website-comapre

stardiviner

2018 年 11 月 23 日

目录

1 INPROGRESS website-compare [37/51]

DEADLINE: *<2018-09-26 Wed>* **SCHEDULED:** *<2018-09-22 Sat>*
website-compare/project.clj

2 新旧网站的基本信息

- 新网站 `http://www.sxszjzx.com`, `http://zjzx.sxsedu.net`
- 旧网站 `http://www.sxti.zj.cn`

内网解析 `http://www.sxszjzx.com/` 是解析到旧网站的。外网解析 `http://www.sxszjzx.com/` 是解析到新网站的。如果你解析到外网，你需要修改 DNS 到 172.18.0.5

2.1 TODO 外网无法访问旧网站

DEADLINE: *<2018-11-19 Mon>*
`http://bbs.360.cn/thread-6776749-1-1.html`

Modify the host file:

```
sudo echo "192.168.1.1 www.sxti.zj.cn" >> /etc/hosts
```

表 1: Clock summary at *[2018-11-22 Thu 21:30]*

Headline	Time			
Total time	**4d 8:53**			
写一个简单的爬虫加交叉对比的脚本来比较新旧网站的内容差别 [37/51]		4d 8:53		
代码架构			0:29	
website 新旧网站使用相同 API，不同的实现 [2/2]			3:55	
website-old.clj HTML elements				
website-new.clj HTML elements				
crawler.clj 获取所有文章的方法 [15/17]			1d 2:11	
benchmarking the HTTP URL request…				
按照网站的导航栏目录入口获取相应类目下的所有文章				1
old				
get navigator bar links				
parse sub-nav page to get all listed…				
iterate all result pages				
new				
get navigator bar links				
parse sub-nav page to get all listed…				
iterate all result pages				
benchmark new website two entry URL				
crawler.clj 爬虫存储提取连接的数据库 [5/5]			1:57	
Redis				
extract all links in nav bar				
save all links to Redis				
crawler.clj 爬取页面的内容 [3/6]			0:49	
scrape the article				
old				
new				
process the crawled article content…				
strip some special characters and spaces				
store.clj 爬虫内容存储到数据库 [3/4]			1:03	
SQLite				
save title and content to SQLite				
比较上传视频			0:50	
检查视频是否可以播放 [4/5]				

3 DONE 代码架构

CLOSED: *[2018-11-16 Fri 15:41]*

4 DONE website 新旧网站使用相同 API，不同的实现 [2/2]

CLOSED: *[2018-11-19 Mon 19:11]*
Use different namespace, same API function names.

4.1 DONE website-old.clj HTML elements

CLOSED: *[2018-11-19 Mon 19:05]*

4.1.1 nav

```clojure
(def selector-nav [:div.head_3 :ul#head_nav])
```

```html
<div class="head_3">
  <ul id="head_nav">

  <li class="nav_active"><a href="/"><span>首页</span></a></li>
  <li class=""><a href="javascript:void(0)"><span>学校概况</span></a>
    <ul>
      <li><a href="/html/school/about.html">学校简介</a></li>
      <li><a href="/e/action/ListInfo/?classid=13">现任领导</a></li>
      <li><a href="/e/action/ListInfo/?classid=283">名优教师</a></li>
      <li><a href="/html/school/office.html">处室联系</a></li>
      <li><a href="/e/action/ListInfo/?classid=14">校园风景</a></li>
      <li><a href="/e/action/ListInfo/?classid=35">楼层分布</a></li>
      <li><a href="/html/school/map.html">交通地图</a></li>
      <li><a href="/e/action/ListInfo/?classid=15">学校荣誉</a></li>
      <li><a href="/e/action/ListInfo/?classid=377">媒体关注</a></li>
    </ul>
  </li>
  <li class=""><a href="javascript:void(0)"><span>新闻频道</span></a>
    <ul>
      <li><a href="/e/action/ListInfo/?classid=33">校园新闻</a></li>
      <li><a href="/e/action/ListInfo/?classid=34">系部新闻</a></li>
      <li><a href="/e/action/ListInfo/?classid=7">对外交流</a></li>
      <li><a href="/e/action/ListInfo/?classid=8">学校荣誉</a></li>
      <li><a href="/e/action/ListInfo/?classid=9">师生荣誉</a></li>
      <li><a href="/e/action/ListInfo/?classid=10">校园视频</a></li>
      <li><a href="/e/action/ListInfo/?classid=11">二周安排</a></li>
      <li><a href="/e/action/ListInfo/?classid=12">每周寄语</a></li>
    </ul>
  </li>
  <li><a href="javascript:void(0)"><span>下载频道</span></a>
    <ul>
      <li><a href="/e/action/ListInfo/?classid=16">党政办</a></li>
      <li><a href="/e/action/ListInfo/?classid=17">教务处</a></li>
      <li><a href="/e/action/ListInfo/?classid=18">德育团委</a></li>
      <li><a href="/e/action/ListInfo/?classid=53">总务处</a></li>
      <li><a href="/e/action/ListInfo/?classid=54">实习处</a></li>
      <li><a href="/e/action/ListInfo/?classid=56">安保处</a></li>
      <li><a href="/e/action/ListInfo/?classid=175">教科室</a></li>
      <li><a href="/e/action/ListInfo/?classid=57">教学软件</a></li>
      <li><a href="/e/action/ListInfo/?classid=58">其它</a></li>
    </ul>
```

```
</li>
<li><a href="javascript:void(0)"><span>处室网站</span></a>
  <ul>
    <li><a href="/e/action/ListInfo/?classid=37">党政办</a></li>
    <li><a href="/e/action/ListInfo/?classid=38">教务处</a></li>
    <li><a href="/e/action/ListInfo/?classid=39">德育团委</a></li>
    <li><a href="/e/action/ListInfo/?classid=40">总务处</a></li>
    <li><a href="/e/action/ListInfo/?classid=41">实习处</a></li>
    <li><a href="/e/action/ListInfo/?classid=43">安保处</a></li>
    <li><a href="/e/action/ListInfo/?classid=104">教科室</a></li>
    <li><a href="/e/action/ListInfo/?classid=44">服务公司</a></li>
  </ul>
</li>
<li><a href="javascript:void(0)"><span>系部网站</span></a>
  <ul>
    <li><a href="/e/action/ListInfo/?classid=46">艺术设计系</a></li>
    <li><a href="/e/action/ListInfo/?classid=47">机械电子系</a></li>
    <li><a href="/e/action/ListInfo/?classid=49">财会信息系</a></li>
    <li><a href="/e/action/ListInfo/?classid=50">商贸旅游系</a></li>
    <li><a href="/e/action/ListInfo/?classid=51">建筑工程系</a></li>
    <li><a href="/e/action/ListInfo/?classid=52">新疆学部</a></li>
  </ul>
</li>
<li><a href="javascript:void(0)"><span style="color: #ff0000">招生宣传</span></a>
  <ul>
    <li><a href="/html/recruit/zsbm.html">招生报名</a></li>
    <li><a href="/html/recruit/plan.html">招生简章</a></li>
    <li><a href="/html/recruit/pro.html">专业介绍</a></li>
    <li><a href="/html/recruit/faq.html">热点问答</a></li>
    <li><a href="/html/recruit/xysh.html">校园生活</a></li>
    </ul>
  </li>
  <li><a href="javascript:void(0)"><span>校务公开</span></a>
    <ul>
      <li><a href="/html/public/org.html">组织架构</a></li>
      <li><a href="/e/action/ListInfo/?classid=19">办学规划</a></li>
      <li><a href="/e/action/ListInfo/?classid=20">管理制度</a></li>
      <li><a href="/e/action/ListInfo/?classid=21">阳光收费</a></li>
      <li><a href="/e/action/ListInfo/?classid=22">评职评优</a></li>
      <li><a href="/e/action/ListInfo/?classid=23">招标公告</a></li>
      <li><a href="/e/action/ListInfo/?classid=390">质量报告</a></li>
    </ul>
  </li>
  <li><a href="javascript:void(0)"><span>专题网站</span></a>
    <ul>
      <li><a href="/html/exemplary/about.html"><span>示范校专题网</span></a></li>
      <li><a href="/htmlhomepage/yiheliangyi/about.html">一核二翼专题网</a></li>
      <li><a href="/e/action/ListInfo/?classid=364">群众路线活动</a></li>
      <li><a href="/e/action/ListInfo/?classid=371">旅游职教集团</a></li>
      <li><a href="/e/action/ListInfo/?classid=39">德 育 品 牌</a></li>
      <li><a href="http://server2.sxszjzx.com/~jwc">精 品 课 程</a></li>
      <li><a href="/e/action/ListInfo/?classid=379">信 息 中 心</a></li>
    </ul>
  </li>
  <li><a target="_blank" href="/e/action/ListInfo/?classid=387"><span style="color: #ff0000">党建工作</span></a></li>

  </ul>
</div>
```

