# 第八届“泰迪杯”全国数据挖掘挑战赛论文报告

所选题目：C 题 ：“智慧政务”与“数据挖掘”有机结合

综合评定成绩：_____

评委评语：

评委签名：

# “智慧政务”与“数据挖掘”有机结合


摘　要

在分析电子政务中数据特点的基础上，概述了电子政务中的数据挖掘的方法和流程，提出需要重点研究的若干关键问题包括框架体系构建、挖掘算法设计、知识管理与分析评价等，最后对数据挖掘在电子政务决策中的应用进行了探讨，分析了电子政务数据的特点，介绍了数据挖掘方法，结合几个可能的应用例子，探讨了电子政务数据挖掘的实现。

信息技术的迅速发展和成熟，使得电子政务应用不断深化。电子政务能够辅助政府更好地为公众服务，也能满足人们对政府和职能部门高效运转的要求。

从全国范围看，经过多年的实践，电子政务建设和应用初见成效。各级政府在不同的层面建设和实现了不同的功能。这其中建立和获取了很多数据。这些数据中具有一些知识。但目前的系统只是实现数据的输入、查询、统计等功能，还没有能够从中挖掘这些知识。充分利用这些数据，挖掘其中的知识，将能够为政府的决策提供更好的支持，能够更好地满足快速有效服务大众的要求。



关键词：电子政务 数据挖掘 决策支持

# "Intelligent government" and "data mining" organic combination


## Abstract

On the basis of analyzing the characteristics of the data in the electronic government affairs, summarizes the method and process of data mining in electronic government affairs, put forward some key problems which should be researched including the construct of framework, the mining algorithm design, knowledge management and the analysis and evaluation, etc., and finally the application of data mining in the electronic government affairs decision-making are discussed in this paper, the authors analyze the characteristics of e-government data, this paper introduces the data mining method, combined with the application of several possible example, probes into the implementation of the e-government data mining.

The rapid development and maturity of information technology make the application of e-government deepen continuously. E-government can assist the government to better serve the public and meet people's requirements for the efficient operation of government and functional departments.

From the national scope, after years of practice, e-government construction and application of initial results. Governments at all levels have built and realized different functions at different levels. A lot of data is built and captured. There is some knowledge in the data. But the current system only realizes the data input, the query, the statistics and so on the function, has not been able to excavate these knowledge. Making full use of the data and mining the knowledge will be able to provide better support for the government's decision-making and better meet the demand of serving the public quickly and effectively.



