# 基于非结构化数据的松江区全域旅游发展监测研究

松江区统计局课题组[1]

[摘要] 作为“一个目标，三大举措”战略布局的重要组成部分，松江区正聚力发展全域旅游，争创国家全域旅游示范区。本文在分析旅游产业发展和旅游数据统计现状的基础上，爬取2011年1季度至2017年2季度国内知名旅游网站的近45万条网民评论数据，使用文本挖掘和机器学习等统计建模方法，从“吸引力”、“满意度”、“负极性”、“主题分类”四个维度对松江、黄浦、青浦、崇明四家国家全域旅游示范区创建单位的全域旅游发展情况进行监测。根据监测结果揭示的短板和不足，对下阶段松江区全域旅游的进一步推进提出若干建议。

关键词：全域旅游 监测体系 文本挖掘 机器学习

## 一、发展全域旅游的必要性

### （一）全域旅游的定义

全域旅游是指在一定区域内，以旅游业为优势产业，通过对区域内经济社会资源，尤其是旅游资源、相关产业、生态环境、公共服务、体制机制、政策法规、文明素质等进行全方位、系统化的优化提升。全域旅游是一种实现区域资源有机整合、产业融合发展、社会共建共享，以旅游业带动和促进经济社会协调发展的新的区域协调发展理念和模式。

国务院总理李克强在2017年政府工作报告中明确提出，要“完善旅游设施和服务，大力发展乡村、休闲、全域旅游。”这是“全域旅游”首次写入政府工作报告，中国政府网将“全域旅游”列为2017年政府工作报告的12个新词之一。

### （二）松江区旅游业发展的现状

**从时点数据看，保持平稳增长。**2016年，松江区实现旅游业收

[1] 课题组组长：潘永俭　课题组成员：宋莉 马一峰

入 84.65 亿元，同比增长 8.8%。其中，旅游景点收入 8.29 亿元，同比增长 4.8%；旅游餐饮收入 19.35 亿元，同比增长 0.8%；旅游住宿收入 11.80 亿元，同比增长 2.6%。接待中外游客 1502.48 万人次，同比增长 5.8%。其中，接待国内游客 1491.42 万人次，同比增长 5.9%；接待国外游客 5.97 万人次，同比下降 13.5%；接待港澳台游客 5.09 万人次，同比下降 14.0%。

表 1 2016 年松江区旅游业基本情况

| 指标名称 | 单位 | 2016 年 | 2015 年 | ±% |
|---|---|---|---|---|
| **接待中外游客** | 万人次 | 1502.48 | 1420.73 | 5.8 |
| #国内游客 | 万人次 | 1491.42 | 1407.92 | 5.9 |
| 国外游客 | 万人次 | 5.97 | 6.90 | -13.5 |
| 港澳台游客 | 万人次 | 5.09 | 5.92 | -14.0 |
| **旅游业收入** | 亿元 | 84.65 | 77.80 | 8.8 |
| #旅游景点营业收入 | 亿元 | 8.29 | 7.91 | 4.8 |
| 旅游餐饮收入 | 亿元 | 19.35 | 19.20 | 0.8 |
| 旅游住宿收入 | 亿元 | 11.80 | 11.50 | 2.6 |

**从时序数据看，增速有所回落。**“十二五”时期，松江区旅游业收入年均增长 9.2%，较“十一五”时期放缓 18.2 个百分点。其中，旅游景点收入年均增长 8.0%，旅游餐饮收入年均增长 11.7%，旅游住宿收入年均增长 9.3%，分别较“十一五”时期放缓 57.1、26.3、29.7 个百分点。接待中外游客年均增长 12.1%，较“十一五”时期放缓 9.6 个百分点。其中，接待国内游客年均增长 12.3%，接待国外游客年均下降 2.4%，接待港澳台游客年均下降 1.2%，分别较“十一五”时期放缓 9.7、14.8、11.8 个百分点。

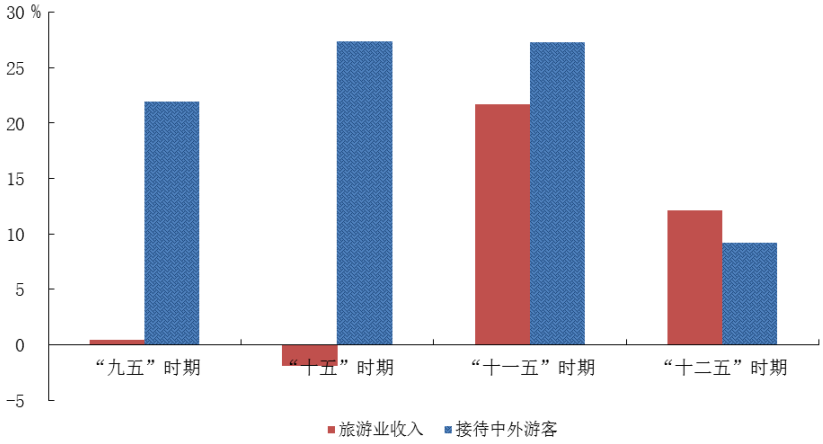


图 1 “九五”—“十二五”时期松江区旅游业收入及接待中外游客年均增速

**从郊区比较看，排名处于中游。**2016 年，松江区旅游业收入同比增长 8.8%，与除闵行外的其他郊区相比排名第四，分别低于嘉定的 44.6%、金山的 10.7%和崇明的 9.1%。接待港澳台游客同比增长 5.8%，与除闵行外的其他郊区相比同样排名第四，分别低于嘉定的 31.7%、金山的 17.8%和宝山的 8.3%。

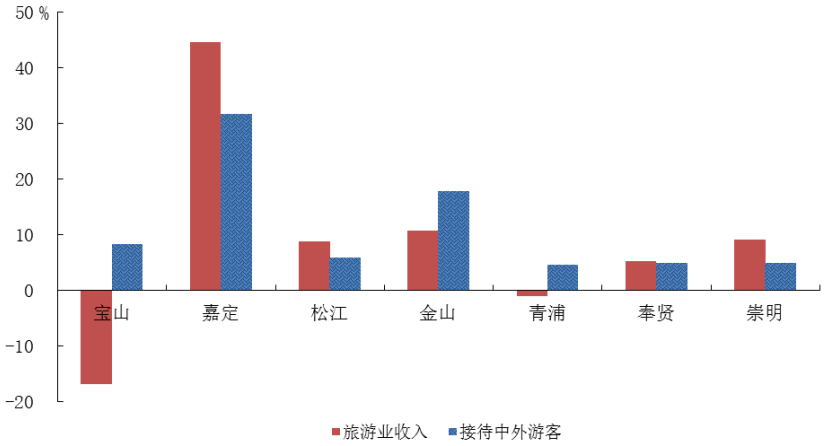


图 2 2016 年松江区及其他郊区旅游业收入及接待中外游客同比增速

**总体来看，现阶段松江区旅游业虽然依然能够保持稳定增长，但增长速度有所放缓，与上海市其他郊区相比也并无明显优势，正因如此，需要突破传统旅游瓶颈，发展全域旅游。**

**（三）松江区发展全域旅游的必要性**

**1、打造地区发展新亮点的客观要求**

作为第三产业的重要组成部分，发展全域旅游是将旅游业打造成为支撑松江区经济平稳、健康、可持续发展新亮点的客观要求，主要体现在三方面。首先，全域旅游是经济转型升级的重要推动力，是生态文明建设的重要引领力，也是展示地区综合实力的重要途径，因此发展全域旅游可作为松江区供给侧结构性改革的重要抓手。其次，全域旅游在改善基础设施、促进产业联动、带动农民就业、建设美丽乡村等方面均具有较为明显的优势，因此发展全域旅游可作为松江区新型城镇化和新农村建设的有效载体。第三，传统旅游业存在旅游产品结构单一，消费链条拓展延伸不足；旅游基础配套薄弱，公共服务体系不健全；旅游品牌形象不突出，市场影响力、竞争力不强等问题，因此发展全域旅游可作为松江区旅游业转型升级的必要方式。

**2、实现地区优势新孕育的客观要求**

松江区被誉为“上海之根、沪上之巅、浦江之首”，是上海历史文化的发祥地，具备良好的旅游基础、特色和优势。“九五”时期是松江区旅游业发展的初创阶段，佘山国家森林公园建立，上海佘山国家旅游度假区建立；“十五”时期是松江区旅游业发展的起步阶段，佘山国家森林公园获评国家 4A 级景区；“十一五”时期是松江区旅游业发展的成形阶段，艾美、开元、索菲特获评五星级饭店，月湖雕塑公园、欢乐谷、辰山植物园相继建成营业；“十二五”时期是松江区旅游业发展的壮大阶段，旅游业进入战略性支柱产业地位的关键培育期，但与此同时也面临着一定困难和挑战。站在新起点上，发展全域旅游

是整合地区旅游资源，突破传统旅游瓶颈，实现“松江旅游”特色更特、优势更优的客观要求。

**3、把握旅游产业新机遇的客观要求**

**从国家层面看，颁布了一系列指导意见，为旅游产业发展提供了完善的顶层设计。**比如国务院下发的《国务院关于印发“十三五”旅游业发展规划的通知》，六部门联合印发的《关于促进交通运输与旅游融合发展的若干意见》，国家旅游局、国家体育总局下发的《关于大力发展体育旅游的指导意见》，十部委联合印发的《关于促进自驾车旅居车旅游发展的若干意见》，国务院办公厅下发的《关于进一步扩大旅游文化体育健康养老教育培训等领域消费的意见》和《关于加快发展健身休闲产业的指导意见》等。