4.1.2 content

```
(def selector-content [:div.page_1])
```

1. sidebar

   http://www.sxti.zj.cn/html/school/about.html

   ```
   (def selector-sidebar [:div.page_left :div.pleft_t3])
   ```

   ```
   <div class="page_left">
     ..

     <div class="pleft_t2">
       <ul class="pleft_t3">

         <li><a href="/html/school/about.html">学校简介</a></li>
         <li><a href="/e/action/ListInfo/?classid=13">现任领导</a></li>
         <li><a href="/e/action/ListInfo/?classid=283">名优教师</a></li>
         <li><a href="/html/school/office.html">处室联系</a></li>
         <li><a href="/e/action/ListInfo/?classid=14">校园风景</a></li>
         <li><a href="/e/action/ListInfo/?classid=35">楼层分布</a></li>
         <li><a href="/html/school/map.html">交通地图</a></li>
         <li><a href="/e/action/ListInfo/?classid=15">学校荣誉</a></li>
         <li><a href="/e/action/ListInfo/?classid=377">媒体关注</a></li>

       </ul>
     </div>

   </div>
   ```

2. article

```clojure
(def selector-article [:div.page_right :div.pright_t3])
```

```html
<div class="page_right">
  <!-- title -->
  <div class="pright_t3">
    <!-- article -->
    <div class="pright_t4">

    </div>
  </div>
</div>
```

