详解crifan的Python库：crifanLib.py

版本：v1.1

Crifan Li

摘要

本文主要介绍了我自己crifan的Python函数库crifanLib.py，包括解释crifanLib.py由来，以及其中各种函数的功能和用法示例。

本文提供多种格式供：

在线阅读	HTML[1]	HTMLs[2]	PDF[3]	CHM[4]	TXT[5]	RTF[6]	WEBHELP[7]
下载（7zip压缩包）	HTML[8]	HTMLs[9]	PDF[10]	CHM[11]	TXT[12]	RTF[13]	WEBHELP[14]

HTML版本的在线地址为：

http://www.crifan.com/files/doc/docbook/crifanlib_python/release/html/crifanlib_python.html

有任何意见，建议，提交bug等，都欢迎去讨论组发帖讨论：

http://www.crifan.com/bbs/categories/crifanlib_python/

修订历史		
修订 1.1	2013-09-29	crl
1. 把crifanLib.py从Python语言总结中整理出来单独成此book 2. 更新xml:id		

[1] http://www.crifan.com/files/doc/docbook/crifanlib_python/release/html/crifanlib_python.html
[2] http://www.crifan.com/files/doc/docbook/crifanlib_python/release/htmls/index.html
[3] http://www.crifan.com/files/doc/docbook/crifanlib_python/release/pdf/crifanlib_python.pdf
[4] http://www.crifan.com/files/doc/docbook/crifanlib_python/release/chm/crifanlib_python.chm
[5] http://www.crifan.com/files/doc/docbook/crifanlib_python/release/txt/crifanlib_python.txt
[6] http://www.crifan.com/files/doc/docbook/crifanlib_python/release/rtf/crifanlib_python.rtf
[7] http://www.crifan.com/files/doc/docbook/crifanlib_python/release/webhelp/index.html
[8] http://www.crifan.com/files/doc/docbook/crifanlib_python/release/html/crifanlib_python.html.7z
[9] http://www.crifan.com/files/doc/docbook/crifanlib_python/release/htmls/index.html.7z
[10] http://www.crifan.com/files/doc/docbook/crifanlib_python/release/pdf/crifanlib_python.pdf.7z
[11] http://www.crifan.com/files/doc/docbook/crifanlib_python/release/chm/crifanlib_python.chm.7z
[12] http://www.crifan.com/files/doc/docbook/crifanlib_python/release/txt/crifanlib_python.txt.7z
[13] http://www.crifan.com/files/doc/docbook/crifanlib_python/release/rtf/crifanlib_python.rtf.7z
[14] http://www.crifan.com/files/doc/docbook/crifanlib_python/release/webhelp/crifanlib_python.webhelp.7z

详解crifan的Python库：crifanLib.py:

Crifan Li

版本：v1.1

出版日期 2013-09-29

[15] http://www.crifan.com/files/doc/docbook/soft_dev_basic/release/html/soft_dev_basic.html#cc_by_nc

目录

范例清单

第 1 章 crifanLib.py简介

1.1. 什么是crifanLib.py

之前在折腾将(新版)百度空间,网易163,新浪sina,QQ空间,人人网,CSDN,搜狐Sohu,博客大巴Blogbus,天涯博客,点点轻博客等博客搬家到WordPress[1]的过程中，先后遇到很多个问题，然后基本上也都自己解决了。对应的也写了相应的代码和函数。

后来就把其中比较常用或通用的功能，整理提取出来，放到一个单独的文件中，即crifanLib.py。

1.2. 到哪里可以下载到crifanLib.py

该文件，之前是以帖子的方式贴出来的：crifan的Python库：crifanLib.py[2]

现在已放到google code中的crifanLib[3]中的crifanLib.py[4]，并且以后会同步保持最新版本的。

[1] http://www.crifan.com/crifan_released_all/website/python/blogstowordpress/
[2] http://www.crifan.com/crifan_python_lib_crifanlib_py/
[3] http://code.google.com/p/crifanlib/
[4] http://code.google.com/p/crifanlib/source/browse/trunk/python/crifanLib.py

第 2 章 crifanLib.py函数及用法详解

下面把所有的函数的用法，都简单解释一下：

crifanLib.py所包含的库

如果你在使用这些函数的遇到说某某函数，类等找不到，那很可能是没有导入对应的库。

所以在介绍之前，先贴出，目前crifanLib.py中所导入的一些库和函数：

```
import os;
import re;
import sys;
import time;
import chardet;
import urllib;
import urllib2;
from datetime import datetime,timedelta;
from BeautifulSoup import BeautifulSoup,Tag,CData;
import logging;
#import htmlentitydefs;
import struct;
import zlib;

# from PIL import Image;
# from operator import itemgetter;
```

2.1. 与时间（time,datetime等）有关的函数

2.1.1. 当前时间转换为时间戳:getCurTimestamp

```
from datetime import datetime,timedelta;

#------------------------------------------------------------------------------
# get current time's timestamp
def getCurTimestamp() :
    return datetimeToTimestamp(datetime.now());

#------------------------------------------------------------------------------
# convert datetime value to timestamp
# from "2006-06-01 00:00:00" to 1149091200
def datetimeToTimestamp(datetimeVal) :
    return int(time.mktime(datetimeVal.timetuple()));
```

例 2.1. getCurTimestamp使用范例

```
curTimestamp = getCurTimestamp();
jsonp = "jsonp" + str(curTimestamp);
```

2.1.2. 将时间戳转换为时间变量:timestampToDatetime

```
from datetime import datetime,timedelta;

#------------------------------------------------------------------------------
# convert timestamp to datetime value
# from 1149091200 to "2006-06-01 00:00:00"
def timestampToDatetime(timestamp) :
    #print "type(timestamp)=",type(timestamp);
    #print "timestamp=",timestamp;
    #timestamp = int(timestamp);
    timestamp = float(timestamp);
    return datetime.fromtimestamp(timestamp);
```

例 2.2. timestampToDatetime使用范例

```
createtimeFloat = float(createtimeMillisecond)/1000;
localTime = timestampToDatetime(createtimeFloat);
```

2.1.3. 计算某段代码执行所消耗的时间:calcTimeStart,calcTimeEnd

```
import time;

#------------------------------------------------------------------------------
#init for calculate elapsed time
def calcTimeStart(uniqueKey) :
    global gVal

    gVal['calTimeKeyDict'][uniqueKey] = time.time();
    return

#------------------------------------------------------------------------------
# to get elapsed time, before call this, should use calcTimeStart to init
def calcTimeEnd(uniqueKey) :
    global gVal

    return time.time() - gVal['calTimeKeyDict'][uniqueKey];
```

例 2.3. calcTimeStart和calcTimeEnd的使用范例

```
calcTimeStart("export_head");
exportHead(blogInfoDic);
gVal['statInfoDict']['exportHeadTime'] = calcTimeEnd("export_head");
```

2.1.4. 将本地GMT8时间转换为GMT标准时间:convertLocalToGmt

```
from datetime import datetime,timedelta;

#------------------------------------------------------------------------------
# convert local GMT8 to GMT time
# note: input should be 'datetime' type, not 'time' type
def convertLocalToGmt(localTime) :
    return localTime - timedelta(hours=8);
```

例 2.4. convertLocalToGmt的使用范例

```
gmtTime = convertLocalToGmt(parsedLocalTime);
```

