2020/02/26

2019传播数据挖掘竞赛优秀作品分析

冠军：网络良性越轨行为研究——以“昆山龙哥反杀案”为例

<https://mp.weixin.qq.com/s/Go2qoM4tzqp1L8h3UkP_-w>

亚军：“昆山反杀案”中微博用户的情绪传播分析<https://weibo.com/ttarticle/p/show?id=2309404397264108912931>

亚军：情感联盟：养成系偶像与粉丝群体关系研究

<https://mp.weixin.qq.com/s/Mm5gZGtXo87p4O1LvY9V3w>

季军：从大数据看基于受众的国产电影社会化营销

<https://mp.weixin.qq.com/s/PCo3-w5NpSHjWUjD8drRuw>

季军：《创造101》中哪位导师最受欢迎？

<https://mp.weixin.qq.com/s/pes-xpFaP3mLstPLsMYRCQ>

季军：网络舆论传播的空间特征研究

<https://mp.weixin.qq.com/s/cLqImFsh-vGXzGYYQ7e-bQ>

# 报告书/ppt框架

## 冠军 -“昆山龙哥反杀案”分析报告 （文字报告书，与APA Style 高度相似）

### 摘要 Abstract

### 引言 Information

### 问题的提出与分析

### 研究过程与方法 Method & Measurement

### 研究发现 Result

### 结论与讨论 Conclusion & Discussion

### \* 附录 Appendix（按所解决问题进行的分类）

#### 对各类型的主题的标注框架

#### 频数分布（数据可视化）

#### 所分析事件的过程表

#### 时间切片

#### 变量top10微博个体

#### 变量相关性分析（数据可视化）

#### 基于时间切片上的各类数据截图

## 亚军 -“昆山反杀案”中微博用户的情绪传播分析 （文字报告书）

### 摘要

### 研究背景与问题的提出

### 研究思路与方法

### 数据处理过程

### 研究发现与结果

### 讨论与结论

### \* 附录 （按具体类型进行的分类）

#### 情感分类+分析流程图(？)

#### 简单的各类数据可视化

#### 参考文献（23篇）

## 亚军 - 情感联盟：养成系偶像与粉丝群体关系研究（ppt）

### 研究背景（包含数据可视化）

### 数据挖掘过程

### 研究发现（包含数据可视化）

### 结果讨论

## 季军 - 从大数据看基于受众的国产电影社会化营销（ppt）

### 技术概况与技术细节

### 微博传播解读（包含数据可视化）

### 大数据视阈下的社会化营销（包含数据可视化）

## 季军 - 《创造101》中哪位导师最受欢迎？（ppt）

### 背景介绍

### 研究假设

### 分析方法

### 分析结果（包含大量数据可视化）

### 研究结论

### 反思总结

## 季军 -网络舆论传播的空间特征研究（ppt）

### 研究背景

### 研究问题

### 研究方法

### 研究发现（包含数据可视化）s

### 总结反思

# 包含的相关数据挖掘处理技术，衡量标准和理论基础

## 冠军 -“昆山龙哥反杀案”分析报告

### LDA主题模型, python的gensim包中的ldamodel函数，把文档—词汇矩阵变成文档—主题矩阵（分布）和主题—词汇矩阵（分布）

#### perplex指标:

#### 语言模型（Language Model）评价指标，PPL是用在自然语言处理领域(NLP)中，衡量语言模型好坏的指标。它主要是根据每个词来估计一句话出现的概率，并用句子长度作标准化。

#### spss软件中的α信度系数法:

#### 克朗巴哈系数（Cronbach's α）是一个统计量，是指量表所有可能的项目划分方法的得到的折半信度系数的平均值，是最常用的信度测量方法。

### DTM（动态主题模型）， gensim库中的LdaSeqModel模型，该模型捕获了顺序组织的文档语料库中主题的演变。

#### Jaccard距离：

#### 杰卡德距离，Jaccard相似指数用来度量两个集合之间的相似性，它被定义为两个集合交集的元素个数除以并集的元素个数。

## 亚军 -“昆山反杀案”中微博用户的情绪传播分析

### 粗粒度和细粒度情感分析：

### 粗粒度主要是基于篇章级和句子级，细粒度一般指词汇级情感分析。

### 网络科学方法（ network?）：

### 将微博系统中用户转发关系抽象成网络科学中的节点和连边，进而构建用户情感传染网络以及单条微博的传播级联树。

### **情感挖掘方法与技术：**

### Python语言中Pandas框架，Matplotlib和Pyecharts数据可视化，基于Anaconda软件中的Spyder进行代码的编写和调试。

## 亚军 - 情感联盟：养成系偶像与粉丝群体关系研究

### Gephi 是一款网络分析领域的数据可视化处理软件。<https://gephi.org/>

#### Force Atlas & Yifan Hu算法布局：Force Atlas为力引导布局，能够产生相当优美的网络布局，并充分展现网络的整体结构及其自同构特征；Yifan Hu多水平布局适用于非常大的图形，特点是粗化图形，减少计算量，运行速度比较快。

