

Project-1 最终汇报文档

准备工作

- 读取csv数据为dataframe，此数据集维度是（89611， 7）。
- 我们的目标是根据content字段的文本判断source字段是否新华社，所以首先我查看content是否有缺失值，content有缺失值的样本直接删除，因为没有文本无法预测而且有缺失值的样本较少删除对整个数据集影响力较小，删去缺失样本后数据集维度是（87054， 7）。

```
news.shape
```

```
(87054, 7)
```

```
news[news.source=='新华社'].shape
```

```
(78661, 7)
```

```
78661/87054
```

```
0.9035805772049532
```

- 观察数据发现新华社文章约占全部样本的90.36%，据此推测模型的准确率应该相当高，比较容易过拟合。
- 定义一个函数首先去掉文本中的特殊字符、中文标点符号，分词后过滤掉停用词最后输出以空格分开的文本，具体效果如下：

```
def split_text(text):return ' '.join([w for w in list(jieba.cut(re.sub('\s|[%s]' % (punctuation), '',text))) if w not in
```

```
split_text(news.iloc[1].content)
```

```
Building prefix dict from the default dictionary ...
```

```
Loading model from cache /var/folders/vr/sknb16y9i0vyq7zp6ub3w76c000qz/2/jieba.cache
```

```
Loading model cost 0.893 seconds.
```

```
Prefix dict has been built successfully.
```

```
'骁龙 835 唯一 Windows10 桌面 平台 认证 ARM 处理器 高通 强调 不会 只 考虑 性能 屏幕掉 小 核心 相反 正 联手 微软 找到 一种 适合 桌面  
平台 兼顾 性能 功耗 完美 方案 报道 微软 已经 拿到 一些 源码 Windows10 更好 理解 big little 架构 资料 显示 骁龙 835 一款 集成 CPU GP  
U 基带 蓝牙 Wi-Fi SoC 传统 Wintel 方案 节省 至少 30% PCB 空间 按计划 今年 Q4 华硕 惠普 联想 首发 骁龙 835win10 电脑 预计 均 二合  
一 形态 产品 当然 高通 骁龙 未来 也许 见到 三星 Exynos 联发科 华为 麒麟 小米 澎湃 进入 Windows10 桌面 平台'
```

- 最后把这个函数应用到所有文本构造出语料库

构造特征和模型训练

- 首先用sklearn的Countvectorizer设置min_df=0.015（忽略文档中词频不足词汇数1.5%）建立词频矩阵，然后再用TfidfTransformer将结果转化成tfidf矩阵shape（87054， 883）。
- 预测值为source = ‘新华社’ 标记为 1， 其它标记为0。
- 用train_test_split函数吧样本分为70%的训练集和30%的测试集，把测试集代入朴素贝叶斯分类器(MultinomialNB)
- 最后用cross-validate计算accuracy， precision， recall和f1score这几个指标，与测试集在这几个指标上的表现做对比。

Cross-validation result

scores

```
{'fit_time': array([0.66403689, 0.431885 , 0.42695594]),  
'score_time': array([0.15592504, 0.15189481, 0.21946597]),  
'test_precision': array([0.95943065, 0.96208067, 0.96164526]),  
'train_precision': array([0.96233212, 0.96128883, 0.96252095]),  
'test_recall': array([0.90745081, 0.91132065, 0.90608307]),  
'train_recall': array([0.91001003, 0.90821138, 0.90775059]),  
'test_accuracy': array([0.8817505 , 0.88746123, 0.88252571]),  
'train accuracy': array([0.8865449 , 0.88405868, 0.88475361]),  
'test_f1': array([0.93271709, 0.93601299, 0.93303772]),  
'train_f1': array([0.93544004, 0.93399664, 0.9343338 ])}
```

Testset result

```
show_test_reslt(y_test,y_predict)
```

```
accuracy: 0.8870467511582494  
precison: 0.9620043806714049  
recall: 0.9110960585919309  
f1_score: 0.9358584101582883
```

这些指标在cross-validation中表现稳定，而且测试集和训练集的数据差别不大，模型拟合效果较好。

检测抄袭

-首先用训练好的模型预测全量样本，把预测为正类但是真实值为负类的样本index拿出来存在一个list里，这里我们获得2813个疑似样本。

- 聚类分析

- 首先用Normalizer将tfidf矩阵将每个属性规则化为单位向量（因为sklearn中的kmeans默认使用欧氏距离，而单位向量组成的矩阵计算欧氏距离与余弦距离有线性关系，这样相当于使用余弦距离度量相似度）
- 然后创建一个kmeans聚类器，在此我分类25个簇。(最佳中心点k的选择可以使用MSE,Silhouette等指标选择elbow确定合适的值,但是因为电脑配置问题跑不出来，暂且自己定一个)
- 最后根据聚类分析的结果构建id-class对应字典和class-id对应字典，这样就能找出疑似抄袭的文本在哪个簇，其中最相似的文本有哪些

聚类的同一类中相似文本效果如下：

19

新华社照片，义乌（浙江），2017年5月26日\第九届中国国际旅游商品博览会开幕\5月26日，来自海南的参展商展示黎族纺织工艺。当日，第九届中国国际旅游商品博览会在浙江义乌国际博览中心开幕。为期4天的展会以“旅游新品·智行天下”为主题，设标准展位2000余个，吸引境内外1300多家企业参展。集中展示旅游装备用品、户外休闲用品、旅游厕所、智慧旅游产品、旅游纪念品等五大类产品。新华社发（黄献明摄）

