Term Extraction

## 介绍

这个工具(java 实现)，主要用来分析文本。已经集成了或者可扩展的算法包括，

* Basic term frequency
* Average term frequency in the corpus
* TF-IDF
* RIDF
* Weirdness
* C-value
* GlossEx
* TermEx

其中，主要使用了三个工具，

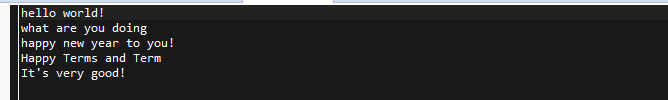
1. OpenNLP：自然语言处理的基础API，网址：<http://opennlp.apache.org/index.html>
2. Dragon Toolkit：在OpenNLP的基础上，进一步实现了自然语言处理的常用操作，网址：<http://dragon.ischool.drexel.edu/default.asp>
3. Jatetoolkit：使用了多种算法来实现自然语言处理，主要运用了OpenNLP和Dragon Toolkit提供的工具，网址：<http://code.google.com/p/jatetoolkit/>

## 使用

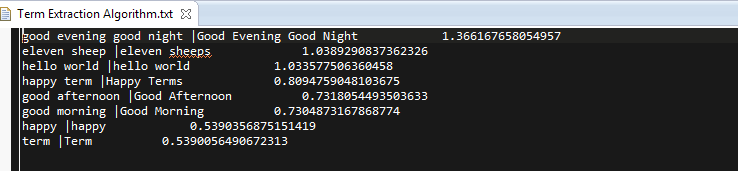
1. 获取代码后，进行调试运行，网址：

<https://144.214.55.106/svn/FinancialMining/TermExtraction/>

1. [**input**] 将待分析的文本，放在’TermExtraction/test/input/’目录下，如



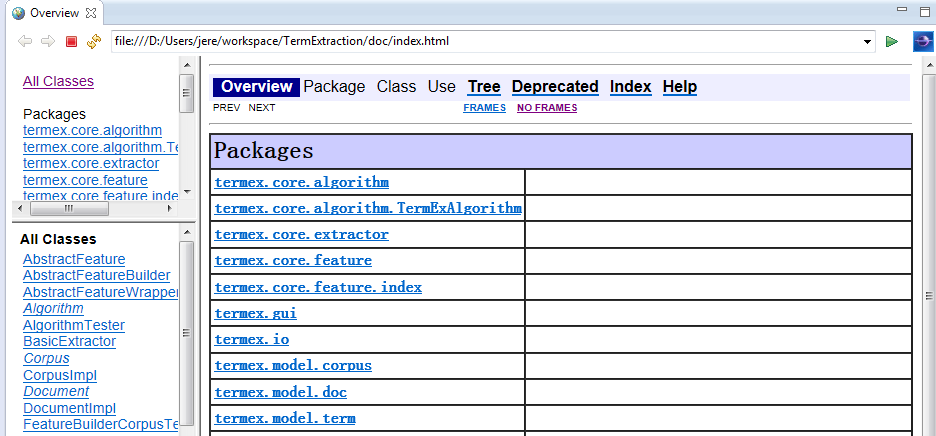
1. [**reference**] 指定模板文件，TermExtraction/res/nlp/bnc\_unifrqs.normal
2. 给定特定的参数，以执行该程序。
   1. 命令行模式下：java van [input] [reference] [output]
   2. Eclipse下：[Run] -> [Run Configurations…] -> [Van] -> [Arguments] -> [Program arguments] 下面指定三个参数
3. [**output**] 即输出文件的路径，Termextraction/test/output/，输出文件如下，



## FAQ

* 如何阅读源代码？

已经生成javadoc文件，TermExtraction/doc/目录下，打开index.html即可查看所有类、方法的名称、作用和注释。如图



* 更多的问题？

[联系：jyfeather88@gmail.com](mailto:联系：jyfeather88@gmail.com)，标题：EpicLab Help