**Keywords: The electronic government affairs Data mining Decision support**

# 目　　录

# 1．挖掘目标

线上：从广大市民投诉的微信、微博、阳光热线等网络问政平台挖掘。

线下：从居民到居委会投诉、市长信箱反映的情况来看

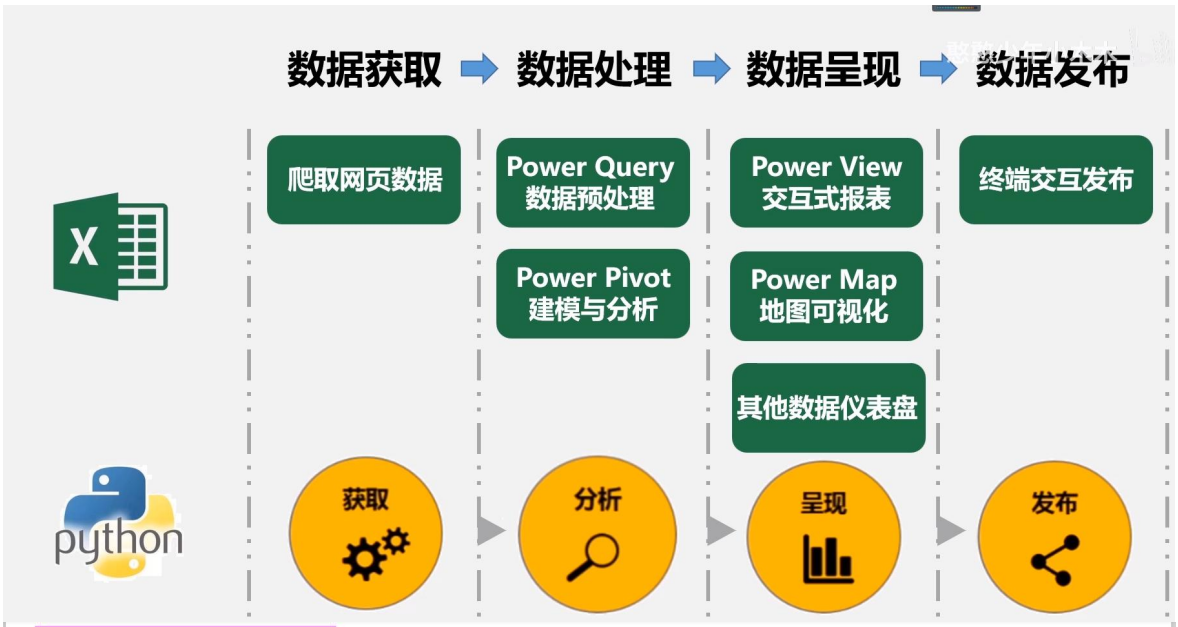
总体来说：挖掘目标群体是居民。

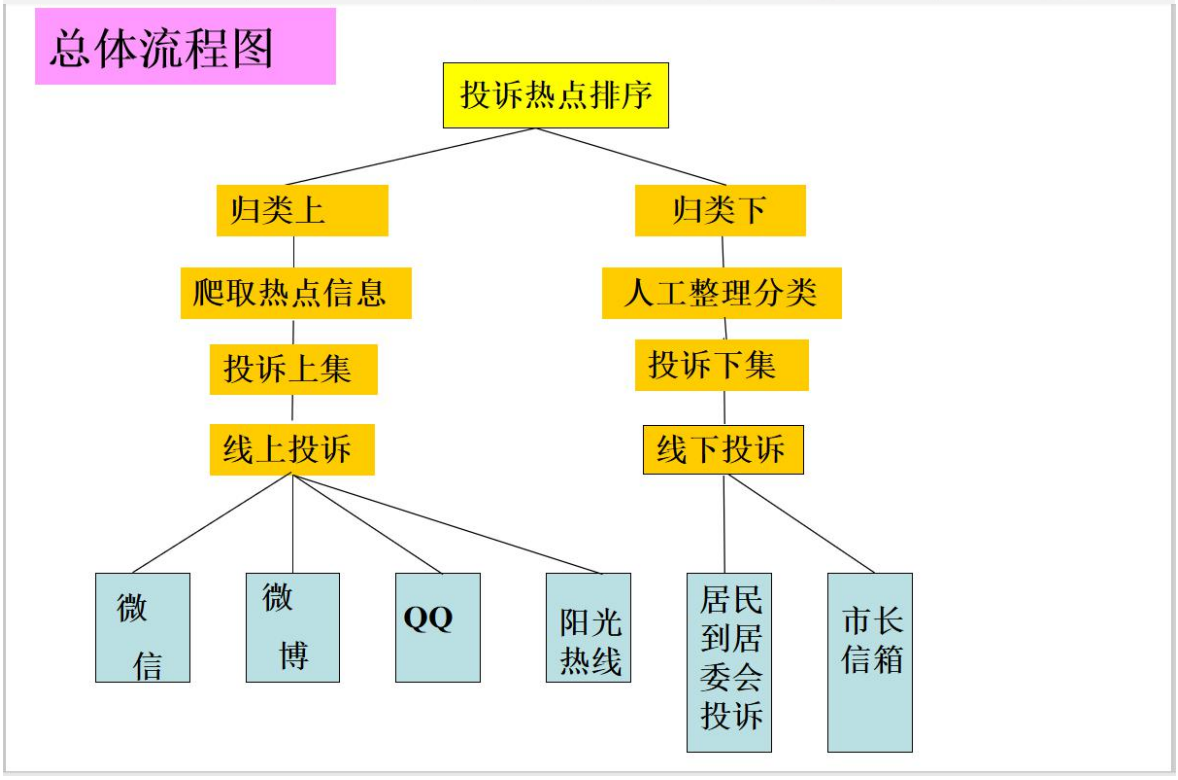
# 2. 分析方法与过程

## 2.1 问题分析

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。智慧政务是推进治理能力现代化的重要技术基础。政府治理工具现代化是国家治理现代化的重要体现。智慧政务既蕴含着技术工具要素，也秉承 着追 逐社会公共利益实现。智慧政务主要依托政务云技术和物联网技术。物联网和云计 算是现代计算机网络智能化发展的新阶段，主要通过数字传感器、电子控制器，将交通物流、环境保护、公共安全、智能消防、工业监测、个人健康等各领域多样性、碎片化的政务信息上传到政务云端然后然后由计算机大数据处理器（云处理器）进行数 据分类 甄 别，再进行网络政务智能化处置或者人工政务处置。说到底，智慧政务就是高科技网络技术与数据处理技术在国家治理过程的综合应用。当然，这不是简单的信息技术在公共部门的简单应用，更多地是涉及政府、政府管理和政府服务范围内的公共管理活动，特别是公共部门职能体系的变革。智慧政务建设实现技术变革和制度变革的协调发展，实现工具理性与价值理性的完美结合，建立与信息社会相适应的政府治理模式，实现善治，促进善政。