集算器

集算器抓取HTML中的内容

润乾软件出品

主要内容

1、简介

2、函数说明

3、抓取工具

1、简介

集算器作为数据计算中间件，可以从数据库或本地文件中取数，承担业务逻辑复杂的计算和外存计算，并将结果返回应用程序。

我们常用的本地文件为txt文件、excel文件等，集算器作为中间件，可以从中抓取所需的内容，如：

	A	
1	=file("D:\\sales.txt").import(#1,#3)	取出sale.txt中的第1列与第3列
2	=file("D:\\emp.xlsx").xlsimport(;"emp",10:20)	取出emp.xlsx名为“emp”的sheet中第10行到第20行

同样的，集算器也可以配合position工具，准确的抓取HTML中的内容。

下面我们来了解一下所需要用到的函数，以及了解使用position工具。

2、函数说明

集算器可以通过使用htmlparse()函数，获取HTML中指定标签的具体数据文本。

函数介绍如下：

s.htmlparse(tag:i:j,...)	获取html文件s中第i个tag下的第j个文本 (i和j从0开始计数)

其中，s 表示html文件的内容，tag 为html中的标签，i 和 j 均为整数。

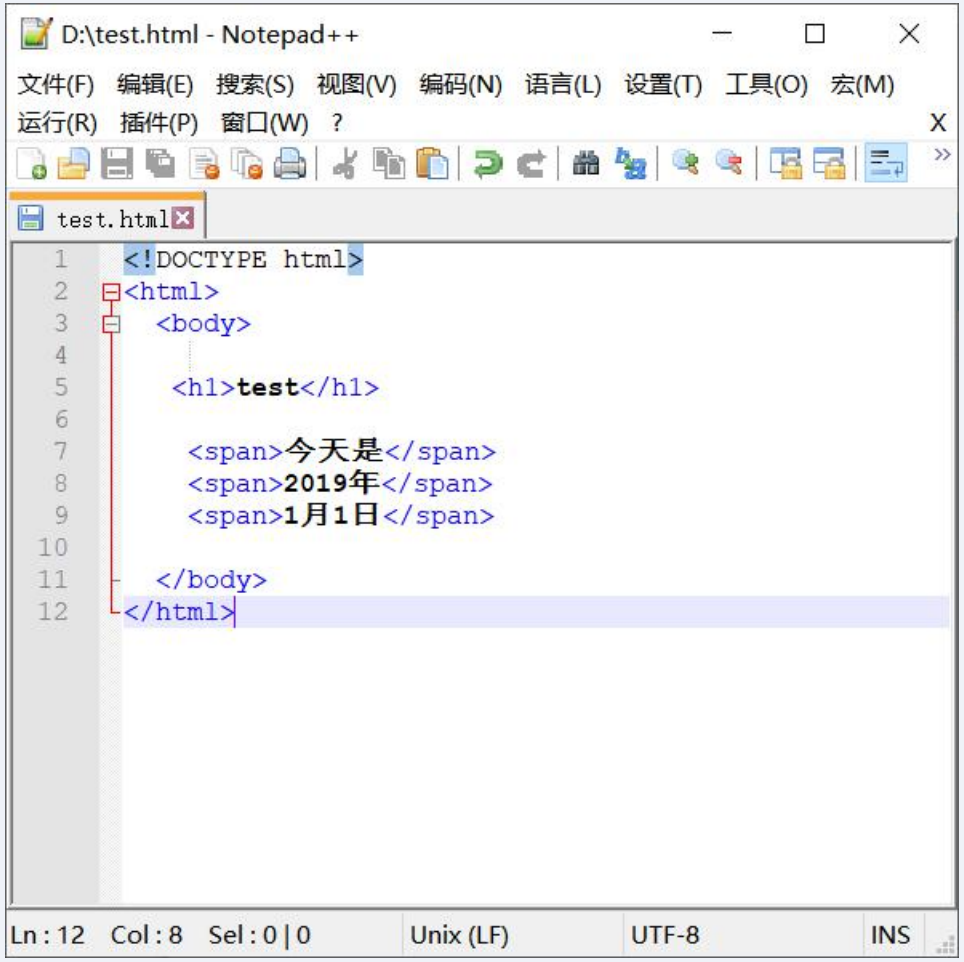
示例1

	A	
1	=file("D:/test.html").read()	读取test.html中的内容
2	=A1.htmlparse("span":1:0)	获取A1中第2个标签下的第1个文本