2.2. 和字符串(str,unicode等)处理有关的函数

2.2.1. 从绝对路径中提取出文件名:extractFilename

```
#------------------------------------------------------------------------------
# got python script self file name
# extract out xxx from:
# D:\yyy\zzz\xxx.py
# xxx.py
def extractFilename(inputStr) :
    argv0List = inputStr.split("\\");
    scriptName = argv0List[len(argv0List) - 1]; # get script file name self
    possibleSuf = scriptName[-3:];
    if possibleSuf == ".py" :
        scriptName = scriptName[0:-3]; # remove ".py"
    return scriptName;
```

例 2.5. extractFilename的使用范例

```
if __name__=="__main__":
    # for : python xxx.py -s yyy    # -> sys.argv[0]=xxx.py
    # for : xxx.py -s yyy           # -> sys.argv[0]=D:\yyy\zzz\xxx.py
    scriptSelfName = extractFilename(sys.argv[0]);
```

2.2.2. 将实体定义替换为字符:repUniNumEntToChar

```
#------------------------------------------------------------------------------
```

```
# replace the &#N; (N is digit number, N > 1) to unicode char
# eg: replace "&amp;#39;" with "'" in "Creepin&#39; up on you"
def repUniNumEntToChar(text):
    unicodeP = re.compile('&#[0-9]+;');
    def transToUniChr(match): # translate the matched string to unicode char
        numStr = match.group(0)[2:-1]; # remove '&#' and ';'
        num = int(numStr);
        unicodeChar = unichr(num);
        return unicodeChar;
    return unicodeP.sub(transToUniChr, text);
```

例 2.6. repUniNumEntToChar的使用范例

```
infoDict['title'] = repUniNumEntToChar(infoDict['title']);
```

2.2.3. 生成全路径的URL地址:genFullUrl

```
#------------------------------------------------------------------------------
# generate the full url, which include the main url plus the parameter list
# Note:
# normally just use urllib.urlencode is OK.
# only use this if you do NOT want urllib.urlencode convert some special chars($,:,{,},...) into
 %XX
def genFullUrl(mainUrl, paraDict) :
    fullUrl = mainUrl;
    fullUrl += '?';
    for i, para in enumerate(paraDict.keys()) :
        if(i == 0):
            # first para no '&'
            fullUrl += str(para) + '=' + str(paraDict[para]);
        else :
            fullUrl += '&' + str(para) + '=' + str(paraDict[para]);
    return fullUrl;
```

例 2.7. genFullUrl的使用范例

```
# Note: here not use urllib.urlencode to encode para,
#       for the encoded result will convert some special chars($,:,{,},...) into %XX
paraDict = {
    'asyn'          :   '1',
    'thread_id_enc' :   '',
    'start'         :   '',
    'count'         :   '',
    'orderby_type'  :   '0',
};
paraDict['thread_id_enc'] = str(threadIdEnc);
paraDict['start'] = str(startCmtIdx);
paraDict['count'] = str(reqCmtNum);
paraDict['t'] = str(cmtReqTime);
```

```
mainUrl = "http://hi.baidu.com/cmt/spcmt/get_thread";
getCmtUrl = genFullUrl(mainUrl, paraDict);
```

2.2.4. 判断两个URL地址是否相似:urlIsSimilar

```
#------------------------------------------------------------------------------
# check whether two url is similar
# note: input two url both should be str type
def urlIsSimilar(url1, url2) :
    isSim = False;

    url1 = str(url1);
    url2 = str(url2);

    slashList1 = url1.split('/');
    slashList2 = url2.split('/');
    lenS1 = len(slashList1);
    lenS2 = len(slashList2);

    # all should have same structure
    if lenS1 != lenS2 :
        # not same sturcture -> must not similar
        isSim = False;
    else :
        sufPos1 = url1.rfind('.');
        sufPos2 = url2.rfind('.');
        suf1 = url1[(sufPos1 + 1) : ];
        suf2 = url2[(sufPos2 + 1) : ];
        # at least, suffix should same
        if (suf1 == suf2) :
            lastSlashPos1 = url1.rfind('/');
            lastSlashPos2 = url2.rfind('/');
            exceptName1 = url1[:lastSlashPos1];
            exceptName2 = url2[:lastSlashPos2];
            # except name, all other part should same
            if (exceptName1 == exceptName2) :
                isSim = True;
            else :
                # except name, other part is not same -> not similar
                isSim = False;
        else :
            # suffix not same -> must not similar
            isSim = False;

    return isSim;
```

例 2.8. urlIsSimilar的使用范例

```
if urlIsSimilar(url, srcUrl) :
    isSimilar = True;
```

2.2.5. 判断一个Url地址是否和一个Url地址列表中的某个Url地址相似:findSimilarUrl

如果相似，返回True和相似的地址；

如果不相似，返回False。

```
#------------------------------------------------------------------------------
# found whether the url is similar in urlList
# if found, return True, similarSrcUrl
# if not found, return False, ''
def findSimilarUrl(url, urlList) :
    (isSimilar, similarSrcUrl) = (False, '');
    for srcUrl in urlList :
        if urlIsSimilar(url, srcUrl) :
            isSimilar = True;
            similarSrcUrl = srcUrl;
            break;
    return (isSimilar, similarSrcUrl);
```

例 2.9. findSimilarUrl的使用范例

```
# to check is similar, only when need check and the list it not empty
if ((gCfg['omitSimErrUrl'] == 'yes') and gVal['errorUrlList']):
    (isSimilar, simSrcUrl) = findSimilarUrl(curUrl, gVal['errorUrlList']);
    if isSimilar :
        logging.warning("  Omit process %s for similar with previous error url", curUrl);
        logging.warning("            %s", simSrcUrl);
        continue;
```

2.2.6. 去除非单词（non-word）的字符:removeNonWordChar

```
#------------------------------------------------------------------------------
# remove non-word char == only retian alphanumeric character (char+number) and
 underscore
# eg:
# from againinput4@yeah to againinput4yeah
# from green-waste to greenwaste
def removeNonWordChar(inputString) :
    return re.sub(r"[^\w]", "", inputString); # non [a-zA-Z0-9_]
```

例 2.10. removeNonWordChar的使用范例

```
wxrValidUsername = removeNonWordChar(gVal['blogUser']);
wxrValidUsername = wxrValidUsername.replace("_", "");
logging.info("Generated WXR safe username is %s", wxrValidUsername);
```

2.2.7. 去除控制字符:removeCtlChr

使得处理后的字符串，在XML都是合法的了。

```
#------------------------------------------------------------------------------
# remove control character from input string
# otherwise will cause wordpress importer import failed
# for wordpress importer, if contains contrl char, will fail to import wxr
# eg:
# 1. http://againinput4.blog.163.com/blog/static/172799491201110111145259/
# content contains some invalid ascii control chars
# 2. http://hi.baidu.com/notebookrelated/blog/item/8bd88e351d449789a71e12c2.html
# 165th comment contains invalid control char: ETX
# 3. http://green-waste.blog.163.com/blog/static/32677678200879111913911/
# title contains control char:DC1, BS, DLE, DLE, DLE, DC1
def removeCtlChr(inputString) :
    validContent = '';
    for c in inputString :
        asciiVal = ord(c);
        validChrList = [
            9, # 9=\t=tab
            10, # 10=\n=LF=Line Feed=换行
            13, # 13=\r=CR=回车
        ];
        # filter out others ASCII control character, and DEL=delete
        isValidChr = True;
        if (asciiVal == 0x7F) :
            isValidChr = False;
        elif ((asciiVal < 32) and (asciiVal not in validChrList)) :
            isValidChr = False;

        if(isValidChr) :
            validContent += c;

    return validContent;
```

