# 1英文分词

<http://itindex.net/detail/48243-%E8%8B%B1%E6%96%87-%E5%88%86%E8%AF%8D-%E7%AE%97%E6%B3%95?utm_source=tuicool&utm_medium=referral>

根据文档相关性计算公式

* TF-IDF: <http://lutaf.com/210.htm>
* BM25: <http://lutaf.com/211.htm>

分词质量对于基于词频的相关性计算是无比重要的

英文(西方语言）语言的基本单位就是单词，所以分词特别容易做，只需要3步：

1. 根据空格/符号/段落 分隔,得到单词组
2. 过滤，排除掉stop word
3. 提取词干

### 第一步：按空格/符号分词

用正则表达式很容易

pattern = r'''(?x) # set flag to allow verbose regexps

([A-Z]\.)+ # abbreviations, e.g. U.S.A.

| \w+(-\w+)\* # words with optional internal hyphens

| \$?\d+(\.\d+)?%? # currency and percentages, e.g. $12.40, 82%

| \.\.\. # ellipsis

| [][.,;"'?():-\_`] # these are separate tokens

'''

re.findall(pattern,待分词文本)

### 第二步：排除stop word

stopword就是类似 a/an/and/are/then 的这类高频词，高频词会对基于词频的算分公式产生极大的干扰，所以需要过滤

### 第三步：提取词干

词干提取( **Stemming**) 这是西方语言特有的处理，比如说英文单词有 单数复数的变形，-ing和-ed的变形，但是在计算相关性的时候，应该当做同一个单词。比如 apple和apples，doing和done是同一个词，提取词干的目的就是要合并这些变态

Stemming有3大主流算法

* [Porter Stemming](http://lutaf.com/j?u=http://www.tartarus.org/~martin/PorterStemmer/index.html)
* [Lovins stemmer](http://lutaf.com/j?u=http://www.cs.waikato.ac.nz/~eibe/stemmers/index.html)
* [Lancaster Stemming](http://lutaf.com/j?u=http://www.comp.lancs.ac.uk/computing/research/stemming/index.htm)

Lucene 英文分词自带了3个stemming算法，分别是

1. EnglishMinimalStemmer
2. 著名的 Porter Stemming
3. KStemmer

词干提取算法并不复杂，要么是一堆规则，要么用映射表，编程容易，但是必须是这种语言的专家，了解构词法才行啊

<http://text-processing.com/demo/stem/> 是一个在线试验词干提取算法的网站

### Lemmatisation

Lemmatisation是和词干提取(Stemming) 齐名的一个语言学名词，中文可以叫做 **词形还原** ,就是通过查询字典，把 "drove" 还原到 "drive"   
而stemming会把单词变短，"apples","apple"处理之后都变成了 "appl"

* [wikipedia关于词形还原的简介](http://lutaf.com/j?u=http://en.wikipedia.org/wiki/Lemmatisation)
* [European languages lemmatizer](http://lutaf.com/j?u=http://lemmatizer.org/) 一个c语言的lib

做计算机语言学研究才会涉及到lemmatization，我个人觉得做搜索完全可以不考虑，Stemming已经可以解决大问题了

**参考**

* [http://text-processing.com/](http://lutaf.com/j?u=http://text-processing.com/)
* [www.nltk.org python的自然语言包，非常有用](http://lutaf.com/j?u=http://www.nltk.org)
* [PYTHON自然语言处理中文版.pdf](http://lutaf.com/j?u=http://vdisk.weibo.com/s/4ffue/1334656530)

# 2中文分词

暂无