然而，虽经经历了电子政务、移动政务二个阶段，却并未实现治理制度的深层变革，仍然停留在“外延式”的技术发展与应用，缺乏“内涵式”的价值增值与转变。智慧政务即将突破原有管制型政府的基本职能，从根本上转变公共管理部门根深蒂固的管理意识和行为习惯，实现从低层次的感官管理方式向高层次信息智能化管理方式的转变，这也是推进治理能力现代化的基础条件。

## 2.2 总体流程图


数据获取
数据处理
数据呈现
数据发布
爬取网页数据
Power Query
数据预处理
Power Pivot
建模与分析
Power View
交互式报表
Power Map
地图可视化
其他数据仪表盘
终端交互发布
python
获取
分析
呈现
发布


总体流程图
投诉热点排序
归类上
爬取热点信息
投诉上集
线上投诉
微信
微博
QQ
阳光热线
归类下
人工整理分类
投诉下集
线下投诉
居民到居委会投诉
市长信箱

# 3.数据预处理

## 3.1 数据筛选

使用 Excel 统计筛选预处理

数据审核

数据审核—原始数据(raw data)

数据审核—二手数据(second hand data)

数据筛选(data filter)

数据排序 (data rank)

数据排序 (方法)

数据透视表（pivot table）

数据透视表(用 Excel 创建数据透视表)

## 3.2 数据统计

| 一级标签 | 计数 |
| --- | --- |
| 城乡建设 | 71 |
| 环境保护 | 57 |
| 交通运输 | 40 |
| 教育文体 | 118 |
| 劳动和社 | 140 |
| 商贸旅游 | 134 |
| 卫生计生 | 41 |

