Atitit 数据挖掘 数据提取 工具 html版nodejs版

C:\0workspace\atiplat\_spider\com.attilax\spider\searchOrgByNameLike.js

*//main***function** *main*(){  
 **var** dir=**"c:\\00orgSave"**;  
 **var** fls=*getFileListV2*(dir);  
 **for**(**idx in** fls)  
 {  
 */\*\*  
 \* { path: 'c:\\00orgSave',  
 filename: 'index\_P57\_1484408092882.html' }  
 \*/* **var** f=fls[**idx**];  
 **var** f\_str= f.**path**+**"/"**+ f.**filename**;  
 **var** data=***fs***.readFileSync(f\_str,**"utf-8"**);  
 *parseByNodeJq*(data);  
  
 *// console.log(f);* }  
}

**function** *parseByNodeJq*(html)  
{  
 **var** jsdom = require(**"jsdom"**);  
 ***$*** = require(**"jquery"**)(jsdom.jsdom().**defaultView**);  
 *//only low ver can this  
 // var $ = require('jquery');* **var** $doc = **$**(html);  
 *// console.log("No. name language star forks ")* $doc.find(**"tr"**).each(**function**(i,project){  
 **var** $project = **$**(project);  
 **var** text=$project.text();  
 *// var forks = $project.find("li.forks").text().trim();* **if**(text.indexOf(**"511"**)>0)  
 {  
 text=text.trim();  
 text=*replaceEnter2Empty*(text);  
 text=*replaceTab2Empty*(text);  
 text= *removeToomanySpacechar*(text,800);  
 *// text=text.replace(" "," ");* text=text.replace(**"\n"**,**" "**);  
 **console**.log(**"--tr line:"**+text);  
 }  
  
 });  
  
}

**function** *getFileListV2*(path) {  
 **var** filesList = [];  
 *readFileList*(path, filesList);  
 **return** filesList;  
}