

期末报告

1. 背景介绍

司马南，独立学者，社会评论家，中国反伪科学代表人物。20 世纪末因揭穿神功骗子而饮誉江湖；21 世纪，因捍卫中国政治制度而再掀风潮。杜子建，作家、企业家、短视频研究者，因最早洞察并提出“微博是可以做营销的”，且雷厉风行地将旗下的选美公司就地改为中国第一家微博营销公司，杜子建又被行业人尊誉为“微博营销教父”。

本工作分析司马南和杜子建微博内容以及其粉丝转发评论内容，试图从中刻画两位名人的用户画像及其粉丝用户的用户画像，并观察二者粉丝用户的区别。

2. 算法介绍

- 本工作选取‘text’（文本）、‘user_description’（用户描述）、‘verified_reason’（认证原因）、‘city’（城市）和‘gender’（性别）作为我们的分析变量。
- 本工作使用 LDA 主题模型对所有司马南和杜子建的微博文本进行主题分类，并进行可视化。
- 通过对所有司马南和杜子建的微博文本进行 jieba 分词并去除停用词，我们绘制出了司马南和杜子建的微博词云图。
- 本工作对司马南和杜子建的微博粉丝性别比例，认证原因和所在城市进行统计绘制柱状图

3. 可视化分析

通过对 LDA 主题模型分类结果进行归纳，本工作将司马南的微博分为 5 个主题，分别为：政治评论、观点表达、讨论交流、练习书法和分享观点。通过这五个主题，我们可以看出司马南的微博刻画出他富有文人气息的公知形象。同时对司马南的微博绘制词云图可以看出司马南关注中美关系，喜欢在微博上发表政治观点，这与他作为社会评论家的身份相吻合。此外他还关心转基因相关问题，这与他作为中国反伪科学代表人物的背景相吻合。

同理，本工作将杜子建的微博分为 3 个主题，分别为：情绪表达、观点表达、讨论交流。可以看出相较于司马南的“公知”形象，杜子建的微博主题较为生活化。同时对杜子建的微博绘制词云图可以看出杜子建经常在微博表达自己对他人的赞美感谢之情，经常谈论“酒”，“市场”等话题，这与他作为企业家和商人的身份相吻合。

最后我们分析两位名人的粉丝用户。杜子建的女性用户比重高于司马南，本工作猜测原因是相较于男性，女性对政治的关注度或者讨论度较低。我们可以将认证原因视为用户职业。从结果可以看出，司马南的粉丝职业比较多元化，而杜子建的粉丝主要来自商业界，这也与杜子建的商人身份相呼应。从粉丝所属省份来看，司马南的粉丝在北上广的较多，本工作推测原因在于司马南作为公知，关注他的粉丝的学历也应该普遍较高，而北上广的平均学历水平较高；杜子建的粉丝有来自河南省的，本工作推测原因在于杜子建本身出身草根家庭，比较容易得到同样来自经济不发达地区人群的关注。