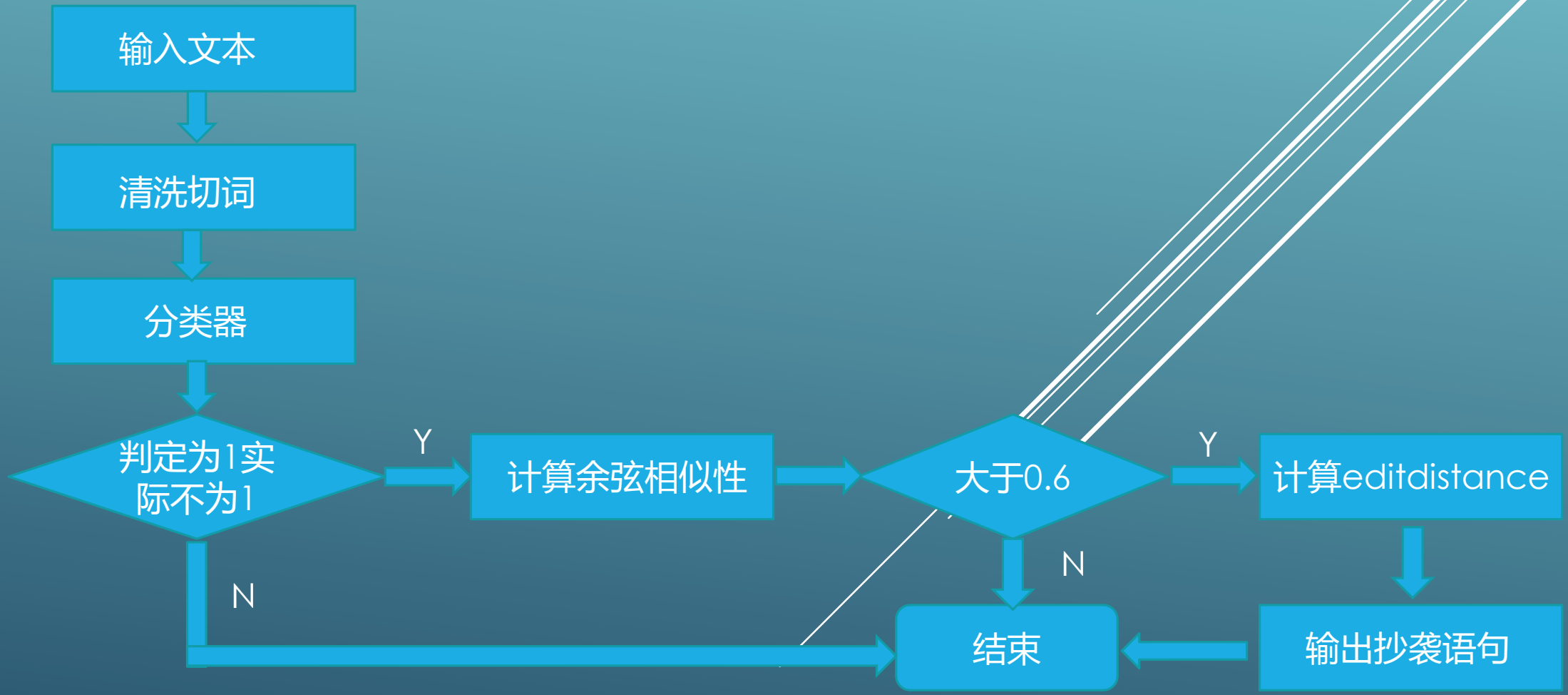
文本相似性检测与抄袭判断

Anan

- 项目目标

对所给文本进行分类，找出新华社所发文章，并在其他文章中找出与新华社文章相似内容并判断是否抄袭。

- 算法流程


输入文本
清洗切词
分类器
判定为1实际不为1
Y
计算余弦相似性
大于0.6
Y
