scws中文分词安装与使用

作者：樱桃 来源：东坡网

本文原地址：http://www.dp1037.com/dpinfo-7-52-0.html

东坡网，为帝国cms加油

由于项目中需要使用分词功能，且需要可以自定义词典，因此使用scws中文分词程序，此程序目前最新版本为1.2.3，本文记录在ubuntu 14环境下安装sws1.2.3和其php扩展过程。

一、下载源码

```
wget http://www.xunsearch.com/scws/down/scws-1.2.3.tar.bz2
tar xvjf scws-1.2.3.tar.bz2
```

二、执行配置脚本和编译

具体选项参数执行 ./configure --help 查看。常用选项为：--prefix=指定安装目录

```
cd scws-1.2.3
./configure --prefix=/usr/local/scws
make && make install
```

检查是否安装成功

```
ls -al /usr/local/scws/lib/libscws.la
/usr/local/scws/bin/scws -h
```

三、下载通用词典

```
cd /usr/local/scws/etc
wget http://www.xunsearch.com/scws/down/scws-dict-chs-gbk.tar.bz2
wget http://www.xunsearch.com/scws/down/scws-dict-chs-utf8.tar.bz2
tar xvjf scws-dict-chs-gbk.tar.bz2
tar xvjf scws-dict-chs-utf8.tar.bz2
```

四、编译PHP扩展

更新PHP扩展需要autoconf、automake及phpize工具，如果没有的话需要新安装。

```
apt-get install autoconf
cd /root/scws-1.2.3/phpext
phpize
./configure --with-scws=/usr/local/scws --with-php-config=/usr/local/bin/php-config
make && make install
```

在 php.ini 中加入以下几行

```
[scws]
extension=scws.so
scws.default.charset=gbk
scws.default.fpath=/usr/local/scws/etc
```

在php程序中用 ini_get('scws.default.fpath') 读取scws.default.fpath目默认字典录设置

使用时指定词典路径和编码：

```
$scws = scws_new();
$scws->set_charset('utf8'); //指定编码
$scws->set_dict('/usr/local/scws/etc/dict.utf8.xdb');//指定词典路径，可以是绝对路径，也可以是相对路径
```

测试分词程序

```
php /root/scws-1.2.3/phpext/scws_test.php
```

如果运行失败，很可能是未正确指定词典路径

五、SCWS使用说明和函数详解

（1）预定义常量

```
SCWS_XDICT_XDB     //词典文件为 XDB
SCWS_XDICT_MEM     //将词典全部加载到内存里
SCWS_XDICT_TXT     //词典文件为 TXT（纯文本）


SCWS_MULTI_NONE    //不进行复合分词
SCWS_MULTI_SHORT   //短词复合
SCWS_MULTI_DUALITY  //散字二元复合
SCWS_MULTI_ZMAIN   //重要单字
SCWS_MULTI_ZALL    //全部单字
```

（2）预定义类

这是一个类似 Directory 的内置式伪类操作，类方法建立请使用 scws_new() 函数，而不能直接用 `new SimpleCWS`。否则不会包含有 handle 指针，将无法正确操作。

```
class SimpleCWS{
```

```
    resource handle;
    bool close(void);
    bool set_charset(string charset)
    bool add_dict(string dict_path[, int mode = SCWS_XDICT_XDB])
    bool set_dict(string dict_path[, int mode = SCWS_XDICT_XDB])
    bool set_rule(string rule_path)
    bool set_ignore(bool yes)
    bool set_multi(int mode)
    bool set_duality(bool yes)
    bool send_text(string text)
    mixed get_result(void)
    mixed get_tops([int limit [, string xattr]])
    bool has_word(string xattr)
    mixed get_words(string xattr)
    string version(void)
}
```

类方法的用与支 scws_xxx_xxx 系列函数用法一致，只不过免去第一参数，参见函数列表。

（3）函数详解

```
mixed scws_new(void)
```

创建并返回一个 SimpleCWS 类操作对象。成功返回类操作句柄，失败返回 false。

```
mixed scws_open(void)
```

创建并返回一个分词操作句柄。成功返回 scws 操作句柄，失败返回 false。

```
bool scws_close(resource scws_handle)
```

关闭一个已打开的 scws 分词操作句柄。参数 scws_handle 即之前由 scws_open 打开的返回值，以下相同。

```
scws_set_charset(resource scws_handle, string charset)
```

