如何免费获得百度文库的收费文档

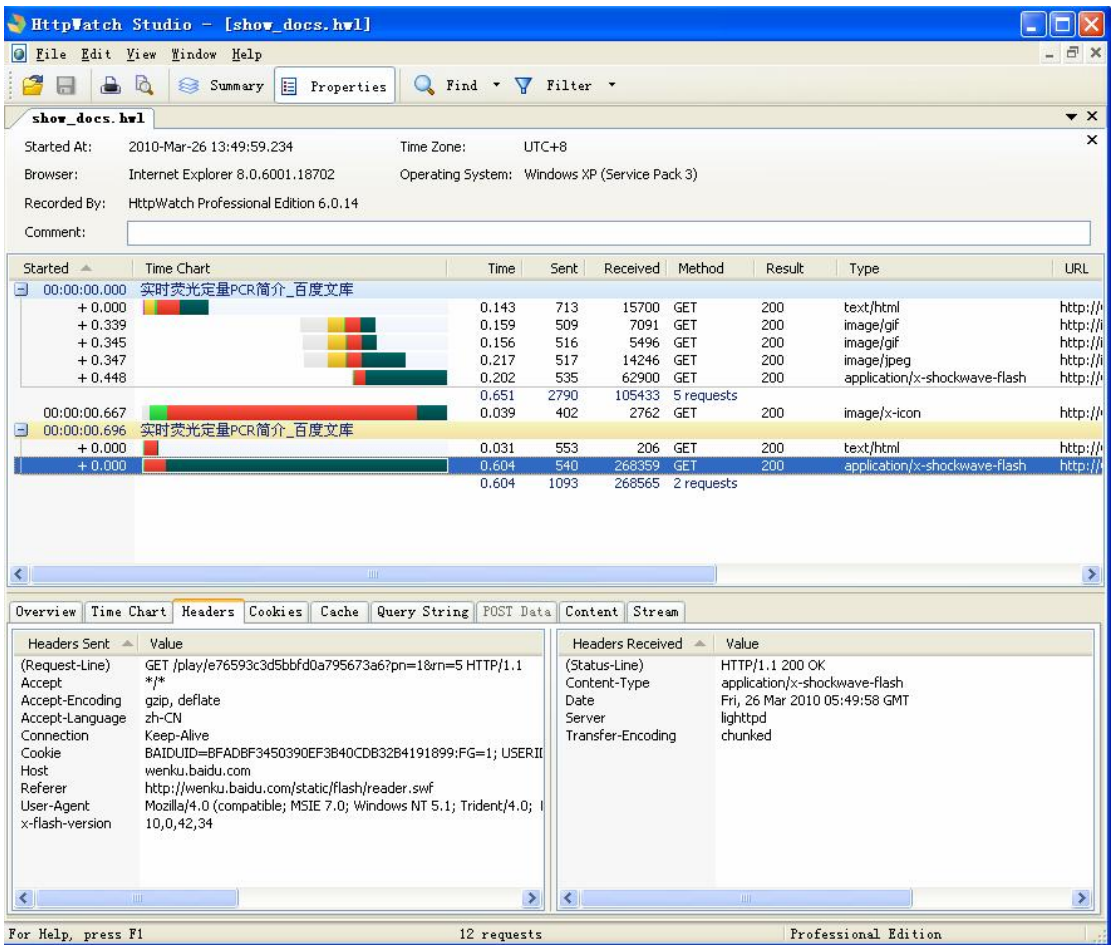
http://wenku.baidu.com
新注册一个用户的的积分是 20。

所以，要想在上面下载许多文档，是不现实的。

对于收费的文档，想想缓存里应该有的吧。
找了，没找到。

感觉挺神奇的，用 flash/flex 显示 word 内容。

用 httpwatch 看一下。

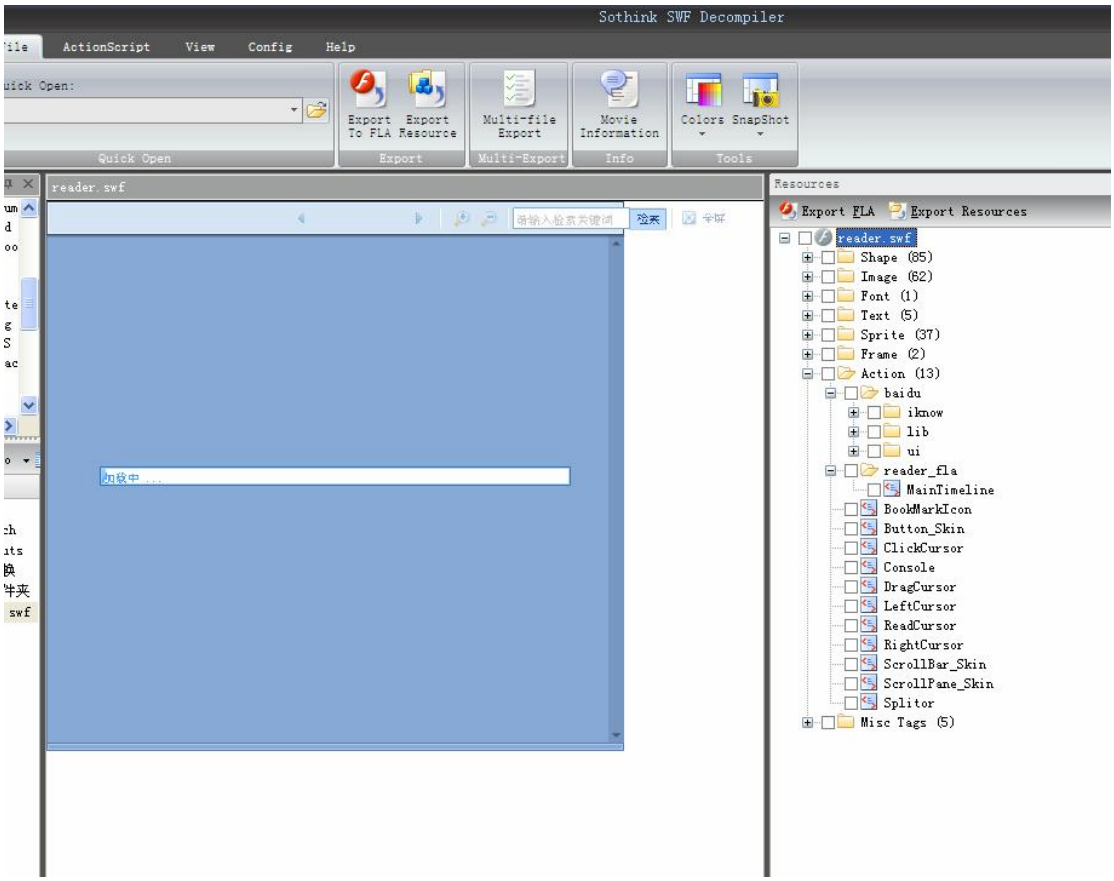


链接：
http://wenku.baidu.com/view/e76593c3d5bbfd0a795673a6.html

找到 2 个跟 flash 相关的链接
http://wenku.baidu.com/static/flash/reader.swf
http://wenku.baidu.com/play/e76593c3d5bbfd0a795673a6?pn=1&rn=5

第一个，是阅读器。第二个是文档内容

从阅读器开始吧。
用 Sothink SWF Decompiler 打开



跟显示的很像，应该就是这个了。

看一下 as 代码吧。

```
package reader_fla
{
	import flash.display.*;
	import flash.events.*;

	dynamic public class MainTimeline extends MovieClip
	{
```

这个类用来，处理进度条的。不管了。

```
package baidu.iknow
{
	import flash.display.*;
	import flash.events.*;

	public class main extends Sprite
	{
```

