



**[作业3]B站弹幕LDA主题调优**

**课程名称 文本信息挖掘概论**

**学生学院 计算机学院**

**专业班 18级软件工程2班**

**学 号 3118005002**

**学生姓名 胡梓泽**

**指导老师 杨易杨**

**日期：2021年6月10日**

目录

[1.数据集说明 4](#_Toc75768372)

[2.LDA主题数可视化调优 4](#_Toc75768373)

[2.1 可视化用到的工具 4](#_Toc75768374)

[2.2 参数准备 4](#_Toc75768375)

[2.3 数据集最佳LDA主题数的确定 4](#_Toc75768376)

[2.3.1 LDA主题数：2 5](#_Toc75768377)

[2.3.2 LDA主题数：3 5](#_Toc75768378)

[2.3.3 LDA主题数：4 6](#_Toc75768379)

[2.3.4 LDA主题数：5 7](#_Toc75768380)

[2.3.5 LDA主题数：6 7](#_Toc75768381)

[2.4 结论：4或5个LDA主题数 8](#_Toc75768382)

[3. UMAP降维过程展示 8](#_Toc75768383)

[3..1 第一步：没有标签的数据展示 8](#_Toc75768384)

[3.2 弹幕数据分类 8](#_Toc75768385)

[3.3 各分类成果展示 9](#_Toc75768386)

[3.4 something more 10](#_Toc75768387)

[附录 10](#_Toc75768388)