例 2.11. removeCtlChr的使用范例

```
# remove the control char in title:
# eg;
# http://green-waste.blog.163.com/blog/static/32677678200879111913911/
# title contains control char:DC1, BS, DLE, DLE, DLE, DC1
infoDict['title'] = removeCtlChr(infoDict['title']);
```

关于控制字符

如果不了解什么是控制字符，请参考：ASCII字符集中的功能/控制字符[1]

[1] http://www.crifan.com/files/doc/docbook/char_encoding/release/html/char_encoding.html#ascii_ctrl_char

2.2.8. 将字符实体替换为Unicode数字实体:replaceStrEntToNumEnt

```
#------------------------------------------------------------------------------
# convert the string entity to unicode unmber entity
# refer: http://www.htmlhelp.com/reference/html40/entities/latin1.html
# TODO: need later use this htmlentitydefs instead following
def replaceStrEntToNumEnt(text) :
    strToNumEntDict = {
        # Latin-1 Entities
        "&nbsp;" :   "&#160;",
        "&iexcl;" :  "&#161;",
        "&cent;"    :  "&#162;",
        "&pound;" :  "&#163;",
        "&curren;" :  "&#164;",
        "&yen;"     :  "&#165;",
        "&brvbar;" :  "&#166;",
        "&sect;" :  "&#167;",
        "&uml;"     :  "&#168;",
        "&copy;" :  "&#169;",
        "&ordf;" :  "&#170;",
        "&laquo;" :  "&#171;",
        "&not;"     :  "&#172;",
        "&shy;"     :  "&#173;",
        "&reg;"     :  "&#174;",
        "&macr;" :  "&#175;",
        "&deg;"     :  "&#176;",
        "&plusmn;" :  "&#177;",
        "&sup2;" :  "&#178;",
        "&sup3;" :  "&#179;",
        "&acute;" :  "&#180;",
        "&micro;" :  "&#181;",
        "&para;" :  "&#182;",
        "&middot;" :  "&#183;",
        "&cedil;" :  "&#184;",
        "&sup1;"    :  "&#185;",
        "&ordm;"    :  "&#186;",
        "&raquo;" :  "&#187;",
        "&frac14;" :  "&#188;",
        "&frac12;" :  "&#189;",
        "&frac34;" :  "&#190;",
        "&iquest;" :  "&#191;",
        "&Agrave;" :  "&#192;",
        "&Aacute;" :  "&#193;",
        "&Acirc;" :  "&#194;",
        "&Atilde;" :  "&#195;",
        "&Auml;" :  "&#196;",
        "&Aring;" :  "&#197;",
        "&AElig;" :  "&#198;",
        "&Ccedil;" :  "&#199;",
        "&Egrave;" :  "&#200;",
        "&Eacute;" :  "&#201;",
        "&Ecirc;" :  "&#202;",
        "&Euml;"    :  "&#203;",
        "&Igrave;" :   "&#204;",
        "&Iacute;" :   "&#205;",
        "&Icirc;" :   "&#206;",
        "&Iuml;"    :   "&#207;",
        "&ETH;"     :   "&#208;",
        "&Ntilde;" :   "&#209;",
        "&Ograve;" :   "&#210;",
        "&Oacute;" :   "&#211;",
        "&Ocirc;" :   "&#212;",
        "&Otilde;" :   "&#213;",
        "&Ouml;" :   "&#214;",
        "&times;" :   "&#215;",
        "&Oslash;" :   "&#216;",
        "&Ugrave;" :   "&#217;",
        "&Uacute;" :   "&#218;",
        "&Ucirc;" :   "&#219;",
        "&Uuml;" :   "&#220;",
        "&Yacute;" :   "&#221;",
        "&THORN;" :   "&#222;",
        "&szlig;" :   "&#223;",
        "&agrave;" :   "&#224;",
        "&aacute;" :   "&#225;",
        "&acirc;" :   "&#226;",
        "&atilde;" :   "&#227;",
        "&auml;" :   "&#228;",
        "&aring;" :   "&#229;",
        "&aelig;" :   "&#230;",
        "&ccedil;" :   "&#231;",
        "&egrave;" :   "&#232;",
        "&eacute;" :   "&#233;",
        "&ecirc;" :   "&#234;",
        "&euml;" :   "&#235;",
        "&igrave;" :   "&#236;",
        "&iacute;" :   "&#237;",
        "&icirc;" :   "&#238;",
        "&iuml;" :   "&#239;",
        "&eth;"     :   "&#240;",
        "&ntilde;" :   "&#241;",
        "&ograve;" :   "&#242;",
        "&oacute;" :   "&#243;",
        "&ocirc;" :   "&#244;",
        "&otilde;" :   "&#245;",
        "&ouml;"  :   "&#246;",
        "&divide;" :   "&#247;",
        "&oslash;" :   "&#248;",
        "&ugrave;" :   "&#249;",
        "&uacute;" :   "&#250;",
        "&ucirc;" :   "&#251;",
        "&uuml;" :   "&#252;",
        "&yacute;" :   "&#253;",
        "&thorn;" :   "&#254;",
        "&yuml;" :   "&#255;",
        # http://www.htmlhelp.com/reference/html40/entities/special.html
        # Special Entities
        "&quot;"    : "&#34;",
        "&amp;"     : "&#38;",
        "&lt;"      : "&#60;",
        "&gt;"      : "&#62;",
        "&OElig;"   : "&#338;",
        "&oelig;"   : "&#339;",
        "&Scaron;"  : "&#352;",
        "&scaron;"  : "&#353;",
        "&Yuml;"    : "&#376;",
        "&circ;"    : "&#710;",
        "&tilde;"   : "&#732;",
        "&ensp;"    : "&#8194;",
        "&emsp;"    : "&#8195;",
        "&thinsp;"  : "&#8201;",
        "&zwnj;"    : "&#8204;",
        "&zwj;"     : "&#8205;",
        "&lrm;"     : "&#8206;",
        "&rlm;"     : "&#8207;",
        "&ndash;"   : "&#8211;",
        "&mdash;"   : "&#8212;",
        "&lsquo;"   : "&#8216;",
        "&rsquo;"   : "&#8217;",
        "&sbquo;"   : "&#8218;",
        "&ldquo;"   : "&#8220;",
        "&rdquo;"   : "&#8221;",
        "&bdquo;"   : "&#8222;",
        "&dagger;"  : "&#8224;",
        "&Dagger;"  : "&#8225;",
        "&permil;"  : "&#8240;",
        "&lsaquo;"  : "&#8249;",
        "&rsaquo;"  : "&#8250;",
        "&euro;"    : "&#8364;",
        }

    replacedText = text;
    for key in strToNumEntDict.keys() :
        replacedText = re.compile(key).sub(strToNumEntDict[key], replacedText);
    return replacedText;
```