**从市级层面看，制定了一系列发展规划，为旅游产业发展提供了明确的战略导向。**《上海市城市总体规划（2016-2040）》要求把松江新城建设成为沪杭轴上的西南门户节点城市，以科教和创新为动力，以服务经济、战略新兴产业和文化创意产业为支撑的现代化宜居新城，具有上海历史文化底蕴和自然山水特色的区域高等教育基地和休闲旅游度假胜地。《上海市旅游业改革发展“十三五”规划》涉及松江旅游的内容包括：“三圈三带一岛”提到“郊区旅游圈中，松江区突出山水生态资源优势，整合浦南‘都市田园’资源，提升佘山国家旅游度假区休闲度假功能，大力发展生态旅游”；“建设 6 个国家级旅游功能区”提到“佘山国家旅游度假区”；“建设 7 个以郊野公园为特色的生态旅游功能区”提到“松江郊野公园”、“广富林郊野公园”；“建

设9条乡村旅游休闲带”提到“浦南乡村旅游休闲带”。

**从区级层面看，实施了一系列重大举措，为旅游产业发展提供了强力的政策保障。**松江区部署了“一个目标、三大举措”的发展战略布局，“一个目标”即“建设‘科创、人文、生态’的现代化新松江”，“三大举措”即“G60科创走廊建设、国家新型城镇化综合试点、旅游产业发展”。在被国家旅游局列为上海四家全域旅游创建示范区之一后，松江区将全域旅游的发展目标设定为“推动松江全面崛起的战略性支柱产业之一、上海全面建成世界著名旅游城市的重要功能区、长三角地区知名旅游休闲城市、国家全域旅游创新示范区”，并提出六大重点任务[1]、八大发展路径[3]和六大工程建设[4]。

## 二、使用非结构化数据监测全域旅游发展的必要性

### （一）非结构化数据的定义

非结构化数据是指数据结构不规则或不完整，没有预定义的数据模型，不方便用数据库二维逻辑表来表现的数据。包括所有格式的办公文档、文本、图片、XML、HTML、各类报表、图像和音频/视频信息等，而其中以文本（如字符、数字、标点、各种可打印的符号等）作为数据形式的非结构化数据就被称为非结构化文本数据。

与传统的结构化数据相比，非结构化数据具有一些特有的优势，主要体现在三方面。一是拥有大量的数据资源，非结构化数据的来源

[2] “六大重点任务”为大力推进松江全域旅游规划编制、大力推进松江全域旅游核心区建设、大力推进水上旅游休闲带建设、大力推进“四大”休闲区建设、大力推荐“旅游+”产业融合发展、大力推进旅游公共服务建设

[3] “八大发展路径”为安全为基、文化为魂、规划为先、项目为王、活动为重、营销为新、服务为本、产业为链

[4] “六大工程建设”为推动项目建设引领工程、推进文化节庆新旅工程、推进贯标创特提升工程、推进公共服务升级工程、推进旅游产业融合工程、推进旅游人才强旅工程

极其广泛,可以是电子邮件或聊天记录,网站咨询或用户评论,国内媒体或国外媒体;二是蕴含丰富的挖掘价值,分析和挖掘非结构化数据可以帮助研究者较为全面地了解到研究对象的现状和趋势,并能及时有效地监测、识别研究对象出现的新问题;三是消耗较少的人财物力,结构化数据的获取通常需要以各类调查为基础,消耗资源较多,非结构化数据的获取则可通过相关数据工程技术,且一般不需要消耗太多人财物力。

**（二）松江区旅游业数据的现状**

区统计局负责统计的与旅游业直接相关的基础数据包括:1、限额以上住宿餐饮业中行业大类为住宿业（61）的 21 家调查单位,主要统计数据涉及财务（月报/季报）、人员及工资（季报）、景气（季报）、信息化(年报)等方面;2、规模以上社会服务业中行业小类为旅行社服务(7271)、旅游管理服务(7272)、电影和影视节目制作(8630)、电影和影视节目发行（8640）、游乐园（8920）的 17 家调查单位,主要统计数据涉及财务（月报）、人员及工资（季报）、景气（季报）、信息化(年报、）创新（年报）等方面。

表 2 松江区统计局负责统计的与旅游业直接相关的调查单位名单

| 住宿餐饮业调查单位名称 | 行业代码 | 社会服务业调查单位名称 | 行业代码 |
|---|---|---|---|
| 上海兰笋山庄 | 6110 | 上海英式风貌投资发展有限公司 | 7212 |
| 上海红楼戴斯宾馆有限公司 | 6110 | 上海九鹿旅行社有限公司 | 7271 |
| 上海佘山森林宾馆有限公司 | 6110 | 上海潘博网络科技有限公司 | 7271 |
| 上海阳坤酒店管理有限公司 | 6110 | 上海松江旅行社有限公司 | 7271 |
| 上海友福旅馆 | 6120 | 上海红森林旅行社有限公司 | 7272 |
| 上海平高酒店有限公司 | 6110 | 上海佘山旅行社有限公司 | 7272 |
| 上海逸崇华酒店管理有限公司 | 6120 | 上海国伟旅行社有限公司 | 7272 |
| 上海松江开元名都大酒店有限公司 | 6110 | 上海春秋包机旅行社有限公司 | 7272 |
| 上海熙庭酒店管理有限公司 | 6120 | 上海之根旅行社有限公司 | 7272 |
| 上海驿居松荣酒店有限公司 | 6120 | 上海云间国际旅行社有限公司 | 7272 |
| 上海维也纳酒店管理有限公司 | 6110 | 上海晨宏旅行社有限公司 | 7272 |
| 上海江南田园休闲会所有限公司 | 6110 | 上海胜强影视基地有限公司 | 8630 |
| 上海港泰酒店投资管理有限公司 | 6120 | 上海雍硕影业有限公司 | 8630 |
| 上海雪浪湖农家乐观光旅游股份有限公司 | 6120 | 上海影视乐园有限公司 | 8640 |
| 上海大众国际会议中心有限公司 | 6110 | 上海华侨城投资发展有限公司 | 8920 |
| 上海松江如家快捷酒店管理有限公司 | 6120 | 上海佘山国家旅游度假区联合发展有限公司 | 7212 |
| 上海世茂庄园置业有限公司佘山茂御酒店 | 6110 | 上海昊浦影视文化有限公司 | 8630 |
| 上海优孚酒店管理有限公司 | 6120 | | |
| 上海宝茸酒店投资管理有限公司 | 6110 | | |
| 上海新晖大酒店有限公司 | 6110 | | |
| 上海如家松卫酒店管理有限公司 | 6120 | | |

区旅游局负责统计的与旅游业直接相关的基础数据包括：1、重点监测的 18 家旅游景点，其中 4A 级景点 5 家、3A 级景点 3 家，主要统计数据包括营收、税金、接待游客数等，按月度调查；2、重点监测的 20 家酒店宾馆，其中 5 星级宾馆 3 家、4 星宾馆 3 家、星宾馆 2 家，主要统计数据包括营收、税金、住宿旅客数、平均客房出租率等，按月度调查；3、重点监测的 45 家旅行社，其中 4A 级旅行社 2 家、3A 级旅行社 11 家，主要统计数据包括营收、税金、组团人数、接待人数等，按月度调查。

表 3 松江区旅游局负责统计的与旅游业直接相关的调查单位名单

| 旅游景点调查单位名称 | 酒店宾馆调查单位名称 | 旅行社调查单位名称 | |
|---|---|---|---|
| 欢乐谷 | 开元名都大酒店 | 君汇 | 优游 |
| 辰山植物园 | 佘山茂御酒店 | 新鑫 | 红森林 |
| 佘山森林公园 | 佘山索菲特大酒店 | 佘山 | 舒艺 |
| 月湖雕塑公园 | 新晖大酒店 | 九鹿 | 两新 |
| 方塔园 | 绿地铂骊酒店 | 通瑞 | 五彩人生 |
| 醉白池 | 红楼宾馆 | 青豆 | 相伴天涯 |
| 影视乐园 | 佘山森林宾馆 | 钟书 | 海角 |
| 雪浪湖 | 兰笋山庄 | 开天 | 惠康 |
| 佘山高尔夫 | 大众国际会议中心 | 小丁丁 | 全联 |
| 天马高尔夫 | 立诗顿宾馆 | 松江 | 红楼 |
| 天马赛车场 | 宝隆花园酒店 | 畅程 | 道坤 |
| 西部渔村 | 学苑宾馆 | 春秋包机 | 晨宏 |
| 青青旅游世界 | 松江假日酒店 | 风度 | 德逸 |
| 博物馆 | 富悦大酒店 | 嘉景 | 茸兴 |
| 西林禅寺 | 锦江之星松江大学城店 | 之根 | 航宇 |
| 泰晤士小镇 | 维也纳国际松江店 | 滨宁 | 天昕 |
| 五厍农业观光园 | 格林豪泰松东店 | 国伟 | 石库门 |
| 新浜旅游公司 | 7 天连锁松江店 | 酷游 | 乐凯 |
| | 如家快捷方舟园店 | 华庭 | 美锦 |
| | 汉庭快捷松江店 | 云间 | 程启 |
| | | 醉白 | 沐名 |
| | | 往来 | 众人 |
| | | 商旅 | |

**总体来看，现阶段松江区与旅游业直接相关的基础数据较少，主要来源于区统计局和区旅游局，且侧重于反映调查单位的经营情况，正因如此，需要在传统的结构化数据基础上，尝试利用资源更为丰富的非结构化数据。**

**（三）使用非结构化数据监测全域旅游发展的必要性**

**1、全域旅游的发展既需要体现经济属性也需要体现民生属性**

全域旅游具有推动地区经济平稳增长和满足人民美好生活需要的双重属性,因此全域旅游的发展既需要侧重体现经济属性,也需要侧重体现民生属性。《国务院关于加快发展旅游业的意见》明确提出

培育旅游业成为“国家经济的战略性支柱产业”和“人民群众更满意的现代服务产业”的双重载体。区统计局、区旅游局联合开展的《上海市松江区全域旅游统计指标体系构建与分析方法研究》课题利用结构化数据从经济指标、就业指标、服务指标三个维度侧重体现全域旅游发展的经济属性；而本课题则利用非结构化数据从吸引力指数、满意度指数、负极性指数、主题分类指数四个维度侧重体现全域旅游发展的民生属性。

