课件地址: <http://git.mashibing.com/>

1. **分词器**
   1. 概念：
   2. 干啥的：
      1. 切词
      2. normalization（提升recall召回率：能搜索到的结果的比率）
   3. 分析器：
      1. character filter（mapping）：分词之前预处理（过滤无用字符、标签等，转换一些&=>and 《Elasticsearch》=> Elasticsearch
         1. [HTML Strip Character Filter](https://www.elastic.co/guide/en/elasticsearch/reference/current/analysis-htmlstrip-charfilter.html" \o "HTML Strip Char Filter)：html\_strip
            1. 参数：escaped\_tags 需要保留的html标签
         2. [Mapping Character Filter](https://www.elastic.co/guide/en/elasticsearch/reference/current/analysis-mapping-charfilter.html" \o "Mapping Char Filter)：type mapping
         3. [Pattern Replace Character Filter](https://www.elastic.co/guide/en/elasticsearch/reference/current/analysis-pattern-replace-charfilter.html" \o "Pattern Replace Char Filter)：type pattern\_replace
      2. tokenizer（分词器）：分词
      3. token filter：停用词、时态转换、大小写转换、同义词转换、语气词处理等。比如：has=>have him=>he apples=>apple the/oh/a=>干掉
   4. ES内置分词器 7.6 15种 自带分词器
      1. standard analyzer：默认分词器，中文支持的不理想，会逐字拆分。
         1. max\_token\_length：最大令牌长度。如果看到令牌超过此长度，则将其max\_token\_length间隔分割。默认为255。
      2. [Pattern Tokenizer](https://www.elastic.co/guide/en/elasticsearch/reference/current/analysis-pattern-tokenizer.html" \o "Pattern Tokenizer)：以正则匹配分隔符，把文本拆分成若干词项。
      3. [Simple Pattern Tokenizer](https://www.elastic.co/guide/en/elasticsearch/reference/current/analysis-simplepattern-tokenizer.html" \o "Simple Pattern Tokenizer)：以正则匹配词项，速度比[Pattern Tokenizer](https://www.elastic.co/guide/en/elasticsearch/reference/current/analysis-pattern-tokenizer.html" \o "Pattern Tokenizer)快。
      4. whitespace analyzer：以空白符分隔 Tim\_cookie

测试：Tim\_cookie received his girlfriend on 2020-05-20

* 1. 自定义分词器：
     1. tokenizer：内置或自定义分词器。（需要）
     2. char\_filter：内置或自定义[字符过滤器](https://www.elastic.co/guide/en/elasticsearch/reference/current/analysis-charfilters.html" \o "字符过滤器参考) 。
     3. filter：内置或自定义token filter 。
     4. position\_increment\_gap：在为文本值数组建立索引时，Elasticsearch在一个值的最后一项和下一个值的第一项之间插入一个假的“空白”，以确保词组查询与来自不同数组元素的两项不匹配。默认为100。查看[position\_increment\_gap](https://www.elastic.co/guide/en/elasticsearch/reference/current/position-increment-gap.html" \o "position_increment_gap)更多。
  2. 中文分词器：
     1. IK分词：ES的安装目录 不要有中文 空格
        1. 下载：<https://github.com/medcl/elasticsearch-analysis-ik>
        2. 创建插件文件夹 cd your-es-root/plugins/ && mkdir ik
        3. 将插件解压缩到文件夹 your-es-root/plugins/ik
        4. 重新启动es
     2. 两种analyzer
        1. ik\_max\_word：细粒度
        2. ik\_smart：粗粒度
     3. IK文件描述
        1. IKAnalyzer.cfg.xml：IK分词配置文件
        2. 主词库：main.dic
        3. 英文停用词：stopword.dic，不会建立在倒排索引中
        4. 特殊词库：
           1. quantifier.dic：特殊词库：计量单位等
           2. suffix.dic：特殊词库：后缀名
           3. surname.dic：特殊词库：百家姓
           4. preposition：特殊词库：语气词
        5. 自定义词库：比如当下流行词：857、emmm...、渣女、舔屏、996
        6. 热更新：
           1. 修改ik分词器源码
           2. 基于ik分词器原生支持的热更新方案，部署一个web服务器，提供一个http接口，通过modified和tag两个http响应头，来提供词语的热更新