Contents

[Target:“PhoneCharge” cik:949341 1](#_Toc28283626)

[Manual flow 1](#_Toc28283627)

[HTML Parsing 4](#_Toc28283628)

[Target:“ Sherman” cik:1287865 4](#_Toc28283629)

[Manual flow 4](#_Toc28283630)

[HTML Parsing 6](#_Toc28283631)

[Target:“Senior Health” cik: 1241199 6](#_Toc28283632)

[Manual Flow 6](#_Toc28283633)

[HTML Parsing 7](#_Toc28283634)

[Target:“G&L” cik: 1109189 7](#_Toc28283635)

[Manual Flow 7](#_Toc28283636)

[HTML Parsing 7](#_Toc28283637)

[Target:“Sirius” cik: 879993 7](#_Toc28283638)

[Manual Flow 7](#_Toc28283639)

[HTML Parsing 9](#_Toc28283640)

[Target:“Visibility” cik: 1073349 9](#_Toc28283641)

[Manual Flow 9](#_Toc28283642)

[HTML Parsing 9](#_Toc28283643)

[Target:“Gas”cik: 726958 9](#_Toc28283644)

[Manual Flow 9](#_Toc28283645)

[HTML Parsing 10](#_Toc28283646)

## Target:“PhoneCharge” cik:949341

### Manual flow

1. 在<https://www.sec.gov/edgar/searchedgar/companysearch.html>网站中，



在fast search中输入excel表中的“cik”列里的，比如：

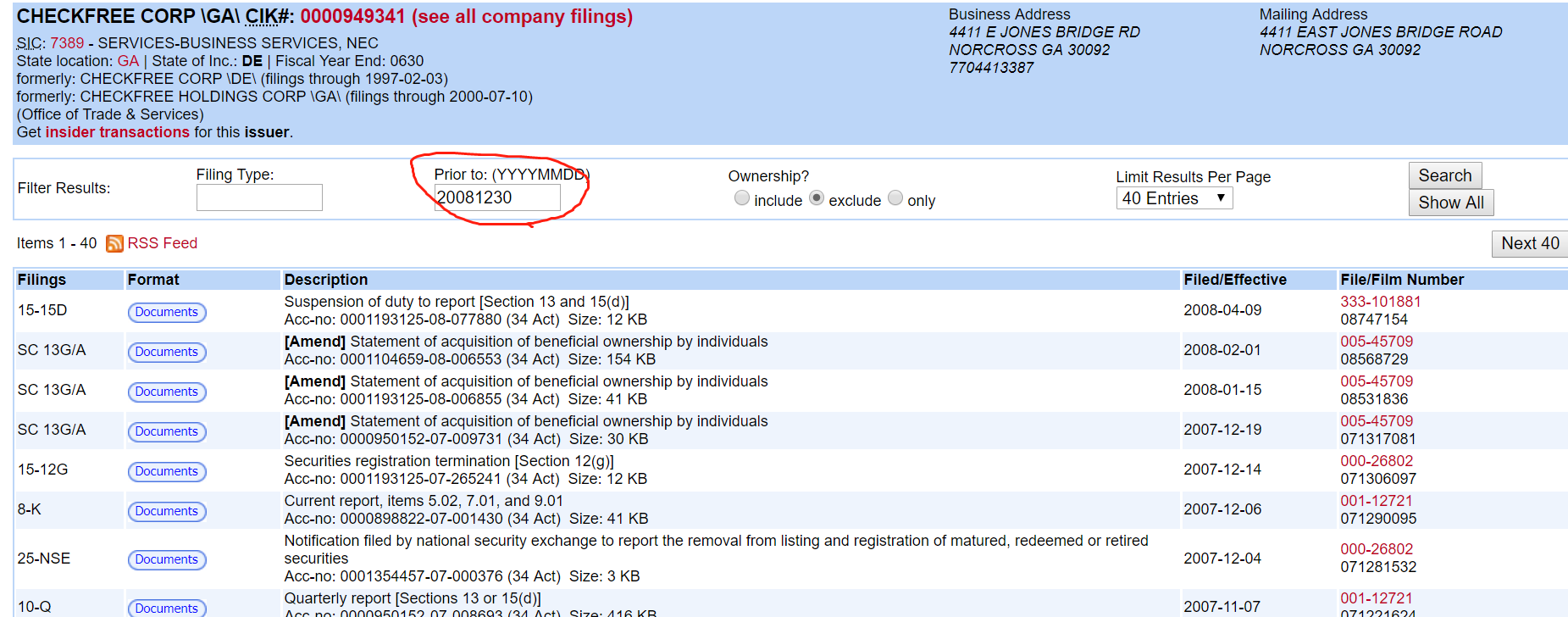
第一行输入 ‘949341’

2. 下面这个网页，显示了‘949341’这个公司的所有报表。我需要在 date\_announced列中（1/3/2006）这一天之后的所有报表中，寻找到第一个出现关于targetname列中的公司 （PhoneCharge）“intangible asset”的信息。 具体操作如下：

