2020-2-1

杨耀飞

灵玖软件

[公司地址]

情报分析平台构架设计草案

# 第一章 需求功能

此平台需求为对于数据库、文本、文档等输入的数据进行存储于分析。并进行搜索、分析、可视化展示更功能。

## 1.1 输入数据来源

数据数据位多元数据包括且不限于如下内容：

1. 互联网采集数据
2. 科技文献数据
3. 已有知识图谱
4. 文档数据

将上述需求进行合并抽象后，输入数据为：

1. 基于文本的文档数据
2. 基于数据库的数据
3. 基于三元组的数据

## 1.2 信息处理与管理

对于输入数据需要进行管理和基本处理，包括：

1. 分用户的数据管理上传
2. 非文本文档解析
3. 内部语料库管理
4. 基础NLP处理

## 1.3 内容搜索与分析

对于数据需要进行搜索与分析， 包括：

1. 文本内容搜索
2. 文本内容的知识图谱抽取
3. 格式化信息中的知识图谱抽取
4. 知识融合以及分析

## 1.4 可视化展示

对于分析内容需要进行可视化展示，包括：

1. 文本搜索内容展示
2. 知识图谱相关分析展示
3. 数据的标注标引交互
4. 定制报告展示和生成
5. 综合统计分析内容展示

## 1.5 后台管理

1. 用户和权限管理
2. 语料库规则库管理
3. 日志系统

# 第二章 整体构架设计

整体构架设计为三层：

1. 应用层： 负责所有功能的展示、操作和业务逻辑

2. 业务层： 负责底层数据的处理、获取、管理、输出

3. 存储与算法层： 负责原始数据存储索引，提供数据分析相关算法

Java Web

React.js

Data Manager

NLPIR-KG

NLPIR-Reporter

NLPIR-Storage

Spider

NLPIR-Cloud

1. Java Web （Java SpringBoot）

用于系统业务逻辑相关的所有数据交互

1. React.js （JavaScript）

用于可视化展示和标注等所有展示的页面编写， 后期利用react.native可扩展为本地化应用。

1. Spider （Python Scrapy Redis FastAPI）

用于从互联网获取相关数据的爬虫集群，具有可扩展可管理能力，在Scrapy框架基础上进行。

1. Data Manager （Java）

对底层数据进行各种操作，并生成JavaWeb需要的数据格式

1. NLPIR-KG （Java GraphDB）

结合特定业务逻辑对底层数据中的三元组进行抽取存入内部图数据库，并提供知识图谱相关的分析功能

1. NLPIR-Reporter （Python FastAPI）

管理用于生成订制化的报告生成脚本，并提供接口供JavaWeb调用。

1. NLPIR-Storage （Python FastAPI）

用于存储索引管理原始数据

1. NLPIR-Cloud （Python FastAPI NLPIR）

用于提供底层算法

图示

描述已自动生成

# 第三章 详细构架设计

本章对未编写和需要进行改进的模块进行阐述

## 3.1 NLPIR-Storage

对于需存储数据类型和对应存储方式见下表：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 数据类型 | 原始信息 | 文本 | Meta | 处理信息 |
| 文档文件 | HDFS/FS | ES | ES | ES |
| DB RDF | ES(origion + DB\_table\_name + 外键) | 无 | ES | 无 |
| 文本 | 无 | ES | ES | ES |

需要进行的开发：

1. 完善现有API
2. DB RDF存储逻辑
3. DB RDF 长程导入后台实现、进度、日志
4. DB RDF 查询和显示逻辑
5. 导出， 全量导出，部分导出
6. 服务状态
7. 提供文件系统存储接口，和HDFS接口进行合并

1. DB

2. RDF

3. EXCEL

4. DOC，PDF 。。。

5. 文本

Elasticsearch Ingest插件

## 3.2 Spider

针对现有的RealEstate Spider 项目进行扩展，使其具有一般网页爬取的能力，并具有一定的管理功能。

1. 动态爬虫脚本加载
2. 爬虫状态监控
3. 关联NLPIR-Storage
4. API调用服务

## 3.3 NLPIR-Reporter

用于报告生成，使用NLPIR-Cloud提供的算法生成报告，内部是一个Python脚本管理服务，进行Python脚本的管理和运行，并将结果交给前台页面。

## 3.4 Data Manager

Java编写，逻辑上的模块，用于Java Web 和 NLPIR-Storage 通讯和数据转换

## 3.5 NLPIR-KG

Java 编写，用于针对业务的知识图谱抽取和管理。 前期与业务逻辑合并处理，后期拆分成模块

# 第四章 可视化

前台JavaScript库选择：

1. Cytoscape.js

2. vis-network

# 第五章

5.1 文档管理系统

图示

描述已自动生成

图示, 示意图

描述已自动生成

1. NLPIR-Storage
   1. 抽象文件IO接口，合并HDFS和FS
   2. 完善RESTful 接口
   3. Elasticsearch Ingest Plugin
      1. LibreOffice解析转换文档
      2. NLPIR组件解析文档格式
      3. 图片OCR内容
2. JavaWeb
3. 硬件装机安装文档规范
   1. 镜像方式：派生龙
   2. 远程访问密码随机生成等脚本

5.2 在线Parser（在线演示系统）

1. NLPIR-Cloud
   1. 增加计费系统
   2. 增加token生成接口
   3. 用户语料转储
   4. 日志存储
2. NLPIR-Storage
   1. 语料存储
   2. 日志存储
   3. 用户信息存储
3. JavaWeb
   1. 用户管理和token存储与管理
   2. 可视化展示
   3. 用户词典管理
   4. 后台查看用户数据
4. 部署运维
   1. SSL证书
   2. 基于Docker的部署NLPI