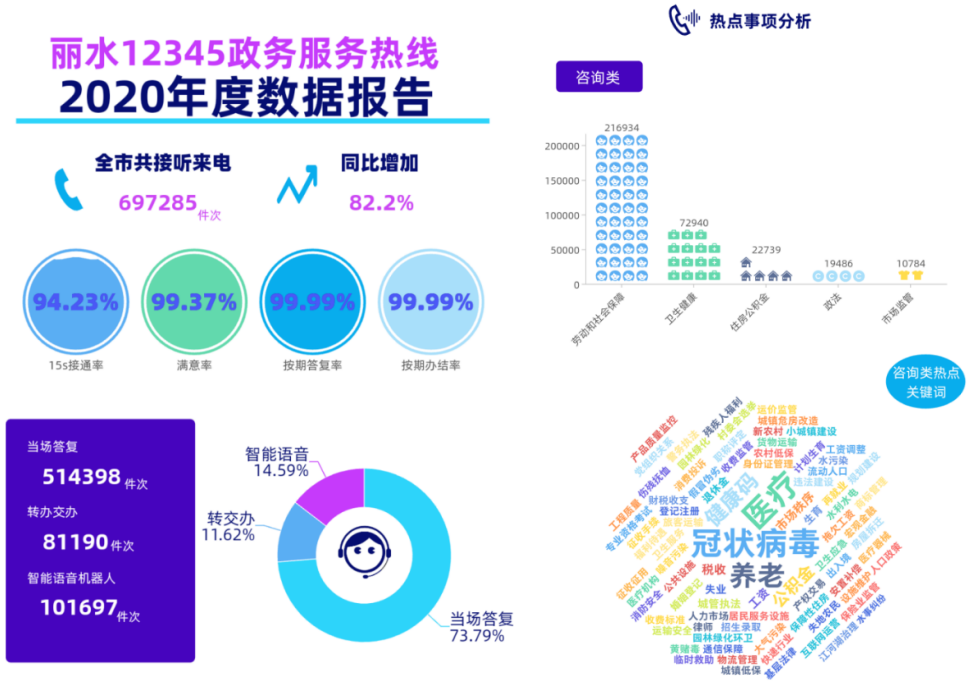
题目：投诉事件文本的分类与摘要生成

1. 问题简介

把一份冗长的文件归纳成一份摘要对人类来说乏味而缓慢，需要阅读整个文档，然后专注于重要的句子，最后，将句子重新写成一个连贯的要点。近年来，随着信息技术的跨越式发展，信息获取的便捷度不断提升，信息量不断增加，使得文本信息呈现爆炸式增长，如何从大量文本信息中快速提取重要的内容，已成为一个迫切需求，因此自动文本摘要生成任务应运而生。如下图(来自 https://new.qq.com/rain/a/20210112A0ANIL00)的公开报告中所示，随着信息技术在信访领域日渐普及，充分推动了信访工作新变化，有助于改善工作中职能宽泛、解决效率低等日益加剧的矛盾。



注：上图中的数据与图例取自《丽水市 12345 政务服务热线 2020 年度数据报告公布》

得益于模型、方法、语料的支撑，文本摘要作为传统的自然语言处理任务，至今依旧有新的发展和创新。机器学习在文本分析中有广泛的应用，特别是在文本的关键字抽取、短文本分类等方面，以及取得了快速发展，但仍有很大的发展空间。摘要生成作为文本生成的一种，除了有着重复、冗余、不连贯、生成较短等问题，还有着摘要任务特定的问题。其中，

最为核心的问题是：如何确定关键信息。

本赛题面向解决社会事件投诉事件文本的智能分析，包括内容分类、事件相关部门推荐以及内容摘要生成等方面。参赛者可以根据自己的情况，分两阶段完成相应的目标任务（详见本文第 3 部分目标任务的详细介绍），研发相关模型，并在给定的测试集上进行比对，最终提交相关的总结报告和模型代码。