例 2.12. replaceStrEntToNumEnt的使用范例

```
line = replaceStrEntToNumEnt(line);
```

2.2.9. 将xxx=yyy转换为元祖（tuple）变量:convertToTupleVal

```
#------------------------------------------------------------------------------
# convert the xxx=yyy into tuple('xxx', yyy), then return the tuple value
# [makesure input string]
# (1) is not include whitespace
# (2) include '='
# (3) last is no ';'
# [possible input string]
# blogUserName="againinput4"
# publisherEmail=""
# synchMiniBlog=false
# publishTime=1322129849397
# publisherName=null
# publisherNickname="\u957F\u5927\u662F\u70E6\u607C"
def convertToTupleVal(equationStr) :
    (key, value) = ('', None);

    try :
        # Note:
        # here should not use split with '=', for maybe input string contains string like this:
        # http://img.bimg.126.net/photo/hmZoNQaqzZALvVp0rE7faA==/0.jpg
        # so use find('=') instead
        firstEqualPos = equationStr.find("=");
        key = equationStr[0:firstEqualPos];
        valuePart = equationStr[(firstEqualPos + 1):];

        # string type
        valLen = len(valuePart);
        if valLen >= 2 :
            # maybe string
            if valuePart[0] == '"' and valuePart[-1] == '"' :
                # is string type
                value = str(valuePart[1:-1]);
            elif (valuePart.lower() == 'null'):
                value = None;
            elif (valuePart.lower() == 'false'):
                value = False;
            elif (valuePart.lower() == 'true') :
                value = True;
            else :
                # must int value
                value = int(valuePart);
        else :
            # len=1 -> must be value
            value = int(valuePart);

        #print "Convert %s to [%s]=%s"%(equationStr, key, value);
    except :
        (key, value) = ('', None);
        print "Fail of convert the equal string %s to value"%(equationStr);

    return (key, value);
```

例 2.13. convertToTupleVal的使用范例

```
# (4) convert to value
for equation in equationList :
    (key, value) = convertToTupleVal(equation);
```

2.2.10. 去除列表（List）中的空值:removeEmptyInList

```
#------------------------------------------------------------------------------
# remove the empty ones in list
def removeEmptyInList(list) :
    newList = [];
    for val in list :
        if val :
            newList.append(val);
    return newList;
```

例 2.14. removeEmptyInList的使用范例

```
# Note: some list contain [u''], so is not meaningful, remove it here
# for only [] is empty, [u''] is not empty -> error while exporting to WXR
infoDict['tags'] = removeEmptyInList(infoDict['tags']);
```

2.2.11. 列表去重（去除重复的值）:uniqueList

```
#------------------------------------------------------------------------------
# remove overlapped item in the list
def uniqueList(old_list):
    newList = []
    for x in old_list:
        if x not in newList :
            newList.append(x)
    return newList
```

例 2.15. uniqueList的使用范例

```
nonOverlapList = uniqueList(matchedList); # remove processed
```

2.2.12. 过滤列表（去除在b中出现的a中的某值）:filterList

```
#------------------------------------------------------------------------------
# for listToFilter, remove the ones which is in listToCompare
# also return the ones which is already exist in listToCompare
def filterList(listToFilter, listToCompare) :
    filteredList = [];
    existedList = [];
    for singleOne in listToFilter : # remove processed
        if (not(singleOne in listToCompare)) :
            # omit the ones in listToCompare
            filteredList.append(singleOne);
        else :
            # record the already exist ones
            existedList.append(singleOne);
    return (filteredList, existedList);
```

例 2.16. filterList的使用范例

```
# remove processed and got ones that has been processed
(filteredPicList, existedList) = filterList(nonOverlapList, gVal['processedUrlList']);
```

2.2.13. 生成随机数的字符串:randDigitsStr

```
#------------------------------------------------------------------------------
# generated the random digits number string
# max digit number is 12
def randDigitsStr(digitNum = 12) :
    if(digitNum > 12):
        digitNum = 12;

    randVal = random.random();
    #print "randVal=",randVal; #randVal= 0.134248340235
    randVal = str(randVal);
    #print "randVal=",randVal; #randVal= 0.134248340235

    randVal = randVal.replace("0.", "");
    #print "randVal=",randVal; #randVal= 0.134248340235

    # if last is 0, append that 0
    if(len(randVal)==11):
        randVal = randVal + "0";
    #print "randVal=",randVal; #randVal= 0.134248340235

    #randVal = randVal.replace("e+11", "");
    #randVal = randVal.replace(".", "");
    #print "randVal=",randVal; #randVal= 0.134248340235
    randVal = randVal[0 : digitNum];
    #print "randVal=",randVal; #randVal= 0.134248340235

    return randVal;
```

例 2.17. randDigitsStr 的使用范例

```
captchaUrl += str(randDigitsStr(6));
```

2.2.14. 将元组列表转换为字典变量:tupleListToDict

```
#------------------------------------------------------------------------------
# convert tuple list to dict value
# [(u'type', u'text/javascript'), (u'src', u'http://partner.googleadservices.com/gampad/
google_service.js')]
# { u'type':u'text/javascript', u'src':u'http://partner.googleadservices.com/gampad/
google_service.js' }
def tupleListToDict(tupleList):
    convertedDict = {};

    for eachTuple in tupleList:
        (key, value) = eachTuple;
        convertedDict[key] = value;

    return convertedDict;
```

例 2.18. tupleListToDict 的使用范例

```
#singleContent: name=script, attrMap=None, attrs=[(u'type', u'text/javascript'), (u'src',
 u'http://partner.googleadservices.com/gampad/google_service.js')]
attrsDict = tupleListToDict(singleContent.attrs);
```

2.3. 文件(file等)方面的函数

2.3.1. 将二进制数据存为文件:saveBinDataToFile

```
#------------------------------------------------------------------------------
# save binary data into file
def saveBinDataToFile(binaryData, fileToSave):
    saveOK = False;
    try:
        savedBinFile = open(fileToSave, "wb"); # open a file, if not exist, create it
        #print "savedBinFile=",savedBinFile;
        savedBinFile.write(binaryData);
        savedBinFile.close();
        saveOK = True;
    except :
        saveOK = False;
    return saveOK;
```

例 2.19. saveBinDataToFile的使用范例

```
# if url is invalid, then add timeout can avoid dead
respHtml = getUrlRespHtml(realUrl, useGzip=False, timeout=gConst['defaultTimeout']);
isDownOK = saveBinDataToFile(respHtml, fileToSave);
```