返回的结果为

序号	成员
1	2019年

2、函数说明



在图中可以看出，第2个标签下的第1个文本内容确实是“2019”

我们还可以在集算器中取多个tag下的内容

	A
1	=file("D:/test.html").read()
2	=A1.htmlparse("span":1:0,"h1":0:0)

返回由“2019年”和“test”组成的序列

序号	成员
1	2019年
2	test

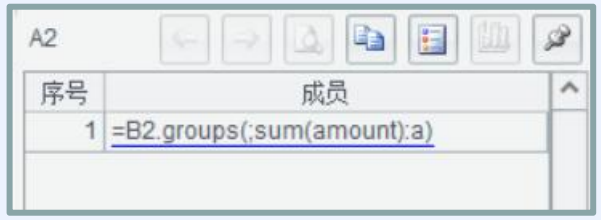
2、函数说明

示例2

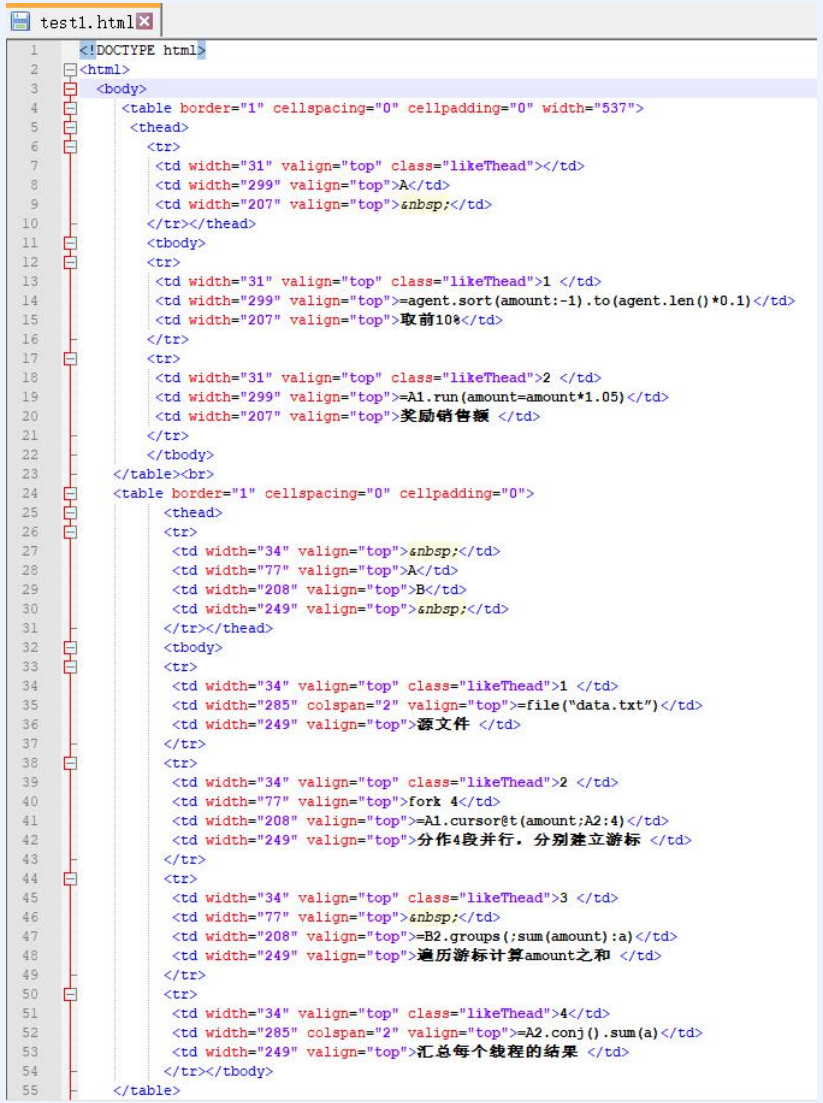
使用htmlparse函数抓取test1.html中的内容

	A
1	=file("D:/test1.html").read()
2	=A1.htmlparse("td":22:0)