2. 数据集

2.1 总体介绍

本竞赛用的事件文本分类总共有 33,323 条数据，来自网络上公开的数据集。数据包括 5 个维度/字段，含分类用的短文本和相关的数值型数据和类别数据等。

2.2 样例

事件文本分类与摘要生成的数据形式如下所示。

	dispute_brief_situation	dispute_eff	involved_person_num	agreement_form	dispute_type
0	婚姻纠纷	简单纠纷	2	书面协议	婚姻家庭纠纷
1	2019年5年1日，当事人赵光和蓝云鹏在碧湖蓝巨星门口引发争吵，互相扭打导致赵光受伤一事申请人民调解。	一般纠纷	2	书面协议	损害赔偿纠纷
2	被申请人拖欠申请人物业费	简单纠纷	2	书面协议	物业纠纷
3	双方当事人为村自来水不能使用发生争吵打架，导致夏勤良受伤逐产生纠纷。	一般纠纷	3	书面协议	损害赔偿纠纷
4	2019年7月至2020年1月叶奇木在物流城D区承包陶志刚的工地钢筋焊接，因双方在报酬上差价有争议，共计15232 元。	一般纠纷	2	书面协议	其他劳动争议纠纷
5	2016年4月13日，被申请人鲍施有由中国建设银行股份有限公司云和支行确认信息真实性，向申请人申请提交《个人开户与银行签约服务申请书》，2017年5月21日、2018年1月6日、2018年2月8日通过互联网分别与申请人签订《中国建设银行借贷通服务协议》，被申请人分别向申请人借款人民币100000元、31100元、48100元并签订了电子合同《中国建设银行"快e贷"个人借款合同》限期1年，还款该卡号：6217001490003192024，还款到期被申请人暂时没能力还款，截止2020年9月10日被申请人尚欠申请人本息合计216687.66元，申请人多次催讨未果，因此产生纠纷，特申请调解。	一般纠纷	2	书面协议	其他合同纠纷
6	当事方系村民委员会代表，现需在本村建设新农村需要本村的土地，该土地有0.26亩属于当事人汤小勇的，现双方因征用土地费用问题引发纠纷。	一般纠纷	2	书面协议	山林土地纠纷
7	2020年7月2日，在莲都区和平路江和平叫了代驾，代驾司机邹智伟说车辆超员不能驾驶，当时江和平因饮酒后对邹智伟进行的殴打，造成邹智伟受伤。	简单纠纷	2	书面协议	损害赔偿纠纷
8	两当事人系同村近邻，双方为了相邻之事引发纠纷，王中伟踢坏了王仁圭家的防盗门一事申请人民调解。	简单纠纷	2	书面协议	其他纠纷
9	马志坤旧房拆建，西边和陶建中交界，陶建中要马志坤在原房基退回1米后再建设。	疑难纠纷	2	书面协议	邻里纠纷

字段说明：

序号	字段	解释	备注
1	dispute_brief_situation	事件概况	
2	dispute_eff	重要程度	可作为因变量
3	involved_person_num	事件涉及人数	
4	agreement_form	事件处理协议方式	可作为因变量
5	dispute_type	事件类别	可作为因变量

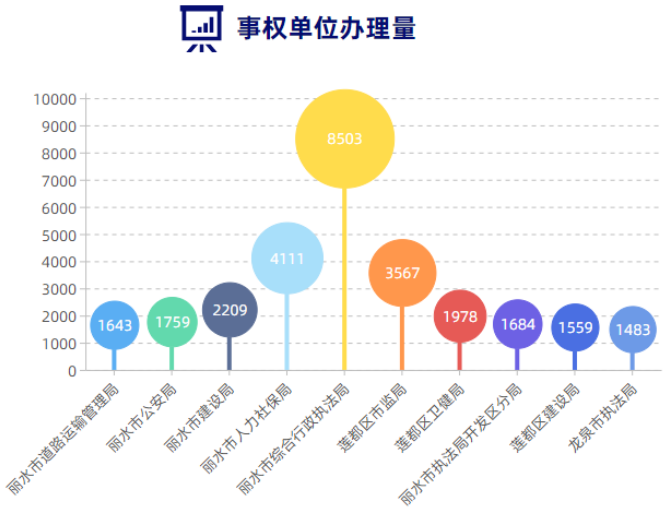
2.1　　获取方式

```
# 导入模块

import pandas as pd

import requests

import json

# 第 1 页数据

url = "http://113.31.111.86:19000/disputes/importData?page={}".format(1)

auth = {"Authorization":"Basic dGVzdDpOZXN0MTIzNDU2"}

data_per = requests.get(url=url, headers=auth)

# json2dataframe

data_df = pd.DataFrame(json.loads(data_per.text)['results'])
```