2.4. 网络方面的函数

2.4.1. 检查/判断/校验网络上某个文件是否有效:isFileValid

```
#------------------------------------------------------------------------------
# check file validation:
# open file url to check return info is match or not
# with exception support
# note: should handle while the file url is redirect
# eg :
# http://publish.it168.com/2007/0627/images/500754.jpg ->
# http://img.publish.it168.com/2007/0627/images/500754.jpg
# other special one:
# sina pic url:
# http://s14.sinaimg.cn/middle/3d55a9b7g9522d474a84d&690
# http://s14.sinaimg.cn/orignal/3d55a9b7g9522d474a84d
# the real url is same with above url
def isFileValid(fileUrl) :
    fileIsValid = False;
    errReason = "Unknown error";

    try :
        #print "original fileUrl=",fileUrl;
        origFileName = fileUrl.split('/')[-1];
        #print "origFileName=",origFileName;

        #old: https://ie2zeq.bay.livefilestore.com/y1mo7UWr-
TrmqbBhkw52I0ii__WE6l2UtMRSTZHSky66-
uDxnCdKPr3bdqVrpUcQHcoJLedlFXa43bvCp_O0zEGF3JdG_yZ4wRT-
c2AQmJ_TNcWvVZIXfBDgGerouWyx19WpA4I0XQR1syRJXjDNpwAbQ/IMG_5214_thumb[1].jpg
        #new: https://kxoqva.bay.livefilestore.com/
y1mQlGjwNAYiHKoH5Aw6TMNhsCmX2YDR3vPKnP86snuqQEtnZgy3dHkwUvZ61Ah8zU3AGiS4whmm_ADrvxd
IMG_5214_thumb%5b1%5d.jpg
        unquotedOrigFilenname = urllib.unquote(origFileName);
        #print "unquotedOrigFilenname=",unquotedOrigFilenname
        lowUnquotedOrigFilename = unquotedOrigFilenname.lower();
        #print "lowUnquotedOrigFilename=",lowUnquotedOrigFilename;

        resp = urllib2.urlopen(fileUrl, timeout=gConst['defaultTimeout']); # note: Python 2.6 has
 added timeout support.
        #print "resp=",resp;
        realUrl = resp.geturl();
        #print "realUrl=",realUrl;
        newFilename = realUrl.split('/')[-1];
        #print "newFilename=",newFilename;

        #http://blog.sina.com.cn/s/blog_696e50390100ntxs.html
        unquotedNewFilename = urllib.unquote(newFilename);
        #print "unquotedNewFilename=",unquotedNewFilename;
        unquotedLowNewFilename = unquotedNewFilename.lower();
        #print "unquotedLowNewFilename=",unquotedLowNewFilename;

        respInfo = resp.info();
        #print "respInfo=",respInfo;
        respCode = resp.getcode();
        #print "respCode=",respCode;

        # special:
        # http://116.img.pp.sohu.com/images/blog/2007/5/24/17/24/11355bf42a9.jpg
        # return no content-length
        #contentLen = respInfo['Content-Length'];

        # for redirect, if returned size>0 and filename is same, also should be considered valid
        #if (origFileName == newFilename) and (contentLen > 0):
        # for redirect, if returned response code is 200(OK) and filename is same, also should be
 considered valid
        #if (origFileName == newFilename) and (respCode == 200):
        if (lowUnquotedOrigFilename == unquotedLowNewFilename) and (respCode == 200):
            fileIsValid = True;
        else :
            fileIsValid = False;

            # eg: Content-Type= image/gif, ContentTypes : audio/mpeg
            # more ContentTypes can refer: http://kenya.bokee.com/3200033.html
            contentType = respInfo['Content-Type'];

            errReason = "file url returned info: type=%s, len=%d, realUrl=%s"%(contentType,
 contentLen, realUrl);
    except urllib2.URLError,reason :
        fileIsValid = False;
        errReason = reason;
    except urllib2.HTTPError,code :
        fileIsValid = False;
        errReason = code;
    except :
        fileIsValid = False;
        errReason = "Unknown error";

    # here type(errReason)= <class 'urllib2.HTTPError'>, so just convert it to str
    errReason = str(errReason);
    return (fileIsValid, errReason);
```

例 2.20. isFileValid的使用范例

```
# indeed is pic, process it
(picIsValid, errReason) = isFileValid(curUrl);
```

2.4.2. 下载网络上某个文件:downloadFile

```
#------------------------------------------------------------------------------
# download from fileUrl then save to fileToSave
# with exception support
# note: the caller should make sure the fileUrl is a valid internet resource/file
def downloadFile(fileUrl, fileToSave, needReport = False) :
    isDownOK = False;
    downloadingFile = '';

    #---------------------------------------------------------------------------
    # note: totalFileSize -> may be -1 on older FTP servers which do not return a file size in response to a retrieval request
    def reportHook(copiedBlocks, blockSize, totalFileSize) :
        #global downloadingFile
        if copiedBlocks == 0 : # 1st call : once on establishment of the network connection
            print 'Begin to download %s, total size=%d'%(downloadingFile, totalFileSize);
        else : # rest call : once after each block read thereafter
            print 'Downloaded bytes: %d' % ( blockSize * copiedBlocks);
        return;
    #---------------------------------------------------------------------------

    try :
        if fileUrl :
            downloadingFile = fileUrl;
            if needReport :
                urllib.urlretrieve(fileUrl, fileToSave, reportHook);
            else :
                urllib.urlretrieve(fileUrl, fileToSave);
            isDownOK = True;
        else :
            print "Input download file url is NULL";
    except urllib.ContentTooShortError(msg) :
        isDownOK = False;
    except :
        isDownOK = False;

    return isDownOK;
```

例 2.21. downloadFile的使用范例

```
if dstPicFile and downloadFile(curUrl, dstPicFile) :
    # replace old url with new url
```

2.4.3. （不用urlretrieve）手动从网络上下载单个文件:manuallyDownloadFile

```
#------------------------------------------------------------------------------
# manually download fileUrl then save to fileToSave
def manuallyDownloadFile(fileUrl, fileToSave) :
    isDownOK = False;
    downloadingFile = '';

    try :
        if fileUrl :
            # 1. find real address
            #print "fileUrl=",fileUrl;
            resp = urllib2.urlopen(fileUrl, timeout=gConst['defaultTimeout']);
            #print "resp=",resp;
            realUrl = resp.geturl(); # not same with original file url if redirect

            # if url is invalid, then add timeout can avoid dead
            respHtml = getUrlRespHtml(realUrl, useGzip=False,
 timeout=gConst['defaultTimeout']);

            isDownOK = saveBinDataToFile(respHtml, fileToSave);
        else :
            print "Input download file url is NULL";
    except urllib.ContentTooShortError(msg) :
        isDownOK = False;
    except :
        isDownOK = False;

    return isDownOK;
```

例 2.22. manuallyDownloadFile的使用范例

```
#if dstPicFile and downloadFile(curUrl, dstPicFile) :
# urlretrieve in downloadFile is too slow while download QQ Space Picture
# so here use manuallyDownloadFile instead
if dstPicFile and manuallyDownloadFile(curUrl, dstPicFile) :
    # replace old url with new url
```