# 1.数据集说明

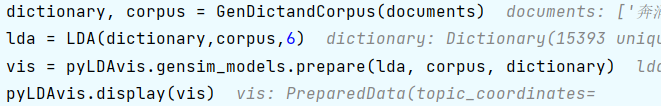
采用作业一二的数据，即经过去掉停用词后的语句。

|  |
| --- |
| 奔涌 后浪 |
| 百大 |
| 愿 青年人 都 往上走 摆脱 冷气 一分 热 发一分光 |
| ………… |

# 2.LDA主题数可视化调优

## 2.1 可视化用到的工具

采用pyLDAvis可视化主题



## 2.2 参数准备

①指定的LDA主题数

②dictionary

③corpus

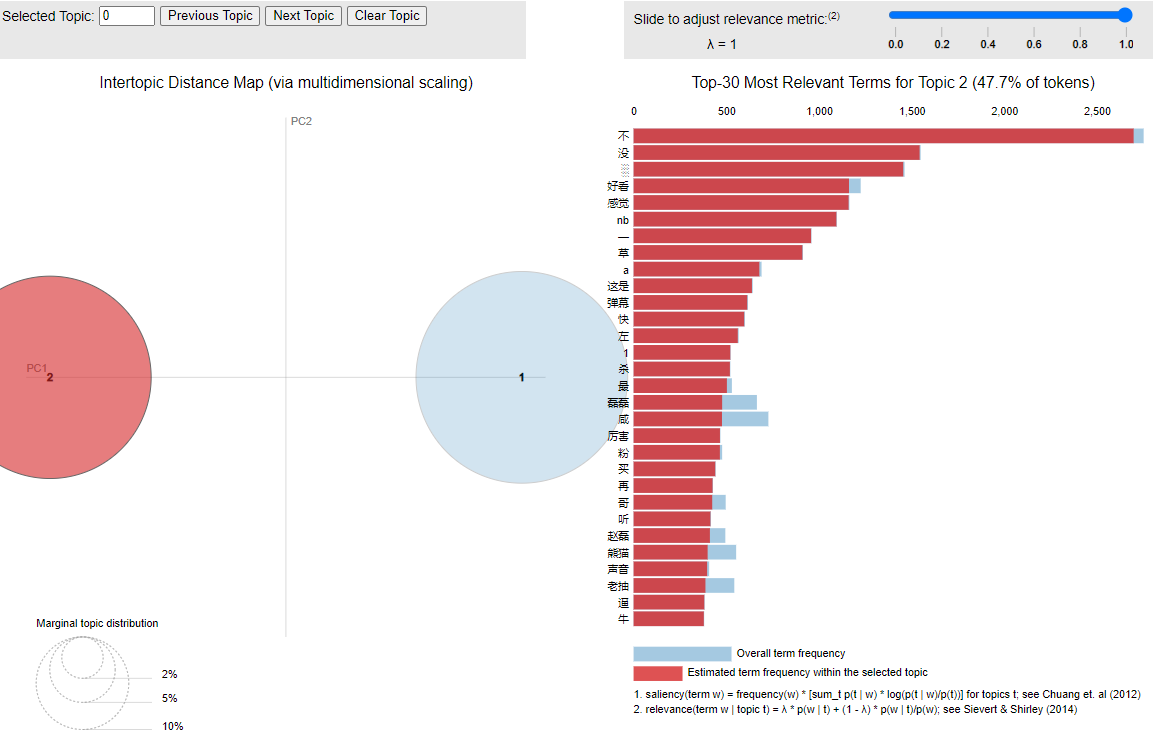
④LDA训练的model

## 2.3 数据集最佳LDA主题数的确定

LDA主题数及其对应前10关键词见附录。