**2、全域旅游的发展既需要测算年度数据也需要进行常规监测**

全域旅游的发展是动态的、循序渐进的，在其发展过程中可能会不断出现值得关注的问题和需要完善的环节，因此全域旅游的发展既需要测算年度数据，科学反映产业发展的阶段性成果，也需要进行常规监测，及时发现产业发展的短板和不足。《上海市松江区全域旅游统计指标体系构建与分析方法研究》课题主要利用 2016 年限额以上住宿餐饮业及规模以上社会服务业年度数据和 2017 年一次性抽样调查结果侧重进行年度数据测算；而本课题则主要利用 2011 年 1 季度至 2017 年 2 季度国内知名旅游网站的近 45 万条网民评论数据侧重进行季度常规监测。

**3、全域旅游的发展既需要反映纵向变动也需要便于横向比较**

上海市的国家全域旅游示范区创建单位除了松江区还有黄浦区、青浦区和崇明区，不管立足本区实际还是借鉴他区经验都具有重要意义，因此全域旅游的发展既需要利用本区数据进行纵向的时序比较，也需要利用他区数据进行横向的截面比较。《上海市松江区全域旅游

统计指标体系构建与分析方法研究》课题使用的基础数据包括规模以上和重点监测调查单位，测算成果为 2016、2017 两年的松江区年度主要指标数据；而本课题使用的基础数据覆盖了全部旅游景点和旅游住宿评论对象，主要监测成果为 2011 年 1 季度至 2017 年 2 季度四家示范区创建单位的季度监测得分，便于进行纵向和横向比较。

## 三、基于非结构化数据的全域旅游发展监测

### （一）基础数据的简要介绍

**本文使用的基础数据主要分为“旅游景点”和“旅游住宿”两类对象，并按“全域旅游”的概念将两类对象的范畴进行扩充。“旅游景点”包括辖区内所有景点、游乐园、动植物园、展览展馆、农家乐等，“旅游住宿”包括辖区内所有星级酒店、经济型酒店、度假村、旅社、民宿等。**

使用“八爪鱼”工具爬取大众点评、去哪儿、携程等多家国内知名旅游网站的网民评论数据 44.53 万条。其中，松江区 399 个爬取对象，13.28 万条评论数据；黄浦区 439 个爬取对象，20.68 万条评论数据；青浦区 492 个爬取对象，7.62 万条评论数据；崇明区 523 个爬取对象，2.95 万条评论数据。

爬取的网民评论数据包含用户名、所在地、性别、日期、评分、评论、获赞数等七个字段，并在爬取时设置正则规则将“所在地”字段内容统一转化为“上海”、“外地”，将“日期”字段内容统一转化为“XX（年）XX（季）”形式，以便后续程序读取。

表 4 上海市四家全域旅游示范区创建单位爬取对象和网民评论数

| | 爬取对象数 | 网民评论数 |
|---|---|---|
| 松江区 | 399 | 132765 |
| 黄浦区 | 439 | 206817 |
| 青浦区 | 492 | 76242 |
| 崇明区 | 523 | 29463 |
| 合计 | 1853 | 445287 |

表 5 上海市四家全域旅游示范区创建单位网民评论数据示例

| 用户名 | 评分 | 评论 | 获赞数 | 性别 | 地区 | 日期 | 评论对象 |
|---|---|---|---|---|---|---|---|
| honey5wu | 5 | 佘山站有短驳车直接到欢乐谷门口，非常方便。园区既有适合大人的游乐设施，也有适合小朋友的室内游乐场，整个园区很有特色，个人感觉并不比迪斯尼差，而且没有很多人，不需要排大长队，小朋友们每次去都玩的很开心～再热些可以去玛雅了～哈哈 | 1 | woman | 上海 | 1702 | 欢乐谷 |
| 团一团一团 | 2 | 体验期去过一次就不会再去了原因如下：人工造浪沙滩不好好开，漂流还要花钱使用救生员竟然要 50 元，乐园里面工作人员脸上一脸丧气连笑都不会！拔草 -.- | 4 | man | 上海 | 1702 | 玛雅海滩水公园 |
| ABClm0707 | 5 | 酒店环境挺不错，有阳台、有山景，而且很安静！泡上一杯茶，坐在阳台上，听听对面山上传出的阵阵鸟叫声，感觉整个心都平静了很多！跟上海市内酒店的喧哗形成了鲜明对比，从市区开车过来大约需要一个小时左右的时间，还好路途不算太远！ | 1 | man | 外地 | 1701 | 佘山茂御臻品之选酒店 |
| 小白兔 2008 | 4 | 索菲特大酒店的位置很不错，佘山地铁站出来走个 1 公里就能到达， 又是以亲子著称的。酒店的几个厅也很大气，可以用来给其他公司搞搞活动，米兰厅的外面的灯光装饰得很不错，酒店里面的内场也很大气，音响效果也非常地不错。 | 0 | woman | 上海 | 1702 | 东方佘山索菲特大酒店 |

## （二）监测体系的总体设计

基于非结构化数据的全域旅游发展监测体系由四项指数组成，分别是全域旅游吸引力指数、全域旅游满意度指数、全域旅游负极性指数、全域旅游主题分类指数。

全域旅游吸引力指数主要用于监测全域旅游示范区创建单位对

市民的吸引程度，全域旅游满意度指数主要用于监测市民对全域旅游示范区创建单位的整体满意程度，全域旅游负极性指数主要用于监测市民对全域旅游示范区创建单位负面评论占比程度，全域旅游主题分类指数主要用于反映在出现“红灯”警报的监测期内市民对全域旅游示范区创建单位的分类负面评论占比程度。

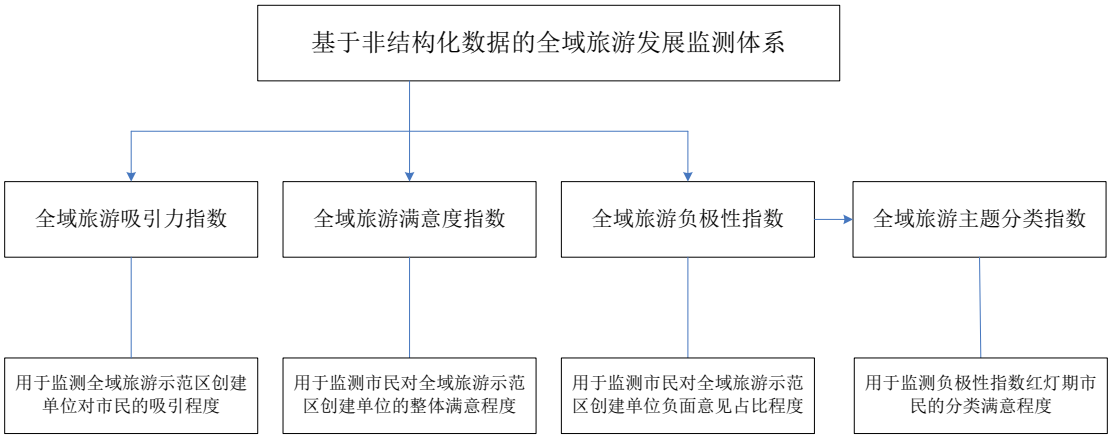


图 3 基于非结构化数据的全域旅游发展监测体系

## （三）全域旅游吸引力指数监测

### 1、指数编制方法

全域旅游吸引力指数主要用于监测全域旅游示范区创建单位对市民的吸引程度，按照公式 $x_i = \frac{n_i}{\max(n_1, n_2 \ldots n_i)}$ 计算评价对象指数得分，式中 $x_i$ 表示第 $i$ 个评价对象的指数得分，$n_i$ 表示第 $i$ 个评价对象的网民评论数，指数得分越高表示评价对象对市民的吸引程度越强。

### 2、指数编制结果

从整体得分看，松江区全域旅游吸引力指数的主要特征包括：**1、处于郊区领先**，黄浦区凭借行政和地理优势，指数得分在大部分季度保持最高，但与其他郊区相比，松江区则具有一定领先优势，指数得分

整体高于青浦区和崇明区。**2、回升趋势明显**，从 2014 到 2015 年，松江区的指数得分完成了“U”筑底，与旅游业收入的增长趋势相符，但从 2016 年 2 季度起开始逐步回升。**3、一季度被青浦赶超**，从 2011 到 2014 年，松江区的指数得分均高于青浦区，且领先优势较大，但 2015 年起连续三年 1 季度的得分均低于青浦区，主要由于青浦区多家地理位置相对集中的草莓采摘园对市民形成了较强的品牌和集聚效应。

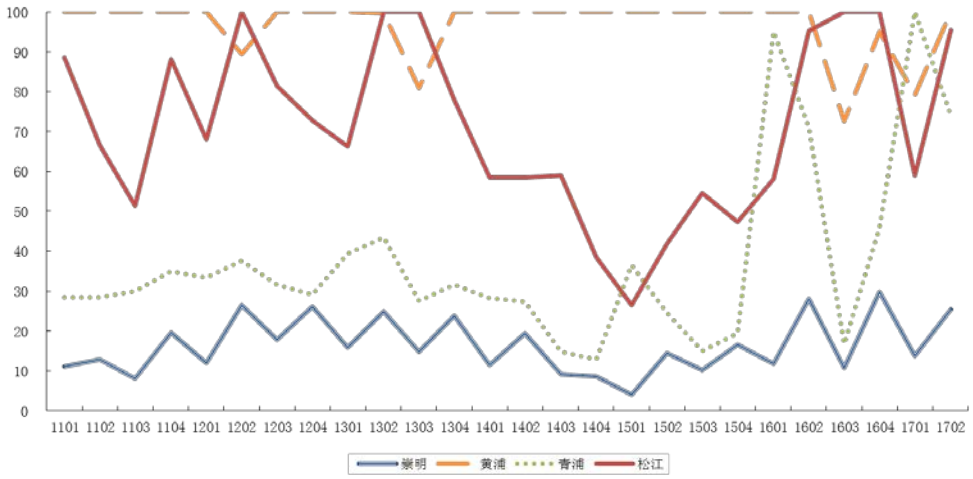


图 4 2011 年 1 季度-2017 年 2 季度全域旅游吸引力指数（总得分）

从分项得分看，松江区全域旅游吸引力指数的主要特征包括：**1、按监测类型分，旅游景点得分高于旅游住宿得分**，松江区旅游景点的指数得分均值为 75.8，高于旅游住宿的指数得分均值 55.8。事实上，就旅游住宿吸引力而言，松江区、青浦区和崇明区暂时都难以对黄浦区构成竞争。**2、按网民地区分，本地网民得分大幅高于外地网民得分**，松江区本地网民的指数得分均值为 78.2，大幅高于外地网民的指数得分均值 22.1。同样的，就对外地网民的吸引力而言，松江区、青浦区和崇明区也都难以对黄浦区构成竞争。**3、按网民性别分，女性网民得分高于男性网民得分**，松江区男性网民和女性网民的指数得

