评论标签提取

赵海臣

背景

- 商品评论页需要聚合出一些褒义类型的评论标签集合，让用户浏览评论时能够方便看到褒义类型的标签。例如浏览“欧莱雅男士洗面奶”的评论区，评论已经聚合出“东西不错”，“效果很好”，“清洁力高”，“气味不错”，“男朋友觉得不错”，“性价比高”等标签，点击即可展开相应的评论。
- 让用户能够更方便浏览众多的评论，提前将有利的信息更多展现给用户。

步骤

- 评论标签提取步骤
  - 评论数据处理
    - 将一个长评论按照标点、空白符分割成小短句
      - 一个小短句往往体现了一个评论的一个情感，例如“可莱丝丽得姿的面膜火到爆~确实很好用，很多人都在推荐，因为有个朋友一个月会去一回韩国，所以小冉经常拖朋友稍~我想说的是聚美跟朋友带回来的真的完全一样，乐天可莱丝一般是有活动什么3+1,就是买的越多越便宜，最重的是用在脸上，脸上啊，保真比较重要~小美的价格也很划算了~”这个评论中，有很多关键词，往往是在单个短句中。
    - 对小短句继续进行分词以及词性标注
  - 词语级别处理与拼接
    - 对短句按词从后往前遍历
      - 否定单词表，若遍历触及则停止，并弃用该短句
      - 单词同义词表，遇到同义词词表中的词，则替换成标准词，以浓缩语义
      - 首次遇到形容词＂ a ＂，保存该形容词
      - 在有形容词的情况下，遇到名词＂ n ＂，保存该名词，并停止遍历
      - 保存“n”+”a”词对，并拼接作为标签词
      - 若短句不满足“n”+”a”,则弃用该短句

步骤

- 评论标签提取步骤(续)
  - 拼接后”n”+”a”标签词的处理
    - 否定标签词表，遇到否定标签词则弃用
    - 标签词同义词表，遇到标签词同义词词表中的词，则替换成标准标签词，以浓缩语义。
  - 标签词降噪处理：我们认为出现频率低的标签词多半是有问题的标签词，放弃这些问题概率大的标签词能够
    - 按category_v3_3统计标签词
      - 若标签词在category_v3_3总体中标签词中出现的次数比例小于0.1%，则过滤掉
      - 若标签词在category_v3_3中出现次数小于10次，则过滤掉
  - 商品标签数量限制
    - 按照对应商品下评论的标签词频率从高到低，取该商品频率最高的8个标签词作展示标签。

词表

- 核心体力活在于4个词表的建立：
  - 同义单词词表：
    - 好 =>不错
    - 很好 => 不错
    - 好用 => 不错
    - ...
  - 否定词词表：
    - 不好
    - 难用
    - 垃圾
    - ...
  - 标签词(n+a)同义词表：
    - 小瓶不错 => 携带方便
    - 小巧不错 => 携带方便
    - 小巧方便 => 携带方便
    - ...
  - 否定标签词(n+a):
    - 质地稀
    - 物流久
    - 味很大
    - ...

THE END

THANK YOU!