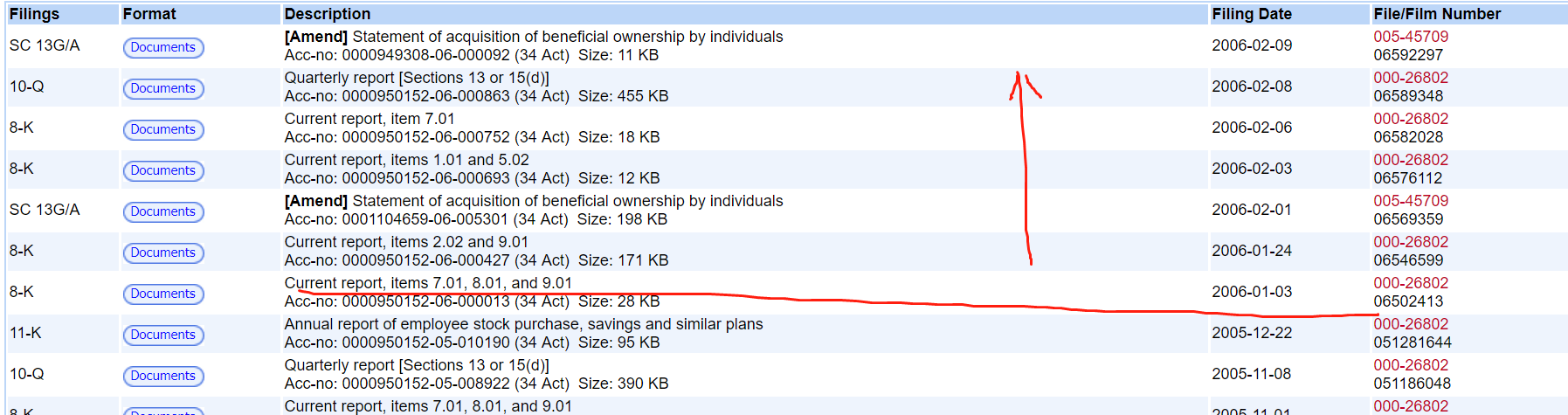


1. 定位到date\_announced列中（1/3/2006）这一天

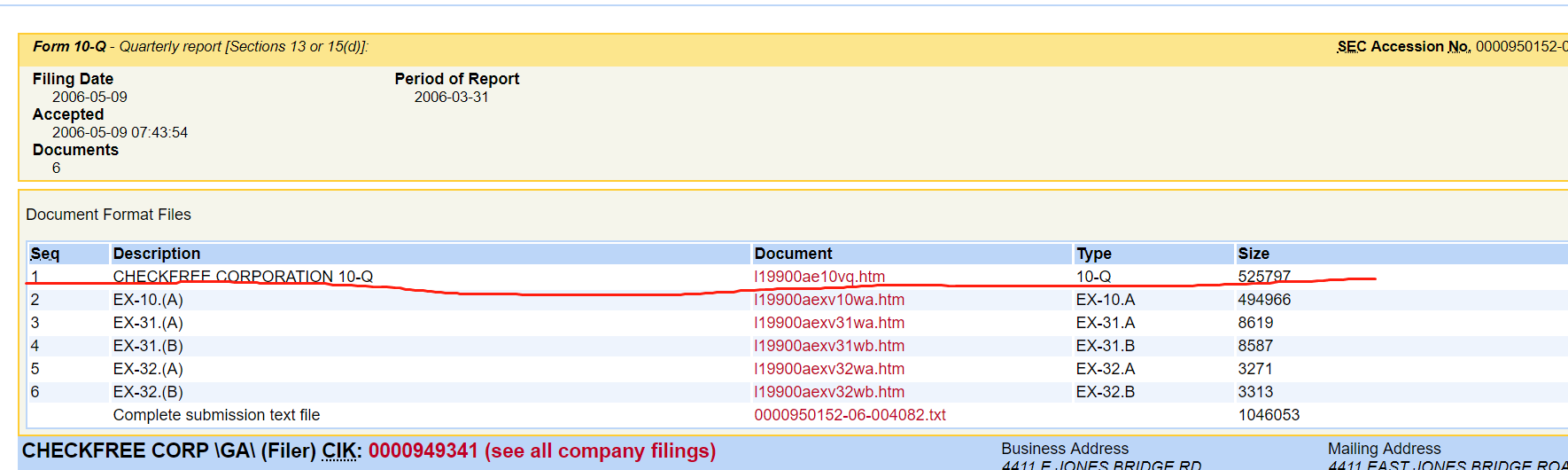
手动定位的最简单方法，是在filter（prior to这里）定位两年后的 某一天， 比如12/30/2008