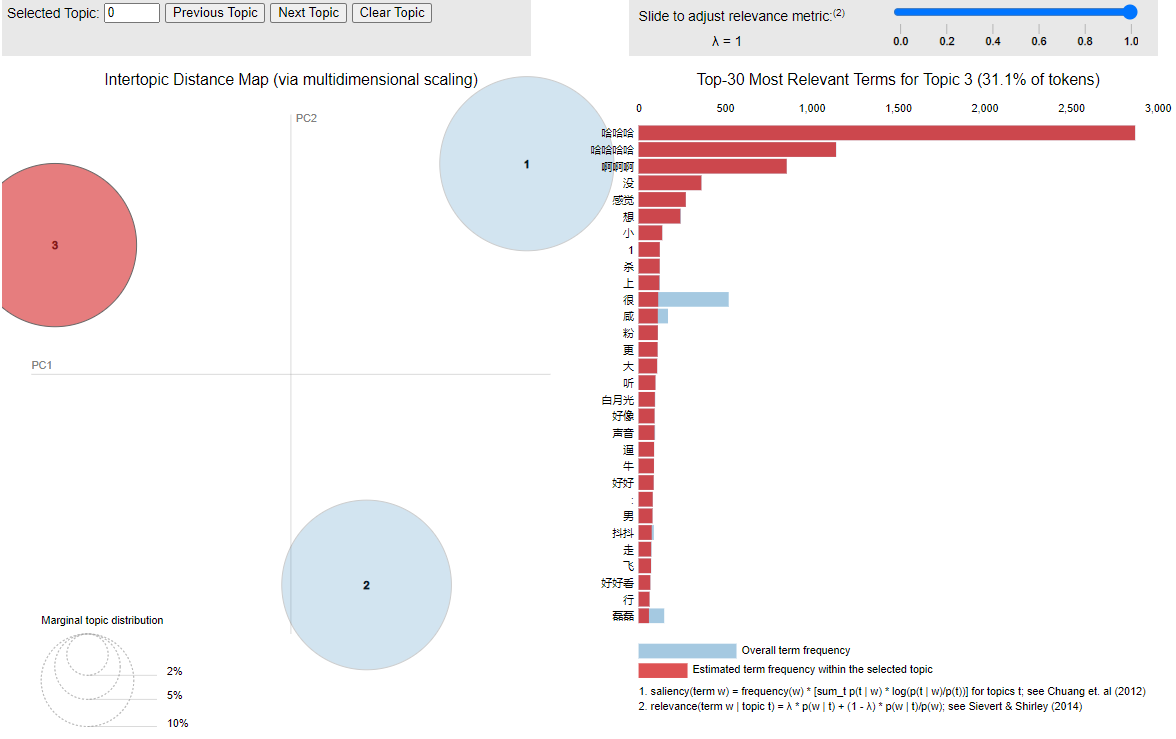
### 2.3.1 LDA主题数：2

如下图，可见分类之间仍有间隙。需要继续增加观察。



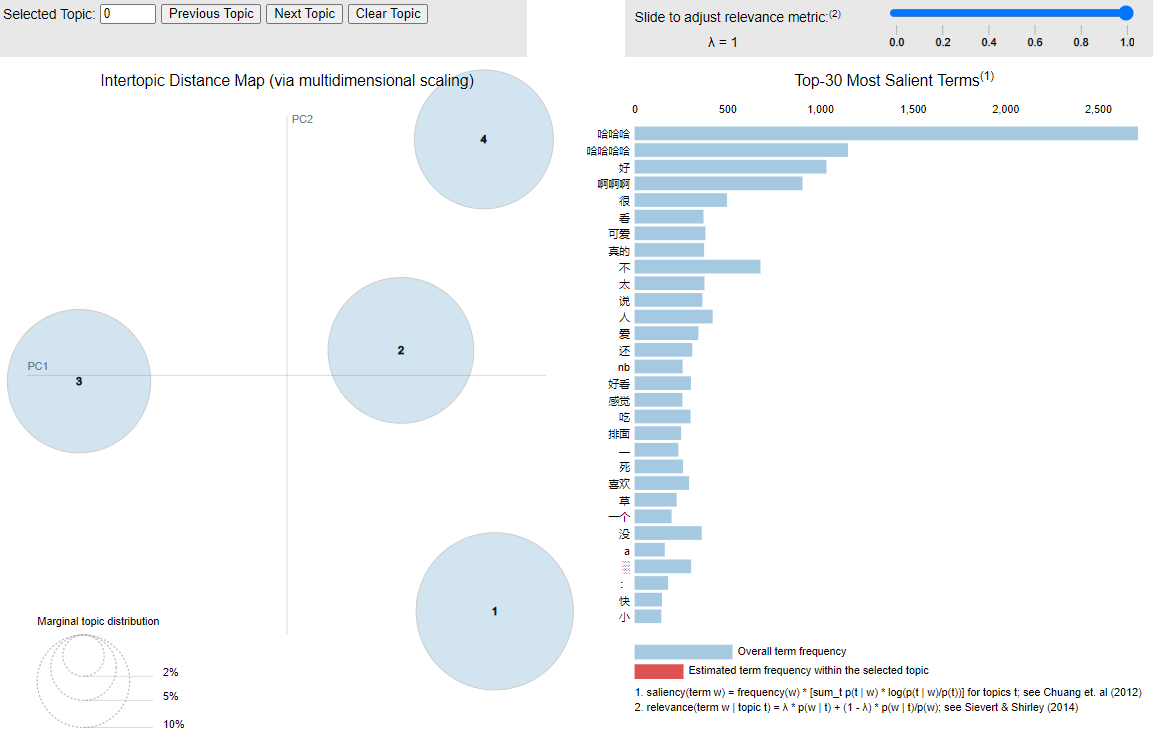
### 2.3.2 LDA主题数：3

如下图，可见分类之间仍有间隙。需要继续增加观察。



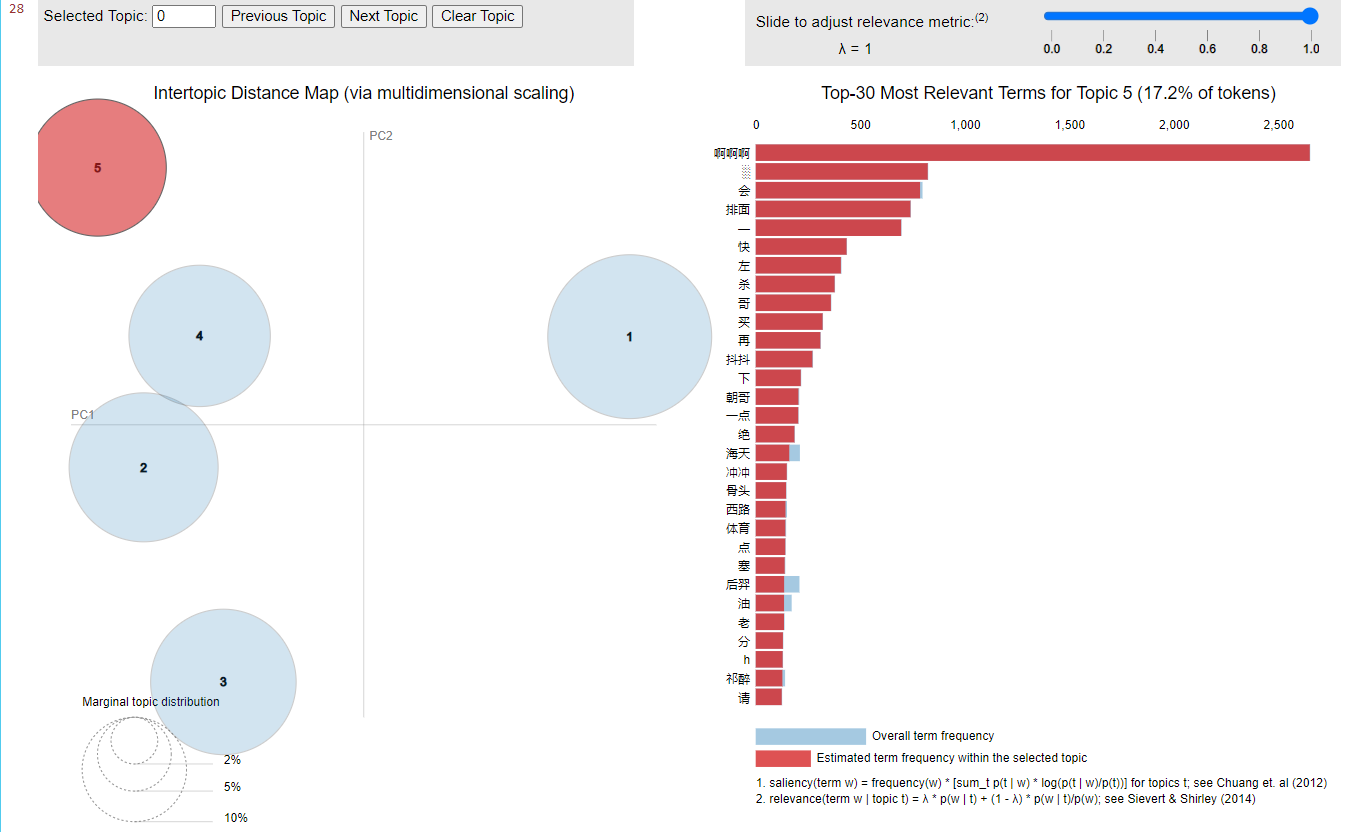
### 2.3.3 LDA主题数：4

如下图，可见分类之间虽然仍有间隙，但相距离并不大。继续增加观察。



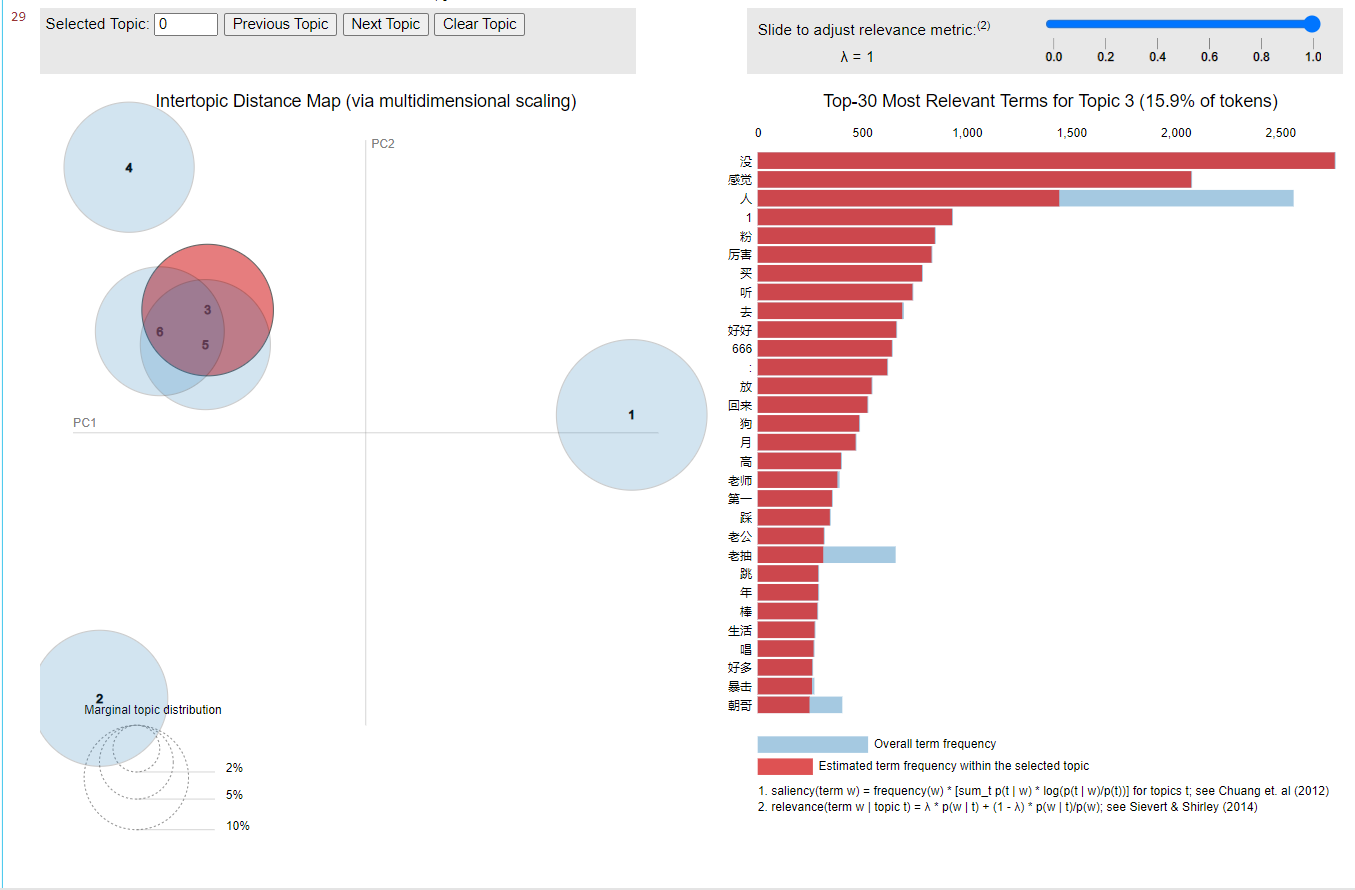
### 2.3.4 LDA主题数：5

如下图，有的主题分类间已经有重合，继续增加观察。