分波动趋势基本一致，只是女性网民的指数得分均值为 72.5，稍高于男性网民的指数得分均值 61.6。

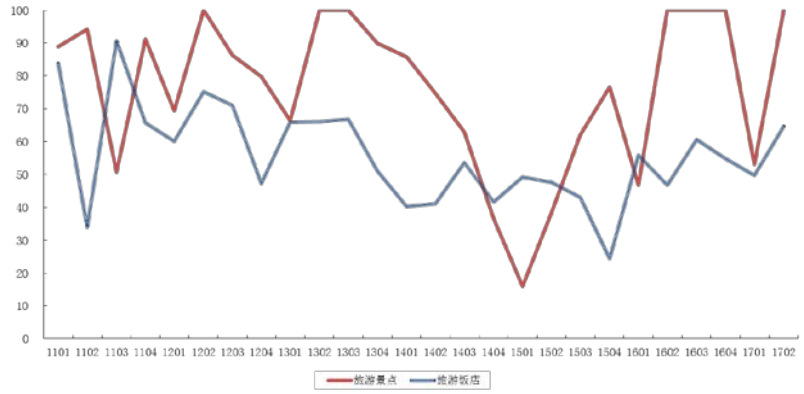


图 5 2011 年 1 季度-2017 年 2 季度松江区全域旅游吸引力指数按监测类型分

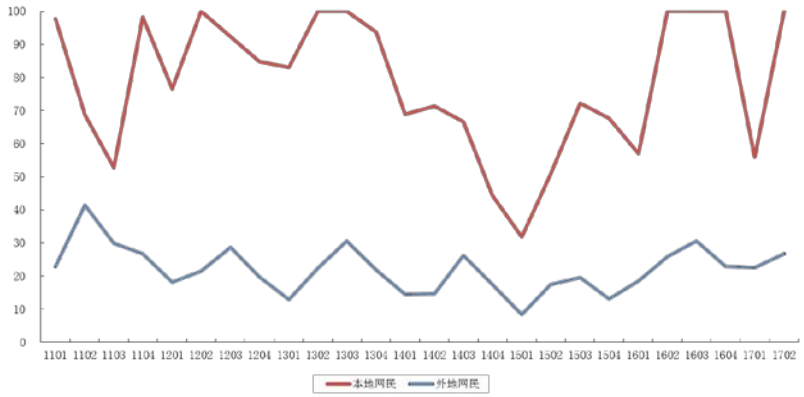


图 6 2011 年 1 季度-2017 年 2 季度松江区全域旅游吸引力指数按网民地区分

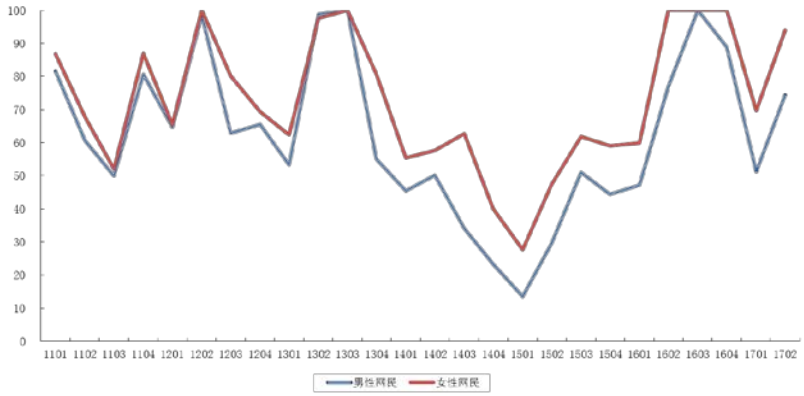


图 7 2011 年 1 季度-2017 年 2 季度松江区全域旅游吸引力指数按网民性别分

## 3、小结

全域旅游吸引力指数的监测结果反映出松江区具有的主要优势包括：第一，处于郊区领先，整体强于青浦区和崇明区。第二，回升趋势明显，在 2014 和 2015 年完成筑底后逐步回升。第三，旅游景点

吸引力较强，以欢乐谷、辰山植物园、泰晤士小镇等为代表。第四，对本地网民吸引力较强，在部分监测期内可与黄浦区形成一定竞争。

然而，需关注的主要问题则包括：第一，对青浦的优势在缩小，年度均值领先幅度从 40 以上降至 30 左右，且一季度被青浦超越。第二，旅游住宿吸引力降低，年度均值从前三年的 60 以上降至近三年的 50 左右。第三，对外地网民吸引力较弱，不仅大幅落后于黄浦区，而且也未出现明显的回升态势。

**（三）全域旅游满意度指数监测**

**1、指数编制方法**

全域旅游满意度指数主要用于监测市民对全域旅游示范区创建单位的满意程度，按照公式 $x_i=\frac{\sum p_{ij}(1+\frac{m_{ij}}{1000})}{n_i+\frac{\sum m_{ij}}{1000}}$ 计算评价对象指数得分，式中 $x_i$ 表示第 $i$ 个评价对象的指数得分，$n_i$ 表示第 $i$ 个评价对象的网民评论数，$p_{ij}$ 表示第 $i$ 个评价对象的第 $j$ 个网民的评分，$m_{ij}$ 表示第 $i$ 个评价对象的第 $j$ 个网民的获赞数，指数得分越高表示市民对评价对象的满意程度越强。

**2、指数编制结果**

从整体得分看，松江区全域旅游满意度指数的主要特征包括：**1、逐年稳定提高**，除 2013 年稍有波动外，松江区的指数得分呈现逐步提高的稳定态势，从 2011 年均值 79.8 提高到 2016 年均值 85.0。**2、全面被青浦超越**，与吸引力指数一季度被青浦区赶超不同，满意度指数则在近年的绝大多数季度内均低于青浦区，且差距有所扩大，2015 至

2016 年均值分别低于青浦区 1.9、3.4。**3、三季度满意度偏低**，按季度均值分析，松江区四个季度的指数均值分别为 81.2、81.5、78.9 和 83.1，三季度的满意度最低，玛雅海滩水公园、上海之根雪浪湖度假村等对指数得分影响较大。

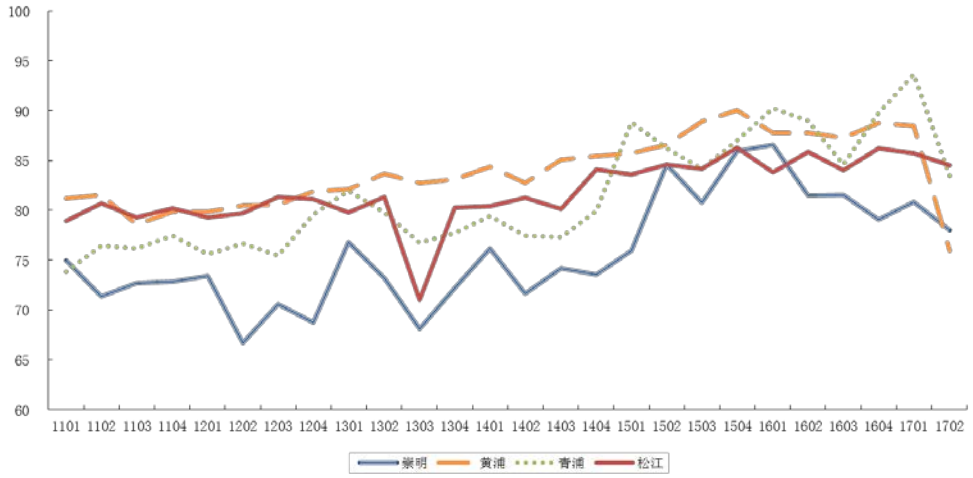


图 8 2011 年 1 季度-2017 年 2 季度全域旅游满意度指数（总得分）

从分项得分看，松江区全域旅游满意度指数的主要特征包括：**1、按监测类型分，旅游景点与旅游住宿得分相关度较弱**，松江区旅游景点与旅游住宿指数得分的波动趋势相互交错，两者间的相关系数仅为 0.205，不同步性较强，且明显低于其他三区。**2、按网民地区分，外地网民得分高于本地网民得分**，松江区外地网民的指数得分均值为 84.4，高于本地网民的指数得分均值 81.6。**3、按网民性别分，女性网民得分高于男性网民得分**，与吸引力指数情况相同，松江区男性网民与女性网民指数得分的波动趋势基本一致，只是女性网民的指数得分均值为 81.9，稍高于男性网民的指数得分均值 80.8。