4.2 DONE website-new.clj HTML elements

CLOSED: *[2018-11-19 Mon 19:11]*

4.2.1 nav

`http://www.sxszjzx.com/`

```clojure
(def selector-nav [:div.nav])
```

```html
<div class="nav">
  <div class="siteWidth">

    <ul id="mainNav" class="mainNav">
```

```
<li class="li1 first1 on1" id="li-home">
  <h3 class="h1">
    <a class="a1" href="/">网站首页</a>
  </h3>
</li>

<li class="li1 hasUl1" id="li-xygk">
  <h3 class="h1">
    <a class="a1" href="/xygk/xyjj">学院概况</a></h3>
  <ul class="ul1" style="display: none;">

    <li class="li2 first2">
      <h3 class="h2"><a class="a2" href="/xygk/xyjj">学院简介</a></h3>
    </li>
    <li class="li2">
      <h3 class="h2"><a class="a2" href="/xygk/xrld">现任领导</a></h3>
    </li>
    <li class="li2">
      <h3 class="h2"><a class="a2" href="/xygk/zzjg">组织架构</a></h3>
    </li>
    <li class="li2">
      <h3 class="h2"><a class="a2" href="/xygk/cslx">处室联系</a></h3>
    </li>
    <li class="li2">
      <h3 class="h2"><a class="a2" href="/xygk/xyfg">校园风光</a></h3>
    </li>
    <li class="li2">
      <h3 class="h2"><a class="a2" href="/xygk/xyry">学院荣誉</a></h3>
    </li>
    <li class="li2">
      <h3 class="h2"><a class="a2" href="/xygk/lsyg">历史沿革</a></h3>
```

```
        </li>
        <li class="li2">
          <h3 class="h2"><a class="a2" href="/xygk/lcfb">楼层分布</a></h3>
        </li>
        <li class="li2 last2">
          <h3 class="h2"><a class="a2" href="/xygk/jtdt">交通地图</a></h3>
        </li>

      </ul>
    </li>
    <li class="li1 hasUl1" id="li-xydt">
      <h3 class="h1">
        <a class="a1" href="/xydt">学院动态</a></h3>
      <ul class="ul1" style="display: none;">

        <li class="li2 first2">
          <h3 class="h2"><a class="a2" href="/xydt/xyxw"
           ↪   target="_blank">学院新闻</a></h3>
        </li>
        <li class="li2">
          <h3 class="h2"><a class="a2" href="/xydt/xbxw">系部新闻</a></h3>
        </li>
        <li class="li2">
          <h3 class="h2"><a class="a2" href="/xydt/mtjj">媒体聚焦</a></h3>
        </li>
        <li class="li2">
          <h3 class="h2"><a class="a2" href="/xydt/xyry1">学院荣誉</a></h3>
        </li>
        <li class="li2">
          <h3 class="h2"><a class="a2" href="/xydt/jsry">教师荣誉</a></h3>
        </li>
        <li class="li2 last2">
          <h3 class="h2"><a class="a2" href="/xydt/xsry">学生荣誉</a></h3>
        </li>

      </ul>
    </li>
    <li class="li1 hasUl1" id="li-xbjs">
      <h3 class="h1">
        <a class="a1" href="/xbjs">系部建设</a></h3>
      <ul class="ul1" style="display: none;">

        <li class="li2 first2 hasUl2">
          <h3 class="h2"><a class="a2" href="/xbjs/yssjx">艺术设计系</a></h3>
        </li>
        <li class="li2 hasUl2">
          <h3 class="h2"><a class="a2" href="/xbjs/jxdzx">机械电子系</a></h3>
        </li>
        <li class="li2 hasUl2">
          <h3 class="h2"><a class="a2" href="/xbjs/chxxx">财会信息系</a></h3>
        </li>
        <li class="li2 hasUl2">
          <h3 class="h2"><a class="a2" href="/xbjs/smlyx">商贸旅游系</a></h3>
        </li>
        <li class="li2 hasUl2">
          <h3 class="h2"><a class="a2" href="/xbjs/jzgcx">建筑工程系</a></h3>
        </li>
        <li class="li2 last2 hasUl2">
          <h3 class="h2"><a class="a2" href="/xbjs/xjxb">新疆学部</a></h3>
        </li>

      </ul>
    </li>
    <li class="li1 hasUl1" id="li-zsjy">
      <h3 class="h1">
        <a class="a1" href="/zsjy/zsbm">招生就业</a></h3>
      <ul class="ul1" style="display: none;">

        <li class="li2 first2">
          <h3 class="h2"><a class="a2" href="/zsjy/zsbm">招生报名</a></h3>
        </li>
        <li class="li2">
          <h3 class="h2"><a class="a2" href="/zsjy/zsjz">招生简章</a></h3>
        </li>
        <li class="li2">
          <h3 class="h2"><a class="a2" href="/zsjy/zyjs6">专业介绍</a></h3>
        </li>
        <li class="li2">
          <h3 class="h2"><a class="a2" href="/zsjy/rdwd">热点问答</a></h3>
        </li>
        <li class="li2">
          <h3 class="h2"><a class="a2" href="/zsjy/xysh">校园生活</a></h3>
        </li>
        <li class="li2">
          <h3 class="h2"><a class="a2" href="/zsjy/zxbm">在线报名</a></h3>
        </li>
        <li class="li2 last2">
          <h3 class="h2"><a class="a2" href="/zsjy/jyxx">就业信息</a></h3>
        </li>

      </ul>
    </li>
    <li class="li1 hasUl1" id="li-ztlm">
      <h3 class="h1">
        <a class="a1" href="/ztlm">专题栏目</a></h3>
      <ul class="ul1" style="display: none;">

        <li class="li2 first2 hasUl2">
          <h3 class="h2"><a class="a2" href="/ztlm/sfxjs">示范校建设</a></h3>
        </li>
        <li class="li2 hasUl2">
          <h3 class="h2"><a class="a2" href="/ztlm/smgc">三名工程</a></h3>
        </li>
        <li class="li2 hasUl2">
          <h3 class="h2"><a class="a2" href="/ztlm/yheyzt">一核二翼专题</a></h3>
        </li>
        <li class="li2 hasUl2">
          <h3 class="h2"><a class="a2" href="/ztlm/qzlxhd">群众路线活动</a></h3>
        </li>
        <li class="li2 hasUl2">
          <h3 class="h2"><a class="a2" href="/ztlm/lyzjjt">旅游职教集团</a></h3>
        </li>
        <li class="li2 hasUl2">
          <h3 class="h2"><a class="a2" href="/ztlm/dypp">德育品牌</a></h3>
        </li>
```

```html
      <li class="li2 hasUl2">
        <h3 class="h2"><a class="a2" href="/ztlm/xysp">校园视频</a></h3>
      </li>
      <li class="li2 last2">
        <h3 class="h2"><a class="a2" href="http://server2.sxszjzx.com/~jwc/" target="_blank">精品课程</a></h3>
      </li>

    </ul>
  </li>
  <li class="li1 hasUl1" id="li-xxgk">
    <h3 class="h1">
      <a class="a1" href="/xxgk">校务公开</a></h3>
    <ul class="ul1" style="display: none;">

      <li class="li2 first2">
        <h3 class="h2"><a class="a2" href="/xxgk/bxgh">办学规划</a></h3>
      </li>
      <li class="li2">
        <h3 class="h2"><a class="a2" href="/xxgk/glzd">公示公告</a></h3>
      </li>
      <li class="li2">
        <h3 class="h2"><a class="a2" href="/xxgk/ygsf">阳光收费</a></h3>
      </li>
      <li class="li2">
        <h3 class="h2"><a class="a2" href="/xxgk/pzpy">评职评优</a></h3>
      </li>
      <li class="li2">
        <h3 class="h2"><a class="a2" href="/xxgk/zbgg">招标公告</a></h3>
      </li>
      <li class="li2">
        <h3 class="h2"><a class="a2" href="/xxgk/zlbg">质量报告</a></h3>
      </li>
      <li class="li2 last2">
        <h3 class="h2"><a class="a2" href="/xxgk/zyxz">资源下载</a></h3>
      </li>

    </ul>
  </li>
  <li class="li1 last1 hasUl1" id="li-djgz">
    <h3 class="h1">
      <a class="a1" href="/djgz">党建工作</a></h3>
    <ul class="ul1" style="display: none;">

      <li class="li2 first2">
        <h3 class="h2"><a class="a2" href="/djgz/djdt">党建动态</a></h3>
      </li>
      <li class="li2">
        <h3 class="h2"><a class="a2" href="/djgz/lzzl">廉政专栏</a></h3>
      </li>
      <li class="li2">
        <h3 class="h2"><a class="a2" href="/djgz/lqhd">亮旗行动</a></h3>
      </li>
      <li class="li2 last2">
        <h3 class="h2"><a class="a2" href="/djgz/xxzl">学习资料</a></h3>
      </li>

    </ul>
  </li>
</ul>