### 2.3.5 LDA主题数：6

如图，主题分类间已严重重合，停止观察。



## 2.4 结论：4或5个LDA主题数

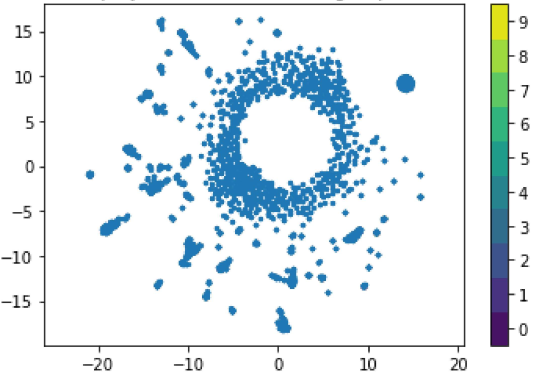
从2.3中的2-6个LDA主题数时的展示图中，可得应选取4或5个作为目标LDA主题数。

# 3. UMAP降维过程展示

## 3..1 第一步：没有标签的数据展示

在2.4中可知标签应有4或5个。如下为打上标签前的数据，经过Umap降维后的二维图。

如图可知，下一步的工作是需要给每一条弹幕纪录打上标签。



## 3.2 弹幕数据分类

分类的方法有很多，在此我将分类的依据判断如下：

①确定LDA主题数

②获取当前主题数下关键词的权重

③对于每条纪录，获取它在每个分类下的权重和，取最高的作为分类结果

## 3.3 各分类成果展示

详细放大图，请见根目录下的图片集。

由以下LDA主题数对应的打上标签的UMAP降维图，可得4个主题数下的颜色分类更加契合结果。

|  |  |
| --- | --- |
| C:\Users\Mr.hu\Desktop\文本挖掘作业\作业三\2个LDA.jpg  LDA：2 | C:\Users\Mr.hu\Desktop\文本挖掘作业\作业三\3个LDA.jpg  LDA：3 |
| C:\Users\Mr.hu\Desktop\文本挖掘作业\作业三\4个LDA.jpg  LDA：4 | C:\Users\Mr.hu\Desktop\文本挖掘作业\作业三\5个LDA.jpg  LDA：5 |