设定分词词典、规则集、欲分文本字符串的字符集。参数 charset 要新设定的字符集，目前只支持 utf8 和 gbk ，默认为 gbk，utf8不要写成utf-8。返回值 始终为 true 。

```
scws_add_dict(resource scws_handle, string dict_path [, int mode])
```

添加分词所用的词典，新加入的优先查找。参数 dict_path 词典的路径，可以是相对路径或完全路径。参数 mode 可选，表示加载的方式。其值有：SCWS_XDICT_TXT 表示要读取的词典文件是文本格式，可以和后2项结合用；SCWS_XDICT_XDB 表示直接读取 xdb 文件，此为默认值； SCWS_XDICT_MEM 表示将 xdb 文件全部加载到内存中，以 XTree 结构存放，可用异或结合另外2个使用。成功返回 true 失败返回 false

```
bool scws_set_dict(resource scws_handle, string dict_path [, int mode])
```

设定分词所用的词典并清除已存在的词典列表。参数设置与scws_add_dict相同。

```
bool scws_set_rule(resource scws_handle, string rule_path)
```

设定分词所用的新词识别规则集（用于人名、地名、数字时间年代等识别）。参数 rule_path 规则集的路径，可以是相对路径或完全路径。

```
bool scws_set_ignore(resource scws_handle, bool yes)
```

设定分词返回结果时是否去除一些特殊的标点符号之类。参数 yes 设定值，如果为 true 则结果中不返回标点符号，如果为 false 则会返回，缺省为 false。

```
bool scws_set_multi(resource scws_handle, int mode)
```

设定分词返回结果时是否复式分割，如“ 中国人 ”返回“ 中国＋人＋中国人 ”三个词。参数 mod 复合分词法的级别，缺省不复合分词。取值由下面几个常量异或组合（也可用 1-15 来表示）：SCWS_MULTI_SHORT (1)短词；SCWS_MULTI_DUALITY (2)二元（将相邻的2个单字组合成一个词）；SCWS_MULTI_ZMAIN (4)重要单字；SCWS_MULTI_ZALL (8)全部单字。

```
bool scws_set_duality(resource scws_handle, bool yes)
```

设定是否将闲散文字自动以二字分词法聚合。参数 yes 设定值，如果为 true 则结果中多个单字会自动按二分法聚分，如果为 false 则不处理，缺省为 false。

```
bool scws_send_text(resource scws_handle, string text)
```

发送设定分词所要切割的文本。参数 text 要切分的文本的内容。返回值 成功返回 true 失败返回 false 。系统底层处理方式为对该文本增加一个引用，故不论多长的文本并不会造成内存浪费；执行本函数时，若未加载任何词典和规则集，则会自动试图在 ini 指定的缺省目录下查找缺省字符集的词典和规则集。

```
mixed scws_get_result(resource scws_handle)
```

根据 send_text 设定的文本内容，返回一系列切好的词汇。返回值 成功返回切好的词汇组成的数组，若无更多词汇，返回 false。返回的词汇包含的键值如下：word _string_ 词本身；idf _float_ 逆文本词频；off _int_ 该词在原文本路的位置；attr _string_ 词性。

```
scws_get_tops(resource scws_handle [, int limit [, string attr]])
```

根据 send_text 设定的文本内容，返回系统计算出来的最关键词汇列表。参数 limit 可选参数，返回的词的最大数量，缺省是 10 。参数 attr 可选参数，是一系列词性组成的字符串，各词性之间以半角的逗号隔开，这表示返回的词性必须在列表中，如果以~开头，则表示取反，词性必须不在列表中，缺省为NULL，返回全部词性，不过滤。 返回值 成功返回统计好的的词汇组成的数组，返回 false。返回的词汇包含的键值如下：word _string_ 词本身；times _int_ 词在文本中出现的次数；weight _float_ 该词计算后的权重；attr _string_ 词性。

```
mixed scws_get_words(resource scws_handle, string attr)
```

根据 send_text 设定的文本内容，返回系统中词性符合要求的关键词汇。参数 attr 是一系列词性组成的字符串，各词性之间以半角的逗号隔开，这表示返回的词性必须在列表中，如果以~开头，则表示取反，词性必须不在列表中，若为空则返回全部词。返回值 成功返回符合要求词汇组成的数组，返回 false，键值与scws_get_result相同。

```
bool scws_has_words(resource scws_handle, string attr)
```