计算editdistance
N
N
结束
输出抄袭语句

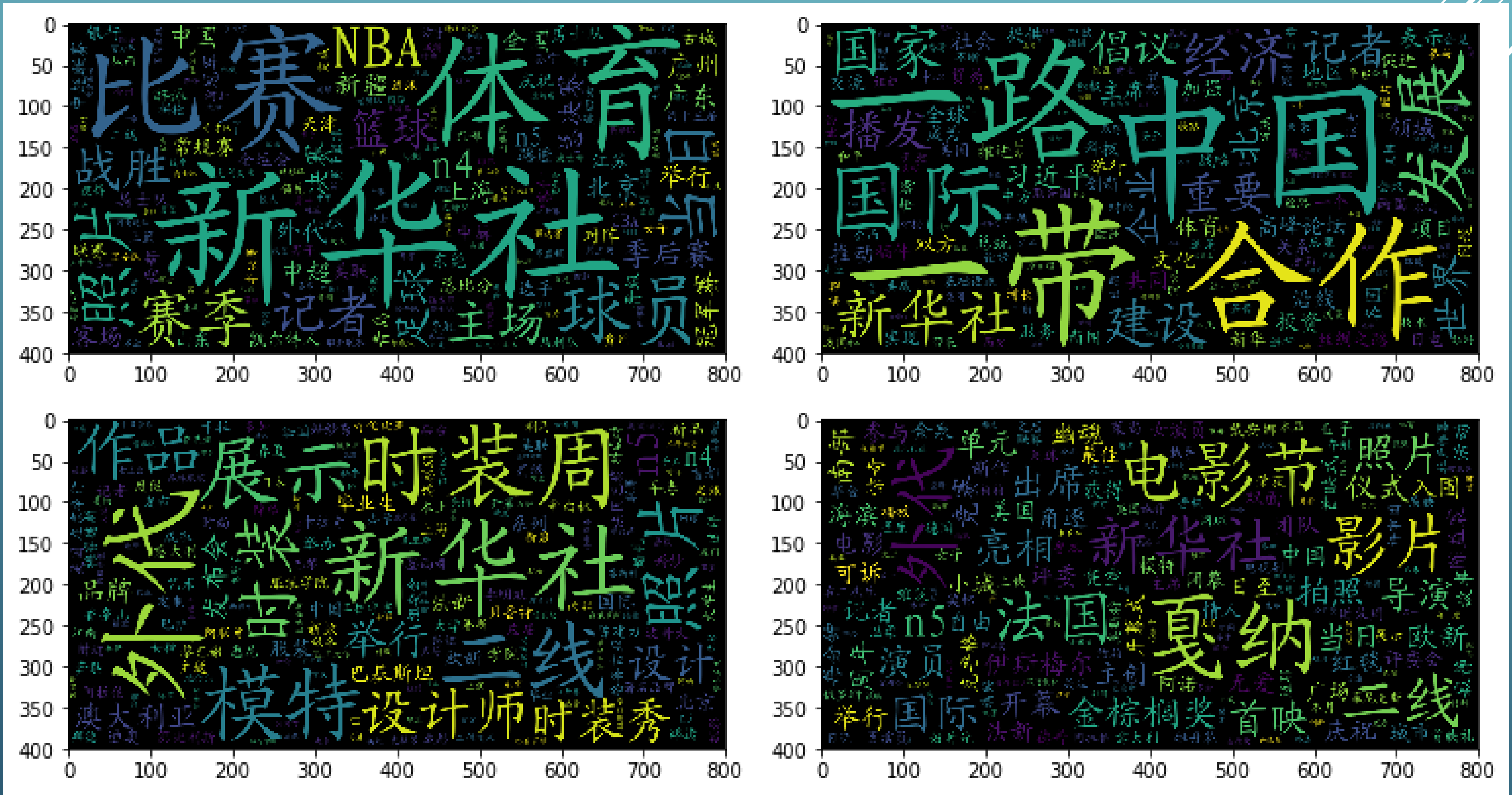
数据基本信息

- 数据长度： 89611
- 数据内容： Index(['id','author','content','feature','title','url'])
- 数据分布：78855篇报道出自新华社

