互联网知识智能感知系统

需求分析

学院名称： 计算机学院

小组成员 王海 吕鑫 马茼茼 胡曦伟 谢欣怡

指导教师： 刘 小 明

时间：2018年1月16日

目录

[1. 编写目的 3](#_Toc19480)

[2. 项目背景 3](#_Toc2480)

[3. 功能需求 3](#_Toc8723)

[3.1 参与者列表 4](#_Toc10137)

[3.2 各模块需求分析 4](#_Toc9321)

[3.2.1 检索模块 4](#_Toc22917)

[3.2.2 文本领域分析模块 4](#_Toc31867)

[3.2.3 文本主题分析模块 4](#_Toc15849)

[3.3具体用例描述 5](#_Toc26399)

[3.4 系统流程分析 5](#_Toc7600)

[3.4.1 系统主要流程 5](#_Toc13139)

[3.5 运行环境描述 7](#_Toc5782)

[3.5.1软件 7](#_Toc20407)

[3.5.2 硬件 7](#_Toc19785)

[3.5.3 网络 7](#_Toc9378)

# 编写目的

本文档适用于基于互联网知识智能感知系统小组的全部成员，根据实训开发要求和管理要求，编写软件工程需求分析说明书，本文档的目的是确定系统开发的范围，对系统结构进行分析，并为开发人员，维护人员提供协议，对软件进行功能性描述，从而作为开发人员编码的基础，且作为小组的协议文档，为最终的测试和验收提供标准。

# 项目背景

#### 在当代社会环境下，信息技术迅猛发展，随着数据量的爆炸式增长，人工智能迅速兴起，而人工智能的定义就是尽可能的让机器去模仿人，从而替代人类的一些无意义操作。而作为人工智能发展最为迅猛的一个分支——自然语言处理，有着非常广泛的应用前景。众多行业领域的研究和应用（如互联舆情分析、网络）都迫切需要一种统一的针对不规范大数据互联网信息的知识库，自然语言处理是信息时代一门非常重要的技术，它可以让机器代替人类处理大量的文本信息，可以让她像人一样分析出文本的内容含义，从而在公共安全、生物、工业、农业、交通、医疗等等方面有广泛应用。

# 功能需求

互联网知识智能感知系统的服务分为三个大部分：ElasticSearch数据检索、文本领域分类、文本主题发现，文本摘要提取。

## 参与者列表

用户：检索信息，对检索出的文本进行领域分类，主题发现，摘要提取

## 各模块需求分析

### 资源获取，存储与检索：分布式互联网通用采集子系统

项目采用分布式搜索引擎Elasticsearch检索服务，我们将Elasticsearch部署在了多台服务器上，我们选择Elasticsearch的原因是因为它在处理大数据要比传统的数据库查询更加快速，在发生错误时，各节点组成对等的网络结构，某些节点出现故障时会自动分配其他节点代替其进行工作。

1. 数据采集：基于采集平台，针对领域的站点数据采集；
2. 学习ES，构建通用检索服务，发布运行到公共平台上。
3. 用户选择查询条件，输入想要查询的关键字。点击查询按钮，进行查询。
4. 将查询结果返回到表格中。
5. 页面实现检索文档，输入文档的可接受分析。

### 文本行业领域分类模块

1.学习svm算法模型软件，学习练习基于svm的互联网文本领域行业的分类模型；用户只需在当前页面进行如下操作即可：

1. 选中上面检索出来的一条数据（或者在文本框中输入自己想要进行文本分类一篇文章）；
2. 选择进行分析；
3. 页面下方将会显示出领域分析出的结果；
4. 学习并运行基于深度学习的文本分类模型；
5. 封装为平台服务，发布运行。

