博物馆新闻采集分析子系统测试报告

**目录**

[一、 导言 1](#_Toc17636)

[1.1 版本修订历史 1](#_Toc27960)

[1.2 目的 1](#_Toc3081)

[1.3 范围 1](#_Toc28958)

[二、 测试时间、地点和人员 2](#_Toc8823)

[三、 测试环境描述 2](#_Toc15630)

[四、 测试执行情况 2](#_Toc12302)

[（1）数据获取时间： 2](#_Toc25268)

[（2）数据入库时间 2](#_Toc27656)

[（3）数据入库更新对于其他子系统的影响 3](#_Toc24489)

[五、 测试结果分析 3](#_Toc15260)

[5.1 测试进度和工作量度量 3](#_Toc4655)

[5.2 综合数据分析 3](#_Toc25188)

[六、 测试评估 4](#_Toc31847)

# 导言

## 版本修订历史

|  |  |  |
| --- | --- | --- |
| **版本号** | **修订内容** | **修订者** |
| 1.0 | 系统测试报告 | 张浩杰 |
| 2.0 | 系统测试报告 | 张浩杰 |

## 目的

该文档的目的是描述子系统测试的总结报告。

本文档的预期读者是：

1. 项目管理人员
2. 测试人员

## 范围

该文档定义了开发端系统测试的结果，主要从数据获取，数据更新入库和情感分析方面进行测试，给出了测试的结论。

## 1.4 缩写说明

1. Scrapy

    Scrapy 是用 Python 实现的一个为了爬取网站数据、提取结构性数据而编写的应用框架。

     Scrapy 常应用在包括数据挖掘，信息处理或存储历史数据等一系列的程序中。

    通常我们可以很简单的通过 Scrapy 框架实现一个爬虫，抓取指定网站的内容或图片。

1. sklearn

    Scikit-learn（也称为sklearn）是Python编程语言的免费软件机器学习库。它具有各种分类、回归和聚类算法，包括支持向量机、随机林、梯度提升、k-均值和DBSCAN，旨在与 Python 数值和科学库NumPy和SciPy进行互操作。

1. Snownlp

SnowNLP是一个python写的类库，可以方便的处理中文文本内容，是受到了TextBlob的启发而写的，由于现在大部分的自然语言处理库基本都是针对英文的，于是写了一个方便处理中文的类库，并且和TextBlob不同的是，这里没有用NLTK，所有的算法都是自己实现的，并且自带了一些训练好的字典。

## 1.5 术语定义

1. 爬虫

网络爬虫（又称为网页蜘蛛，网络机器人，在FOAF社区中间，更经常的称为网页追逐者），是一种按照一定的规则，自动地抓取万维网信息的程序或者脚本。另外一些不常使用的名字还有蚂蚁、自动索引、模拟程序或者蠕虫。

1. 机器学习

机器学习是一门多领域交叉学科，涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。专门研究计算机怎样模拟或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构使之不断改善自身的性能。

它是人工智能的核心，是使计算机具有智能的根本途径。

1. 文本分析

文本分析是对文本进行解析，以便从中提取机器可读的事实。文本分析的目的是用自由文本内容创建结构化数据。这个过程可以看作是将一堆非结构化、异构的文档分割成易于管理和解释的数据块。

### 1.6 参考资料

1. python参考资料：

[https://github.com/wistbean/learn\_python3\_spider](https://github.com/wistbean/learn\_python3\_spider)

1. 《Python网络数据采集》
2. Scrapy文档

[Scrapy官方文档](https://docs.scrapy.org/en/latest/intro/tutorial.html)

[Scrapy教程](https://www.runoob.com/w3cnote/scrapy-detail.html)

1. 机器学习

[sklearn官方文档](https://sklearn.apachecn.org/)

1. snownlp文档

[snownlp教程](https://blog.csdn.net/qq\_34337272/article/details/79577288)

# 测试时间、地点和人员

1. 测试时间：2021-5-10至2020-5-16，基本按照计划进行。
2. 地点：小组成员都在学校，线下沟通。
3. 人员：全体小组成员。

# 测试环境描述

测试的目的是验证数据获取和入库的时效性。测试机带有虚拟环境连接远程服务器的PC机。

PC机配置如下：

1. 操作系统：win10
2. CPU：I7，2.60GHz
3. 内存：16GB

# 测试执行情况

1.新闻入库时间

1. 新闻相关信息爬取

此部分爬取中为了防止网站反爬虫设置了一个随机延时，每爬取一条随机延时0~1s，爬取1000条数据的时间是200s。如果除去随机延时，时间会更快。

1. 新闻内容爬取

1000条新闻爬取加入库用时370s。

所以，入库1000条数据的总时间为（200+370）=570s

2.情感分析部分

测试数据1000条，用时180s.

3.数据入库更新对于其他子系统的影响

数据入库更新采用添加没有出现过的数据，忽略已存在的数据。所以在数据更新期间，如果其他子系统想要获取博物馆的相关新闻，会出现暂时只有前段时间的新闻，但是过半小时左右（数据入库之后），会看到最新的新闻。除非该博物馆本来就没有新闻，否则不会出现大面积空白的停滞情况。

# 测试结果分析

## 5.1 测试评估

功能方面，基本实现了需求中提到的所有功能；

性能方面，爬虫项目实现了功能的完成，还设置了UA伪装、自动更新；

API实现正常，返回结果正确；

安全性方面，对于开发人员有着较为清晰的机制与逻辑，同时能够通过远程调用服务器程序进行远程调用，有着不错的安全性。

## 5.2 综合数据分析

数据库初次更新：入库约3万条新闻

总时间=平均入库一条新闻的时间\*新闻条数

=0.57s\*29214

=16652s

数据库再次更新：

总时间=平均入库一条新闻的时间\*新闻条数

=0.57s\*每天的新新闻

# 测试评估

项目仍有较高提升空间，望后期再接再厉，使项目日益完善。