱的判断，虽然学习效果不好，但是问题就简化成搜集数据集了，大大降低了开发难度。权衡之下两者取这个的可能性较大。

假设用第二种方法，由于细分好的数据集难找，而正负面的数据集却很多，要么考虑把现有数据集组内分工人工匹配我们设定好的等级，可行度较低，因为要训练比较好的机器学习数据量实在太大。要么考虑降低制作质量，能够达到很好的训练水平，缺点是分析范围就窄了许多，从原来的喜怒哀惧变成只有积极消极，但降低了难度，可行性较高，并且甚至有情感分析API可以直接用。具体怎么做还待定…

写了第一次迭代部分的测试报告和用例。