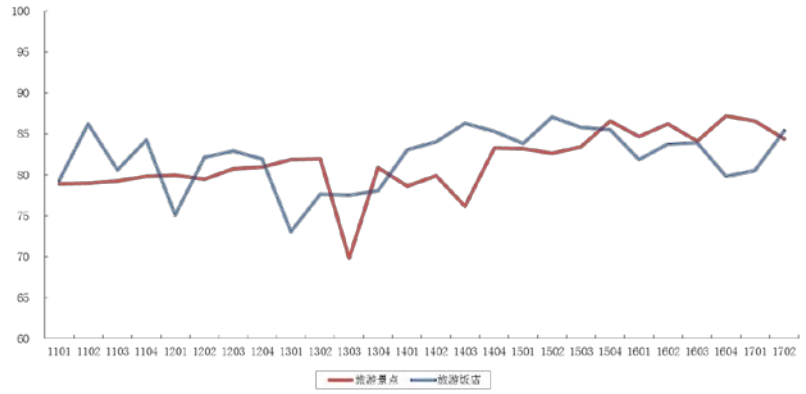


图 9 2011 年 1 季度-2017 年 2 季度松江区全域旅游满意度指数按监测类型分

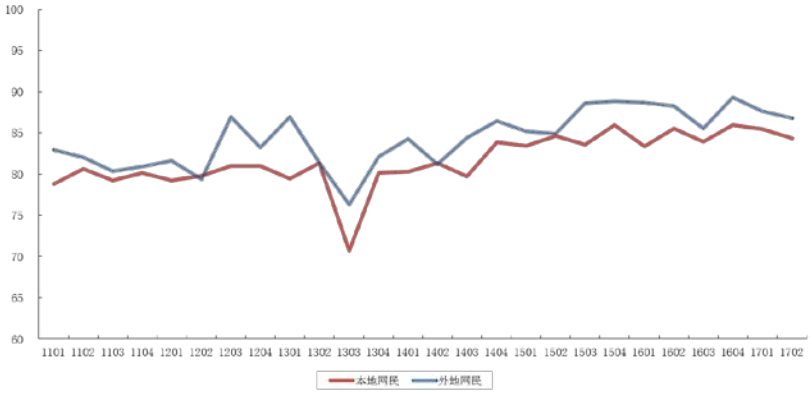


图 10 2011 年 1 季度-2017 年 2 季度松江区全域旅游满意度指数按网民地区分

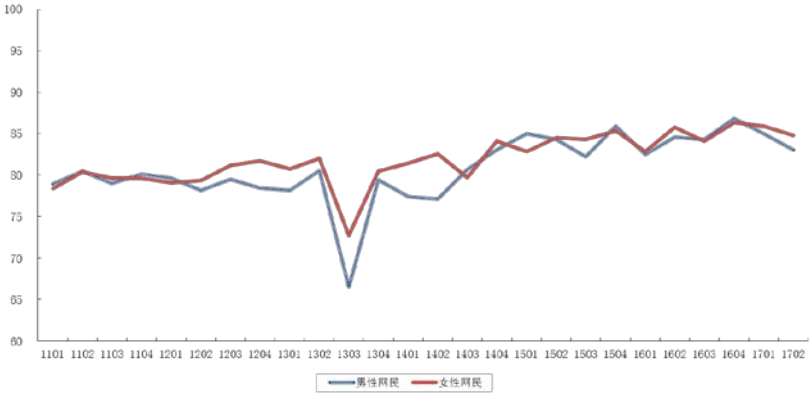


图 11 2011 年 1 季度-2017 年 2 季度松江区全域旅游满意度指数按网民性别分

## 3、小结

全域旅游满意度指数的监测结果反映出松江区具有的主要优势包括：第一，逐年稳定提高，年度均值从 79.8 提高到 85.0。第二，外地网民满意度较高，明显高于青浦区和崇明区，仅略低于黄浦区。

然而，存在的主要问题则包括：第一，近年被青浦超越，从 2011 年领先 3.8 到 2016 年落后 3.4。第二，三季度得分偏低，玛雅海滩

水公园、上海之根雪浪湖度假村等重点监测对象对三季度满意度得分的影响较大。第三，旅游景点与旅游住宿相关度较弱，市民对两者的满意度存在一定差异，同步性和协调性较差。

**（四）全域旅游负极性指数监测**

**1、指数编制方法**

全域旅游负极性指数主要用于监测市民对全域旅游示范区创建单位负面评论的占比程度，使用文本情感分析模型对评价对象进行评分，指数得分越高表示市民对评价对象负面评论的占比程度越高。

文本情感分析就是使用自然语言处理等统计建模技术来识别和提取原始语料中的观点、态度等主观信息，可以基于情感词典，也可以基于机器学习。

指数编制的主要步骤:**第一步，语料清洗。**使用 UFT8 编码按行分别读取松江区、黄浦区、青浦区、崇明区的网民评论数据，去除标点、符号及空格，删除字符数为 0 的评论记录。**第二步，初步分词。**调用 JiebaR 包的 worker 函数作为分词引擎对初步清洗后的语料按照词进行切分，并进行词性标注。**第三步，停用词处理。**先通过词性标注去除停用词，将代词、助词、连词、介词、语气词去除，再通过停用词词典去除停用词，本文使用的是哈工大、川大、百度等相关资源的集成版。**第四步，扩充情感词典。**把《知网》的 HowNet 情感词典和台湾大学中文情感词典合并去重后作为种子词，先获得与种子词余弦值较大的相关词，再计算种子词与相关词的语义相似度，保留语义相似度较大的相关词，最终得到扩充后的情感词典。**第五步，文本极性分**

**析。**使用词频和逆向文档频率计算特征权重,使用向量空间模型将每条评论记录转化为特征向量，并使用基于情感词典和基于机器学习两种方法将评论记录分为正面评论和负面评论两类。**第六步，定义负面评论记录。**同时满足下列两条及以上规则，即被标注为负面评论记录。“评分”字段小于等于 2；基于情感词典的评论记录得分小于 1；被 LibSVM 分类器归类于负面评论。**第七步，计算评价对象得分。**按照公式 $x_i=\frac{n_i^{'}+\frac{\sum m_i^{'}}{100}}{n_i+\frac{\sum m_i}{100}}$ 计算评价对象指数得分，式中 $x_i$ 表示第 $i$ 个评价对象的指数得分，$n_i$ 表示第 $i$ 个评价对象的网民评论数，$n_i^{'}$ 表示第 $i$ 个评价对象的网民负面评论数，$m_i$ 表示第 $i$ 个评价对象的网民评论数获赞数，$m_i^{'}$ 表示第 $i$ 个评价对象的网民负面评论数获赞数。

**2、扩充情感词典的统计建模**

**（1）发现新词并再次分词**

在对语料进行初步分词后会发现一些专有名词往往被切分成了多个字符,一般的解决方法是将其人工识别并添加进自定义词典后再重新分词，但当语料规模较大、专有名词较复杂时，这种人工识别的方法就较为低效。本文设定字符长度为 6，词频阈值为 10,使用互信息值和信息熵作为判断字符串能否成为新词的评价指标,将新词添加进自定义词典后,对语料进行再次分词。

本文使用的新词评价指标的计算公式为 $value=\log(MI+1)+\log(\min(HL,HR)+1)$，式中，$MI$ 表示字符串的互信息值，$HL$、$HR$ 表示字符串的左邻字集和右邻字集的信息熵，该指标取值越

大表示特定字符串成为新词的可能性越大。

互信息值的计算公式为 $MI(t)=\log(\frac{p(t)}{p(x)p(y)})$，式中，$p(t)=n_t/N$，$p(x)=n_x/N$，$p(y)=n_y/N$，$n_t$、$n_x$、$n_y$分别表示字符串$t$、$x$、$y$在语料中出现的频次，$N$表示语料中字符长度满足阈值的字符串总数，该指标取值越大表示字符串内部凝聚程度越强。信息熵的计算公式为 $HL(t)=-\sum_x p(x|t)\log p(x|t)$，$HR(t)=-\sum_x p(y|t)\log p(y|t)$，式中，$p(x|t)$表示字符串$x$是字符串$t$左邻字符的概率，$p(y|t)$表示字符串$y$是字符串$t$右邻字符的概率，该指标取值越大表示字符串边界划分能力越强。

**（2）扩充情感词典**

第一步，构建基础情感词典。对《知网》的 HowNet 情感词典和台湾大学的中文情感词典进行合并。其中，HowNet 情感词典包含积极情感词 4566 个，消极情感词 4370 个；中文情感词典包含积极情感词 2810 个，消极情感词 8276 个。经过合并去重后，得到的基础情感词典包含积极情感词 6506 个，消极情感词 11184 个，将该基础词典作为种子词，并也转换为词向量。

第二步，初次扩充情感词典。使用转化好的词向量模型寻找与种子词向量夹角较小，余弦值较大的词集合，设定余弦值阈值为 0.6，删除余弦值小于 0.6 的词，保留余弦值大于等于 0.6 的词，完成对基础情感词典的初次扩充。

第三步，计算语义相似度，再次扩充情感词典。哈工大的《同义词词林（扩展版）》包含 77343 个词，按五级分类结构编排，把存在于《词林》中的词称为原子术语（PT），把不存在于《词林》中的词

称为组合术语（CT）。对于两个原子术语语义相似度的计算公式为 $Sim(PT_1,PT_2)=\max_{c_1\in Code(PT_1)}\max_{c_2\in Code(PT_2)}Sim(c_1,c_2)$，式中，$c_1$、$c_2$表示原子术语的编码，$Sim(c_1,c_2)=\frac{Spd(c_1,c_2)}{5}$，$Spd(c_1,c_2)$表示$c_1$、$c_2$的重合度。对于其他情况语义相似度的计算公式为 $Sim(T_1,T_2)=\alpha\times(\frac{1}{m}+\frac{1}{n})\times\sum_{i=1}^{m}Sim(PT_{1,i},PT_{1,j_i})+(0.5-\alpha)\times\frac{m}{n}\times\sum_{i=1}^{m}\left\{\left[\frac{R(T_1,PT_{1,i})}{\sum_{i'=1}^{m}i'}+\frac{R(T_2,PT_{2,j_i})}{\sum_{j'=1}^{m}j'}\right]\times Sim(PT_{1,i},PT_{1,j_i})\right\}$，式中，$PT_{1,i}$表示术语$T_1$中第$i$个原子术语，$PT_{2,j_i}$表示$T_2$中与$PT_{1,i}$语义相似度最大的原子术语，$m$、$n$分别表示术语$T_1$、$T_2$包含原子术语的个数，$\frac{R(T_1,PT_{1,i})}{\sum_{i'=1}^{m}i'}$、$\frac{R(T_2,PT_{2,j_i})}{\sum_{j'=1}^{m}j'}$分别表示$PT_{1,i}$、$PT_{2,j_i}$在术语$T_1$、$T_2$中所处位置的权重之和（根据词汇“重心后移”思想赋权），并令$\alpha$取值为 0.3。计算种子词与初次扩充词的语义相似度，保留语义相似度大于等于 0.6 的词，完成对基础情感词典的再次扩充。