2.4.4. 获得Url地址的响应:getUrlResponse

```
#------------------------------------------------------------------------------
# get response from url
# note: if you have already used cookiejar, then here will automatically use it
# while using rllib2.Request
def getUrlResponse(url, postDict={}, headerDict={}, timeout=0, useGzip=False) :
    # makesure url is string, not unicode, otherwise urllib2.urlopen will error
    url = str(url);

    if (postDict) :
        postData = urllib.urlencode(postDict);
        req = urllib2.Request(url, postData);
        req.add_header('Content-Type', "application/x-www-form-urlencoded");
    else :
        req = urllib2.Request(url);

    if(headerDict) :
        #print "added header:",headerDict;
        for key in headerDict.keys() :
            req.add_header(key, headerDict[key]);

    defHeaderDict = {
        'User-Agent'    : gConst['userAgentIE9'],
        'Cache-Control' : 'no-cache',
        'Accept'        : '*/*',
        'Connection'    : 'Keep-Alive',
    };

    # add default headers firstly
    for eachDefHd in defHeaderDict.keys() :
        #print "add default header: %s=%s"%(eachDefHd,defHeaderDict[eachDefHd]);
        req.add_header(eachDefHd, defHeaderDict[eachDefHd]);

    if(useGzip) :
        #print "use gzip for",url;
        req.add_header('Accept-Encoding', 'gzip, deflate');

    # add customized header later -> allow overwrite default header
    if(headerDict) :
        #print "added header:",headerDict;
        for key in headerDict.keys() :
            req.add_header(key, headerDict[key]);

    if(timeout > 0) :
        # set timeout value if necessary
        resp = urllib2.urlopen(req, timeout=timeout);
    else :
        resp = urllib2.urlopen(req);

    return resp;
```

例 2.23. getUrlResponse的使用范例

```
resp = getUrlResponse(url, postDict, headerDict, timeout, useGzip);
respHtml = resp.read();
```

2.4.5. 获得Url返回的HTML网页（源码）内容:getUrlRespHtml

```
#------------------------------------------------------------------------------
# get response html==body from url
#def getUrlRespHtml(url, postDict={}, headerDict={}, timeout=0, useGzip=False) :
def getUrlRespHtml(url, postDict={}, headerDict={}, timeout=0, useGzip=True) :
    resp = getUrlResponse(url, postDict, headerDict, timeout, useGzip);
    respHtml = resp.read();
    if(useGzip) :
        #print "---before unzip, len(respHtml)=",len(respHtml);
        respInfo = resp.info();

        # Server: nginx/1.0.8
        # Date: Sun, 08 Apr 2012 12:30:35 GMT
        # Content-Type: text/html
        # Transfer-Encoding: chunked
        # Connection: close
        # Vary: Accept-Encoding
        # ...
        # Content-Encoding: gzip

        # sometime, the request use gzip,deflate, but actually returned is un-gzip html
        # -> response info not include above "Content-Encoding: gzip"
        # eg: http://blog.sina.com.cn/s/comment_730793bf010144j7_3.html
        # -> so here only decode when it is indeed is gziped data
        if( ("Content-Encoding" in respInfo) and (respInfo['Content-Encoding'] == "gzip")) :
            respHtml = zlib.decompress(respHtml, 16+zlib.MAX_WBITS);
            #print "+++ after unzip, len(respHtml)=",len(respHtml);

    return respHtml;
```

例 2.24. getUrlRespHtml的使用范例：不带额外参数

```
respHtml = getUrlRespHtml(url);
```

例 2.25. getUrlRespHtml的使用范例：带额外参数

```
modifyUrl = gVal['blogEntryUrl'] + "/blog/submit/modifyblog";
#logging.debug("Modify Url is %s", modifyUrl);

#http://hi.baidu.com/wwwhaseecom/blog/item/79188d1b4fa36f068718bf79.html
foundSpBlogID = re.search(r"blog/item/(?P<spBlogID>\w+?).html", url);
if(foundSpBlogID) :
    spBlogID = foundSpBlogID.group("spBlogID");
    logging.debug("Extracted spBlogID=%s", spBlogID);
else :
    modifyOk = False;
    errInfo = "Can't extract post spBlogID !";
    return (modifyOk, errInfo);

newPostContentGb18030 = newPostContentUni.encode("GB18030");
categoryGb18030 = infoDict['category'].encode("GB18030");
titleGb18030 = infoDict['title'].encode("GB18030");

postDict = {
    "bdstoken"      : gVal['spToken'],
    "ct"            : "1",
    "mms_flag"      : "0",
    "cm"            : "2",
    "spBlogID"      : spBlogID,
    "spBlogCatName_o": categoryGb18030, # old catagory
    "edithid"       : "",
    "previewImg"    : "",
    "spBlogTitle"   : titleGb18030,
    "spBlogText"    : newPostContentGb18030,
    "spBlogCatName" : categoryGb18030, # new catagory
    "spBlogPower"   : "0",
    "spIsCmtAllow"  : "1",
    "spShareNotAllow":"0",
    "spVcode"       : "",
    "spVerifyKey"   : "",
}

headerDict = {
    # 如果不添加Referer，则返回的html则会出现错误："数据添加的一般错误"
    "Referer" : gVal['blogEntryUrl'] + "/blog/modify/" + spBlogID,
    }
respHtml = getUrlRespHtml(modifyUrl, postDict, headerDict);
```

2.4.6. 检查（所返回的）cookieJar中，是否所有的cookie都存在:checkAllCookiesExist

因为成功登录某网页后，一般都会有对应的cookie返回，所以常用此函数去判断是否成功登录某网页。

```
#------------------------------------------------------------------------------
# check all cookies in cookiesDict is exist in cookieJar or not
def checkAllCookiesExist(cookieNameList, cookieJar) :
    cookiesDict = {};
    for eachCookieName in cookieNameList :
        cookiesDict[eachCookieName] = False;

    allCookieFound = True;
    for cookie in cookieJar :
        if(cookie.name in cookiesDict) :
            cookiesDict[cookie.name] = True;

    for eachCookie in cookiesDict.keys() :
        if(not cookiesDict[eachCookie]) :
            allCookieFound = False;
            break;

    return allCookieFound;
```

例 2.26. checkAllCookiesExist的使用范例

```
#http://www.darlingtree.com/wordpress/archives/242
gVal['cj'] = cookielib.CookieJar();

opener = urllib2.build_opener(urllib2.HTTPCookieProcessor(gVal['cj']));
urllib2.install_opener(opener);
resp = urllib2.urlopen(baiduSpaceEntryUrl);

loginBaiduUrl = "https://passport.baidu.com/?login";
#username=%D0%C4%C7%E9%C6%DC%CF%A2%B5%D8&password=xxx&mem_pass=on
postDict = {
    'username'  : username,
    'password'  : password,
    'mem_pass'  : 'on',
    };
resp = getUrlResponse(loginBaiduUrl, postDict);
```

```
# check whether the cookie is OK
cookieNameList = ["USERID", "PTOKEN", "STOKEN"];
loginOk = checkAllCookiesExist(cookieNameList, gVal['cj']);
if (not loginOk) :
    logging.error("Login fail for not all expected cookies exist !");
    return loginOk;
```

2.5. 字符编码相关的函数

2.5.1. 判断字符串是否只包含ASCII字符:strIsAscii

```
#------------------------------------------------------------------------------
# depend on chardet
# check whether the strToDect is ASCII string
def strIsAscii(strToDect) :
    isAscii = False;
    encInfo = chardet.detect(strToDect);
    if (encInfo['confidence'] > 0.9) and (encInfo['encoding'] == 'ascii') :
        isAscii = True;
    return isAscii;
```

例 2.27. strIsAscii的使用范例

```
if(not strIsAscii(extractedBlogUser)) :
    # if is: http://hi.baidu.com/资料收集
    # then should quote it, otherwise later output to WXR will fail !
    extractedBlogUser = urllib.quote(extractedBlogUser);
```