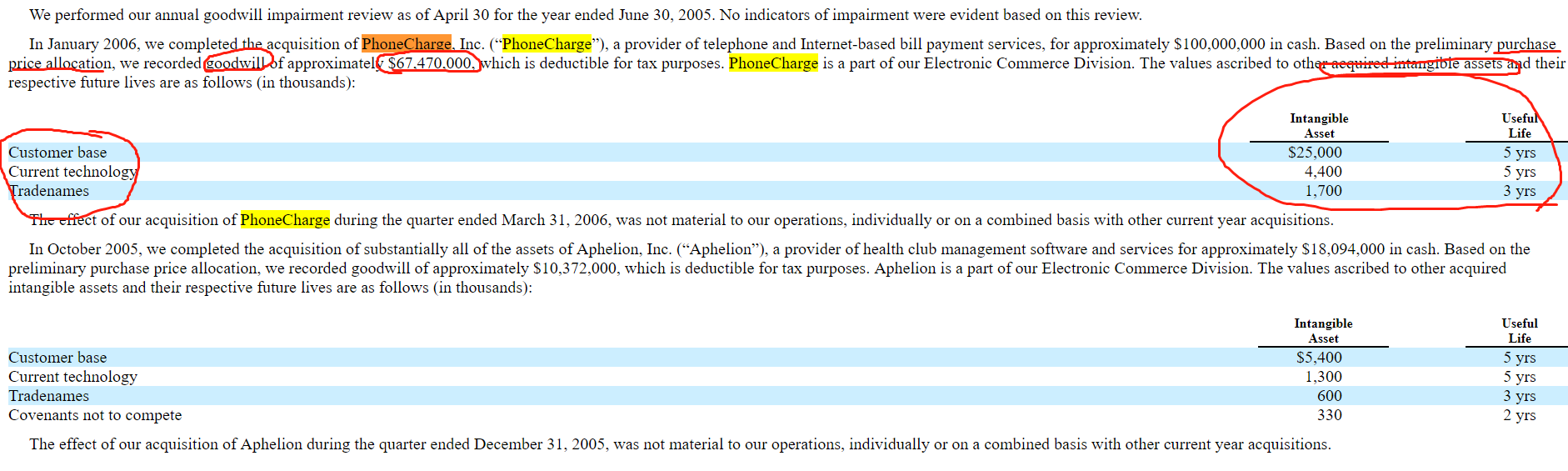
2) 点开1/3/2006之后的每一个报表， 在每一个报表中寻找



3）点击“documents”后，打开每个报表，是这样的页面，一般选择第一个文件链接打开



4）当寻找到第一次出现关于targetname列中公司（PhoneCharge）以下信息时停止

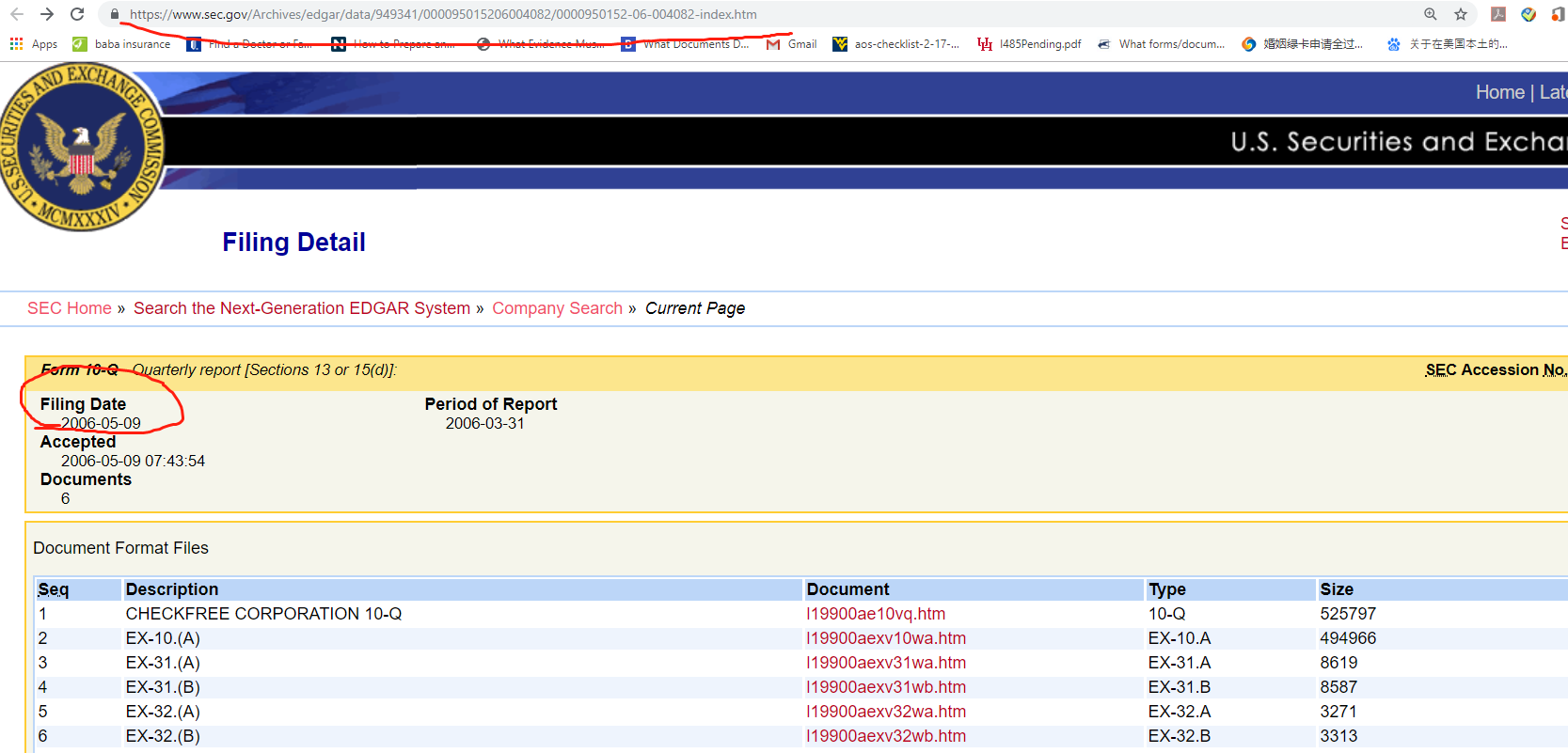


这一个例子中， 首先我们需要的是关于“phonecharge”公司的“purchase price allocation”信息。包括“goodwill”的“67,470,000”这个数字；

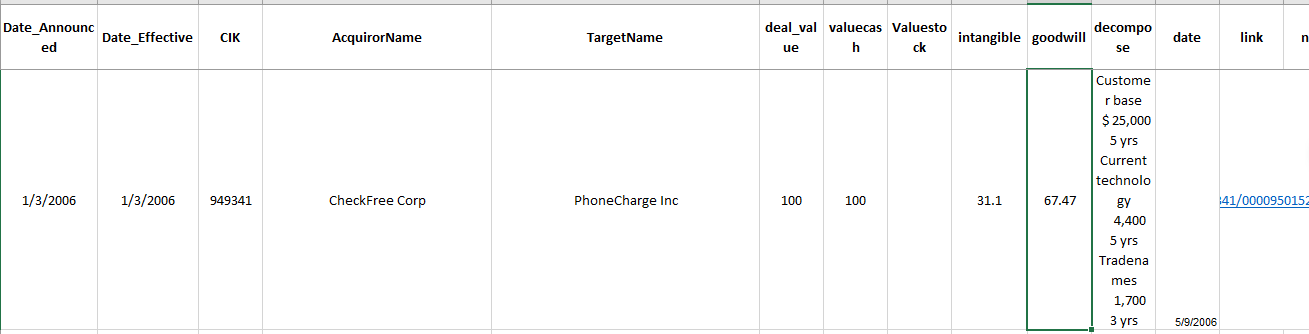
“acquired intangible assets”之下的这个表格；

表格中的数字加总（25,000+4,400+1,700）；

以及这个报表的时间，和链接（如下所示）



1. 存入excel表中：

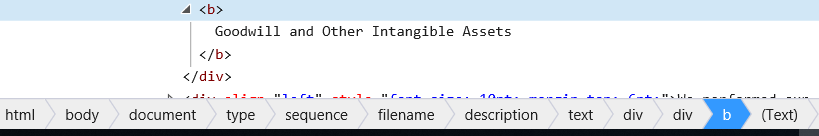


### HTML Parsing

It is well structed. Title is the first sub <div>, text is the next <div>, then table in the subsequent <div>. the information about next company follows at another <div>…

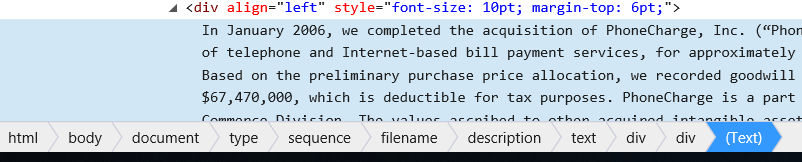
Title

div/div1/b



Text

div/div2



Table

div/div3/….