图 3.2-1


图 3.2-2

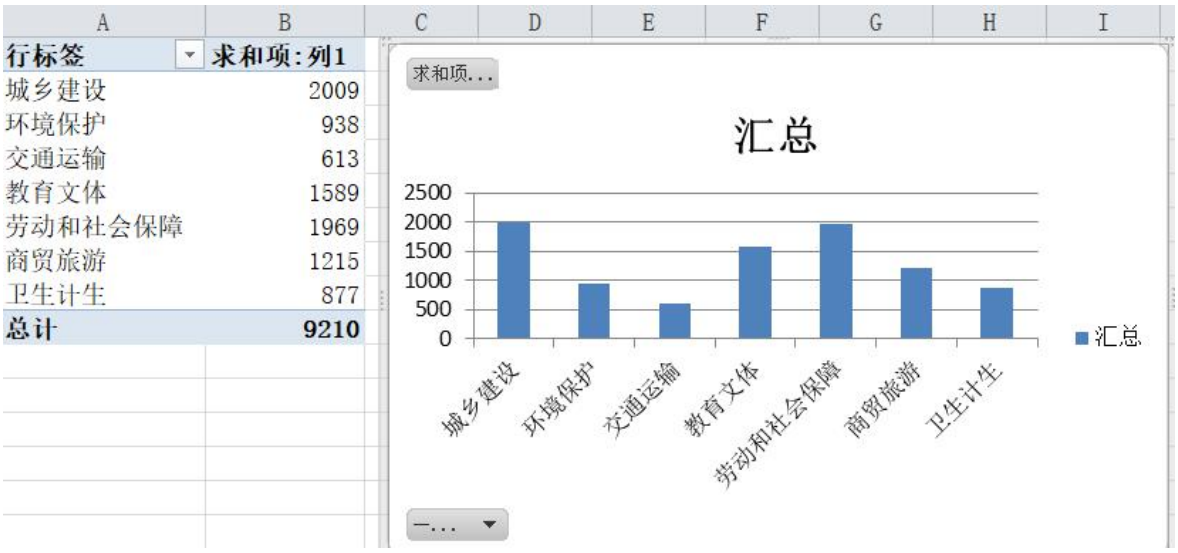



图 3.2-3

# 4、数据优化及建模

## 4.1 数据优化

首先使用 python 优化附加 2 的留言主题的数据冗余，用 python 爬虫爬取一样的关键词并且归为一类，比如爬取留言主题的“施工”“烂尾”“物业”“水箱”“停水”“费”.等等，只有有任何一个关键字里的字，就归为城乡建设问题。其他类似。1 个 Excel 文件，数据内容不多，但文件打开却需要 20 多秒。通过资源管理器查看文件属性，发现体积竟有十几兆。而类似情况下，这种 Excel 通常不过几十 KB。这种情况很多网友都曾遇过，一般都是 Excel 中暗含了某些特殊元素。那么如何快速地让“巨婴 Excel”苗条下来呢？今天小编就给大家提供几条思路。

1. 清理无用对象

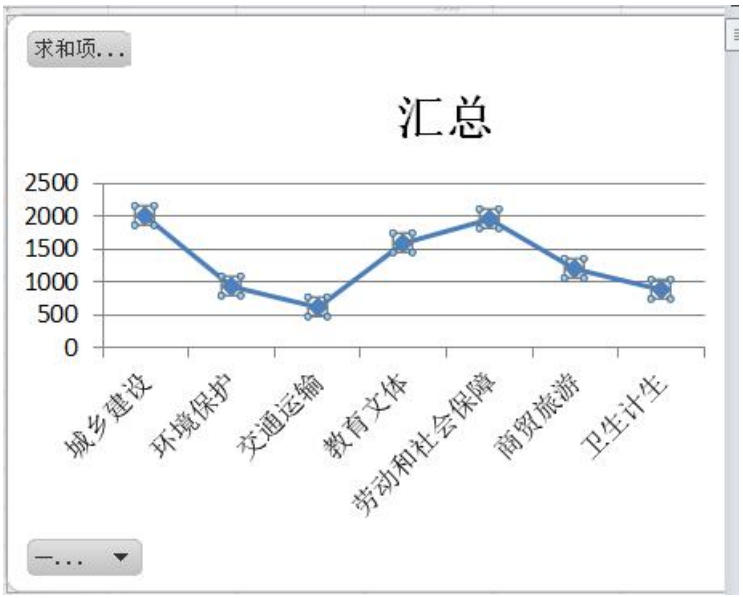
一般来说，遇到这种体积巨大，表面又看不出什么异常的 Excel 文件。首先要考虑的，就是里面是否夹杂了一些不可见的自选图形。为了避免手动清除清理不干净的情况，遇到这种问题，我们一般通过 Excel 的“定位”功能实现。

2. 清理冗余公式
3. 隐藏 Sheet

4.2 建模

采用回归模型与时序预测模型

| 回归模型 | 回归方程 |
| --- | --- |
| 一元线性 | y=β0+β1x |
| 多元线性 | y=β0+β1x1+…+βkxk |
| 二次曲线 | y=β0+β1x+β2x2 |
| 复合曲线 | y=β0βx |
| 增长曲线 | y=eβ0+β1x |




# 5、结果分析

查虫率还可以，

In [1]:
```
import xlrd
```

In [2]:
```
bok = xlrd.open_workbook(r'D:\工作簿2.xlsx')

sht = bok.sheets()[0]

row1 = sht.row_values(0)
row1
```

Out[2]: ['留言编号', '留言用户', '留言主题', '留言时间', '留言详情', '一级标签']

