# 项目实践

## 需求说明

假定在你电脑的文件系统某个目录下或其子目录下存放有大量的文档，文档格式有doc、docx、ppt、pptx、txt、pdf等等。仿照 百度文库，实现一个文件检索系统。需求：

1、能够对文件名进行搜索

2、能够对文件内容进行搜索

3、能够下载检索的文件

4、在搜索列表上展示一小段文件内容，如这段内容中包含搜索关键字，需要高亮。

## 界面原型

首页



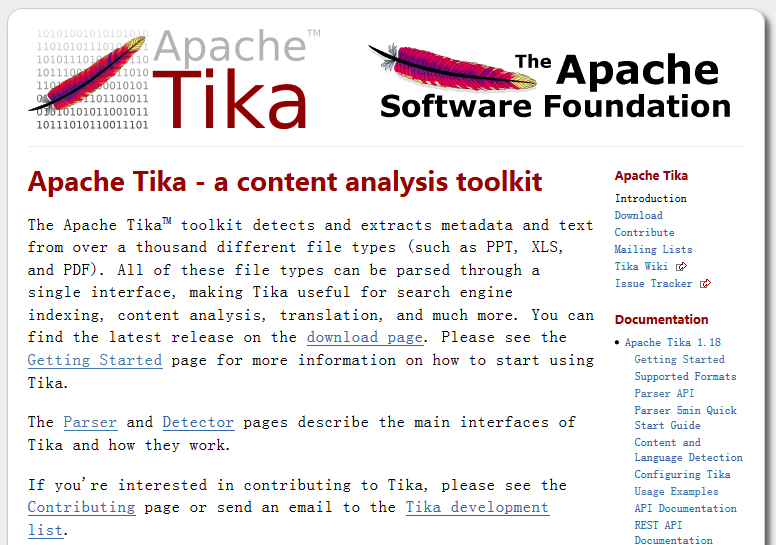
搜索结果页



## 内容提取工具

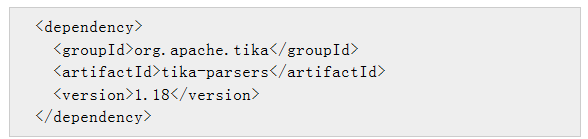
Word、ppt、pdf等文档的内容需要提取出来才能索引，提取工具 tika：

<http://tika.apache.org/>

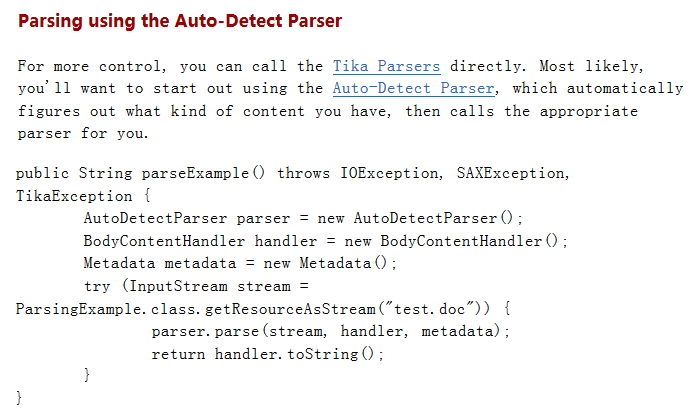


使用示例：

<http://tika.apache.org/1.18/gettingstarted.html> 找到要maven引入什么



<http://tika.apache.org/1.18/examples.html> 找到自动解析文档的样例



搞清楚 Metadata中都可以取到些什么文档的元数据。如何取。

用 tika把存放目录下的文档的内容提取出来，进行索引。