读书笔记（四）

黄怀宇

2021 年 1 月 22 日

1 各个工具的特点和原理(续)

1.1 doc2vec

1.2 word2vec

1.3 one-hotm

1.4 BOW

1.5 CBOW

1.6 Skip-Gram

2 工作进度

2.1 前后端数据传输

数据初步处理后利用 CORS 实现跨域数据传输。想通过 node.js 和 cors 通过 POST 传输数据实现。

2.2 基于业务特点的设计

微博的评论是随着页面滚动而加载出来的，数据是实时生成的，服务端的 js 脚本应该实时增量式地将文本发给服务器处理，服务器根据已经训练好的算法实时返回（判断刚加载出来的评论是否应该删除）。

中途不需要人工操作。

2.3 服务端算法模型训练

需要爬取尽可能多的微博评论，进行无监督聚类，将训练好了之后的模型用于判断新来的评论属于哪个类别。

2.4 第二阶段再做

加上人工优化，持续反馈

2.5 还需做的

1. 预先加载评论好(滚动滑条就会加载很多内容，或其它方法)再返回，因为大部分都被删除了。脚本优化。

2. 爬取微博尽可能多的评论，选择合适的算法组合进行聚类。

2.6 项目难点

参考文献