Another acquisition of C Company

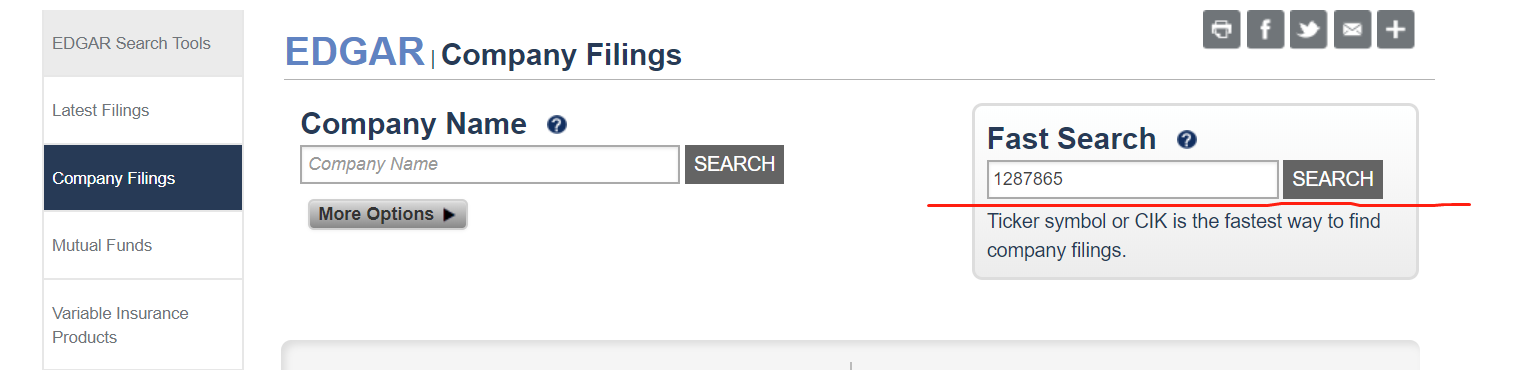
div/div4 …

## Target:“ Sherman” cik:1287865

### Manual flow

第二个例子：

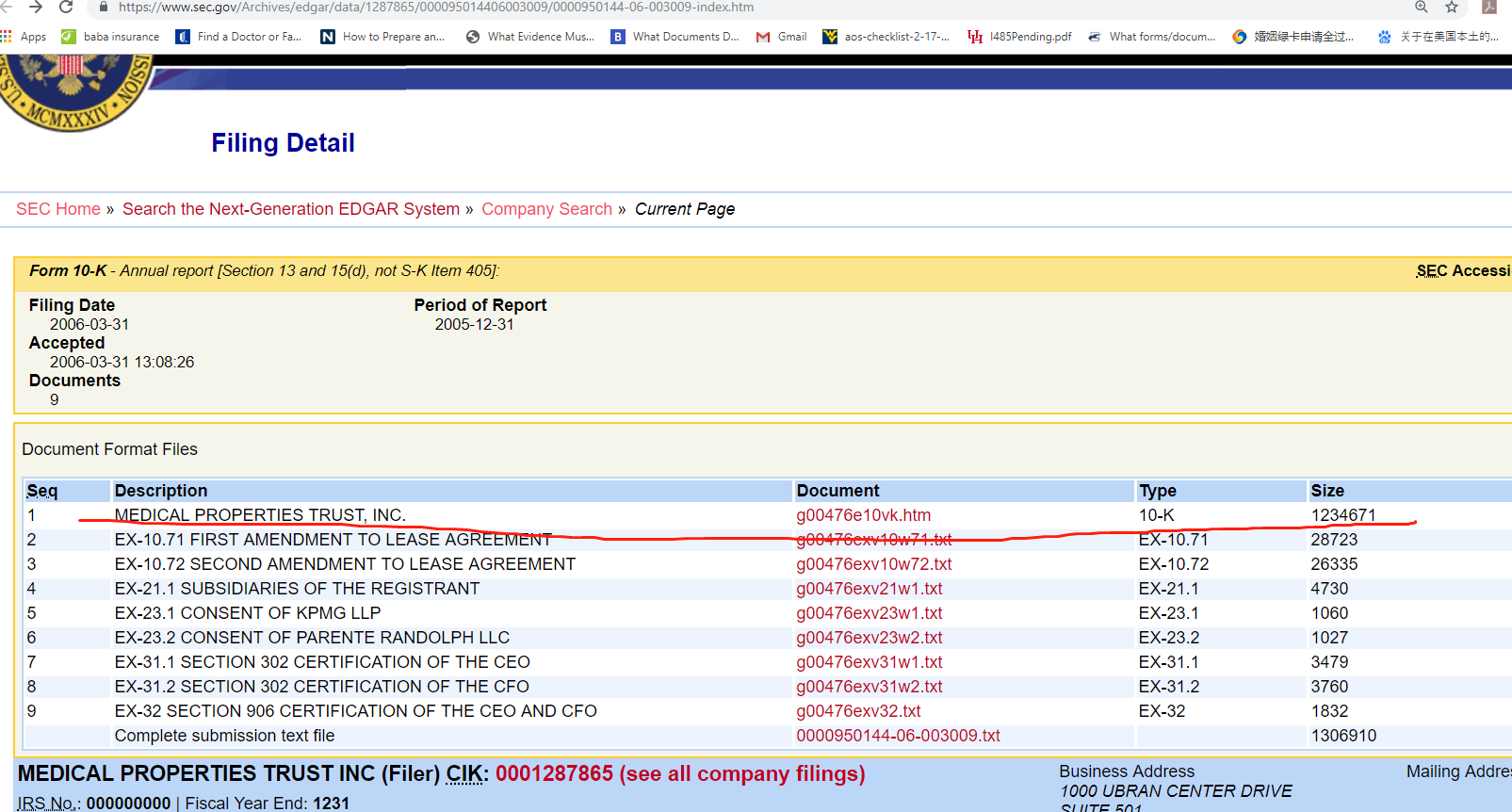
1. 输入cik列 1287865



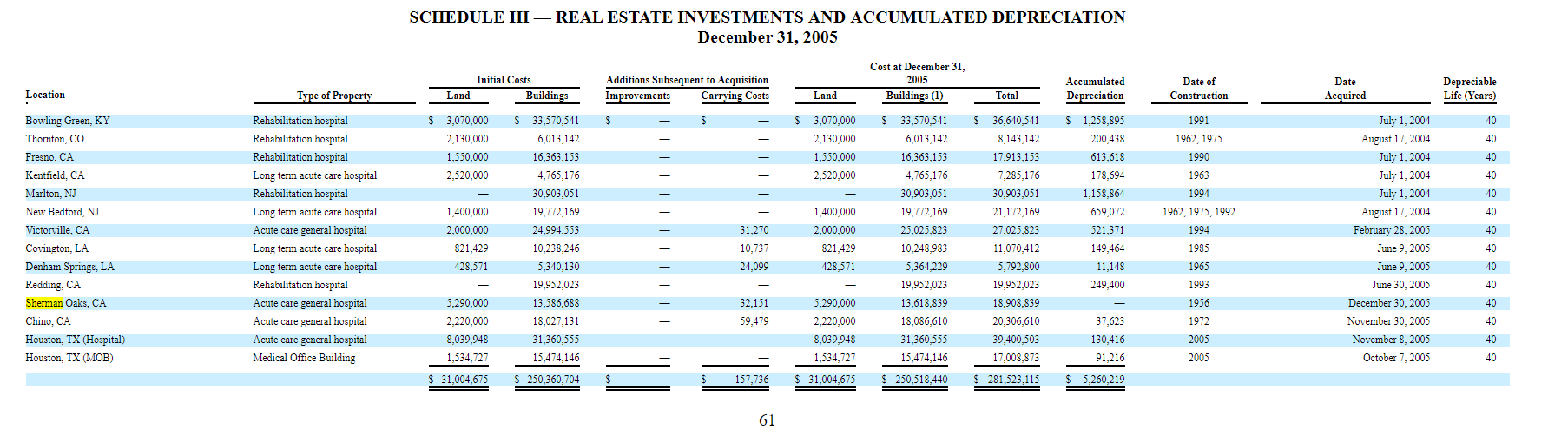
1. 寻找date\_announced列日期之前所有报表 （1/3/2006）