找到加载的主函数了。

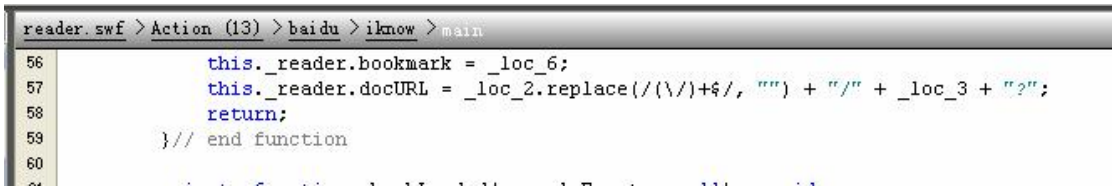
```
package baidu.iknow
{
	import flash.display.*;
	import flash.events.*;

	public class Reader extends Sprite
	{
		private var _firstPagesNum:int;
		private var _normalPageNum:int;
		private var _bookmark:String;
		private var _toolBar:ToolBar;
		private var _docURL:String;
		private const BG_COLOR:int = 8890837;
		private var _bg:Sprite;
		private const BORDER_COLOR:int = 5668272;
		private var _docViewer:DocViewer;
		private var _loading:Loading;
```

上面的读取类，明白了不？

_docURL，文档地址
BG_COLOR，背景色出来了

下面分析下 DocViewer，以及这个 url，怎么个展示法了。。。

在 DocViewer 中找到这样一段。。。正好证明了上面的第 2 个跟 flash 相关的链接



reader.swf > Action (13) > baidu > iknow > DocViewer

```
247             var tmpURL:String;
248             if (this._firstPagesNum == -1)
249             {
250                 tmpURL = this._docURL + "pn=" + (this._pagesLoaded + 1) + "&rn=" + this._r
251             }
252             else
253             {
254                 tmpURL = this._docURL + "pn=1&rn=" + this._firstPagesNum;
255                 this._firstPagesNum = -1;
256             }
257             var binaryRequest:* = new URLRequest(tmpURL);
258             try
259             {
```

（这样写，是为了防止文档页数越界，到最后一页，自动返回第一页）

http://wenku.baidu.com/play/e76593c3d5bbfd0a795673a6?pn=1&rn=5

上面的这个地址，就是由 main.as 和 DocViewer.as 两个类生成出来的。

e76593c3d5bbfd0a795673a6，文档编号
pn=1，已经加载了第 1 页，当前显示的是第 1 页
rn=5，一共 5 页

注意，pn<=5

读下来，怎么处理的呢？才成功显示成我们需要的文档的呢？

```
var binaryLoader:* = new URLLoader();
binaryLoader.dataFormat = "binary";
binaryLoader.addEventListener(ProgressEvent.PROGRESS, this.binaryLoading);
binaryLoader.addEventListener(Event.COMPLETE, this.binaryLoadComplete);
binaryLoader.addEventListener(IOErrorEvent.IO_ERROR, this.binaryLoadError);
数据处理

var binaryRequest:* = new URLRequest(tmpURL);
读取数据

binaryLoader.load(binaryRequest);
if (this._showLoading)
{
    dispatchEvent(new Event("SHOW_LOADING", true));
}
this._inLoading = true;
```