### NLPIR汉语分词系统：<http://ictclas.nlpir.org/>

### 主要功能包括中文分词；英文分词；词性标注；命名实体识别；新词识别；关键词提取；支持用户专业词典与微博分析。

### 霍斯提（Holsti）公式：

### 假定两个编码员分别同时做了m 1 和m 2 个单位的编码 , 其中一致的编码数为m，则编码者间信度 = 2m/(m1+m2)。

### spss数据线性回归分析

### Python调取SnowNLP中文情感分析包

## 季军 - 从大数据看基于受众的国产电影社会化营销

### 中文分词：HMM模型

### 隐马尔可夫模型(Hidden Markov model, HMM)是关于时序的概率模型，描述由一个隐藏的马尔可夫链随机生成不可观测的状态随机序列，再由各个状态生成一个可观测的随机序列的过程。

### Viterbi算法

### Viterbi算法其实就是多步骤每步多选择模型的最优选择问题，依次计算完所有步骤后，通过回溯的方法找到最优选择路径。

### 开源框架Jieba分词 <https://github.com/fxsjy/jieba>

### TF-IDF算法

### TF-IDF是一种统计方法，用以评估一字词对于一个文件集或一个语料库中的其中一份文件的重要程度。

### 聚类算法

### D3.js <https://github.com/d3/d3/wiki/Tutorials>

### D3是这样一个特殊的JavaScript库，它利用现有的Web标准，通过更简单的（数据驱动）方式来制作炫目的可视化效果。

## 季军 - 《创造101》中哪位导师最受欢迎？

### “使用与满足”理论，镜像互动与自我投射 ，视听符号

### KeyCounter聚合

### SnowNLP的sentiments功能 <https://github.com/isnowfy/snownlp>

## 季军 - 网络舆论传播的空间特征研究

### 流动空间理论

### Pandas，Matplotlib和Gepi

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 冠军 | 亚军 | | 季军 | | |
| 网络良性越轨行为研究——以“昆山龙哥反杀案”为例 | “昆山反杀案”中微博用户的情绪传播分析 | 情感联盟：养成系偶像与粉丝群体关系研究 | 从大数据看基于受众的国产电影社会化营销 | 《创造101》中哪位导师最受欢迎？ | 网络舆论传播的空间特征研究 |
| 报告形式与框架 | 报告书（四颗心）❤❤❤❤ | 报告书（三颗心）❤❤❤ | ppt（两颗半心）❤❤\* | ppt（两颗心）❤❤ | ppt（两颗半心）❤❤\* | ppt（三颗心）❤❤❤ |
| \*说明 | 框架经典规范，个人建议增加参考文献部分 | 框架规范有逻辑 | 框架合理有序，但存在一定数量不相关元素，态度显得稍为随意 | 框架随意不够系统全面 | 目录部分的分类成熟，但实际实施时混乱 | 框架规范有逻辑 |
| 内容类型 | 数据挖掘处理衡量标准（四颗心）❤❤❤❤ | 数据挖掘处理理论基础（三颗心）❤❤❤ | 数据挖掘处理衡量标准（三颗心）❤❤❤ | 数据挖掘处理（两颗半心）❤❤\* | 数据挖掘处理理论基础（两颗心）❤❤ | 数据挖掘处理理论基础（两颗星）❤❤ |
| \*说明 | 运用多样成熟的数据处理方式与高认可度衡量标准 | 运用多样数据处理方式且引入理论支撑 | 运用外部软件协助数据处理，但提及的衡量标准较冷门 | 运用多种程序语言处理数据，但缺乏相关理论支撑和衡量标准 | 较少的数据处理方式与较充分理论支撑 | 较少的数据处理方式与理论支撑 |
| 亮点 | 报告形式成熟，文案分析角度多元，行文充实严谨，大量的数据处理过程与结果 | 有大量理论依据支撑，报告流程清晰 | 数据处理过程有分类且说明充分 | 运用的数据处理方式多样且例举清晰，ppt布局合理并配有充足的文字解释数据 | 数据处理过程仔细且清晰呈现 | ppt框架清晰，研究角度和问题具有理论支撑，研究分析有条理 |
| 不足 | 并未引用相关文献理论基础，数据呈现上多是平淡的表格形式，可视化部分偏少 | 附录中出现不必要因素，数据可视化效果有待提升 | 报告中插入许多过于随意的配图，数据文字分析过于简单，可视化效果有待提升 | 报告框架问题大且缺少理论支撑 | 报告前后条理没有对应，过多分页，放置过多的数据图而未加文字说明，数据可视化效果过于单一平淡 | 数据可视化较单一 |