Atitit.分词引擎设计与 分词结果预处理

### ****第二步：将原文档传给分次组件(Tokenizer)。****

**分词组件(Tokenizer)会做以下几件事情(此过程称为Tokenize)：**

**1. 将文档分成一个一个单独的单词。**

**2. 去除标点符号。**

**3. 去除停词(Stop word)。**

所谓停词(Stop word)就是一种语言中最普通的一些单词，由于没有特别的意义，因而大多数情况下不能成为搜索的关键词，因而创建索引时，这种词会被去掉而减少索引的大小。

英语中挺词(Stop word)如：“the”,“a”，“this”等。

对于每一种语言的分词组件(Tokenizer)，都有一个停词(stop word)集合。

**经过分词(Tokenizer)后得到的结果称为词元(Token)。**

在我们的例子中，便得到以下词元(Token)：

“Students”，“allowed”，“go”，“their”，“friends”，“allowed”，“drink”，“beer”，“My”，“friend”，“Jerry”，“went”，“school”，“see”，“his”，“students”，“found”，“them”，“drunk”，“allowed”。

### ****第三步：将得到的词元(Token)传给语言处理组件(Linguistic Processor)。****

语言处理组件(linguistic processor)主要是对得到的词元(Token)做一些同语言相关的处理。

**对于英语，语言处理组件(Linguistic Processor)一般做以下几点：**

**1. 变为小写(Lowercase)。**

**2. 将单词缩减为词根形式，如“cars”到“car”等。这种操作称为：stemming。**

**3. 将单词转变为词根形式，如“drove”到“drive”等。这种操作称为：lemmatization。**

**Stemming 和 lemmatization的异同：**

* 相同之处：Stemming和lemmatization都要使词汇成为词根形式。
* 两者的方式不同：
  + Stemming采用的是“缩减”的方式：“cars”到“car”，“driving”到“drive”。
  + Lemmatization采用的是“转变”的方式：“drove”到“drove”，“driving”到“drive”。
* 两者的算法不同：
  + Stemming主要是采取某种固定的算法来做这种缩减，如去除“s”，去除“ing”加“e”，将“ational”变为“ate”，将“tional”变为“tion”。
  + Lemmatization主要是采用保存某种字典的方式做这种转变。比如字典中有“driving”到“drive”，“drove”到“drive”，“am, is, are”到“be”的映射，做转变时，只要查字典就可以了。
* Stemming和lemmatization不是互斥关系，是有交集的，有的词利用这两种方式都能达到相同的转换。

**语言处理组件(linguistic processor)的结果称为词(Term)。**

在我们的例子中，经过语言处理，得到的词(Term)如下：

“student”，“allow”，“go”，“their”，“friend”，“allow”，“drink”，“beer”，“my”，“friend”，“jerry”，“go”，“school”，“see”，“his”，“student”，“find”，“them”，“drink”，“allow”。

也正是因为有语言处理的步骤，才能使搜索drove，而drive也能被搜索出来