1. 每个报表打开后，类似这样，打开第一个document

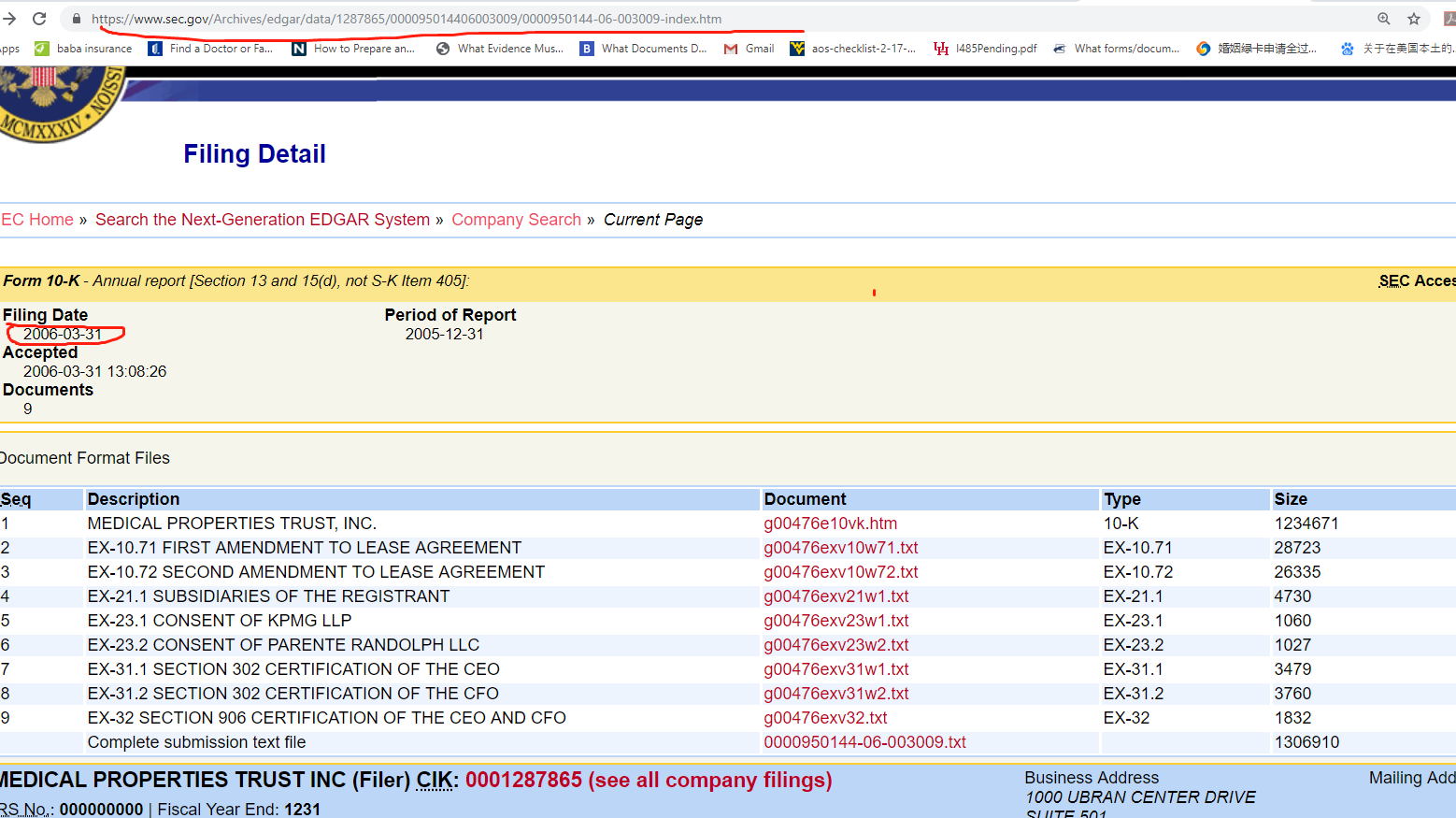


1. 找到如下信息关于targetname列名字 （Sherman）

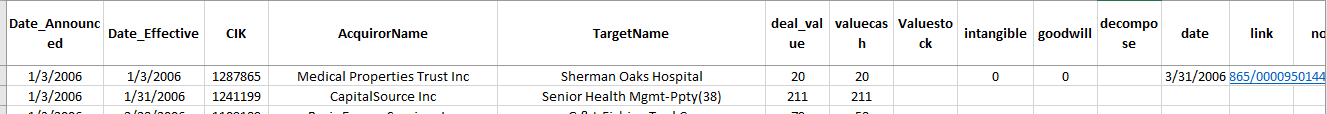
这里是由关键字“initial costs”“acquisition”“date acquired”和“sherman”定位到此表

然后发现“goodwill”和“intangible assets”未出现，所以都为0，填入excel表中。

并且将此表日期和链接填入excel



1. 结果如下

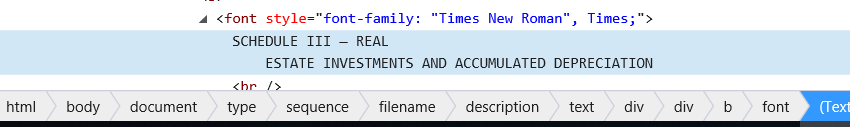


### HTML Parsing

Title

Div/div1/b

Bold title div1 is the one before the table!

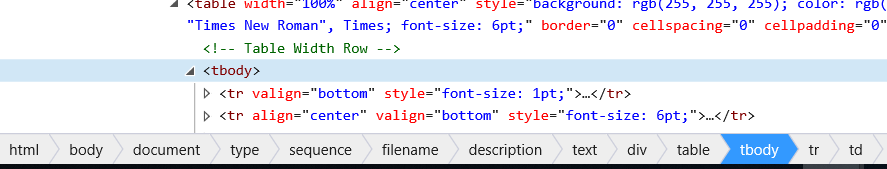


Text

None

Table

Div/table/…



More Company?

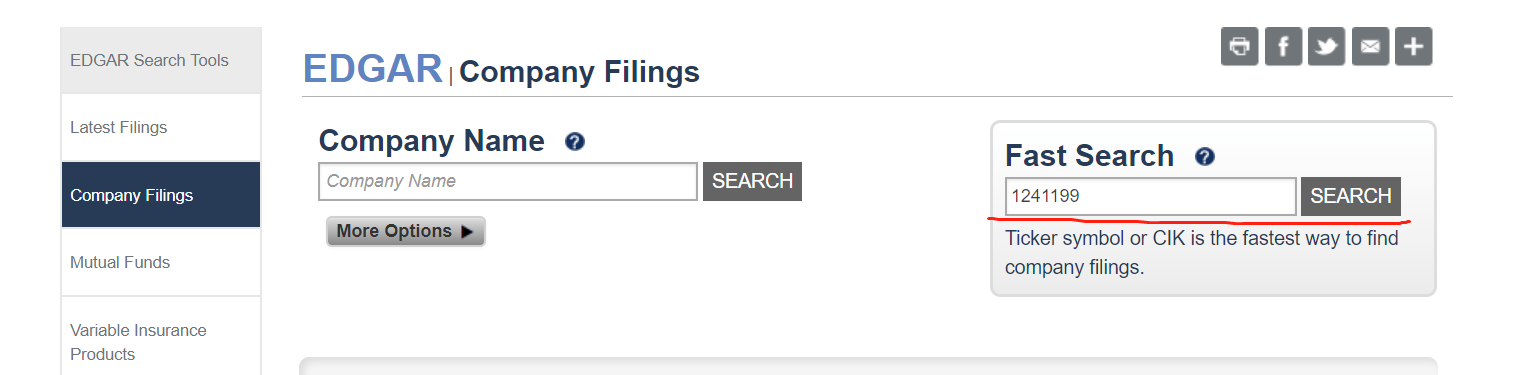
No, no more sub div.