经过对情感词典的两次扩充，得到的扩充积极情感词 104 个，扩充消极情感词 191 个。由此，最终得到的情感词典包含积极情感词 6610 个，消极情感词 11375 个。

### 3、文本极性分析的统计建模

#### （1）计算特征权重

使用词频和逆向文档频率计算特征权重，其是一种可用于文档检索和文本挖掘的权重计算基数，能够评价出一个词在一个文档集合中的重要性，若某词在某篇文档中出现次数较多而在其他文档出现次数较少，那么该词就是特征词。

对于文档$d_j$中的词$\omega_i$，其 TFIDF 值定义为$TFIDF_{w_i d_j} = TF_{w_i d_j} \times IDF_{w_i}$，TFIDF 值越大表示该词越重要。TF 词频，是指词$w_i$在文档 $d_j$中出现的频数除以所有词在文档$d_j$出现的频数和，代表词出现的频繁程度，公式为$TF_{w_i d_j} = \frac{I(w_i, d_j)}{\sum_i I(w_i, d_j)}$。IDF 逆向文档频率，是指文档总数$|D|$除以包含词$w_i$的文档数加 1 的对数，代表词的普遍程度，公式为$IDF_{w_i} = \log(\frac{|D|}{1+|\{j: w_i \in d_j\}|})$。

**（2）转化特征向量**

使用向量空间模型将文本转化为计算机能够处理的结构化数据。向量空间模型的基本思想是，把每条评论记录看作一个词袋，把每个词看作一个特征项，从而使得每条评论记录可以用由词组成的特征向量来表示。

首先，使评论记录的所有词组成特征空间$\Omega=\{t_1, t_2, \cdots, t_n\}$，式中，$t_i$表示所有词的集合，n表示特征空间$\Omega$的维度。然后，用特征空间$\Omega$中的一组向量来表示一条评论记录$d=\{x_1, x_2, \cdots, x_n\}$，式中，$x_i$表示评论记录d中第i个词的权重。

**（3）基于情感词典的文本极性分析**

第一步，导入各类词典，导入扩充后的情感词典、否定词词典、程度副词词典。第二步，定义情感词组，两情感词间的所有否定词和程度副词与这两情感词中的后一情感词构成一个情感词组。第三步，计算情感词组得分。情感词组得分=否定词分值 × 程度副词分值 × 情感词分值。第四步，计算评论记录得分。记录得分=Σ情感词组得分，评论记录得分大于 1 时标注为正面评论，评论记录得分小于 1 时标注

为负面评论。

**（4）基于机器学习的文本极性分析**

使用支持向量机作为文本极性分类器，其可将低维度无法进行线性分开的问题，映射到高维空间，在高维空间中找到一个最优决策平面，即最优超平面，这个平面能最好地分割两个分类中的数据点。

第一步，构建训练集。从全部评论记录中选取 5 万条正面评论，选取 5 万条负面评论构成训练集，由于数量较多，不便于人为挑选，因此定义选取规则。正面评论在“评分”字段大于等于 4 且基于情感词典极性分析得分大于 1 的评论记录中随机挑选;负面评论在“评分”字段小于等于 2 且基于情感词典分析得分小于 1 的评论记录中随机挑选。第二步，设置模型参数。调用 WEKA 软件的 LibSVM 工具包作为分类器，选择径向基函数为核函数，并进行十折交叉验证。第三步，训练并评价模型。分别在 200 维、400 维、600 维、800 维、1000 维特征水平下训练模型，使用准确率、召回率、F1 值评价模型性能。在 1000 维特征水平下，三项指标值均接近 90%，效果基本令人满意，可用于分类全部评论记录。

表 6 不同特征维度模型性能对比

| | Precision(%) | Recall(%) | F1(%) |
|---|---|---|---|
| 200 维特征 | 76.9 | 76.8 | 76.7 |
| 400 维特征 | 82.3 | 81.7 | 81.6 |
| 600 维特征 | 86.1 | 85.7 | 85.7 |
| 800 维特征 | 88.0 | 88.0 | 88.0 |
| 1000 维特征 | 89.1 | 89.1 | 89.0 |

## 4、指数编制结果

从整体得分看，松江区全域旅游负极性指数的主要特征包括：1、

**负面评论占比波动程度最大**，按历史数据分析，松江区指数得分的最低值是 2011 年 4 季度的 10.8，最高值则是 2013 年 3 季度的 31.6，离散系数达到 0.253，高于其他三区。**2、负面评论占比近年偏高**，按年度均值分析，松江区 2015、2016 两年的指数得分均值已达到 20.4 和 21.5，不仅高于其他三区，而且差距有所扩大。**3、三季度负面评论占比最高**，按季度均值分析，松江区四个季度的指数得分均值分别为 16.9、16.2、21.4 和 16.9，三季度的负面评论占比最高，与满意度指数的主要特征一致。

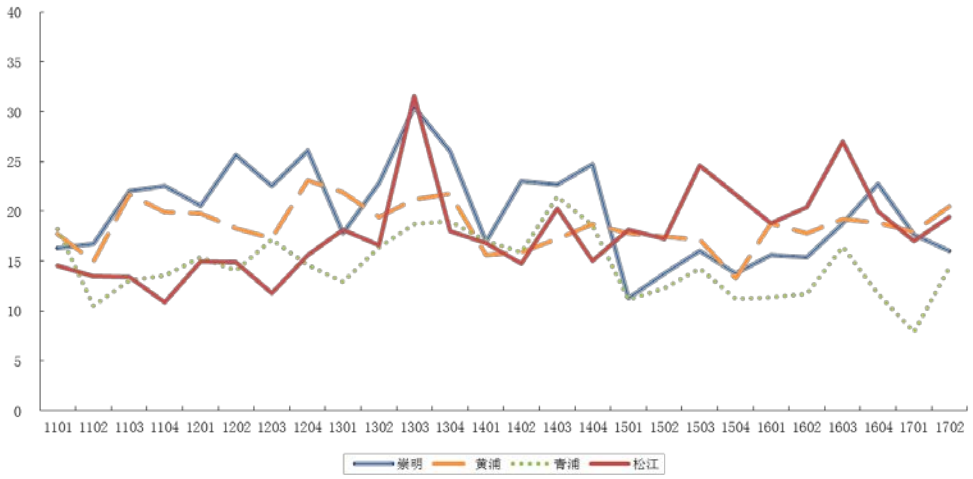


图 12 2011 年 1 季度-2017 年 2 季度全域旅游负极性指数（总得分）

从分项得分看，松江区全域旅游负极性指数的主要特征包括：**1、按监测类型分，旅游景点负面评论占比高于旅游住宿**，松江区旅游景点的指数得分均值为 19.8，高于旅游住宿的指数得分均值 12.4。**2、按网民地区分，本地网民负面评论占比高于外地网民得分**，松江区本地网民的指数得分均值为 17.9，稍高于外地网民的指数得分均值 17.1。**3、按网民性别分，女性网民负面评论占比高于男性网民得分**，松江区男性网民的指数得分均值为 20.3，高于女性网民的指数得分均值 16.9。

