第八届“泰迪杯”数据挖掘挑战赛——

C 题：“智慧政务”中的文本挖掘应用

一、问题背景

近年来，随着微信、微博、市长信箱、阳光热线等网络问政平台逐步成为政府了解民意、汇聚民智、凝聚民气的重要渠道，各类社情民意相关的文本数据量不断攀升，给以往主要依靠人工来进行留言划分和热点整理的相关部门的工作带来了极大挑战。同时，随着大数据、云计算、人工智能等技术的发展，建立基于自然语言处理技术的智慧政务系统已经是社会治理创新发展的新趋势，对提升政府的管理水平和施政效率具有极大的推动作用。

附件给出了收集自互联网公开来源的群众问政留言记录，及相关部门对部分群众留言的答复意见。请利用自然语言处理和文本挖掘的方法解决下面的问题。

二、解决问题

1、群众留言分类

在处理网络问政平台的群众留言时，工作人员首先按照一定的划分体系（参考附件 1 提供的内容分类三级标签体系）对留言进行分类，以便后续将群众留言分派至相应的职能部门处理。目前，大部分电子政务系统还是依靠人工根据经验处理，存在工作量大、效率低，且差错率高等问题。请根据附件 2 给出的数据，建立关于留言内容的一级标签分类模型。

通常使用 F-Score 对分类方法进行评价：

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i + R_i},$$

其中 P_i 为第 i 类的查准率，R_i 为第 i 类的查全率。

2、热点问题挖掘

某一时段内群众集中反映的某一问题可称为热点问题，如“XXX 小区多位业主多次反映入夏以来小区楼下烧烤店深夜经营导致噪音和油烟扰民”。及时发现热点问题，有助于相关部门进行有针对性地处理，提升服务效率。请根据附件 3 将某一时段内反映特定地点或特定人群问题的留言进行归类，定义合理的热度评价指标，并给出评价结果，按表 1 的格式给出排名前 5 的热点问题，并保存为文件“热点问题表.xls”。按表 2 的格式给出相应热点问题对应的留言信息，并保存为“热点问题留言明细表.xls”。

表 1-热点问题表

热度排名	问题 ID	热度指数	时间范围	地点/人群	问题描述
1	1	…	2019/08/18 至 2019/09/04	A 市 A5 区魅力之城小区	小区临街餐饮店油烟噪音扰民
2	2	…	2017/06/08 至 2019/11/22	A 市经济学院学生	学校强制学生去定点企业实习
…	…	…	…	…	…

表 2-热点问题留言明细表

问题ID	留言编号	留言用户	留言主题	留言时间	留言详情	点赞数	反对数
1	360104	A012417	A 市魅力之城商铺无排烟管道，小区内到处油烟味	2019/08/18 14:44:00	A 市魅力之城小区自打交房入住后，底层商铺无排烟管道，经营餐馆导致大量油烟排入小区内，每天到凌晨还在营业……	0	0
1	360105	A120356	A5 区魅力之城小区一楼被搞成商业门面，噪音扰民严重	2019/08/26 08:33:03	我们是魅力之城小区居民，小区朝北大门两侧的楼栋下面一楼，本来应是架空层，现搞成商业门面，嗓声严重扰民，有很大的油烟味往楼上窜，没办法居住……	1	0
1	360106	A235367	A 市魅力之城小区底层商铺营业到凌晨，各种噪音好痛苦	2019/08/26 01:50:38	2019 年 5 月起，小区楼下商铺越发嚣张，不仅营业到凌晨不休息，各种烧烤、喝酒的噪音严重影响了小区居民休息……	0	0
…	…	…	…	…	…	…	…
1	360109	A0080252	魅力之城小区底层门店深夜经营，各种噪音扰民	2019/09/04 21:00:18	您好：我是魅力之城小区的业主，小区临街的一楼是商铺，尤其是餐馆夜宵摊等，每到凌晨都还在营业，每到晚上睡觉耳边都充斥着吆喝……	0	0
2	360110	A110021	A 市经济学院寒假过年期间组织学生去工厂工作	2019/11/22 14:42:14	西地省 A 市经济学院寒假过年期间组织学生去工厂工作，过年本该是家人团聚的时光，很多家长一年回来一次，也就过年和自己孩子见一次面，可是这样搞……	0	0
2	360111	A1204455	A 市经济学院组织学生外出打工合理吗？	2019/11/5 10:31:38	学校组织我们学生在外边打工，在东莞做流水线工作，还要倒白夜班。本来都在学校好好上课，十月底突然说组织到外省打工……	1	0
…	…	…	…	…	…	…	…
2	360114	A0182491	A 市经济学院变相强制实习	2017/06/08 17:31:20	系里要求我们在实习前分别去指定的不同公司实训，我这的公司的工作内容和老师之前介绍以及我们专业几乎不对口，不做满 6 个月不给实训分，不能毕业……	9	0

3、答复意见的评价

针对附件 4 相关部门对留言的答复意见，从答复的相关性、完整性、可解释性等角度对答复意见的质量给出一套评价方案，并尝试实现。

三、数据说明

附件 1 提供了一种内容分类三级标签体系，样例如下：

表 3-内容分类三级标签体系

一级分类	二级分类	三级分类
城乡建设	安全生产	事故处理
城乡建设	安全生产	安全生产管理
城乡建设	安全生产	安全隐患
...	...	...

附件 2、附件 3、附件 4 的数据来源于互联网公开渠道，具体表结构如下：

表 4-附件 2 表结构

留言编号	留言用户	留言主题	留言时间	留言详情	一级分类
744	A089211	建议增加 A 小区快递柜	2019/10/18 14:44	我们是 A 小区居民...	交通运输

表 5-附件 3 表结构

留言编号	留言用户	留言主题	留言时间	留言详情	点赞数	反对数
744	A089211	建议增加 A 小区快递柜	2019/10/18 14:44	我们是 A 小区居民...	100	2

表 6-附件 4 表结构

留言编号	留言用户	留言主题	留言时间	留言详情	答复意见	答复时间
744	A089211	建议增加 A 小区快递柜	2019/10/18 14:44	我们是某小区居民...	网民‘A089211’你好...	2019/10/19 8:40

附录：

请仔细阅读以下说明：

1、关于赛题数据

(1) 示例数据：2020 年 3 月 1 日 9:00:00 随赛题公布。

(2) 全部数据：2020 年 4 月 11 日 9:00:00 公布。

(3) 测试数据：2020 年 4 月 25 日 9:00:00 公布。

2、提交作品

(1) 命名方式：论文命名为“C 题”，附件命名为“作品附件”，测试结果命名为“作品测试结果”。

(2) 论文及附件内请勿出现队号、学校、学院、队员以及指导老师相关任何信息，否则视该作品为无效作品。

(3) 请参赛队于 2020 年 4 月 24 日 16:00:00 之前在竞赛官网“提交作品”处提交论文（PDF 版，大小不超过 50M）及附件（包含论文正文（Word 版）、过程数据、程序、热点问题表、热点问题留言明细表的压缩包，大小不超过 200M）。

3、公布测试数据，提交测试结果

2020 年 4 月 25 日 9:00:00 准时公布测试数据，请在“赛题与数据”页面对应的题目右下方下载测试数据，并于 2020 年 4 月 26 日 9:00:00 前在“提交测试结果”页面提交测试结果。