根据 send_text 设定的文本内容，返回系统中是否包括符合词性要求的关键词。参数 attr 是一系列词性组成的字符串，各词性之间以半角的逗号

隔开，这表示返回的词性必须在列表中，如果以~开头，则表示取反，词性必须不在列表中，若为空则返回全部词。返回值 如果有则返回 true，没有就返回 false。

六、两个例子

（1）使用类方法分词

```
<?php
$so = scws_new();
$so->set_charset('gbk');
$so->set_dict($sh, '/usr/local/scws/etc/dict.xdb');
$so->set_rule($sh, '/usr/local/scws/etc/rules.ini');
// 这里没有调用 set_dict 和 set_rule 系统会自动试调用 ini 中指定路径下的词典和规则文件
$so->send_text("我是一个中国人,我会C++语言,我也有很多T恤衣服");
while ($tmp = $so->get_result())
{
 print_r($tmp);
}
$so->close();
?>
```

（2）使用函数提取高频词

```
<?php
$sh = scws_open();
scws_set_charset($sh, 'gbk');
scws_set_dict($sh, '/usr/local/scws/etc/dict.xdb');
scws_set_rule($sh, '/usr/local/scws/etc/rules.ini');
$text = "我是一个中国人，我会C++语言，我也有很多T恤衣服";
scws_send_text($sh, $text);
$top = scws_get_tops($sh, 5);
print_r($top);
?>
```

注意：输入的文字，词典、规则文件这三者的字符集必须统一，如果不是默认的 gbk 字符集请调用 SimpleCWS::set_charset 或 scws_set_charset 来设定，否则可能出现意外错误。

七、北大词性标注

- Ag —— 形语素，形容词性语素。形容词代码为a，语素代码ｇ前面置以A。
- a —— 形容词，取英语形容词adjective的第1个字母。
- ad —— 副形词，直接作状语的形容词。形容词代码a和副词代码d并在一起。
- an —— 名形词，具有名词功能的形容词。形容词代码a和名词代码n并在一起。
- b —— 区别词，取汉字“别”的声母。
- c —— 连词，取英语连词conjunction的第1个字母。
- Dg —— 副语素，副词性语素。副词代码为d，语素代码ｇ前面置以D。
- d —— 副词，取adverb的第2个字母，因其第1个字母已用于形容词。
- e —— 叹词，取英语叹词exclamation的第1个字母。
- f —— 方位词，取汉字“方”
- g —— 语素，绝大多数语素都能作为合成词的“词根”，取汉字“根”的声母。
- h —— 前接成分，取英语head的第1个字母。
- i —— 成语，取英语成语idiom的第1个字母。
- j —— 简称略语，取汉字“简”的声母。
- k —— 后接成分
- l —— 习用语，习用语尚未成为成语，有点“临时性”，取“临”的声母。
- m —— 数词，取英语numeral的第3个字母，n，u已有他用。
- Ng —— 名语素，名词性语素。名词代码为n，语素代码ｇ前面置以N。
- n —— 名词，取英语名词noun的第1个字母。
- nr —— 人名，名词代码n和“人(ren)”的声母并在一起。
- ns —— 地名，名词代码n和处所词代码s并在一起。
- nt —— 机构团体，“团”的声母为t，名词代码n和t并在一起。
- nz —— 其他专名，“专”的声母的第1个字母为z，名词代码n和z并在一起。
- o —— 拟声词，取英语拟声词onomatopoeia的第1个字母。
- ba —— 介词 把、将
- bei —— 介词 被
- p —— 介词，取英语介词prepositional的第1个字母。
- q —— 量词，取英语quantity的第1个字母。
- r —— 代词，取英语代词pronoun的第2个字母,因p已用于介词。

- s —— 处所词，取英语space的第1个字母。
- Tg —— 时语素，时间词性语素。时间词代码为t,在语素的代码g前面置以T。
- t —— 时间词，取英语time的第1个字母。
- dec —— 助词 的、之
- deg —— 助词 得
- di —— 助词 地
- etc —— 助词 等、等等
- as —— 助词 了、着、过
- msp —— 助词 所
- u —— 其他助词，取英语助词auxiliary
- Vg —— 动语素，动词性语素。动词代码为v。在语素的代码g前面置以V。
- v —— 动词，取英语动词verb的第一个字母。
- vd —— 副动词，直接作状语的动词。动词和副词的代码并在一起。
- vn —— 名动词，指具有名词功能的动词。动词和名词的代码并在一起。
- w —— 其他标点符号
- x —— 非语素字，非语素字只是一个符号，字母x通常用于代表未知数、符号。
- y —— 语气词，取汉字“语”的声母。
- z —— 状态词，取汉字“状”的声母的前一个字母。