用 binaryLoader 加载读过来的数据，并设置，正在加载的提示状态。

下面，主要看 binaryLoader 的几个监听事件了。

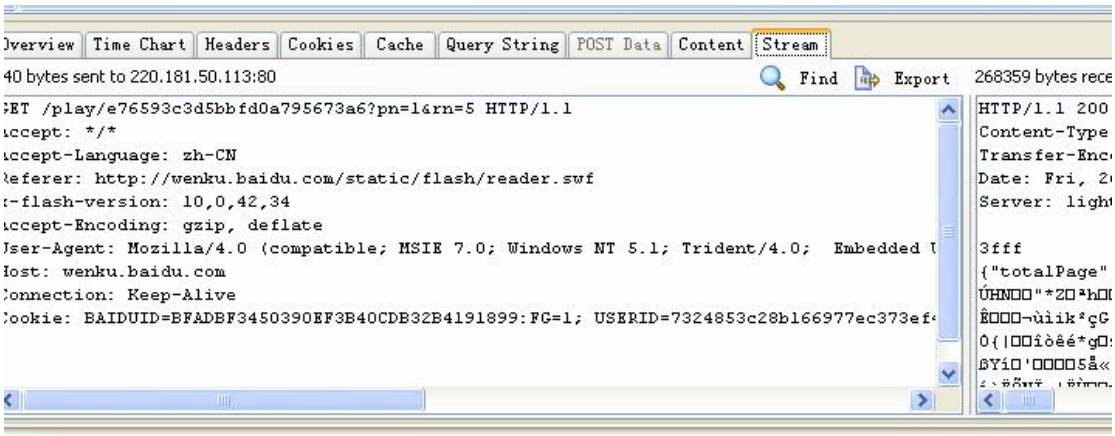
ProgressEvent.PROGRESS, this.binaryLoading
正在加载的处理事件
Event.COMPLETE, this.binaryLoadComplete
加载完成的处理事件
IOErrorEvent.IO_ERROR, this.binaryLoadError
加载出错的处理事件

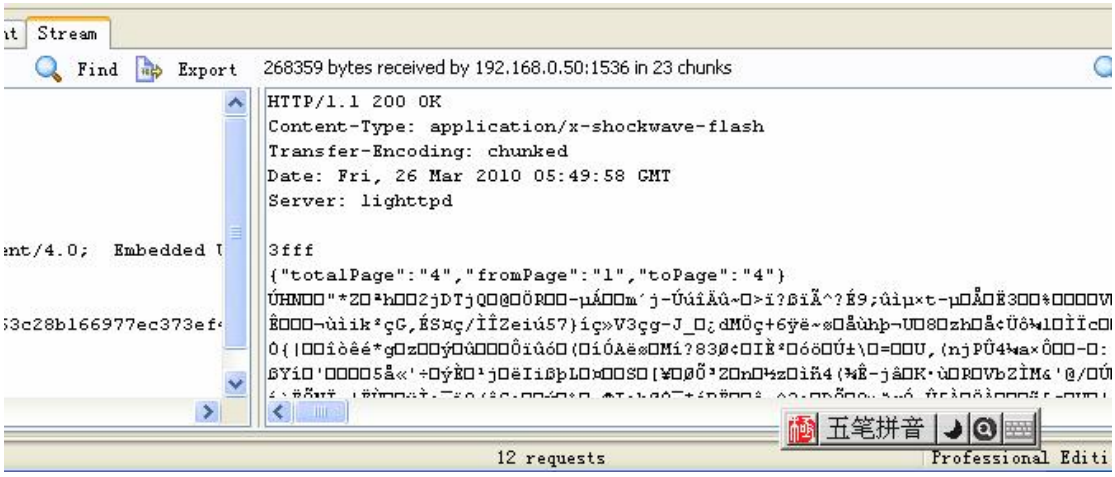
```
        private function binaryLoading(param1:ProgressEvent = null) : void
        {
            var _loc_2:* = param1.bytesTotal;
            var _loc_3:* = param1.bytesLoaded;
            this._loadPercent = _loc_3 / _loc_2;
            if (this._showLoading)
            {
                dispatchEvent(new Event("SHOW_LOADING", true));
            }
            return;
        }// end function
```