2.5.2. 获得（最有可能的）字符串的字符编码类型:getStrPossibleCharset

此代码中是判断是否大于0.5来决定是否是可能的字符串类型。使用者可根据自己需要，改为自己想要的概率，比如0.8等。

```
#------------------------------------------------------------------------------
# get the possible(possiblility > 0.5) charset of input string
def getStrPossibleCharset(inputStr) :
    possibleCharset = "ascii";
    #possibleCharset = "UTF-8";
    encInfo = chardet.detect(inputStr);
    #print "encInfo=",encInfo;
    if (encInfo['confidence'] > 0.5):
        possibleCharset = encInfo['encoding'];
    return possibleCharset;
    #return encInfo['encoding'];
```

例 2.28. getStrPossibleCharset的使用范例

```
validCharset = getStrPossibleCharset(dataJsonStr);
logging.debug("Now try use the detected charset %s to decode it again", validCharset);
```

2.6. 语言翻译方面的函数

2.6.1. 翻译（中文）字符串（为英文字符串）:translateString

此函数支持多种语言。

如无额外参数，则默认是将中文翻译为英文。

```
#------------------------------------------------------------------------------
# depend on BeautifulSoup
# translate strToTranslate from fromLanguage to toLanguage
# return the translated unicode string
# some frequently used language abbrv:
# Chinese Simplified:   zh-CN
# Chinese Traditional:  zh-TW
# English:              en
# German:               de
# Japanese:             ja
# Korean:               ko
# French:               fr
# more can be found at:
# http://code.google.com/intl/ru/apis/language/translate/v2/using_rest.html#language-params
def translateString(strToTranslate, fromLanguage="zh-CN", toLanguage="en"):
    transOK = False;
    translatedStr = strToTranslate;
    transErr = '';

    try :
        # following refer: http://python.u85.us/viewnews-335.html
        postDict = {'hl':'zh-CN', 'ie':'UTF-8', 'text':strToTranslate, 'langpair':"%s|%s"%(fromLanguage, toLanguage)};
        googleTranslateUrl = 'http://translate.google.cn/translate_t';
        resp = getUrlRespHtml(googleTranslateUrl, postDict);
        #logging.debug("---------------google translate resp html:\n%s", resp);
    except urllib2.URLError,reason :
        transOK = False;
        transErr = reason;
    except urllib2.HTTPError,code :
        transOK = False;
        transErr = code;
    else :
        soup = BeautifulSoup(resp);
        resultBoxSpan = soup.find(id='result_box');
        if resultBoxSpan and resultBoxSpan.span and resultBoxSpan.span.string :
            transOK = True;
            #translatedStr = resultBoxSpan.span.string.encode('utf-8');
            googleRetTransStr = resultBoxSpan.span.string;
            translatedStr = unicode(googleRetTransStr);

            # just record some special one:
            # from:
            #【转载】[SEP4020  u-boot]  start.s  注释
            # to:
            # The 【reserved] [the SEP4020 u-boot] start.s comment
        else :
            transOK = False;
            transErr = "can not extract translated string from returned result";

    transErr = str(transErr);

    if transOK :
        return (transOK, translatedStr);
    else :
        return (transOK, transErr);
```

例 2.29. translateString的使用范例

```
(transOK, translatedStr) = translateString(strToTrans, "zh-CN", "en");
```

2.6.2. 将中文字符串翻译为英文字符串:transZhcnToEn

```
#------------------------------------------------------------------------------
# translate the Chinese Simplified(Zh-cn) string to English(en)
def transZhcnToEn(strToTrans) :
    translatedStr = strToTrans;
    transOK = False;
    transErr = '';

    if strIsAscii(strToTrans) :
        transOK = True;
        translatedStr = strToTrans;
    else :
        (transOK, translatedStr) = translateString(strToTrans, "zh-CN", "en");

    return (transOK, translatedStr);
```

例 2.30. transZhcnToEn的使用范例

```
(transOK, translatedName) = transZhcnToEn(nameUtf8);
```