3. 目标任务

任务一：分类与推荐

要求仅根据事件描述的短文本，判断事件文本所属的类别，并给出相关的职能部门的推荐。这个任务将使用分类和推荐这两类基本的机器学习算法。旨在解决实际管理中案件错综复杂带来的困难，当事件到达时，可以辅助相关工作人员做部门推荐判断。事件分类的结果和相关职能部门的推荐结果（如下图），将分别以 top1，top3 和 top5 准确率进行衡量。



注：上图中的数据与图例取自《丽水市 12345 政务服务热线 2020 年度数据报告公布》

示例 1：

输入：2020 年 10 月 16 日，在莲都区万地广场蕾特恩专业祛痘国际连锁店，双方因祛痘经费发生纠纷。

输出参考：

类别 top5：其他消费纠纷；损害赔偿纠纷；民间借贷纠纷；银行业纠纷；其他纠纷

部门：莲都区市监局

示例 2：

输入：2020 年 8 月 13 日 11 时 50 分许，在松阳县桃花源后门对面路段发生交通事故。

输出参考：

类别 top5：道路交通事故纠纷；其他纠纷；损害赔偿纠纷；其他劳动争议纠纷；邻里纠纷
部门：丽水市道路运输管理局

任务二：摘要生成

要求根据事件描述的短文本，探索不同类别事件的标准叙述方式，给出内容摘要。该任务的目的旨在规范化文本格式，可以使得事件描述更规范合理，便于数据管理和数据统计，提高工作人员归纳总结的效率。

常用的摘要生成方法中，按照输出类型可分为抽取式摘要和生成式摘要。抽取式方法从原文中选取关键词、关键句组成摘要。这种方法天然的在语法、句法上错误率低，保证了一定的效果。抽取式摘要在语法、句法上有一定的保证，但是也面临了一定的问题，例如：内容选择错误、连贯性差、灵活性差等问题。生成式摘要允许摘要中包含新的词语或短语，灵活性高，随着近几年神经网络模型的发展，序列到序列（Seq2Seq）模型被广泛的用于生成式摘要任务，并取得一定的成果。

当下的文本摘要更关注什么是真正的摘要，而不仅仅是简单地句子压缩。利用外部知识，利用关键词信息等方式来更好的辅助摘要的生成。各个摘要模型各有优点，在实验结果上各有优势。

示例 1：

输入：申请人颜先平是被申请人蓝忠土的员工，在丽水市金利亚轴承制造有限公司上班，在2020年10月18日傍晚，颜先平在工作期间右脚受伤。立即送往张村卫生院进行救治，现在双方对于赔偿问题无法达成一致，特申请调解。

输出参考：申请人颜先平在丽水市金利亚轴承制造有限公司上班，工作期间右脚受伤，对于赔偿问题申请调解。

示例 2：

输入：2019年2月28日，王温生驾驶三轮电动车沿云和县复兴路由西向东行驶至云和县和信路与复兴路交叉路口左转弯时，与冯红波驾驶的车牌号为浙K01033的直行的小型轿车发生碰撞，致王温生受伤及两车损坏的交通事故。当事人王温生负事故的主要责任；当事人冯红波负事故的次要责任。

输出参考：王温生驾驶三轮电动车与冯红波驾驶小型轿车发生碰撞，致王温生受伤及两车损坏。当事人王温生负事故的主要责任；当事人冯红波负事故的次要责任。

4.其它说明

参赛队伍在通过接口（API）获取源数据时，由于服务器性能限制，建议不要频繁请求接口，以及最好将请求数据存储到本地电脑中。

若接口请求数据过程发生异常或限制时，建议次日重新获取资源。