## Target:“Senior Health” cik: 1241199

### Manual Flow

另一个例子：

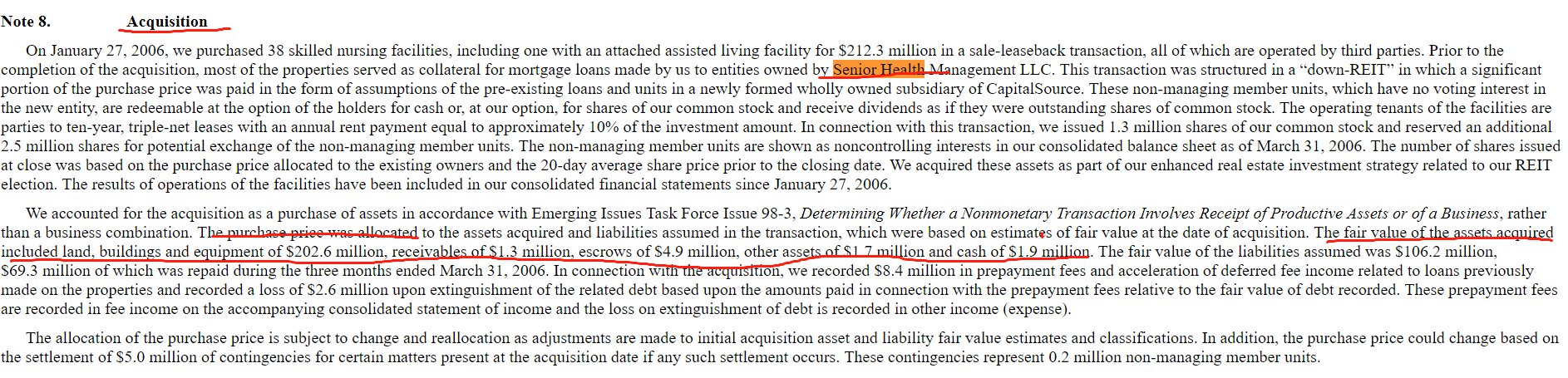
1. Cik为1241199



1. 1/3/2006 之前的报表



1. 寻找到关于targetname列“Senior Health”有关的



定位依据“acquisition”“senior health”（这个信息从excel表中的targetname列得到）“purchase price was allocated to”

1.）发现goodwill和intangible assets在此处未出现，所以填入0在excel中

2） 然后摘抄 The fair value of the assets acquired included land, buildings and equipment of $202.6 million, receivables of $1.3 million, escrows of $4.9 million, other assets of $1.7 million and cash of $1.9 million.

这句话在excel里的decompose列

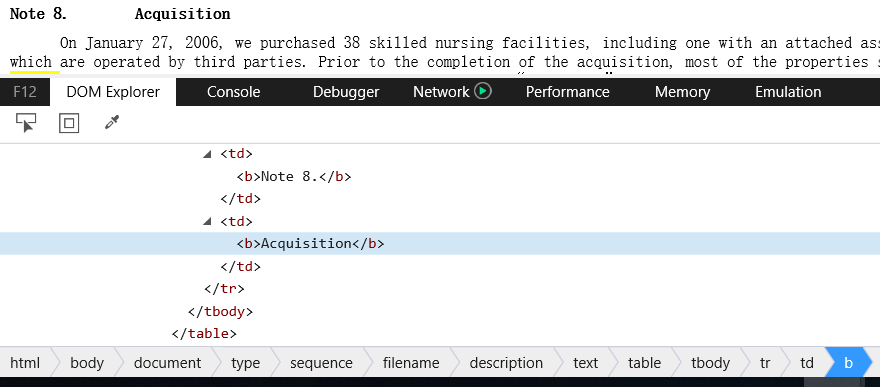
1. 然后摘抄这个报表时间和链接



### HTML Parsing

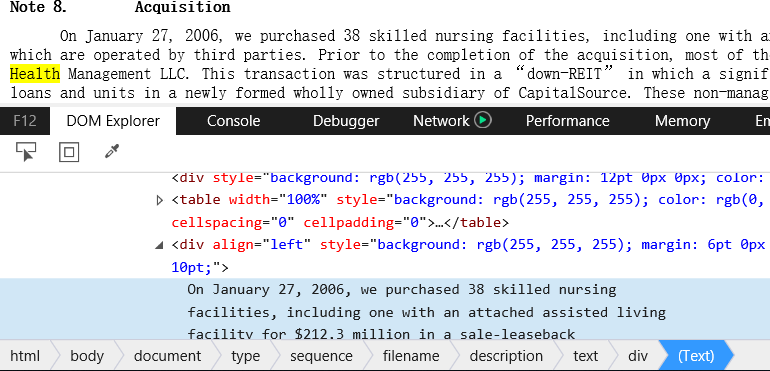
Title

Title is a table item!! We can ignore the title and search keywords in the text instead…



Text

Text’s <div\> is next to the title’s <table\>



Table

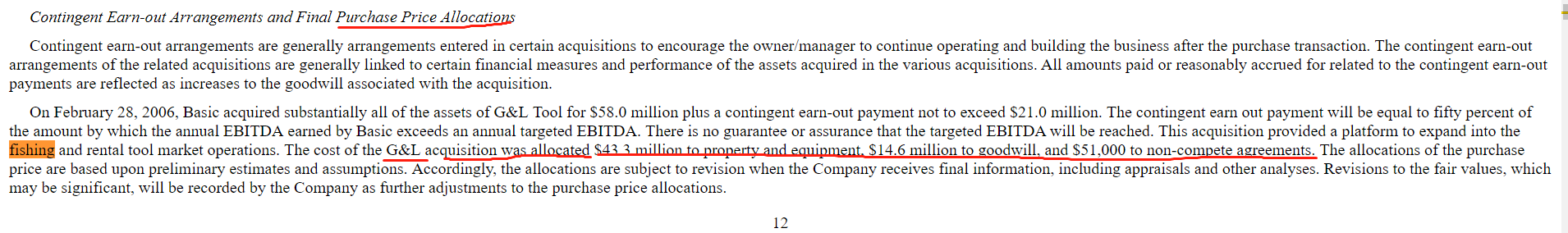
No data table at all!

## Target:“G&L” cik: 1109189

### Manual Flow

另一个例子

1. Cik=1109189
2. 找到报表出现这样的信息：



定位信息“G&L”（来自excel 的targetname列），“purchase price allocation”“acquisition”“was allocated”

其中goodwill为14.6， intangible assets为51

将The cost of the G&L acquisition was allocated $43.3 million to property and equipment, $14.6 million to goodwill, and $51,000 to non-compete agreements.

