点评情感分析算法说明

1.情感分析介绍

1.1 情感分析的目的

目前所上线的点评标签不含有情感信息（如图一所示），为用户所提供的信息量非常有限，而且有可能会对用户产生误导，所以我们需要对标签标注情感信息，也就是说具体的分析出该标签是正面的还是负面的，从而更好的展示出来。

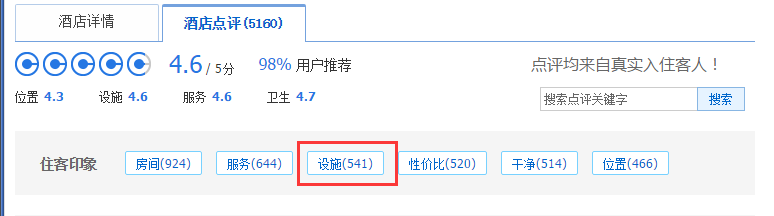
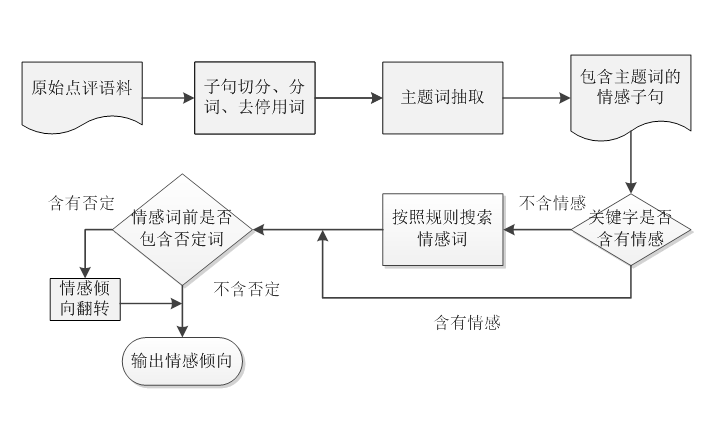


图1.目前的点评标签

1.3核心思想

根据创建的情感语料库，在关键字前后开窗口，利用位置信息得到情感倾向

1.2 情感分析的流程图



1.3 情感分析流程说明

1）从Hive表Olap\_HtlbiModelDB.FactCommHotelValidComment抽取点评

2）分词去停用词，分词工具：jieba分词，词典使用se词典+所有情感词

3）主题词抽取，将分词后结果与已知的主题词表进行匹配

4）使用正则的方法将包含主题词的子句提取出来

5）以主题词为中心，将子句划分为窗口（窗口大小可自定义），判断主题词是否含有情感，如果主题词包含情感，比如，物美价廉，则直接查找前面窗口中是否含有否定词，如果不含否定词则直接输出该主题词的情感，否则将该主题词的情感反转并输出。

5）在窗口中匹配情感词，情感词的权重由句子的左到右依次提高，并选取最高句子权重的情感词作为该子句的情感词。

1.4 情感分析的应用

1）为点评标签提供情感属性，可以判断判断用户该条点评对该主题的情感倾向是正面的还是负面的。

2）为点评转义提供技术基础，点评转义可以根据主题与情感查找相似点评并进行替换。

2.情感分析细节说明

1. 需要人工check关键字的情感倾向并进行分类。
2. 将与主题词共现非常高（阈值与Top数都是可控的）的词人工check并加入情感词典。
3. 动态情感词问题，使用规则词典定义动态情感词，例如当主题词为“价格”，情感词为“高”的时候则情感倾向是需要反转的。
4. 关键字需要人工check并赋默认情感值，即在子句中没有找到情感信息的时候，就使用默认情感值。

3.情感分析测试结果

1）点评句级别测试效果：对5万句点评进行情感分析，按原始比例抽取1000句进行人工标注再与模型的预测结果比较，最终得到平均准确率为95%，召回率90%。

2）酒店标签级别测试效果：上海、丽江、三亚，每个城市在点评数为10-50、 50-100、100-500、 500-1000的酒店中每个阶段各取20家酒店 总共240家酒店进行情感倾向分析并打标签，若某个标签同时满足以下条件则打上标签（正面点评数-负面点评数\*3）/总点评数>threshold1（10%）& 该标签的正面点评数>=threshold2(为5条) 。最终，打上标签的酒店共214家。