```

```
<script type="text/javascript">
  (function () {
  var navST;
  var navST1;
  var name = 'mainNav';
  var t = 200;
  var type = 1;
  var removeOn = 'False';
  var effect = 'slideDown';
  var appendItem = '#';
  var li = "#" + name + " li";
  var index = 0;

  if (!$("#" + name + " .li1").hasClass("on1")) {
  $("#" + name + " .li1").first().addClass("on1");
  } //默认第一个加.on1 类
  index = $("#" + name + " .li1").index($("#" + name + " .on1"));

  //鼠标离开导航后，回复默认.on1 类位置
  $("#" + name)
  .hover(
  function () {
  if (navST1 != null) {
  clearTimeout(navST1);
  }
  },
  function () {
  navST1 = setTimeout(function () {
  $("#" + name + " .li1").removeClass("on1").eq(index).addClass("on1");
  },
  500);
```

```
}
);

if (type == '1') {
li = "#" + name + " .li1";
}
if (appendItem != '#') { //插入内容
var appendHtml = $(appendItem).html();
$(li).first().append(appendHtml);
$(appendItem).remove();
}

if (type == '3') {
$("#" + name + " .on1").find("ul").first().show();
}

$(li)
.hover(function () {
var curItem = $(this);
var onNum = (curItem.attr("class").split(" "))[0].replace("li", "");
$(li).removeClass("on" + onNum);
curItem.addClass("on" + onNum);
navST = setTimeout(function () { //延时触发

if ($("ul:first", curItem).css("display") != "block") {
$(li + " .ul" + onNum).hide();
if (effect == 'fade') {
$("ul:first", curItem).fadeIn(t);
} else {
$("ul:first", curItem).slideDown(t);
}
```

```
    };
    navST = null;
    },
    t);
    },
    function () {
    if (navST != null) {
    clearTimeout(navST);
    }
    if (type == '1' || type == '2') {
    if (effect == 'fade') {
    $(this).find("ul").first().fadeOut(t);
    } else {
    $(this).find("ul").first().slideUp(t);
    }
    }
    if (removeOn == 'True') {
    $(this).removeClass("on1");
    }
    },
    t); //end hover
    })()
  </script>

</div>
</div>
```

4.2.2 content

```
(def selector-content [:div#content])
```

```
<div id="content">
  ....
</div>
```

1. sidebar

   `http://www.sxszjzx.com/xygk/xyjj`

```
(def selector-sidebar [:aside.side])
```

```
<aside class="side">

  <div id="sideMenu">
    <div class="hd">
      <h3>学院概况</h3>
    </div>
    <div class="bd">
      <ul class="menuList">
        <li class="on"><a href="/xygk/xyjj">学院简介</a></li>
        <li><a href="/xygk/xrld">现任领导</a></li>
        <li><a href="/xygk/zzjg">组织架构</a></li>
        <li><a href="/xygk/cslx">处室联系</a></li>
        <li><a href="/xygk/xyfg">校园风光</a></li>
        <li><a href="/xygk/xyry">学院荣誉</a></li>
        <li><a href="/xygk/lsyg">历史沿革</a></li>
        <li><a href="/xygk/lcfb">楼层分布</a></li>
        <li><a href="/xygk/jtdt">交通地图</a></li>
      </ul>
    </div>
  </div>

  <div id="sideRmph" class="sideBox">
    <div class="hd">
      <h3>热点资讯</h3>
    </div>
    <div class="bd">
      <ul class="infoListB">

        <li class="noData">暂无资料</li>
      </ul>
    </div>
  </div>

</aside>
```

