Apache Lucene 基本概念：

- 文档(Document):索引和搜索时使用的主要数据载体，包含一个或多个存有数据的字段。
- 字段(field):文档的一部分，包含名称和值两部分。
- 词(term):一个搜索单元，表示文本的一个词。
- 标记(token):表示在字段文本中出现的词，由这个词的文本，开始和结束偏移量以及类型组成。

Lucene将所有的信息写成一个称为倒排索引(inverted index)的结构中，倒排索引建立索引中词和文档之间的映射，数据是面向词而不是文档的。

英文分词效果：

- ❑ Elasticsearch Server 1.0 (document 1)；
- ❑ Mastering Elasticsearch (document 2)；
- ❑ Apache Solr 4 Cookbook (document 3)。

那么，简化版的索引可以看成是这样的：

词	计　数	文　档
1.0	1	<1>
4	1	<3>
Apache	1	<3>
Cookbook	1	<3>
Elasticsearch	2	<1>,<2>
Mastering	1	<2>
Server	1	<1>
Solr	1	<3>

Lucene过滤器分析：

- 小写过滤器(lowercase filter)把所有的标记变成小写
- 同义词（synonyms filter）基于基本的同义词规则，把一个标记换成另一个同义的标记
- 多语言词干提取过滤器（multiple language stemming filter）减少标记（文本），得到词根或者基本形式，即词干。
- 可以选择全文检索的分析，也可以不使用分析，例如查询LightRed，标准分析器分析查询后，会去查询Light与red。若不经分析，就会直接精确查询LightRed这个单词。**应明确索引和查询词匹配否则什么都不会返回。**
- ASCII 过滤器:移除词条中所有非ASCII字符.