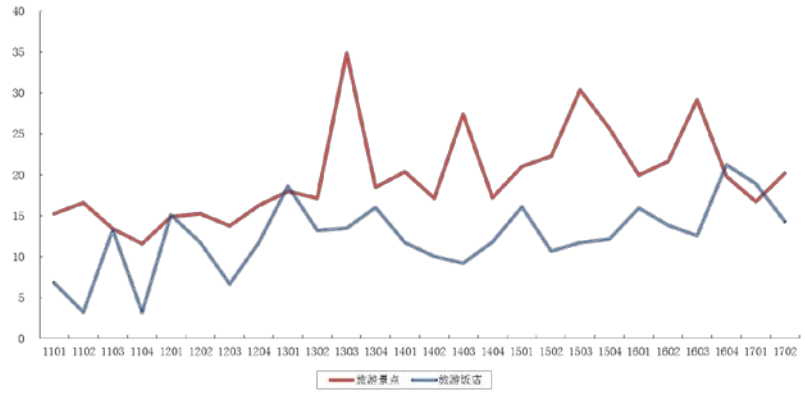


图 13 2011 年 1 季度-2017 年 2 季度松江区全域旅游负极性指数按监测类型分

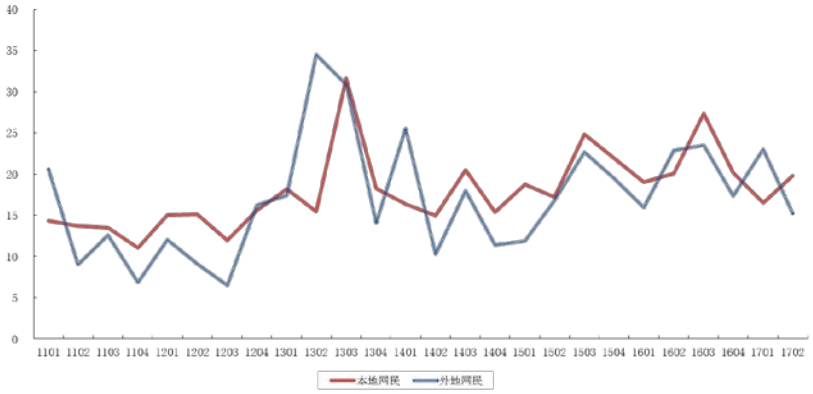


图 14 2011 年 1 季度-2017 年 2 季度松江区全域旅游负极性指数按网民地区分

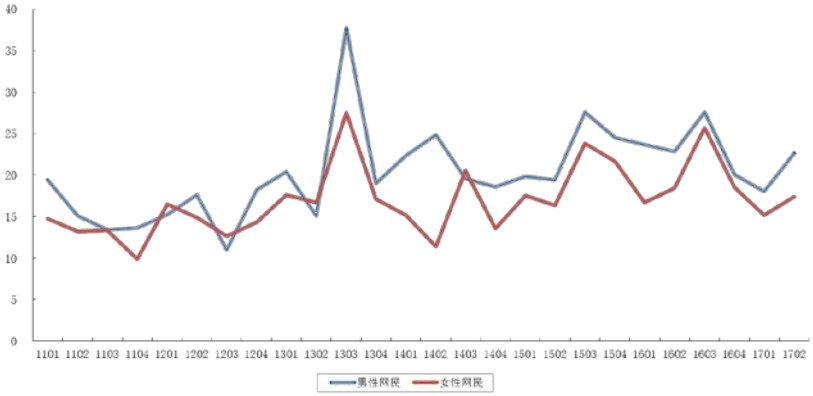


图 15 2011 年 1 季度-2017 年 2 季度松江区全域旅游负极性指数按网民性别分

## 5、小结

负极性指数则侧重揭示松江区全域旅游发展的短板和不足，监测结果主要反映三方面信息。

第一，负面评论占比波动大，离散系数高于其他三区，表示受季节性或随机性因素影响较大。第二，负面评论占比走势弱，指数得分年度均值从 2011 年的 13.1 提高至 2016 年的 21.5，表示存在的问题

不仅较为复杂，而且有加重趋势。第三，各个类别均不具优势，特别在 2015 年以来，不论是旅游景点还是旅游住宿，不论是本地网民还是外地网民，不论是男性网民还是女性网民，负面评论占比均高于其他三区。

**（五）全域旅游主题分类指数监测**

**1、指数编制方法**

全域旅游负极性指数的监测结果侧重反映松江区全域旅游发展的短板和不足,而主题分类指数则进一步揭示短板和不足的主要方面，本文使用 LDA 模型对指数得分偏高的监测期的负面评论进行隐含主题挖掘。

LDA 是一种文档主题生成模型，包含词、主题和文档三层结构，可以用来识别大规模文档集或语料库中隐含的主题信息,文档的每个词都是通过“以一定概率选择了某个主题，并从这个主题中以一定概率选择某个词”这样一个过程得到，文档到主题服从多项式分布，主题到词服从多项式分布。

指数编制的主要步骤:**第一步,设定“红灯”警报线。**设定各监测期内四家国家全域旅游示范区创建单位负极性指数均值加上一个偏离度作为“红灯”警报线。**第二步,构建 LDA 模型。**对出现“红灯”警报监测期内的负面评论初步清洗后按分隔符号或空格切割,并进行分词和去停用词操作，再调用 topicmodels 包构建 LDA 模型，设定 $\alpha = 0.1$，$\beta = 0.02$，根据五折交叉检验的 Perplexity 统计量值设定主题数量为 20。**第三步，主题归类。**对 LDA 模型输出的 20 个主题结

果进一步归整为“客流/排队耗时”、“服务/设施/环境”、“饮食便捷度/性价比”、“交通/停车”、“票价/房价/租金”五类综合主题。**第四步，计算评价对象得分。**对于某个综合主题，按照公式 $x_i = \frac{n_i^{'} + \frac{\sum m_i^{'}}{100}}{n_i + \frac{\sum m_i}{100}}$ 计算评价对象指数得分，式中 $x_i$ 表示第 $i$ 个评价对象的指数得分，$n_i$ 表示第 $i$ 个评价对象的网民评论数，$n_i^{'}$ 表示第 $i$ 个评价对象的网民负面评论数，$m_i$ 表示第 $i$ 个评价对象的网民评论数获赞数，$m_i^{'}$ 表示第 $i$ 个评价对象的网民负面评论数获赞数。特别的，由于一条负面评论经切割后可能涉及多个综合主题，因此出现“红灯”警报的监测期内的各主题分类指数之和往往大于负极性指数。

**2、指数编制结果**

**按警报分布情况分析：从警报数量看，**松江区出现“红灯”警报的有 6 期，占全部监测期的 23.1%，低于崇明区的 50.0%，但高于黄浦区的 11.5%和青浦区的 3.8%。**从警报时期看，**松江区出现“红灯”警报的监测期均在 2013 年以后，且半数发生在三季度。反映出随着对游客吸引力的逐步增强，松江区全域旅游发展的短板和不足也经受着越来越严峻的考验，并可能借由网络等信息传播渠道被不断放大。

表 7 2011 年 1 季度-2017 年 2 季度全域旅游“红灯”警报对象

| 监测期 | “红灯”警报对象 | 监测期 | “红灯”警报对象 |
|---|---|---|---|
| 1101 | 青浦 | 1402 | 崇明 |
| 1102 | 崇明 | 1403 | 崇明 |
| 1103 | 崇明 | 1404 | 崇明 |
| 1104 | 崇明 | 1501 | 松江 |
| 1201 | 崇明 | 1502 | 黄浦 |
| 1202 | 崇明 | 1503 | 松江 |
| 1203 | 崇明 | 1504 | 松江 |
| 1204 | 崇明 | 1601 | 无 |
| 1301 | 黄浦 | 1602 | 松江 |
| 1302 | 崇明 | 1603 | 松江 |
| 1303 | 松江 | 1604 | 崇明 |
| 1304 | 崇明 | 1701 | 无 |
| 1401 | 无 | 1702 | 黄浦 |

**按分类指数情况分析：1、旅游旺季“客流/排队耗时”负面评论占比最高。**2013 年 3 季度，该主题的负面评论有 1152 条，占当期全部评价的比重为 19.1%；2015 年 3 季度，该主题的负面评论有 1404 条，占当期全部评价的比重为 14.6%；2015 年 4 季度，该主题的负面评论有 1512 条，占当期全部评价的比重为 20.0%；2016 年 3 季度，该主题的负面评论有 2385 条，占当期全部评价的比重为 21.3%。**2、旅游淡季“服务/设施/环境”负面评论占比较高。**2015 年 1 季度，该主题的负面评论有 756 条，占当期全部评价的比重为 17.8%；2016 年 2 季度，该主题的负面评论有 630 条，占当期全部评价的比重为 9.1%。**3、“交通/停车”负面评论占比较稳定。**即便在旅游旺季该主题的负面评论占比也处于相对稳定的较低水平，主要由于选择公共交通出行的暑期学生客流占比较大，且部分景点提供的免费接驳服务比