A2 获取第23个<td>标签下的第1个文本内容



标签中存在属性时，也可以获取到文本内容。



2、函数说明

示例3

使用htmlparse函数抓取tableTest.html中的内容

	A
1	=file("D:/tableTest.html").read()
2	=A1.htmlparse("table":7)

A2 获取第7个<table>标签下的所有数据

A2

序号	成员
1	[,A]
2	[1,=file("/home/ctx/lineitem_xh.ctx").create().cursor(L_ORDE...
3	[2,=A1.switch(L_ORDERKEY,orders:#;L_PARTKEY,part:#;L...
4	[3,=A2.select(L_ORDERKEY.O_TOTALPRICE>0 && L_PAR...
5	[4,=A3.groups(year(L_SHIPDATE):l_year; sum(L_EXTENDE...

tableTest.html

```
653 <table width="563"> BEL
654  <tbody>
655   <tr>
656    <td width="33" style="vertical-align: top;"><strong><span>1
       </span></strong></td>
657    <td width="350" style="vertical-align: top;">
       =file("/home/ctx/lineitem_xh.ctx").create().cursor(L_ORDERKEY,L_PAR
       TKEY,   L_SUPPKEY,L_EXTENDEDPRICE,L_DISCOUNT,L_SHIPDATE)</td>
658   </tr>
659   <tr>
660    <td width="33" style="vertical-align: top;"><strong><span>2
       </span></strong></td>
661    <td width="350" style="vertical-align: top;">
       =A1.switch(L_ORDERKEY,orders:#;L_PARTKEY,part:#;L_SUPPKEY,supplier:
       #)</td>
662   </tr>
663   <tr>
664    <td width="33" style="vertical-align: top;"><strong><span>3
       </span></strong></td>
665    <td width="350" style="vertical-align: top;">
       =A2.select(L_ORDERKEY.O_TOTALPRICE&gt;0   &amp;&amp;
       L_PARTKEY.P_SIZE&gt;0 &amp;&amp; L_SUPPKEY.S_ACCTBAL&lt;999999)
       </td>
666   </tr>
667   <tr>
668    <td width="33" style="vertical-align: top;"><strong><span>4
       </span></strong></td>
669    <td width="350" style="vertical-align: top;">
       =A3.groups(year(L_SHIPDATE):l_year;   sum(L_EXTENDEDPRICE *
       (1 - L_DISCOUNT)):revenue)</td>
670   </tr>
671  </tbody>
672 </table>
```

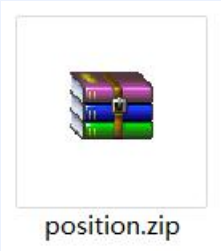
若tag为table标签，则会抓取出table标签下的所有数据

3、抓取工具

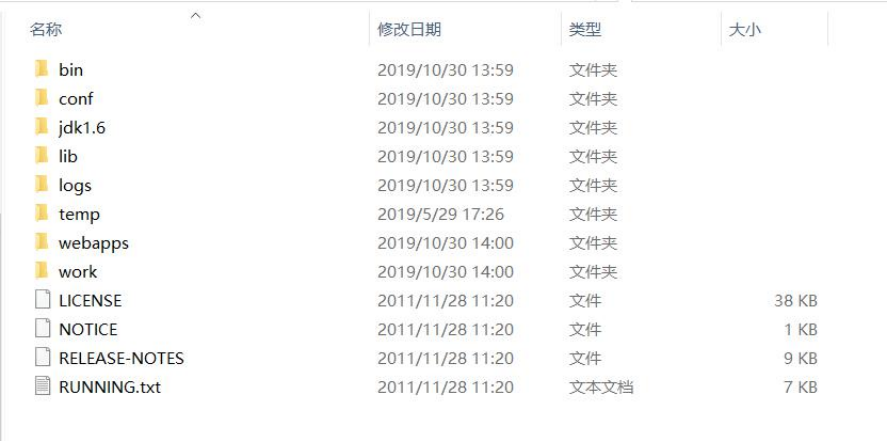
然而，HTML文件通常比上述示例中的内容复杂，自己去数某个位置的标签序号显然很困难，所以我们提供了position工具，可以直观定位到html中具体标签的位置。