2. article

`http://www.sxszjzx.com/xygk/xyjj`

```
(def selector-article [:div.mainContent])
```

```
<div class="mainContent">

  <div class="mHd">
    <div class="path">

      <em>您的位置：</em><a href="/">首页</a>
      &gt;<a href="/xygk/xyjj">学院概况</a>&gt;<a href="/xygk/xyjj">学院简介</a></div>
    <h3>学院简介</h3>
```

```
</div>
<div class="mBd">
  <!-- 正文内容 S -->
  <div class="articleCon">
    <div class="printArea" data-power-area="content">
      <!-- 标题 -->
      <h3 class="title">学院简介</h3>
      <div class="property">
        <span>【字体：<a href="javascript:;"
         ↪  data-power-command="reducefont">小</a> <a
         ↪  href="javascript:;"
         ↪  data-power-command="enlargefont">大</a>】</span>
      </div>
      <!-- 正文 -->
      <div class="conTxt" data-power-defaultfontsize="16"
       ↪  data-power-defaultlineheight="2"
       ↪  data-power-imgmaxwidth="800">
```

```
<div><strong style="padding: 0px; margin: 0px; outline:
↪ none; color: rgb(51, 51, 51); font-family:
↪ &quot;Microsoft Yahei&quot;; font-size: 14px;
↪ white-space: normal; background-color: rgb(255, 255,
↪ 255)">绍兴技师学院（筹）绍兴市职教中心</strong><span
↪ style="color: #333333; font-family: &quot;Microsoft
↪ Yahei&quot;; font-size: 14px; background-color:
↪ #FFFFFF">创办于</span><strong style="padding: 0px;
↪ margin: 0px; outline: none; color: rgb(51, 51, 51);
↪ font-family: &quot;Microsoft Yahei&quot;; font-size:
↪ 14px; white-space: normal; background-color: rgb(255,
↪ 255, 255)">1958</strong><span style="color: #333333;
↪ font-family: &quot;Microsoft Yahei&quot;; font-size:
↪ 14px; background-color: #FFFFFF">年，是以培养现代服务业和
↪ 先进制造业技能人才为主，集学历教育、职业培训、技能鉴定为一
↪ 体的综合性职业学校。</span><strong style="padding: 0px;
↪ margin: 0px; outline: none; color: rgb(51, 51, 51);
↪ font-family: &quot;Microsoft Yahei&quot;; font-size:
↪ 14px; white-space: normal; background-color: rgb(255,
↪ 255, 255)">1996</strong><span style="color: #333333;
↪ font-family: &quot;Microsoft Yahei&quot;; font-size:
↪ 14px; background-color: #FFFFFF">年被评为首批国家级重点职
↪ 业学校，</span><strong style="padding: 0px; margin: 0px;
↪ outline: none; color: rgb(51, 51, 51); font-family:
↪ &quot;Microsoft Yahei&quot;; font-size: 14px;
↪ white-space: normal; background-color: rgb(255, 255,
↪ 255)">2013</strong><span style="color: #333333;
↪ font-family: &quot;Microsoft Yahei&quot;; font-size:
↪ 14px; background-color: #FFFFFF">年被确定为首批国家中等职
↪ 业教育改革发展示范学校。</span><br style="padding: 0px;
↪ margin: 0px; outline: none; color: rgb(51, 51, 51);
↪ font-family: &quot;Microsoft Yahei&quot;; font-size:
↪ 14px; white-space: normal; background-color: rgb(255,
↪ 255, 255)"><span style="color: #333333; font-family:
↪ &quot;Microsoft Yahei&quot;; font-size: 14px;
↪ background-color: #FFFFFF">&nbsp;    学校实行
↪ “</span><strong style="padding: 0px; margin: 0px;
↪ outline: none; color: rgb(51, 51, 51); font-family:
↪ &quot;Microsoft Yahei&quot;; font-size: 14px;
↪ white-space: normal; background-color: rgb(255, 255
          </div>
        </div>
        <div class="userControl">


          <div class="bdsharebuttonbox"><a href="#" class="bds_more" data-cmd="more"></a><a href="#" class="bds_qzone" data-cmd="qzone" title="分享到QQ空间"></a><a href="#" class="bds_tsina" data-cmd="tsina" title="分享到新浪微博"></a><a href="#" class="bds_tqq" data-cmd="tqq" title="分享到腾讯微博"></a><a href="#" class="bds_renren" data-cmd="renren" title="分享到人人网"></a><a href="#" class="bds_weixin" data-cmd="weixin" title="分享到微信"></a></div>
          <script>
            window._bd_share_config = {
            "common": {
            "bdSnsKey": {},
            "bdText": "",
            "bdMini": "2",
            "bdMiniList": false,
            "bdPic": "",
            "bdStyle": "1",
            "bdSize": "24"
            },
            "share": {}
            };
            with (document) {
            0[(getElementsByTagName('head')[0] || body)
            .appendChild(createElement('script'))
            .src =
                ↪  'http://bdimg.share.baidu.com/static/api/js/share.js?v=89860593.js?cdnve
                ↪  +
            ~(-new Date() / 36e5)];
            }
          </script>
        </div>
      </div>
      <!-- 正文内容 E -->
    </div>

  </div>
```

5 TODO crawler.clj 获取所有文章的方法 [15/17]

5.1 DONE benchmarking the HTTP URL request speed [2/2]

CLOSED: *[2018-11-19 Mon 18:04]*

- ☒ record to Org Enlive.org
- ☒ optimize code in here

5.1.1 clj-http + enlive/html-snippet

```
(require '[clj-http.client :as http])
(require '[net.cgrand.enlive-html :as html])
(use 'criterium.core)

(pr (quick-bench
     (-> (http/get "https://www.baidu.com")
         :body
         html/html-snippet)))
```

```
Evaluation count : 6 in 6 samples of 1 calls.
             Execution time mean : 260.639936 ms
    Execution time std-deviation : 74.939971 ms
   Execution time lower quantile : 161.264308 ms ( 2.5%)
   Execution time upper quantile : 337.518891 ms (97.5%)
                   Overhead used : 36.702468 ns
nil
```

5.1.2 enlive/html-resource + URL

```
(require '[net.cgrand.enlive-html :as html])
(import 'java.net.URL)
(use 'criterium.core)

(pr (quick-bench
      (html/html-resource (URL. "https://www.baidu.com"))))
```

```
Evaluation count : 18 in 6 samples of 3 calls.
             Execution time mean : 64.137600 ms
    Execution time std-deviation : 13.476096 ms
   Execution time lower quantile : 49.122882 ms ( 2.5%)
   Execution time upper quantile : 79.881576 ms (97.5%)
                   Overhead used : 36.702468 ns
nil
```