2.7. Beautifulsoup相关的函数

2.7.1. 从soup的Contents中移除某个（带某种属性的）标签: removeSoupContentsTagAttr

```
#------------------------------------------------------------------------------
#remove specific tag[key]=value in soup contents (list of BeautifulSoup.Tag/
BeautifulSoup.NavigableString)
# eg:
# (1)
# removeSoupContentsTagAttr(soupContents, "p", "class", "cc-lisence")
# to remove <p class="cc-lisence" style="line-height:180%;">......</p>, from
# [
# u'\n',
# <p class="cc-lisence" style="line-height:180%;">......</p>,
# u'\u5bf9......\u3002',
#  <p>跑题了。......我争取。</p>,
#  <br />,
#  u'\n',
#  <div class="clear"></div>,
# ]
# (2)
#contents = removeSoupContentsTagAttr(contents, "div", "class", "addfav", True);
# remove <div class="addfav">.....</div> from:
# [u'\n',
# <div class="postFooter">......</div>,
# <div style="padding-left:2em">
   # ...
   # <div class="addfav">......</div>
   # ...
# </div>,
 # u'\n']
def removeSoupContentsTagAttr(soupContents, tagName, tagAttrKey, tagAttrVal="",
 recursive=False) :
   global gVal;

   #print "in removeSoupContentsClass";

   #print "[",gVal['currentLevel'],"] input tagName=",tagName," tagAttrKey=",tagAttrKey,"
 tagAttrVal=",tagAttrVal;

   #logging.debug("[%d] input, %s[%s]=%s, soupContents:%s",
 gVal['currentLevel'],tagName,tagAttrKey,tagAttrVal, soupContents);
   #logging.debug("[%d] input, %s[%s]=%s", gVal['currentLevel'],tagName, tagAttrKey,
 tagAttrVal);

   filtedContents = [];
   for singleContent in soupContents:
       #logging.debug("current singleContent=%s",singleContent);

       #logging.info("singleContent=%s", singleContent);
       #print "type(singleContent)=",type(singleContent);
       #print "singleContent.__class__=",singleContent.__class__;
```

```
        #if(isinstance(singleContent, BeautifulSoup)):
        #if(BeautifulSoup.Tag == singleContent.__class__):
        #if(isinstance(singleContent, instance)):
        #if(isinstance(singleContent, BeautifulSoup.Tag)):
        if(isinstance(singleContent, Tag)):
            #print "isinstance true";

            #logging.debug("singleContent: name=%s, attrMap=%s, attrs=
%s",singleContent.name, singleContent.attrMap, singleContent.attrs);
            # if( (singleContent.name == tagName)
                # and (singleContent.attrMap)
                # and (tagAttrKey in singleContent.attrMap)
                # and ( (tagAttrVal and (singleContent.attrMap[tagAttrKey]==tagAttrVal)) or (not
tagAttrVal) ) ):
                # print "++++++++found tag:",tagName,"[",tagAttrKey,"]=",tagAttrVal,"\n
in:",singleContent;
                # #print "dir(singleContent)=",dir(singleContent);
                # logging.debug("found %s[%s]=%s in %s", tagName, tagAttrKey, tagAttrVal,
singleContent.attrMap);

            # above using attrMap, but attrMap has bug for:
            #singleContent: name=script, attrMap=None, attrs=[(u'type', u'text/javascript'), (u'src',
u'http://partner.googleadservices.com/gampad/google_service.js')]
            # so use attrs here
            #logging.debug("singleContent: name=%s, attrs=%s", singleContent.name,
singleContent.attrs);
            attrsDict = tupleListToDict(singleContent.attrs);
            if( (singleContent.name == tagName)
                and (singleContent.attrs)
                and (tagAttrKey in attrsDict)
                and ( (tagAttrVal and (attrsDict[tagAttrKey]==tagAttrVal)) or (not tagAttrVal) ) ):
                #print "++++++++found tag:",tagName,"[",tagAttrKey,"]=",tagAttrVal,"\n
in:",singleContent;
                #print "dir(singleContent)=",dir(singleContent);
                logging.debug("found %s[%s]=%s in %s", tagName, tagAttrKey, tagAttrVal,
attrsDict);
            else:
                if(recursive):
                    #print "-----sub call";
                    gVal['currentLevel'] = gVal['currentLevel'] + 1;
                    #logging.debug("[%d] now will filter %s[%s=]%s, for singleContent.contents=%s",
gVal['currentLevel'], tagName,tagAttrKey,tagAttrVal, singleContent.contents);
                    #logging.debug("[%d] now will filter %s[%s=]%s", gVal['currentLevel'],
tagName,tagAttrKey,tagAttrVal);
                    filteredSingleContent = singleContent;
                    filteredSubContentList =
removeSoupContentsTagAttr(filteredSingleContent.contents, tagName, tagAttrKey, tagAttrVal,
recursive);
                    gVal['currentLevel'] = gVal['currentLevel'] -1;
                    filteredSingleContent.contents = filteredSubContentList;
                    #logging.debug("[%d] after filter, sub contents=%s", gVal['currentLevel'],
filteredSingleContent);
                    #logging.debug("[%d] after filter contents", gVal['currentLevel']);
                    filtedContents.append(filteredSingleContent);
                else:
                    #logging.debug("not recursive, append:%s", singleContent);
                    #logging.debug("not recursive, now append singleContent");
                    filtedContents.append(singleContent);
            # name = singleContent.name;
            # if(name == tagName):
                # print "name is equal, name=",name;

                # attrMap = singleContent.attrMap;
                # print "attrMap=",attrMap;
                # if attrMap:
                    # if tagAttrKey in attrMap:
                        # print "tagAttrKey=",tagAttrKey," in attrMap";
                        # if(tagAttrVal and (attrMap[tagAttrKey]==tagAttrVal)) or (not tagAttrVal):
                            # print "++++++++found tag:",tagName,"[",tagAttrKey,"]=",tagAttrVal,"\n
in:",singleContent;
                            # #print "dir(singleContent)=",dir(singleContent);
                            # logging.debug("found tag, tagAttrVal=%s, %s[%s]=%s", tagAttrVal,
tagName, tagAttrVal, attrMap[tagAttrKey]);
                        # else:
                            # print "key in attrMap, but value not equal";
                            # if(recursive):
                                # print "-----sub call 111";
                                # gVal['currentLevel'] = gVal['currentLevel'] + 1;
                                # singleContent = removeSoupContentsTagAttr(singleContent.contents,
tagName, tagAttrKey, tagAttrVal, recursive);
                                # gVal['currentLevel'] = gVal['currentLevel'] -1;
                            # filtedContents.append(singleContent);
                    # else:
                        # print "key not in attrMap";
                        # if(recursive):
                            # print "-----sub call 222";
                            # gVal['currentLevel'] = gVal['currentLevel'] + 1;
                            # singleContent = removeSoupContentsTagAttr(singleContent.contents,
tagName, tagAttrKey, tagAttrVal, recursive);
                            # gVal['currentLevel'] = gVal['currentLevel'] -1;
                        # filtedContents.append(singleContent);
                # else:
                    # print "attrMap is None";
                    # if(recursive):
                        # print "-----sub call 333";
                        # gVal['currentLevel'] = gVal['currentLevel'] + 1;
                        # singleContent = removeSoupContentsTagAttr(singleContent.contents,
tagName, tagAttrKey, tagAttrVal, recursive);
                        # gVal['currentLevel'] = gVal['currentLevel'] -1;
                    # filtedContents.append(singleContent);
            # else:
                # print "name not equal, name=",name," tagName=",tagName;
                # if(recursive):
                    # print "-----sub call 444";
                    # gVal['currentLevel'] = gVal['currentLevel'] + 1;
                    # singleContent = removeSoupContentsTagAttr(singleContent.contents,
tagName, tagAttrKey, tagAttrVal, recursive);
                    # gVal['currentLevel'] = gVal['currentLevel'] -1;
                # filtedContents.append(singleContent);
        else:
            # is BeautifulSoup.NavigableString
            #print "not BeautifulSoup instance";
            filtedContents.append(singleContent);

    #print "filterd contents=",filtedContents;
    #logging.debug("[%d] before return, filtedContents=%s", gVal['currentLevel'], filtedContents);

    return filtedContents;
```

例 2.31. removeSoupContentsTagAttr 的使用范例

```
foundPostbody = soup.find(attrs={"class":"postBody"});
contents = foundPostbody.contents;
contents = removeSoupContentsTagAttr(contents, "p", "class", "cc-lisence", True); #版权声明
contents = removeSoupContentsTagAttr(contents, "div", "class", "relpost", True); #历史上的今天, 相关帖子
contents = removeSoupContentsTagAttr(contents, "div", "class", "addfav", True); #收藏到
```

2.7.2. 查找contents中第一个NavigableString: findFirstNavigableString

```
#------------------------------------------------------------------------------
# find the first BeautifulSoup.NavigableString from soup contents
def findFirstNavigableString(soupContents):
    firstString = None;
    for eachContent in soupContents:
        # note here must import NavigableString from BeautifulSoup
        if(isinstance(eachContent, NavigableString)):
            firstString = eachContent;
            break;

    return firstString;
```

2.7.3. 将soup的contents转换为Unicode字符串: soupContentsToUnicode

```
#------------------------------------------------------------------------------
# convert soup contents into unicode string
def soupContentsToUnicode(soupContents) :
    #method 1
    mappedContents = map(CData, soupContents);
    #print "mappedContents OK";
    #print "type(mappedContents)=",type(mappedContents); #type(mappedContents)= <type 'list'>
    contentUni = ''.join(mappedContents);
    #print "contentUni=",contentUni;

    # #method 2
    # originBlogContent = "";
    # logging.debug("Total %d contents for original soup contents:", len(soupContents));
    # for i, content in enumerate(soupContents):
        # if(content):
            # logging.debug("[%d]=%s", i, content);
            # originBlogContent += unicode(content);
        # else :
            # logging.debug("[%d] is null", i);

    # logging.debug("---method 1: map and join---\n%s", contentUni);
    # logging.debug("---method 2: enumerate   ---\n%s", originBlogContent);

    # # -->> seem that two method got same blog content

    #logging.debug("soup contents to unicode string OK");
    return contentUni;
```

例 2.32. soupContentsToUnicode 的使用范例

```
postmetadataUni = soupContentsToUnicode(foundPostmetadata.contents);
```

参考书目

[1] 将(新版)百度空间,网易163,新浪sina,QQ空间,人人网,CSDN,搜狐Sohu,博客大巴Blogbus,天涯博客,点点轻博客等博客搬家到WordPress[1]

[1] http://www.crifan.com/crifan_released_all/website/python/blogstowordpress/