➢ 下载链接: http://c.raqsoft.com.cn/article/1572574420058

将position.zip解压到本地，并点击打开position下的bin目录，运行startup.bat即可。

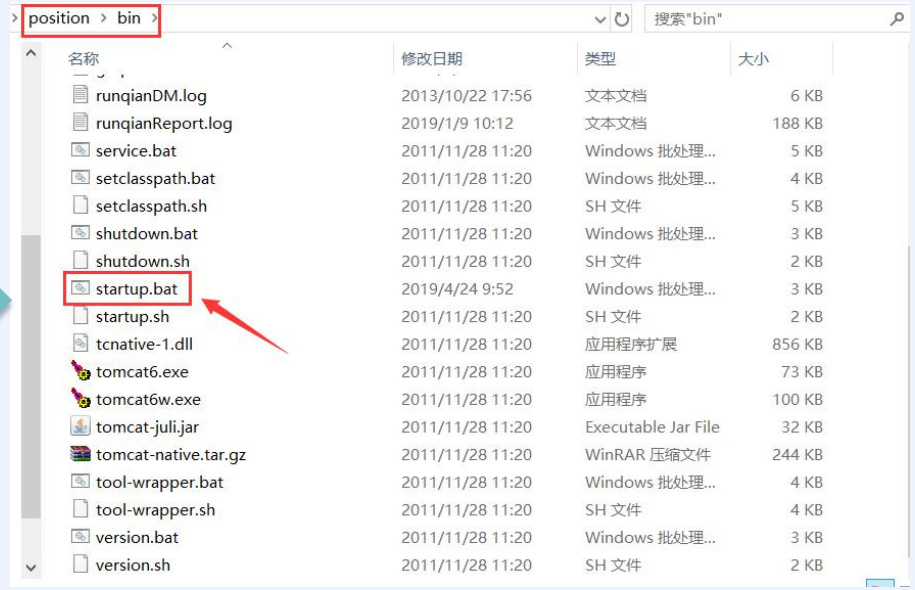


解压到本地



名称	修改日期	类型	大小
bin	2019/10/30 13:59	文件夹	
conf	2019/10/30 13:59	文件夹	
jdk1.6	2019/10/30 13:59	文件夹	
lib	2019/10/30 13:59	文件夹	
logs	2019/10/30 13:59	文件夹	
temp	2019/5/29 17:26	文件夹	
webapps	2019/10/30 14:00	文件夹	
work	2019/10/30 14:00	文件夹	
LICENSE	2011/11/28 11:20	文件	38 KB
NOTICE	2011/11/28 11:20	文件	1 KB
RELEASE-NOTES	2011/11/28 11:20	文件	9 KB
RUNNING.txt	2011/11/28 11:20	文本文档	7 KB

解压后



position › bin

搜索"bin"

名称	修改日期	类型	大小
runqianDM.log	2013/10/22 17:56	文本文档	6 KB
runqianReport.log	2019/1/9 10:12	文本文档	188 KB
service.bat	2011/11/28 11:20	Windows 批处理...	5 KB
setclasspath.bat	2011/11/28 11:20	Windows 批处理...	4 KB
setclasspath.sh	2011/11/28 11:20	SH 文件	5 KB
shutdown.bat	2011/11/28 11:20	Windows 批处理...	3 KB
shutdown.sh	2011/11/28 11:20	SH 文件	2 KB
startup.bat	2019/4/24 9:52	Windows 批处理...	3 KB
startup.sh	2011/11/28 11:20	SH 文件	2 KB
tcnative-1.dll	2011/11/28 11:20	应用程序扩展	856 KB
tomcat6.exe	2011/11/28 11:20	应用程序	73 KB
tomcat6w.exe	2011/11/28 11:20	应用程序	100 KB
tomcat-juli.jar	2011/11/28 11:20	Executable Jar File	32 KB
tomcat-native.tar.gz	2011/11/28 11:20	WinRAR 压缩文件	244 KB
tool-wrapper.bat	2011/11/28 11:20	Windows 批处理...	4 KB
tool-wrapper.sh	2011/11/28 11:20	SH 文件	4 KB
version.bat	2011/11/28 11:20	Windows 批处理...	3 KB
version.sh	2011/11/28 11:20	SH 文件	2 KB