文本自动聚类

- 准备数据 （数据清洗、切词）
- 建立TFIDF向量
- 用K-Means聚类
- 建立 id==>kinds 和 kind==>ids 映射

• 词云展示聚类结果


比赛
NBA
体育
新华社
赛季
记者
球员
战胜
篮球
一路
中国
国际
一带
合作
新华社
建设
经济
记者
倡议
发展
国家
作品
展示
时装周
新华社
模特
设计师
时装秀
二线
举行
电影节
照片
影片
新华社
夏纳
法国
导演
金棕榈奖
首映
二线
亮相
出席

文本分类

- 准备数据 （数据打标签、随机选取训练和测试数据、数据清洗、切词）
- 建立TFIDF向量
- 使用分类器分类
- 性能评估

```
the shape of tfidf_train is (71688, 223582)
the shape of tfidf_test test is (17923, 223582)
************************
LR
************************
predict info:
accuracy:0.961
precision:0.962
recall:0.996
f1-score:0.979
************************
NaiveBayes
************************
predict info:
accuracy:0.919
precision:0.916
recall:1.000
f1-score:0.956
************************
************************
SVM
************************
predict info:
accuracy:0.983
precision:0.986
recall:0.995
f1-score:0.990
************************
KNN
************************
predict info:
accuracy:0.912
precision:0.912
recall:0.990
f1-score:0.949
```

找出可能抄袭的文章

- 按照朴素贝叶斯分类器结果，选取出分类器判断为新华社的文章但实际不是的文章，认为这些文章有可能抄袭。

```
result = []
result = [test_index_list[i] for i in range(len(pred)) if list(pred)[i]==1 and list(test_class_list)[i]!=1]
np.save('result.npy', result)
result
```

```
[3031,
 970,
 3554,
 5356,
 4183,
 10960,
 4205,
 5922,
 2204,
 5583,
 2257,
 1647,
```

抄袭判定

- 计算可能抄袭的文本和所有新华社文章的余弦相似性，并选出最大值
- 确定阈值，选出相似性较高的样本（> 0.6）

```
Distance[1:10]
```

```
[[0.6494384882555035, 970, 11195],
 [0.5733220400657106, 3554, 87171],
 [0.39922290906550567, 5356, 816],
 [0.43151980935653134, 4183, 49637],
 [0.29234808464247675, 10960, 44527],
 [0.3418259037335308, 4205, 77091],
 [0.42617012704034163, 5922, 11647],
 [0.6053822503822367, 2204, 70773],
 [0.3907711306469972, 5583, 34016]]
```

```
[element for element in Distance if element[0] > 0.6]
```

```
[[0.6494384882555035, 970, 11195],
 [0.6053822503822367, 2204, 70773],
 [0.6840092672418868, 1647, 80129],
 [0.7517655210783512, 4336, 42684],
 [0.7431984858582031, 2742, 70855],
 [0.8723397885775069, 4299, 41565],
 [0.6863448253959228, 3337, 13792],
 [0.6948898542065315, 10971, 75478],
 [0.6468489977022723, 1048, 80858],
 [0.7979725661626185, 8000, 19414],
 [0.6877479392918336, 5428, 541],
 [0.631429454935902, 3869, 28519],
```

• 精确定位

➢ 利用editdistance精确定位

- ✓ 文本序号7525和29376的余弦相似性为0.7923
- ✓ 找出这两个文本中完全相同的句子（即editdistance为0）
- ✓ 结果打印如下

```
ed = []
for k in range(len(string1)):
    for i in range(len(string2)):
        if get_edit_distance(string1[k], string2[i]) == 0:
            ed.append([k,i])
```

```
ed
```

```
[[12, 26], [27, 37], [28, 38], [40, 51]]
```

旅行社 旅行社
将所有旅游购物企业纳入社会普通商品零售企业进行统一监管 将所有旅游购物企业纳入社会普通商品零售企业进行统一监管
严禁变相安排和诱导购物 严禁变相安排和诱导购物
旅游巡回法庭 旅游巡回法庭