设置进度条，没什么好看的。

```
        private function binaryLoadComplete(param1:Event = null) : void
        {
            var _loc_11:ByteArray;
            var _loc_12:int;
            trace("binaryLoadComplete...");
            Console.log("binaryLoadComplete...");
            if (this._noDoc)
            {
                if (this._noDoc.parent)
                {
                    this._noDoc.parent.removeChild(this._noDoc);
                }
                this._noDoc = null;
            }
            this._loadPercent = 0;
            var _loc_2:* = URLLoader(param1.target);
            var _loc_3:* = _loc_2.data;
            this._byteArray = [];
            var _loc_4:Array;
            var _loc_5:int;
            var _loc_6:* = _loc_3.length;
            while (_loc_5 < _loc_6)
            {
```

有点长，下面还有。。不拿过来了。。
这里就是处理下载的内容的。
看一下，下载的内容什么样。。先。。在 httpwatch 中看一下，content 和 stream

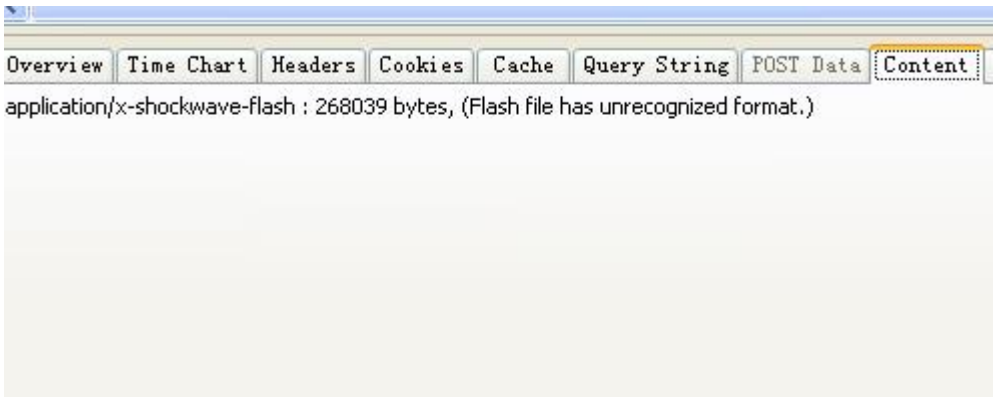




Server:Lighttpd，好熟悉。。先不谈这个，均衡用的。

{"totalPage":"4","fromPage":"1","toPage":"4"} .页数出来了

下面应该是内容。



Content 里显示不是可显示的 flash 格式，肯定不是了。(268039bytes)，不用说了，这么大 200 多 KB，差不多了，写到这，我看了一下，我写的这个文档大小。