|  |
| --- |
| C:\Users\Mr.hu\Desktop\文本挖掘作业\作业三\6个.jpg  LDA：6 |

## 3.4 something more

得到的分类结果，我们可用于为B站弹幕作若干分类。

在此我提出以下可扩展的应用：

①对于视频UP主而言，可以有更多的形式认识关注者的特征，从而调整视频内容。

②在分类基础上，识别恶意弹幕，减轻筛选人员的工作强度

# 附录

LDA主题数及其对应前10关键词

|  |  |
| --- | --- |
| 2 | [(0,  '0.022\*"好" + 0.015\*"不" + 0.013\*"都" + 0.011\*"很" + 0.008\*"没" + 0.008\*"看" + '  '0.008\*"说" + 0.008\*"真的" + 0.007\*"还" + 0.006\*"喜欢"'),  (1,  '0.079\*"哈哈哈" + 0.030\*"哈哈哈哈" + 0.023\*"啊啊啊" + 0.010\*"可爱" + 0.010\*"太" + '  '0.009\*"爱" + 0.008\*"░" + 0.008\*"好看" + 0.007\*"nb" + 0.007\*"死"')] |
| 3 | [(0,  '0.019\*"很" + 0.011\*"好看" + 0.009\*"—" + 0.009\*"nb" + 0.008\*"没" + 0.007\*"排面" + '  '0.007\*"真的" + 0.006\*"弹幕" + 0.005\*"哭" + 0.005\*"上"'),  (1,  '0.036\*"好" + 0.021\*"都" + 0.015\*"人" + 0.013\*"说" + 0.013\*"太" + 0.011\*"喜欢" + '  '0.010\*"░" + 0.010\*"感觉" + 0.009\*"会" + 0.009\*"吃"'),  (2,  '0.107\*"哈哈哈" + 0.041\*"哈哈哈哈" + 0.032\*"啊啊啊" + 0.015\*"不" + 0.013\*"看" + '  '0.013\*"可爱" + 0.012\*"还" + 0.010\*"爱" + 0.010\*"死" + 0.006\*"："')] |
| 4 | …………………… |
| 5 | [(0,  '0.204\*"哈哈哈" + 0.064\*"哈哈哈哈" + 0.015\*"排面" + 0.011\*"弹幕" + 0.011\*"a" + '  '0.009\*"最" + 0.008\*"杀" + 0.008\*"逼" + 0.008\*"牛" + 0.007\*"哥"'),  (1,  '0.039\*"不" + 0.021\*"很" + 0.021\*"好看" + 0.017\*"真的" + 0.010\*"都" + 0.010\*"快" + '  '0.009\*"哭" + 0.008\*"上" + 0.007\*"赵磊" + 0.006\*"贵州"'),  (2,  '0.028\*"人" + 0.025\*"说" + 0.024\*"太" + 0.019\*"nb" + 0.016\*"░" + 0.015\*"1" + '  '0.010\*"小" + 0.008\*"做" + 0.008\*"再" + 0.008\*"厉害"'),  (3,  '0.022\*"爱" + 0.019\*"都" + 0.018\*"没" + 0.017\*"会" + 0.015\*"哈哈哈哈" + 0.012\*"吃" + '  '0.012\*"：" + 0.010\*"真" + 0.009\*"感觉" + 0.009\*"这是"'),  (4,  '0.057\*"好" + 0.048\*"啊啊啊" + 0.022\*"看" + 0.020\*"可爱" + 0.017\*"喜欢" + 0.015\*"死" + '  '0.015\*"—" + 0.014\*"想" + 0.013\*"草" + 0.013\*"还"')] |
| 6 | ……………… |