5.2 DONE 按照网站的导航栏目录入口获取相应类目下的所有文章 [9/9]

CLOSED: *[2018-11-20 Tue 13:40]*

5.2.1 DONE old

CLOSED: *[2018-11-20 Tue 20:13]*

1. **DONE** get navigator bar links **CLOSED:** *[2018-11-20 Tue 10:50]*
   http://www.sxszjzx.com/html/school/about.html

```html
<body>
  <div class="page_all">
    <div class="head_1">
      <div class="head_2">
        <div class="head_4">
          <div class="page_1">
            <div class="page_left">
              <div class="page_right">
                <div class="pright_t3">
                  <div class="pright_t4">
```

```clojure
(def website-old-url "http://www.sxti.zj.cn")
(def website-old-html (get-html "http://www.sxti.zj.cn/"))

(defn get-html
  "Get HTML string as result."
  [url]
  (-> (http/get url {:as "GB2312"})
      :body
      html/html-snippet))

(defonce nav-bar
  (html/select
   (drop 1
         (first (map #(html/select % [:li])
                     ;; nav bar
                     (html/select
                      website-old-html
                      [:html :body :div.page_all :div.head_2
                      ↪   :div.head_3 :ul#head_nav]))))
   [:a]))

(defonce nav-bar-links-map
  (map #(let [link  (str website-old-url
                         ;; :attrs nil (:href does not exist)
                         (if (nil? (:attrs %))
                           nil
                           ;; :href "javascript:void(0)"
                           (if (= (first (html/attr-values % :href))
                           ↪   "javascript:void(0)")
                             nil
                             ;; :href "/..."
                             (first (html/attr-values % :href)))))
              title (html/text %)]
          {title link})
       nav-bar))

(pprint nav-bar-links-map)
```

```
({"学校概况" "http://www.sxti.zj.cn"}
 {"学校简介" "http://www.sxti.zj.cn/html/school/about.html"}
 {"现任领导" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=13"}
 {"名优教师" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=283"}
```

```
{"处室联系" "http://www.sxti.zj.cn/html/school/office.html"}
{"校园风景" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=14"}
{"楼层分布" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=35"}
{"交通地图" "http://www.sxti.zj.cn/html/school/map.html"}
{"学校荣誉" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=15"}
{"媒体关注" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=377"}
{"学校简介" "http://www.sxti.zj.cn/html/school/about.html"}
{"现任领导" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=13"}
{"名优教师" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=283"}
{"处室联系" "http://www.sxti.zj.cn/html/school/office.html"}
{"校园风景" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=14"}
{"楼层分布" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=35"}
{"交通地图" "http://www.sxti.zj.cn/html/school/map.html"}
{"学校荣誉" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=15"}
{"媒体关注" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=377"}
{"新闻频道" "http://www.sxti.zj.cn"}
{"校园新闻" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=33"}
{"系部新闻" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=34"}
{"对外交流" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=7"}
{"学校荣誉" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=8"}
{"师生荣誉" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=9"}
{"校园视频" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=10"}
{"二周安排" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=11"}
{"每周寄语" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=12"}
{"校园新闻" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=33"}
{"系部新闻" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=34"}
{"对外交流" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=7"}
{"学校荣誉" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=8"}
{"师生荣誉" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=9"}
{"校园视频" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=10"}
{"二周安排" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=11"}
```

```
{"每周寄语" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=12"}
{"下载频道" "http://www.sxti.zj.cn"}
{"党政办" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=16"}
{"教务处" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=17"}
{"德育团委" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=18"}
{"总务处" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=53"}
{"实习处" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=54"}
{"安保处" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=56"}
{"教科室" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=175"}
{"教学软件" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=57"}
{"其它" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=58"}
{"党政办" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=16"}
{"教务处" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=17"}
{"德育团委" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=18"}
{"总务处" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=53"}
{"实习处" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=54"}
{"安保处" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=56"}
{"教科室" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=175"}
{"教学软件" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=57"}
{"其它" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=58"}
{"处室网站" "http://www.sxti.zj.cn"}
{"党政办" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=37"}
{"教务处" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=38"}
{"德育团委" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=39"}
{"总务处" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=40"}
{"实习处" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=41"}
{"安保处" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=43"}
{"教科室" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=104"}
{"服务公司" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=44"}
{"党政办" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=37"}
{"教务处" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=38"}
```

```
{"德育团委" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=39"}
{"总务处" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=40"}
{"实习处" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=41"}
{"安保处" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=43"}
{"教科室" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=104"}
{"服务公司" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=44"}
{"系部网站" "http://www.sxti.zj.cn"}
{"艺术设计系" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=46"}
{"机械电子系" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=47"}
{"财会信息系" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=49"}
{"商贸旅游系" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=50"}
{"建筑工程系" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=51"}
{"新疆学部" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=52"}
{"艺术设计系" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=46"}
{"机械电子系" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=47"}
{"财会信息系" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=49"}
{"商贸旅游系" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=50"}
{"建筑工程系" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=51"}
{"新疆学部" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=52"}
{"招生宣传" "http://www.sxti.zj.cn"}
{"招生报名" "http://www.sxti.zj.cn/html/recruit/zsbm.html"}
{"招生简章" "http://www.sxti.zj.cn/html/recruit/plan.html"}
{"专业介绍" "http://www.sxti.zj.cn/html/recruit/pro.html"}
{"热点问答" "http://www.sxti.zj.cn/html/recruit/faq.html"}
{"校园生活" "http://www.sxti.zj.cn/html/recruit/xysh.html"}
{"招生报名" "http://www.sxti.zj.cn/html/recruit/zsbm.html"}
{"招生简章" "http://www.sxti.zj.cn/html/recruit/plan.html"}
{"专业介绍" "http://www.sxti.zj.cn/html/recruit/pro.html"}
{"热点问答" "http://www.sxti.zj.cn/html/recruit/faq.html"}
{"校园生活" "http://www.sxti.zj.cn/html/recruit/xysh.html"}
{"校务公开" "http://www.sxti.zj.cn"}
```

```
{"组织架构" "http://www.sxti.zj.cn/html/public/org.html"}
{"办学规划" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=19"}
{"管理制度" "http://www.sxti.zj.cn/e/action/ListInfo/?classid=20"}
...)
```