5.2 ……

# 6.结论

## 6.1 不能完全大范围的爬，有权限问题，在一个爬的是用户访问的数据，有的数据没有用户访问就没有的爬，也就是预测不了

# 7.优缺点分析

## 7.1 模型的优点

1.建立的模型能与实际紧密联系，结合实际情况对所提出的问题进行求解，使模型更贴近实际，通用性、推广性较强。 2.基于-----的模型算法新颖，且计算方便；基于------的模型考虑相对全面，仿真结果合理性较强；基于-----算子和-------的评价模型比较精确，得到的因素权重可信度比较高。3.-------的可视化界面形象逼真，操作简便，便于推广； 4.--个模型通过对实验数据的分析不仅使问题得到了一定程度上的解决，而且还能迅速掌握了实验数据的特点为建立更合理的模型提供了参考经验。 5.------模型对于数据分布及样本量、指标多少无严格限制，既适于小样本资料，也适于多评价单元、多指标的大系统，较为灵活、方便。 6.模型---可操作性强，适用范围广泛，基于可能度的-----模型比较精准，得到的因素权重可信度比较高。模型---安排方案具体，在模型---的基础上进一步细分，提出了较为精细的方案。模型---提出了一个通用指标，可广泛应用于其它领域。 7.模型---可靠性高，所采用的研究方法移植性强，但所求得的估计值可能存在一定偏差。模型----对----函数的

构思存在一定的独到之处，引入了非线性规划，但是模型检验方式较为复杂。1.模型建立的合理性，模型的建立是在对样本数据进行充分挖掘的基础之上的，通过数据之间的内在关系观察计算，提炼出各个指标之间的关系，建立起模型； 2.对众多指标用科学的方法进行选取，同时对一些未量化的指标建立模型，进行科学合理的量化，由这些指标建立----指标体系； 3.模型的建立是按照问题问题的解决思路进行的，首先分析和发现现有规律，然后对现有的规律进行评价，根据评价标准建立新模型，层次渐进易于理解； 4.使用 SPSS 统计软件和 excel 进行统计，大大减少了计算量，同时应用---和 matlab 进行优化，得出理想结果。 缺点： 由于所给数据的自身存在某些局限性，我们对模型进行了简化，即假设-------，这样简易的处理会影响到我们后面出院时间的计算，关键指标选取时，舍去了一些相关指标，这降低评价指标体系的完善性。

### 7.2 模型的缺点

模型的缺点： 1.基于----的预测模型运算过程比较麻烦，数据多，运算过程庞大，编程以及程序运行耗时比较多。 2.基于（模糊多目标的学费标准）模型中的参数确定的（模糊性）决定了其推广的相对难度，需要经过更加专业的处理。 3.（如学费标准）制定过程中的随机因素较多，使得模型不能将其准确地反应出来。 4.模型复杂因素较多，不能对其进行全面的考虑，造成与实际有一定的不相符之处。 模型的改进： 模型一考虑了两个一级指标共六个二级指标构成的评价指标体系，来评价病床的合理安排。这主要是从处理上来考虑的，可以尝试采用更多更有效的指标来评价模型，从而让模型达到达到更加优化的目的。 模型的推广： 本文构建了基于----算子的（病床合理安排模糊综合评价模型，解决了排队模型的评价问题，采用（模糊数）的形式表示相关变量，具有一定的合理性，可以用于各种不确定性评价问题。本文提出的基于模糊线性规划的病床合理安排模型具有良好的应用前景，可以和排队论的基本模型相结合，得出更加优化的结果。本文提出的基于----算子的----模型，解决了----问题，可以用于其它不确定性多属性决策问题中。本文建立的----模型可以用于其它的比例分配问题中，而且简便易行，效果显著。

# 8.算法（程序）的改进与推广

### 8.1 算法的改进

直接爬网页

### 8.2 程序的推广

差不多了，贴贴写了 2 个小时的引入随机模型的矩阵匹配算法，要解决的问题见这里。 算法的优点是迭代 5 万次只要 1 分 21s 缺点是对于每一个 MxN Traf 矩阵，要构造其任意元素不同行不同列的有 max(M,N)个 1 的 0-1 矩阵; 当 MxN 很大时，以排列数级数增加匹配矩阵个数。