### 文本主题分析：自动主题发现，面向主题分类与聚类

1.学习LDA算法模型，练习训练主题发现模型；

2.基于公共平台，封装为平台服务，发布运行；

用户只需在当前页面进行如下操作即可：

1. 选中上面检索出来的一条数据（或者在文本框中输入自己想要进行文本主题摘要的一篇文章）；
2. 选择进行分析；
3. 页面下方将会显示出文本主题分析出的结果。

3.实现对所采集数据的特定主题，热点，焦点发现。

### **自动摘要生成：**

1.基于LDA模型，实现对特定主题的单文档，多文档等抽取式摘要生成；

2.学习并运行现有基于深度学习的文本自动生成模型，实现面向主题的生成式摘要；

3.基于公共平台开发，封装为平台服务，发布运行：

用户只需在当前页面进行如下操作即可：

1. 选中上面检索出来的一条数据（或者在文本框中输入自己想要进行文本主题摘要的一篇文章）；
2. 选择进行分析；
3. 页面下方将会显示出文本摘要分析出的结果。

## 3.3具体用例描述

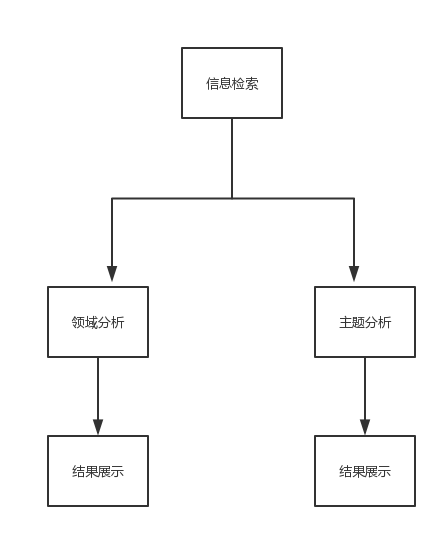
用例规约如下：

|  |  |  |  |
| --- | --- | --- | --- |
| 用例名称： | 互联网智能知识感知系统使用用例 | | |
| 用例ID： | 1 | | |
| 角色： | 用户 | | |
| 启用阶段： | 整个网站过程 | | |
| 用例说明： | 用户进行检索，把检索数据进行分析 | | |
| 前置条件： | **无** | | |
| 基本事件流： | 参与者动作 | | 系统响应 |
| 1、用户选择查询条件，输入关键字，点击按钮进行查询  2、用户选中要进行分析的数据点击按钮，进行分析 | | 1、系统返回数据到页面上  2、返回分析结果到页面上  （领域 主题发现 摘要提取） |
| 其它事件流： | 无 | | |
| 异常事件流： | 参与者动作 | 系统响应 | |
| 无休止点击发送 | Net error | |
| 后置条件： | 无 | | |