哈哈 ,, 有可能，是把文档内容直接给下载完了哦。。
不说了。。
直接分析下载结束后的处理代码。

代码，还是贴一下吧。。好写注释。

```
private function binaryLoadComplete(param1:Event = null) : void
        {
            var _loc_11:ByteArray;
            var _loc_12:int;
            trace("binaryLoadComplete...");
            Console.log("binaryLoadComplete...");
            if (this._noDoc)
            {
                if (this._noDoc.parent)
                {
                    this._noDoc.parent.removeChild(this._noDoc);
                }
                this._noDoc = null;
            }
            this._loadPercent = 0;
            var _loc_2:* = URLLoader(param1.target);
            var _loc_3:* = _loc_2.data;
            this._byteArray = [];
            var _loc_4:Array;
            var _loc_5:int;
            var _loc_6:* = _loc_3.length;
            while (_loc_5 < _loc_6)
            {
```

```
                    if (_loc_5 + 3 < _loc_6)
                    {
                        if (_loc_3[_loc_5] == 67 || _loc_3[_loc_5] == 70 && _loc_3[_loc_5 + 1] == 87 && _loc_3[_loc_5 + 2] == 83 && _loc_3[_loc_5 + 3] == 9 || _loc_3[_loc_5 + 3] == 10)
                        {
                            _loc_4.push(_loc_5);
                        }
                    }
                    else
                    {
                        _loc_4.push(_loc_6);
                        break;
                    }
                    _loc_5++;
                }
                var _loc_7:* = _loc_3.readMultiByte(_loc_4[0], "utf-8");
                trace(_loc_7);
                Console.log(_loc_7);
                var _loc_8:* = JSON.decode(_loc_7);
                this._pagesAll = Number(_loc_8["totalPage"]);
                if (!this._pagesLoaded)
                {
                    this._pagesLoaded = 0;
                }
                this._pagesLoaded = this._pagesLoaded + (Number(_loc_8["toPage"]) - Number(_loc_8["fromPage"]) + 1);
                this._fromPage = Number(_loc_8["fromPage"]);
                this._toPage = Number(_loc_8["toPage"]);
                trace("pagesall:" + this._pagesAll + " , frompage:" + this._fromPage + " , topage:" + this._toPage);
                Console.log("pagesall:" + this._pagesAll + " , frompage:" + this._fromPage + " , topage:" + this._toPage);
                this._pagethLoading = Number(_loc_8["fromPage"])--;
                var _loc_9:int;
                while (_loc_9 < _loc_4.length--)
                {

                    _loc_11 = new ByteArray();
                    _loc_12 = _loc_4[_loc_9 + 1] - _loc_4[_loc_9];
                    _loc_3.readBytes(_loc_11, 0, _loc_12);
                    this._byteArray.push(_loc_11);
                    _loc_9++;
```

```
                }
                trace("..............这一次加载了多少页：" + this._byteArray.length);
                Console.log("..............这一次加载了多少页：" + this._byteArray.length);
                this._hasConvertPages = 0;
                this._allConvertPages = Math.min(this._byteArray.length, this._toPage -
this._fromPage + 1);
                if (this._allConvertPages > 0)
                {
                        this.byteArr2DisplayObj(this._hasConvertPages);
                }
                else
                {
                        trace("blank document ...");
                        this.processNoDoc();
                        this._loadPercent = 0;
                        dispatchEvent(new Event("STOP_LOADING", true));
                        this._inLoading = false;
                }
                return;
        }// end function
```