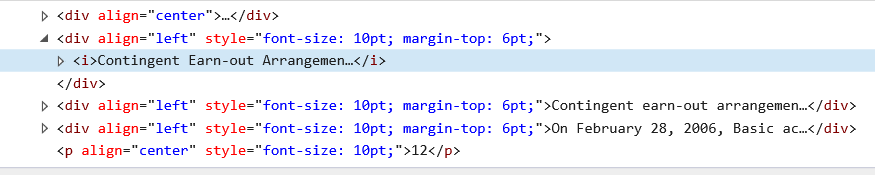
摘入到excel的decomposezhong

### HTML Parsing

It is also well structed, the italic title is the first <div>, texts are in subsequent <div>, no corresponding table and the text happens to be the last <div>.

Title: *…Purchase Price Allocations*

Italic instead of bold in the sub <div>



Text:

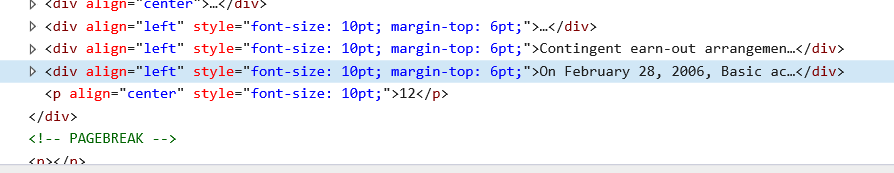


Table:

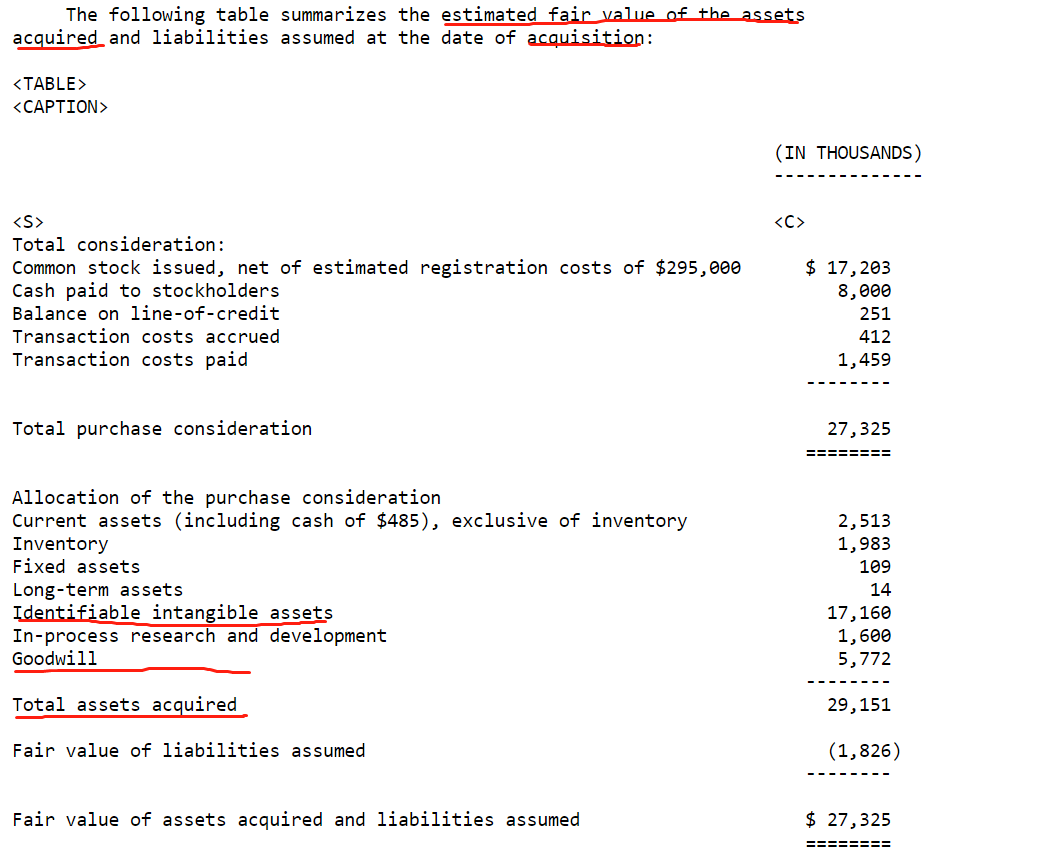
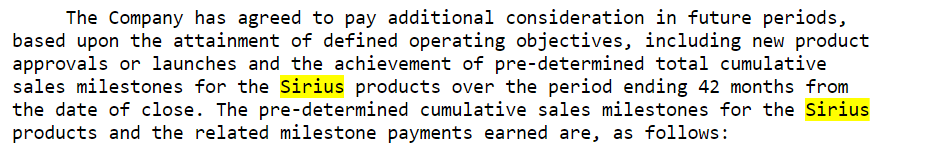
None. The text happens to be the last <div> in parent div.

## Target:“Sirius” cik: 879993

### Manual Flow

另一个例子

Cik=879993



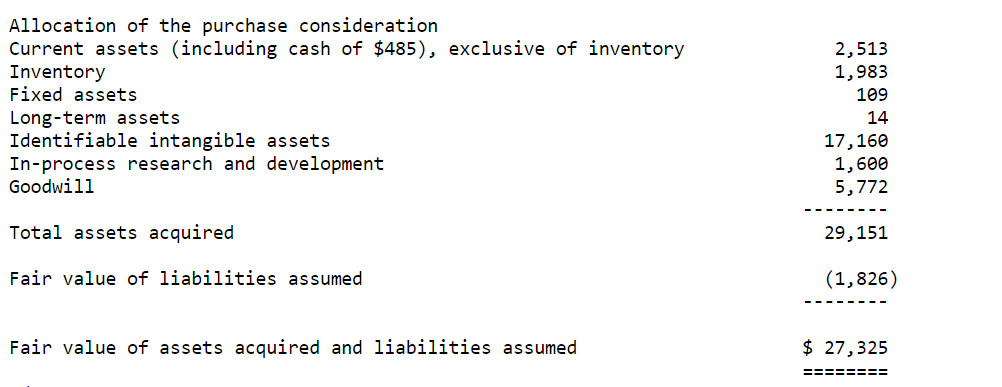
定位信息“sirius”（来自excel的targetname列）“estimated fair value of the assets

Acquired”“intangible assets”“total assets acquired”

所以goodwill是5,772

Intangible 是17，160

将以下表摘入excel的decompose中

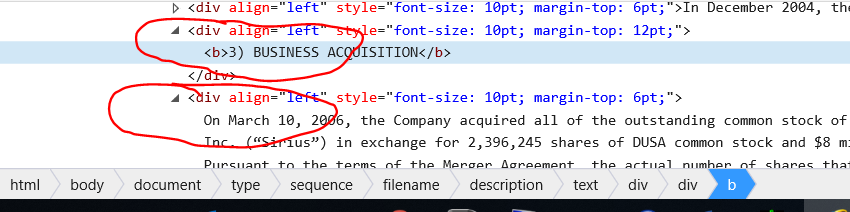


### HTML Parsing

20061106\_y26727e10vq.htm:

Title and Text

Well structed…



Table

Well structed, can stop at next bold <div>

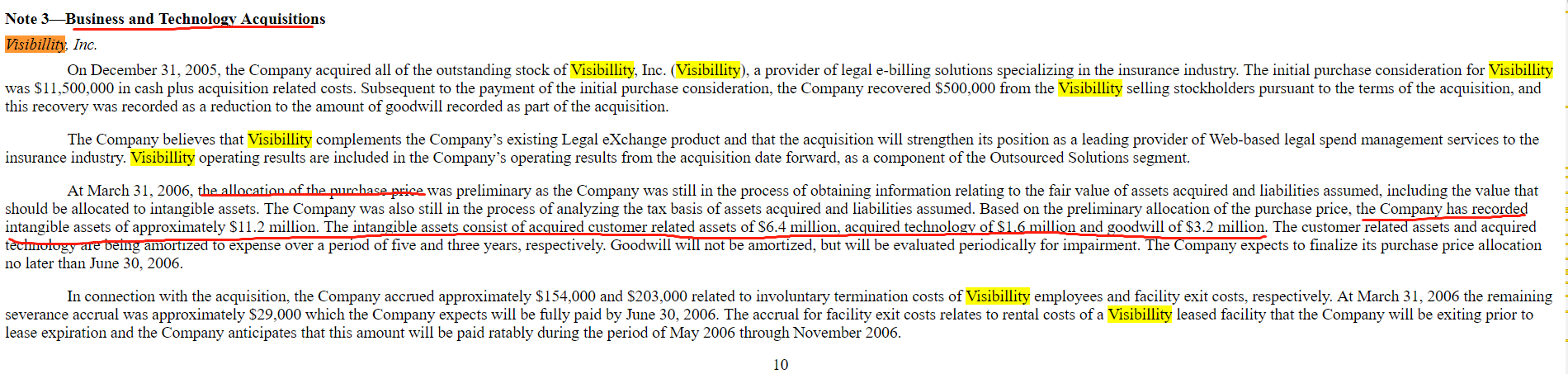


## Target:“Visibility” cik: 1073349

### Manual Flow

另一个例子

Cik=1073349



定位信息“visibility”从excel中的targetname中得到，“business and acquisitions”“the allocation of the purchase price”

所以intangible assets 是11.2， goodwill是3.2

然后把这句话“The intangible assets consist of acquired customer related assets of $6.4 million, acquired technology of $1.6 million and goodwill of $3.2 million.”也摘下来

### HTML Parsing

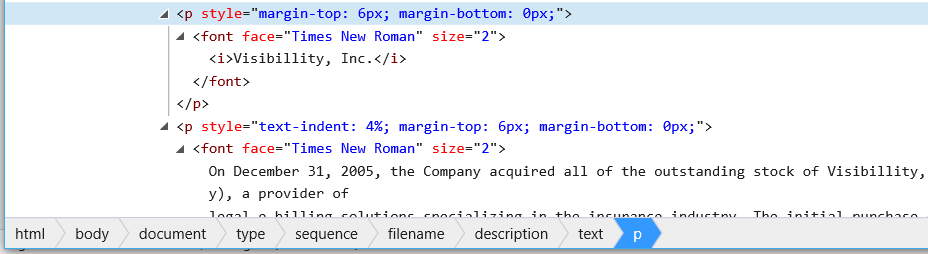
Golden: 20060509\_d10q.htm

Title:

/<text>/<p>/<b>

Text:

/<text>/<p>

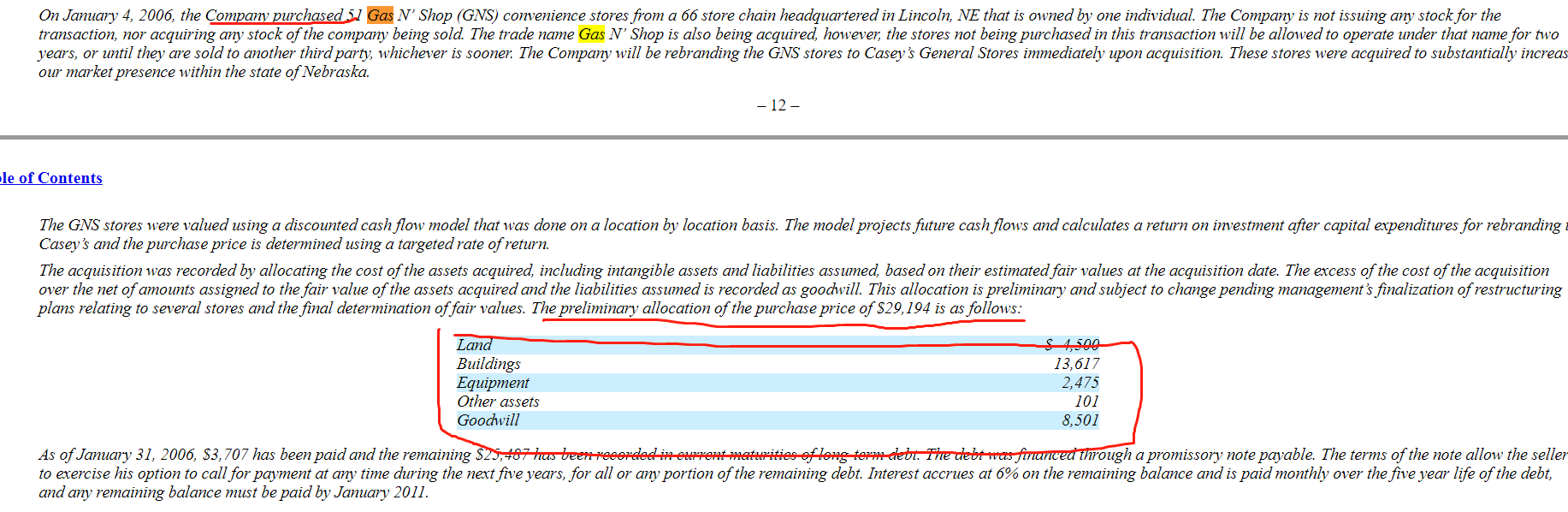


## Target:“Gas”cik: 726958

### Manual Flow

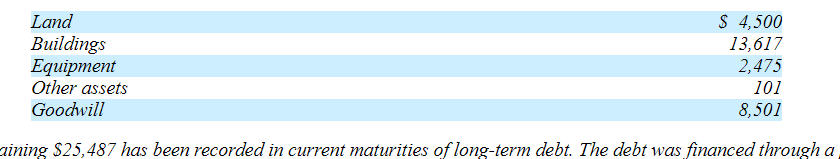
另一个例子

Cik=726958



定位信息“Gas”来自excel，“purchased 。。。 Gas”“the allocation of the purchase price”

并且需要摘下这个表



### HTML Parsing

C:\nobackup\Research\Python\edgar\726958\20060313\_d10q.htm