较到位。

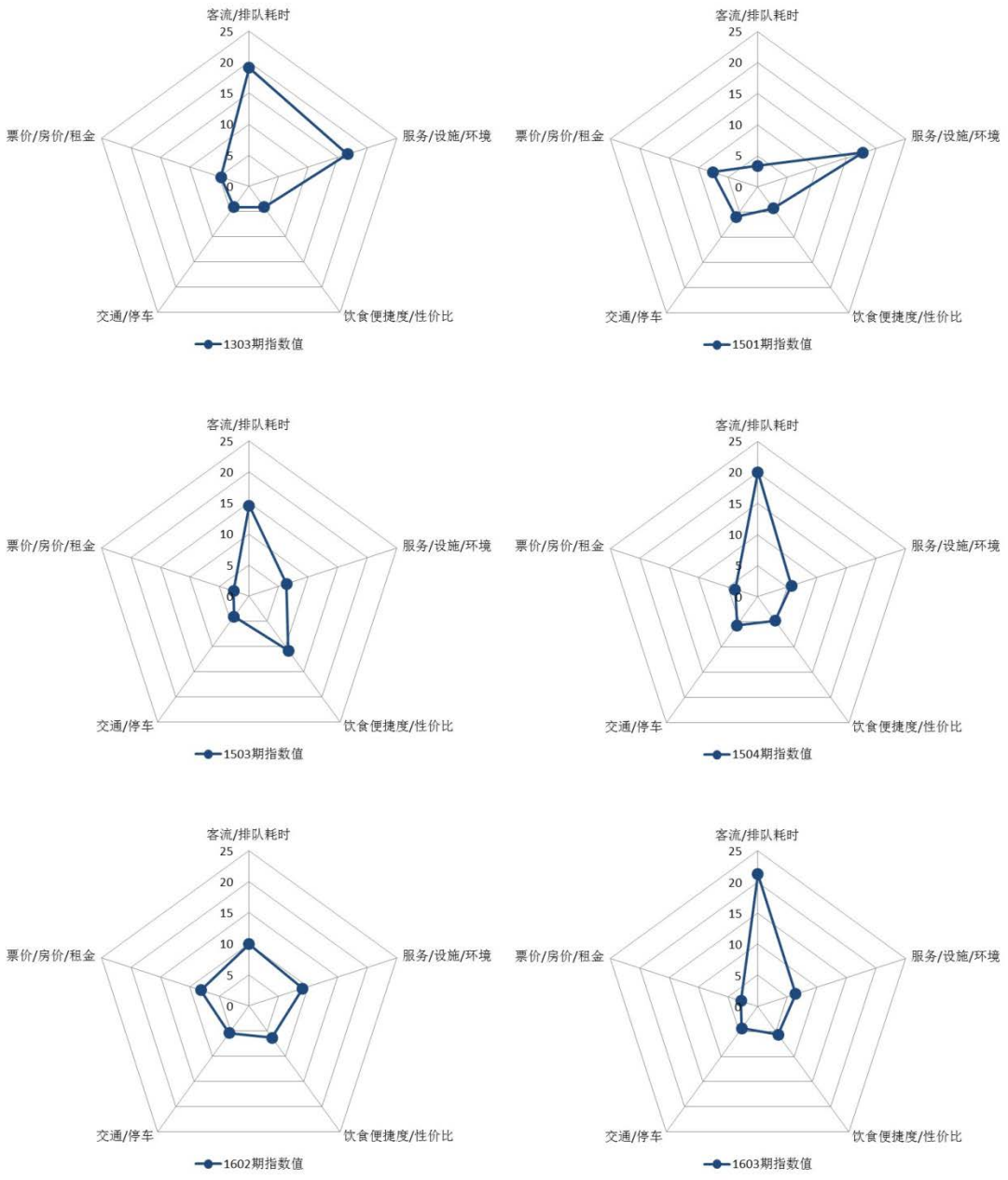


图 16 2011 年 1 季度-2017 年 2 季度松江区全域旅游“红灯”警报期主题分类指数

## 3、小结

主题分类指数在负极性指数的基础上进一步分析了在出现“红灯”警报的监测期内市民对各综合主题的负面评论占比程度，监测结果主要反映三方面信息。

第一，“红灯”警报出现的频率在提高，松江区共出现六次“红

灯”警报，2015 年以前仅 1 次，2015 年以后有 5 次。第二，两项分类主题影响“红灯”警报出现，旅游旺季的“红灯”警报主要受“客流/排队耗时”主题负面评论占比较高影响，旅游淡季的“红灯”警报主要受“服务/设施/环境”主题负面评论占比较高影响。第三，个别特殊事件导致“红灯”警报出现，大部分“红灯”警报的出现与该监测期内的个别特殊事件密切相关，比如 1303 期“红灯”警报对应的特殊事件是“玛雅海滩水公园”开园，1504 期“红灯”警报对应的特殊事件是“欢乐谷”举办“万圣节”活动等。

## 四、下阶段，松江区全域旅游发展的对策建议

本文通过爬取 2011 年 1 季度至 2017 年 2 季度国内知名旅游网站的近 45 万条网民评论数据，使用文本挖掘和机器学习等统计建模方法，从“吸引力”、“满意度”、“负极性”、“主题分类”四个维度对黄浦、青浦、松江、崇明四家国家全域旅游示范区创建单位的全域旅游发展情况进行监测。

根据监测结果反映的主要问题，下阶段，松江区需要着重处理好三项矛盾，以实现全域旅游发展的三个“全”。即处理好旅游旺季和旅游淡季不协调的矛盾，以实现全域旅游季节维度的“全”；处理好旅游景点与旅游住宿不融合的矛盾，以实现全域旅游类型维度的“全”；处理好部门数据信息与产业发展信息不对称的矛盾，以实现全域旅游管理维度的“全”。

### （一）处理好旅游旺季和旅游淡季不协调的矛盾，以实现全域旅游季节维度的“全”

监测结果反映的松江区全域旅游发展的第一项矛盾是旅游旺季和旅游淡季的不协调。

一方面，旅游淡季的吸引力指数偏低。一季度吸引力指数各年均值仅为 61.0，而其他三季度均在 70 以上。更需要注意的是，一季度吸引力指数从 2015 年起连续三年被青浦区超越，且差距被不断拉大。另一方面，旅游旺季的满意度指数偏低、负极性指数偏高。三季度满意度指数各年均值仅为 80.0，而其他三季度均在 81 以上；负极性指数各年均值为 21.4，而其他三季度均在 17 以下。此外，主题分类指数显示，旅游旺季的“红灯”警报主要受“客流/排队耗时”主题负面评论占比较高影响，旅游淡季的“红灯”警报主要受“服务/设施/环境”主题负面评论占比较高影响。

建议通过以下“两个提供”以实现全域旅游季节维度的“全”。

**一是提供更丰富的旅游产品。**松江区全域旅游的发展既需要“欢乐谷”、“辰山植物园”、“佘山国家森林公园”、“泰晤士小镇”等核心旅游产品的引领，也需要“醉白池”、“方塔园”等老牌旅游产品的支撑，还需要农家乐、采摘园、郊野公园等潜力旅游产品的助推，更可以设计“亲子游”、“怀旧游”、“科创游”等特色旅游产品。青浦区正是依靠多家地理位置相对集中的草莓采摘园形成集聚效应和品牌优势，从而实现一季度吸引力指数对松江区的超越。

**二是提供更优质的旅游服务。**旅游住宿方面，尤其是度假式、经济型的宾馆、旅店，需要注意旅游淡季，特别是春节长假期间，由于员工配备不足、节约经营成本导致的服务设施不完备、服务质量不到

位的问题。旅游景点方面，尤其是热门景点，可以尝试通过优惠非高峰时段门票价格、引入游玩项目预登记和游玩人数提醒等微信推送服务、增设等候区纳凉和遮阴设施等多种措施分流游客数量、减少游客投诉。

**（二）处理好旅游景点和旅游住宿不融合的矛盾，以实现全域旅游类型维度的“全”**

监测结果反映的松江区全域旅游发展的第二项矛盾是旅游景点和旅游住宿的不融合。

“吸引力”指数方面，旅游景点和旅游住宿的监测得分相关系数仅为 0.185，而本地网民和外地网民，以及男性网民和女性网民的监测得分相关系数分别为 0.411 和 0.941。“满意度”指数方面，旅游景点和旅游住宿的监测得分相关系数仅为 0.205，而本地网民和外地网民，以及男性网民和女性网民的监测得分相关系数分别为 0.813 和 0.889。“负极性”指数方面，旅游景点和旅游住宿的监测得分相关系数仅为 0.143，而本地网民和外地网民，以及男性网民和女性网民的监测得分相关系数分别为 0.594 和 0.778。

建议通过以下“两个联动”以实现全域旅游类型维度的“全”。

**一是联动三次产业与传统旅游。**从国民经济行业分类来看，传统旅游仅局限于第三产业，而全域旅游则可扩展至全部三次产业。相关职能部门需要进一步促进旅游人才、旅游资本、旅游科技等生产要素在不同产业间的集聚、溢出和流动，充分发挥“旅游+”的综合带动功能，形成全产业化的旅游产品和业态，推动传统旅游与其他产业共

生共荣，形成相关产业全域联动的大格局。

**二是联动旅游景点与旅游住宿、餐饮、零售。**如果说三次产业与传统旅游是产业间的横向联动，那么旅游景点与旅游住宿、餐饮、零售就是产业内的纵向联动。相关职能部门需要进一步统筹、整合区内旅游资源，增强旅游相关企业的信息交流，推动旅游相关企业的战略合作，塑造和延伸旅游相关产业链，为游客提供观光娱乐、住宿餐饮、休闲购物为一体的全方位旅游服务，实现政府、企业、市民共享全域旅游创造的发展红利。

**（三）处理好部门数据信息与产业发展信息不对称的矛盾，以实现全域旅游管理维度的“全”。**

监测结果反映的松江区全域旅游发展的第三项矛盾是部门数据信息与产业发展信息不对称的矛盾。

“吸引力”指数方面，对青浦优势缩小、旅游住宿吸引力降低、对外地网民吸引力较弱等发展问题无法由现行的部门数据发掘。“满意度”指数方面，近年被青浦超越、三季度得分偏低、旅游景点与旅游住宿相关度度较弱等发展问题无法由现行的部门数据发掘。“负极性”指数方面，负面评论占比波动大、负面评论占比走势弱、各个类别负面评论占比均不具优势等发展问题无法由现行的部门数据发掘。“主题分类”指数方面，“红灯”警报出现的频率在提高、两项分类主题影响“红灯”警报出现、个别特殊事件导致“红灯”警报出现等发展问题无法由现行的部门数据发掘。

建议通过以下“两个探索”以实现全域旅游管理维度的“全”。

**一是探索数据统计的部门合作机制。**松江区旅游产业正处于由传统旅游向全域旅游转型升级的关键期，而统计数据是体现产业发展阶段性成果的重要参考和依据，相关职能部门需要协同合作，进一步充实产业数据资源。建议由区统计局负责统计规模以上全域旅游相关法人单位数据，由区旅游局负责统计不到入统标准但仍具有一定规模的全域旅游相关法人单位数据，由区统计局、区旅游局共同推算其他规模以下全域旅游相关法人单位数据。

**二是探索产业发展的舆情监测机制。**统计数据能够侧重体现全域旅游的经济属性和发展成果，而舆情监测则能够侧重体现全域旅游的民生属性和发展短板，相关职能部门不仅需要关注数据也需要关注舆情。建议相关职能部门充分运用大数据思维、大数据资源和大数据技术，开发多维度、高频率、可视化的全域旅游舆情监测系统，及时发现发展问题、及时完善发展短板、及时应对特殊事件，有效提高松江区全域旅游的发展质量和发展水平，顺利完成示范区创建。

**参考文献**

[1] 常国珍,曾珂,朱江. 用商业案例学 R 语言数据挖掘[M]. 北京:电子工业出版社，2017.

[2] 隋浩. 基于 Word2Vec 的微博情感新词识别与倾向判断研究[D]. 广西:广西大学,2016.

[3] 袁磊. 评论文本情感分析算法的研究[D]. 安徽:合肥工业大学,2016.

[4] 韦强申. 领域关键词抽取:结合 LDA 与 Word2Vec[D]. 贵州:贵州师范大学,2016.

[5] 陈晓东. 基于情感词典的中文微博情感倾向分析研究[D]. 湖北:华中科技大学,2012.

[6] 李晓,解辉,李立杰. 基于 Word2vec 的句子语义相似度计算研究[J].计算机科学,2017,(9)：256-260.

[7] 徐硕,朱礼军,乔晓东,薛春香. 基于双序列比对的中文术语语义相似度计算的新方法[J]. 情报学报，2010，(4)：701-708.

[8] 夏天. 汉语词语语义相似度计算研究[J]. 计算机工程，2007，(6)：191-194.

[9] 章成志. 一种基于语义体系的同义词识别研究[J].淮阴工学院学报,2004,(1):59-67.

[10] 山文彩. 松江全域旅游发展路径探析[R]. 上海:上海市松江区旅游局，2017.

**课题组组长：潘永俭**
**课题组成员：宋莉 马一峰**
**联系人：马一峰 37735623**