## 3.4 系统流程分析

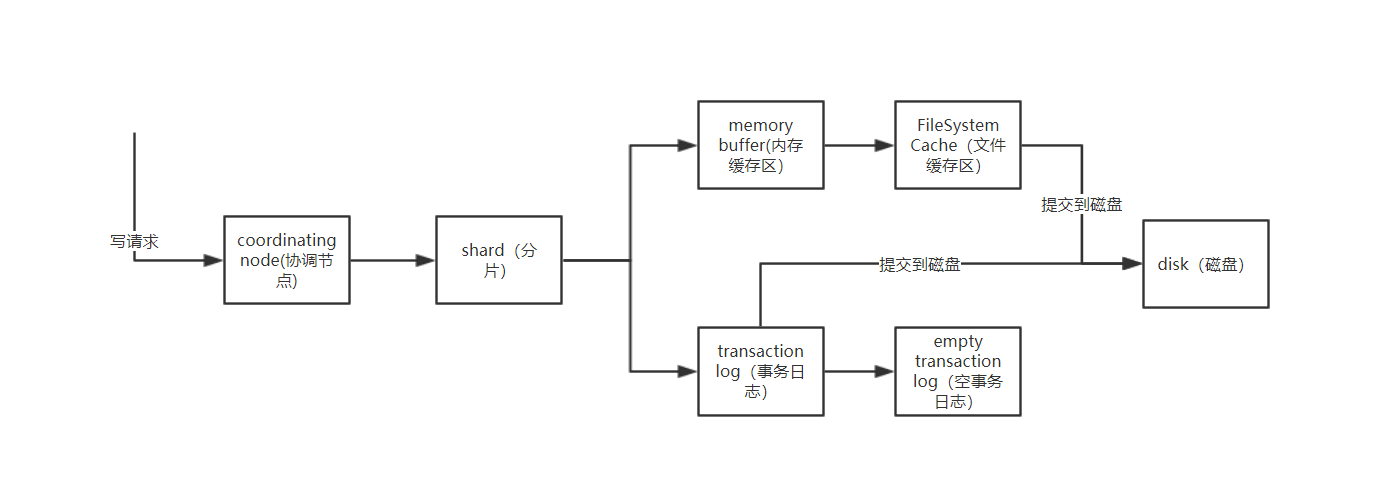
### 3.4.1 系统主要流程

系统模块图如下所示：



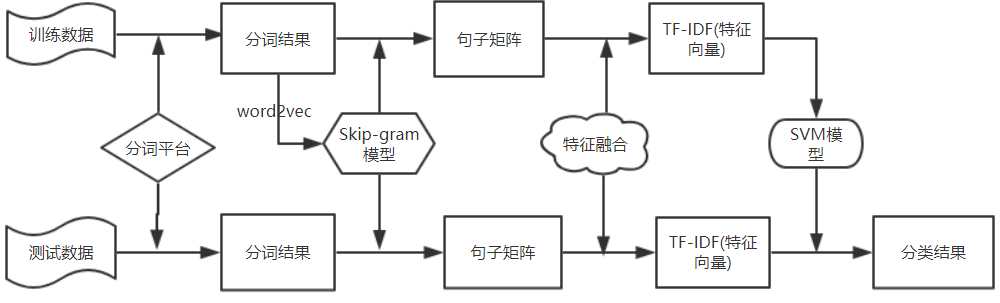
### 3.4.2 检索流程

检索模块图如下所示：

检索流程图

### 3.4.3 领域分类流程

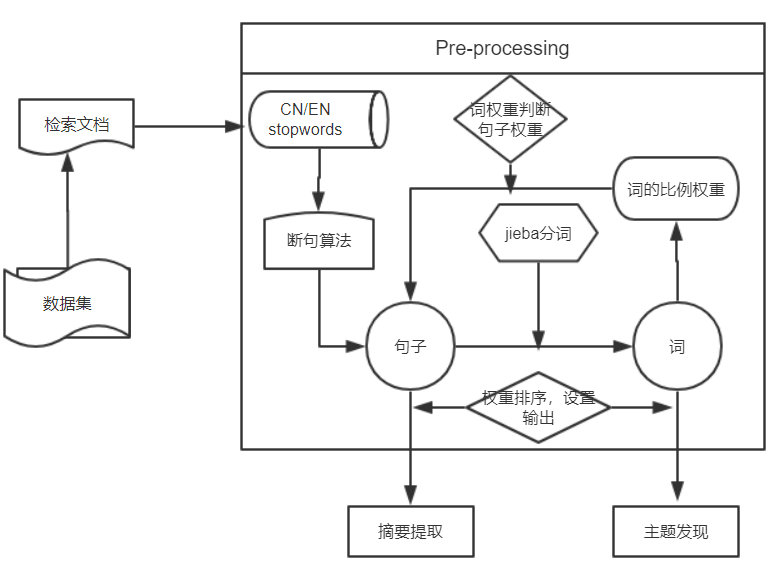
文本领域分类模块图如下所示：



领域分类模块图

### 3.4.4主题分析分类流程

文本主题分析模块图如下所示：



主题分析模块图

## 3.5 运行环境描述

### 3.5.1软件

操作系统：win7、win8，在这两种比较常用的操作系统上进行测试。服务器端操作系统为windows server 2008。

使用的浏览器：谷歌浏览器、火狐浏览器、360浏览器、不同版本的IE浏览器。

开发平台：PyCharm , WebStorm , idea, postman等软件

### 3.5.2 硬件

计算机：最低配置1G内存，单核，50G硬盘。

显卡：最低配置512M

显示器：台式机17英寸分辨率1024\*768、台式机19英寸分辨率1440\*900；笔记本14.1寸分辨率1366\*768，笔记本15.4寸分辨率1280\*800。

服务器：至少两台，服务器端最低配置8G内存，双核，50G硬盘。

### 3.5.3 网络

保证计算机能连接到校园网即可。