# 9.参考文献


[1] 作者，书名，出版地：出版社，出版年。
[2] 作者，论文名，杂志名，卷期号：起止页码，出版年。
[3] 作者，资源标题，网址，访问时间（年月日）。

1 江必新：《推进国家治理体系和治理能力现代化》，载于《红旗文稿》2013 年第 11 期。

2 陈桂龙：《智慧政务 2.0 模式》,载于《中国建设信息》2014 年第 5 期。

3 俞可平：《治理与善治》，社会科学文献出版社 2000 年版，第 4 页。

4 程又中、张勇：《城乡基层治理：使之走出困境的政府责任》，载于《社会主义研究》2009 年第 4 期。

5 薄贵利:《推进政府治理现代化》，载于《中国行政管理》2014 年第 5 期。

附录 1

```
# coding=UTF-8
'''
  function:爬取附件 2.xlsx 的关键信息，并写入新的 Excel 文件
'''
import requests
import re
from openpyxl import workbook  # 写入 Excel 表所用
from openpyxl import load_workbook  # 读取 Excel 表所用
from bs4 import BeautifulSoup as bs
import os
```

```
os.chdir(r'C:\Users\Administrator\Desktop')  # 更改工作目录为桌面


def getHtml(src):
    html = requests.get(src).content
    getData(text, src)
    urls = re.findall('href="(.*filter=?)', text)
    for u in range(len(urls) - 2):
        next_url = 'https://movie.douban.com/top250' + urls[u]
        html = requests.get(next_url).content
        getData(html, next_url)


def getData(html, num_url):
    global ws  # 全局工作表对象
    Name = []  # 存储留言编号
    Dr = []  # 存储留言用户
    Ma = []  # 存储留言主题
```

```
        Si = []  # 存储留言时间
        R_score = []  # 存储留言评分
        R_count = []  # 存储评论人数
        R_year = []  # 存储留言评分时间
        R_area = []  # 存储地区
        R_about = []  # 存储一级标签
        soup = bs(html, 'lxml')
        for n in soup.find_all('div',
class_='hd'):
            # ts = n.contents[1].text
            ts =
n.contents[1].text.strip().split('/')[0]
            Name.append(ts)
        for p in soup.find_all('p',
class_=''):
            infor =
p.text.strip().encode('utf-8')
            ya = re.findall('[0-9]+.*\/?',
infor)[0]  # re 得到年份和地区
            R_year.append(ya.split('/')[0])
# 得到年份
            R_area.append(ya.split('/')[1])
# 得到地区
```

```
R_about.append(infor[infor.rindex('/')
+ 1:])   # rindex 函数取最后一个/下标，得到类型
            try:
                sub = infor.index('留言用户')
# 取得留言用户下标

Dr.append(infor[0:sub].split(':')[1])
# 得到留言主题信息
                mh =
infor[sub:].split(':')[1]   # 得到留言评分后面的信息

Ma.append(re.split('[1-2]+', mh)[0])   #
正则切片得到留言评分时间
            except:
                print
                'no find'

Dr.append(infor.split(':')[1].split('/')
[0])
                Ma.append('无介绍...')
```

```
    for r in soup.find_all('div', class_='star'):
        rs = r.contents  # 得到该 div 的子节点列表
        R_score.append(rs[3].text)
        R_count.append(rs[7].text)
    for s in soup.find_all('span', 'inq'):
        Si.append(s.text)
    if len(Si) < 25:
        for k in range(25 - len(Si)):
            Si.append('本页有的电影没简介，建议查看核对，链接:' + num_url)

    for i in range(25):  # 每页25条数据,写入工作表中
        ws.append([Name[i], R_year[i], R_area[i], R_about[i],
                   Dr[i], Ma[i], R_score[i], R_count[i], Si[i]])


if __name__ == '__main__':
```

```
    #    读取存在的Excel表测试
    #        wb=load_workbook('test.xlsx')
#加载存在的Excel表
    #        a_sheet =
wb.get_sheet_by_name('Sheet1') #根据表
名获取表对象
    #        for row in a_sheet.rows: #遍历
输出行数据
    #            for cell in row: #每行的每
一个单元格
    #                print cell.value,

    #  创建Excel表并写入数据
    wb = workbook.Workbook()   # 创建
Excel对象
    ws=wb.active   # 获取当前正在操作的
表对象
    # 往表中写入标题行,以列表形式写入!
  留言用户  留言主题  留言时间  留言详情
  一级标签
    ws.append(['留言编号','留言用户','
留言主题','留言时间','留言详情','一级标
签']
```

```
', '年份', '地区' '留言评分时间'',])
    src =
'https://movie.douban.com/top250'
    getHtml(src)
    wb.save('附件 2.0.xlsx')
```