2. **DONE** parse sub-nav page to get all listed result articles **CLOSED:** *[2018-11-20 Tue 19:56]*

```
<a href="/e/action/ListInfo/index.php?page=109&amp;classid=33&amp;totalnum=1648">尾页</a>
```

```
http://www.sxti.zj.cn/e/action/ListInfo/index.php?page=15&classid=33&totalnum=1648
```

```
(defn- get-mainContent-html
  "Get the <div class=\"page_right\" element."
  [html]
  (html/select html [:html :body :div.page_all :div.page_1 :div.page_right]))

(map #(let [link  (str website-old-url (first (html/attr-values % :href)))
            title (html/text %)]
        {title link})
     (html/select
      (get-mainContent-html
       (get-html "http://www.sxti.zj.cn/e/action/ListInfo/?classid=33" {:as "GB2312"}))
      [:ul.newsList1 :li :a]))
```

```clojure


(defn- get-mainContent-html
  "Get the <div class=\"page_right\" element."
  [html]
  (html/select html [:html :body :div.page_all :div.page_1
  ↪  :div.page_right]))

(defn get-page-article-links
  "Get articles list's every article link and title."
  [nav-link]
  (map #(let [link  (str website-old-url (first (html/attr-values %
  ↪  :href)))
              title (html/text %)]
          {title link})
       (html/select
        (get-mainContent-html
         (get-html nav-link {:as "GB2312"}))
        [:ul.newsList1 :li :a])))
```

3. **DONE** iterate all result pages **CLOSED:** *[2018-11-20 Tue 20:13]*

```clojure
(defn get-total-result-pages
  "How much result pages?"
  [nav-link]
  (Integer.
   ((keyword (str "/e/action/ListInfo/index.php?" "page"))
    (clojure.walk/keywordize-keys
     (ring.util.codec/form-decode
      (first
       (html/attr-values
        (last
         (html/select
          (get-mainContent-html (get-html nav-link {:as "GB2312"}))
          [:div.yema1 :a]))
        :href)))))))

(comment
  (get-total-result-pages
   ↪   "http://www.sxti.zj.cn/e/action/ListInfo/?classid=33"))

(defn get-all-article-links
  "Get a nav's all articles link and title with map as return."
  [nav-link]
  (for [n (range 1 (inc (get-total-result-pages nav-link)))]
    (let [url (str nav-link "&page=" n)]
      (get-page-article-links url))))

;;
 ↪   "http://www.sxti.zj.cn/e/action/ListInfo/index.php?classid=33&page=1&totalnum=1648

(get-all-article-links
 ↪   "http://www.sxti.zj.cn/e/action/ListInfo/?classid=33")
```

5.2.2 DONE new

CLOSED: *[2018-11-20 Tue 16:50]*

- `http://zjzx.sxsedu.net/`

1. **DONE** get navigator bar links **CLOSED:** *[2018-11-20 Tue 13:40]*

- http://zjzx.sxsedu.net/xygk/xyjj

```html
<body>
  <header id="header">
    <div class="nav">
      <div id="content">
        <aside class="side">
          <div class="mainContent">
            <div class="mHd">
              <h3>学院简介</h3>
              <div class="mBd">
                <div class="articleCon">
                  <div class="printArea">
                    <h3 class="title">学院简介</h3>
                    <div class="conTxt">
                      <div>
                        <footer id="footer">
```

```clojure
(require '[clj-http.client :as http]
         '[net.cgrand.enlive-html :as html])
(import 'java.net.URL)

(defn get-html
  "Get HTML string as result."
  [url]
  ;; faster, use tagsoup internal. But I don't know how to specify encoding.
  (html/html-resource (URL. url)))

(def website-new-url "http://www.sxszjzx.com")
```

```
(def website-new-html (get-html website-new-url))

;;; Nav sections
(def nav-bar
  (drop 1 (html/select
           website-new-html
           [:html :body :div.wrap :div.nav
            :div.siteWidth :ul#mainNav.mainNav
            :li])))

(def nav-links
  (map #(let [nav   (-> (html/select % [:a])
                        first)
              link  (str website-new-url (first (html/attr-values nav :href)))
              title (html/text nav)]
          {title link})
       nav-bar))

(pprint nav-links)
```

```
({"学院概况" "http://www.sxszjzx.com/xygk/xyjj"}
 {"学院简介" "http://www.sxszjzx.com/xygk/xyjj"}
 {"现任领导" "http://www.sxszjzx.com/xygk/xrld"}
 {"组织架构" "http://www.sxszjzx.com/xygk/zzjg"}
 {"处室联系" "http://www.sxszjzx.com/xygk/cslx"}
 {"校园风光" "http://www.sxszjzx.com/xygk/xyfg"}
 {"学院荣誉" "http://www.sxszjzx.com/xygk/xyry"}
 {"历史沿革" "http://www.sxszjzx.com/xygk/lsyg"}
 {"楼层分布" "http://www.sxszjzx.com/xygk/lcfb"}
```

```
{" 交通地图" "http://www.sxszjzx.com/xygk/jtdt"}
{" 学院动态" "http://www.sxszjzx.com/xydt"}
{" 学院新闻" "http://www.sxszjzx.com/xydt/xyxw"}
{" 系部新闻" "http://www.sxszjzx.com/xydt/xbxw"}
{" 媒体聚焦" "http://www.sxszjzx.com/xydt/mtjj"}
{" 学院荣誉" "http://www.sxszjzx.com/xydt/xyry1"}
{" 教师荣誉" "http://www.sxszjzx.com/xydt/jsry"}
{" 学生荣誉" "http://www.sxszjzx.com/xydt/xsry"}
{" 系部建设" "http://www.sxszjzx.com/xbjs"}
{" 艺术设计系" "http://www.sxszjzx.com/xbjs/yssjx"}
{" 机械电子系" "http://www.sxszjzx.com/xbjs/jxdzx"}
{" 财会信息系" "http://www.sxszjzx.com/xbjs/chxxx"}
{" 商贸旅游系" "http://www.sxszjzx.com/xbjs/smlyx"}
{" 建筑工程系" "http://www.sxszjzx.com/xbjs/jzgcx"}
{" 新疆学部" "http://www.sxszjzx.com/xbjs/xjxb"}
{" 招生就业" "http://www.sxszjzx.com/zsjy/zsbm"}
{" 招生报名" "http://www.sxszjzx.com/zsjy/zsbm"}
{" 招生简章" "http://www.sxszjzx.com/zsjy/zsjz"}
{" 专业介绍" "http://www.sxszjzx.com/zsjy/zyjs6"}
{" 热点问答" "http://www.sxszjzx.com/zsjy/rdwd"}
{" 校园生活" "http://www.sxszjzx.com/zsjy/xysh"}
{" 在线报名" "http://www.sxszjzx.com/zsjy/zxbm"}
{" 就业信息" "http://www.sxszjzx.com/zsjy/jyxx"}
{" 专题栏目" "http://www.sxszjzx.com/ztlm"}
{" 示范校建设" "http://www.sxszjzx.com/ztlm/sfxjs"}
{" 三名工程" "http://www.sxszjzx.com/ztlm/smgc"}
{" 一核二翼专题" "http://www.sxszjzx.com/ztlm/yheyzt"}
{" 群众路线活动" "http://www.sxszjzx.com/ztlm/qzlxhd"}
{" 旅游职教集团" "http://www.sxszjzx.com/ztlm/lyzjjt"}
{" 德育品牌" "http://www.sxszjzx.com/ztlm/dypp"}
{" 校园视频" "http://www.sxszjzx.com/ztlm/xysp"}
```

```
{"精品课程" "http://www.sxszjzx.comhttp://server2.sxszjzx.com/~jwc/"}
{"校务公开" "http://www.sxszjzx.com/xxgk"}
{"办学规划" "http://www.sxszjzx.com/xxgk/bxgh"}
{"公示公告" "http://www.sxszjzx.com/xxgk/glzd"}
{"阳光收费" "http://www.sxszjzx.com/xxgk/ygsf"}
{"评职评优" "http://www.sxszjzx.com/xxgk/pzpy"}
{"招标公告" "http://www.sxszjzx.com/xxgk/zbgg"}
{"质量报告" "http://www.sxszjzx.com/xxgk/zlbg"}
{"资源下载" "http://www.sxszjzx.com/xxgk/zyxz"}
{"党建工作" "http://www.sxszjzx.com/djgz"}
{"党建动态" "http://www.sxszjzx.com/djgz/djdt"}
{"廉政专栏" "http://www.sxszjzx.com/djgz/lzzl"}
{"亮旗行动" "http://www.sxszjzx.com/djgz/lqhd"}
{"学习资料" "http://www.sxszjzx.com/djgz/xxzl"})
```