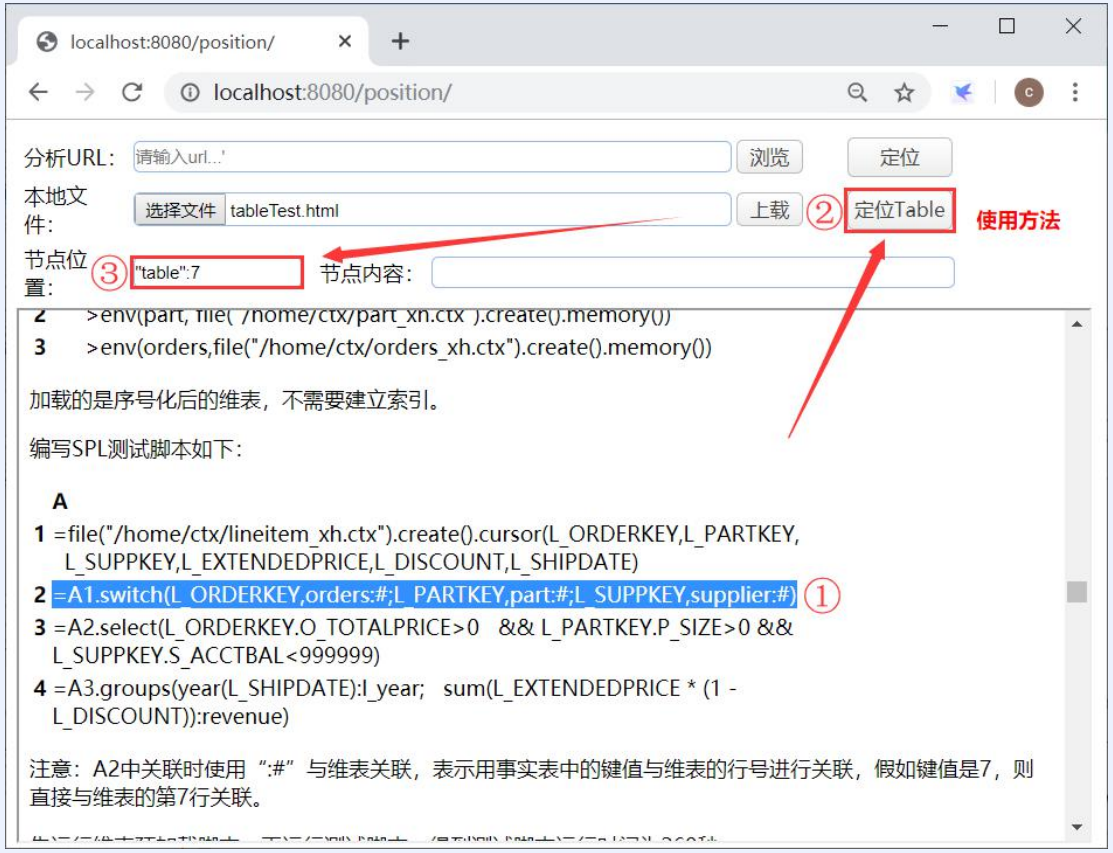
双击startup.bat

3、抓取工具

- 成功启动startup.bat后，打开浏览器并输入http://localhost:8080/position/，出现如下界面;
- 点击“选择文件”，选中本地的HTML文件并“上载”（如图1）；
- 在图中选取需要抓取的文本，然后点击“定位”（如图2），图中“节点位置”的内容【“h3”：1,0 】可直接代入s.htmlparse(tag:i:j,...)中，作为tag:i:j参数使用。

图1

图2

3、抓取工具

- position工具同样可以定位table标签的位置（如图）
- 上载本地文件后，选中table标签中的数据，点击【定位Table】按钮，节点位置处显示的参数可代入函数中使用