新华社照片，义乌（浙江），2017年5月26日\第九届中国国际旅游商品博览会开幕\5月26日，参展客商选购非洲绿松石饰品。当日，第九届中国国际旅游商品博览会在浙江义乌国际博览中心开幕。为期4天的展会以“旅游新品·智行天下”为主题，设标准展位2000余个，吸引境内外1300多家企业参展。集中展示旅游装备用品、户外休闲用品、旅游厕所、智慧旅游产品、旅游纪念品等五大类产品。新华社发（吕建福）

新华社照片，义乌（浙江），2017年5月26日\第九届中国国际旅游商品博览会开幕\5月26日，观众在拍摄可移动展展。当日，第九届中国国际旅游商品博览会在浙江义乌国际博览中心开幕。为期4天的展会以“旅游新品·智行天下”为主题，设标准展位2000余个，吸引境内外1300多家企业参展。集中展示旅游装备用品、户外休闲用品、旅游厕所、智慧旅游产品、旅游纪念品等五大类产品。新华社发（黄献明摄）

新华社照片，义乌（浙江），2017年5月26日\第九届中国国际旅游商品博览会开幕\5月26日，参展商调试可对话送餐机器人。当日，第九届中国国际旅游商品博览会在浙江义乌国际博览中心开幕。为期4天的展会以“旅游新品·智行天下”为主题，设标准展位2000余个，吸引境内外1300多家企业参展。集中展示旅游装备用品、户外休闲用品、旅游厕所、智慧旅游产品、旅游纪念品等五大类产品。新华社发（黄献明摄）

- 根据以上的结果，我们只需要遍历疑似抄袭文档和它所在簇中的新闻文本，计算edit distance就定位可以找出抄袭文档（或者大篇幅引用）和原文。

大致效果如下图：

怀疑抄袭：

中国5月份46座城市新建商品住宅价格环比上涨，4月份为48座上涨，5月份15个一线城市和热点二线城市房地产市场基本稳定，5月份房地产调控政策效果继续显现。

统计局：15个一线城市和热点二线城市房价同比涨幅全部回落

国家统计局城市司高级统计师刘建伟解读5月份房价数据

5月份一二线城市房价平均涨幅继续回落

国家统计局今日发布了2017年5月份70个大中城市住宅销售价格统计数据。对此，国家统计局城市司高级统计师刘建伟进行了解读。

一、15个一线城市和热点二线城市新建商品住宅价格同比涨幅全部回落，9个城市环比下降或持平

5月份，因地制宜、因城施策的房地产调控政策效果继续显现，15个一线城市和热点二线城市房地产市场基本稳定，从环比看，15个城市新建商品住宅价格涨幅均比上月回落，回落幅度在0.5至6.4个百分点之间。从环比看，9个城市新建商品住宅价格下降或持平；5个城市涨幅在0.5%以内。

二、70个大中城市中一二线城市房价同比涨幅持续回落

5月份，70个大中城市中一二线城市房价同比涨幅比上月回落的城市分别有29和18个。其中，一二线城市同比涨幅回落尤其明显。据测算，一线城市新建商品住宅和二手住宅价格同比涨幅均连续8个月回落，5月份比4月份分别回落2.2和1.7个百分点；二线城市新建商品住宅和二手住宅价格同比涨幅分别连续6个月和4个月回落，5月份比4月份分别回落0.8和0.5个百分点。

三、70个大中城市中房价环比下降及涨幅回落城市个数均有所增加

5月份，70个大中城市中新建商品住宅价格环比下降的城市有9个，比上月增加1个；涨幅回落的城市有26个，比上月增加3个。二手住宅价格环比下降的城市有7个，比上月增加2个；涨幅回落的城市有30个，比上月增加8个。

相似原文：

国家统计局19日发布数据，5月份，15个一线城市和热点二线城市新建商品住宅价格同比涨幅全部回落，其中9个城市环比下降或持平。这9个价格环比下降或持平的城市为：北京、上海、南京、杭州、合肥、福州、郑州、深圳、成都。

“5月份，因地制宜、因城施策的房地产调控政策效果继续显现，15个一线城市和热点二线城市房地产市场基本稳定。”国家统计局城市司高级统计师刘建伟说，从环比看，15个城市新建商品住宅价格涨幅均比上月回落，回落幅度在0.5至6.4个百分点之间。从环比看，9个城市新建商品住宅价格下降或持平；5个城市涨幅在0.5%以内。

国家统计局当天还发布了5月份70个大中城市住宅销售价格统计数据。刘建伟介绍，5月份，70个大中城市中新建商品住宅和二手住宅价格同比涨幅比上月回落的城市分别有29和18个，其中，一二线城市同比涨幅回落尤其明显。据测算，一线城市新建商品住宅和二手住宅价格同比涨幅连续8个月回落，5月份比4月份分别回落2.2和1.7个百分点；二线城市新建商品住宅和二手住宅价格同比涨幅分别连续6个月和4个月回落，5月份比4月份分别回落0.8和0.5个百分点。

此外，70个大中城市中房价环比下降及涨幅回落城市个数均有所增加。刘建伟显示，5月份，70个大中城市中新建商品住宅价格环比下降的城市有9个，比上月增加1个；涨幅回落的城市有26个，比上月增加3个。二手住宅价格环比下降的城市有7个，比上月增加2个；涨幅回落的城市有30个，比上月增加8个。

editdistance: 305

- 但是最后发现，疑似抄袭文本中有些有的词语或者主题雷同，实际上不是抄袭或者无关系的新闻，相似度也很高，但是找不到改动。