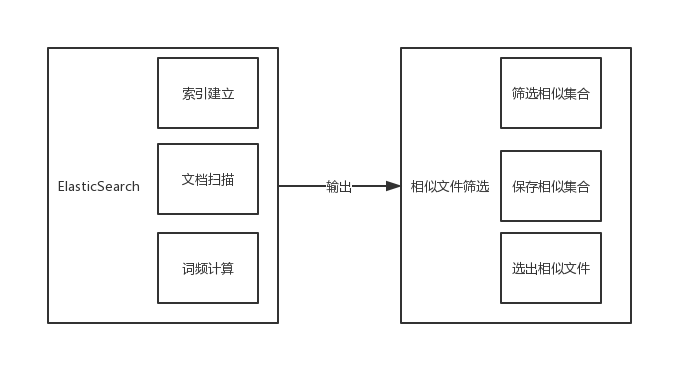
# 代码重复度计算算法说明文档

本算法目标：

* 得到基准项目和待测项目的词频向量集合
* 计算待测项目的每个词频向量与基准项目的余弦相似度集合
* 选出相似文件

## 主要模块及关系



## 相关函数描述

* Elasticsearch索引建立

参数：整体项目

返回：文档索引

接收：Elasticsearch

功能：索引建立

* Elasticsearch文档扫描

参数：文档索引

返回：文档集合

接收：Elasticsearch

功能：文档扫描

* 剩余文件判断

参数：整体项目

返回：bool

接收：无

功能：剩余文件判断

* TF词频计算

参数：文件

返回：tf:float

接收：无

功能：TF词频计算

* IDF逆向文件频率

参数：文件

返回：idf:float

接收：无

功能：IDF逆向文件频率

* TF-IDF计算

参数：tf:float, idf:float

返回：tf\_idf:float

接收：无

功能：TF-IDF计算

* 余弦相似度集合

参数：文件集

返回：map集合

接收：无

功能：返回余弦相似度集合

* 集合存储

参数：文件集

返回：bool

接收：无

功能：保存文件集合

* 相似文件筛选

参数：文件集

返回：相似文件集:map

接收：无

功能：从文件集中筛选并加入map中并返回

## 主要实现过程