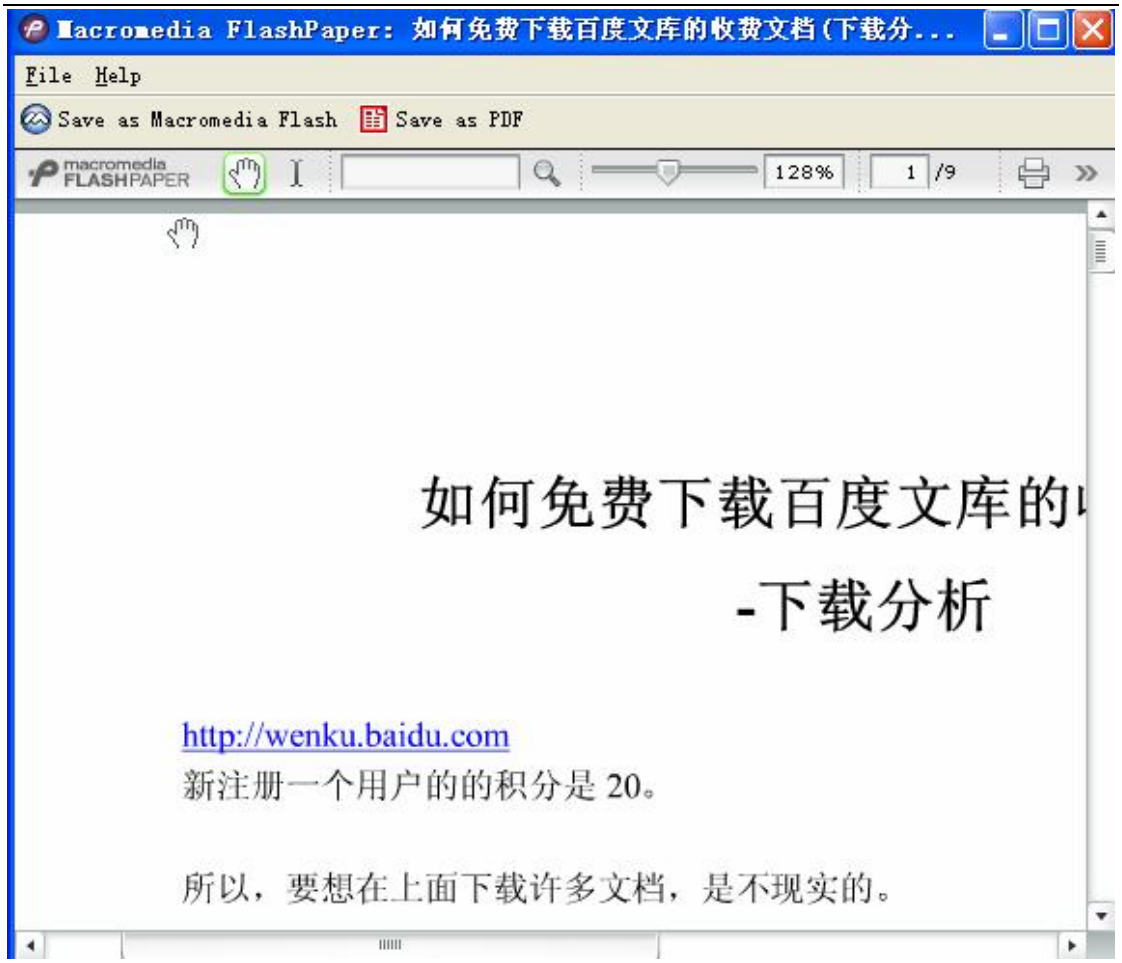
先留着，哪位有兴趣，写个程序哦？

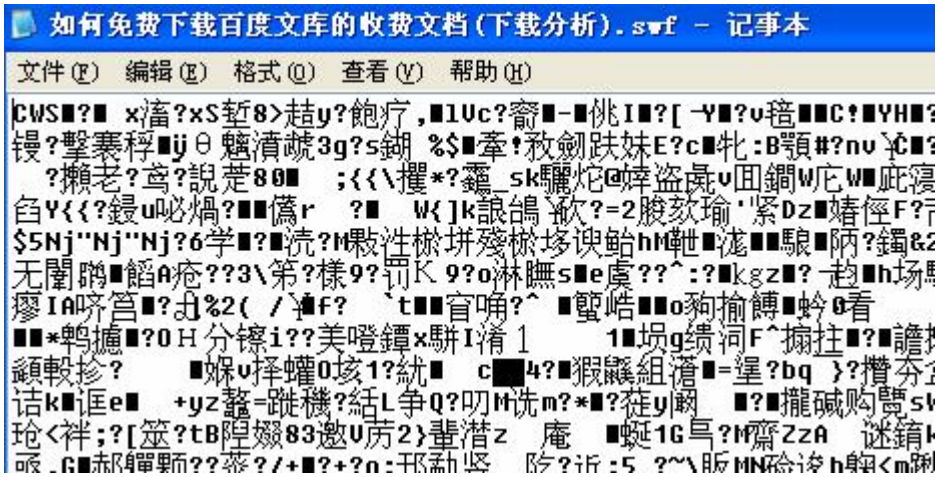
提示：
模拟请求后，打开这个页，将内容，下载。
Httpclient+ selenium 或 htmlutil，可以完成收费文档下载了。。。
别忘了，互相转换 cookies，session 什么的。。待续。。。

Mail:
gaoqs1984@gmail.com

上官伯阳
2010-3-26

下载 flashpaper，自己转换一下文档成 swf





说明了，是由 flashpaper 做成的 swf，再显示的