2. **DONE** parse sub-nav page to get all listed result articles **CLOSED:** *[2018-11-20 Tue 15:44]*

```
<div class="mainContent">

  <!-- nav name -->
  <div class="mHd">
    <div class="path">

      <em>您的位置：</em><a href="/">首页</a>
      &gt;<a href="/xydt">学院动态</a>&gt;<a href="/xydt/xyxw" target="_blank">学院新闻</a></div>
    <h3>学院新闻</h3>
  </div>
```

```

<div class="mBd">
  <!-- 正文内容 S -->
  <ul class="pageTPList">

    <!-- article -->
    <li class="first">

      <div class="title">
        <a target="_blank" class="tit"
        ↪  href="/xydt/xyxw/content_39935" title="六十载同心同德建
        ↪  名校　一甲子匠智匠力创品牌">六十载同心同德建名校　一甲子匠
        ↪  智匠力创品牌</a>
      </div>

      <div class="pic">
        <a target="_blank" href="/xydt/xyxw/content_39935">
          <img alt="六十载同心同德建名校　一甲子匠智匠力创品牌"
          ↪  src="/upload/sxszjzx/contentmanage/article/image/2018/11/09/20a47dec28
        </a>
      </div>

      <div class="con">

        <div class="intro">
          &nbsp;薪火相承，一甲子春华秋实。2018 年，绍兴技师学院（筹）、
          ↪  绍兴市职业教育中心迎来了建校六十周年华诞。学校始建于
          ↪  1958 年，前身为鉴湖公社初级中学，1985 年更名为绍兴市第一
          ↪  职业技术中学，1995 年市一职中与市中兴职中和市二职中合并
          ↪  为绍兴市职教中心，1997 年市树人...
        </div>

        <div class="others">
          <span class="date">2018-11-09</span>
        </div>

      </div>

    </li>



    <li>

      <div class="title">
        <a target="_blank" class="tit" href="/xydt/xyxw/content_39927" title=" 你的样子，我刚好喜欢 ——我校开展“仪容仪表示范班”评比活动">你的样子，我刚好喜欢 ——我校开展“仪容仪表示范班”评比活动</a>
      </div>


      <div class="con">

        <div class="intro">
          &nbsp;&nbsp;为进一步加强学生的文明礼仪教育，强化学校的常规管理，使学生养成良好的行为习惯，促进优良的校风、班风的形成，近期德育处开展了“仪容仪表示范班”的评选活动。&nbsp; 根据学校的实际情况，学校制定了仪容仪表示范班的评选条件。要求男生不烫发染发，前面头发不盖...
        </div>

```

```
  <div class="others">
   <span class="date">2018-11-08</span>
  </div>

 </div>

</li>



<li>

 <div class="title">
  <a target="_blank" class="tit"
  ↪  href="/xydt/xyxw/content_39923" title=" 当快闪遇上诗歌
  ↪  ——我们一起告白我的国">当快闪遇上诗歌——我们一起告白我的
  ↪  国</a>
 </div>

 <div class="pic">
  <a target="_blank" href="/xydt/xyxw/content_39923">
   <img alt=" 当快闪遇上诗歌——我们一起告白我的国"
   ↪  src="/upload/sxszjzx/contentmanage/article/image/2018/11/08/399e2baba1
  </a>
 </div>

 <div class="con">

  <div class="intro">
```

```
      &nbsp;&nbsp; &nbsp;近日，德育处、团委举办了一场“颂中华诗词，
       ↪  寻文化基因”的诗歌快闪活动。新颖的活动形式吸引了不少师生
       ↪  的关注，得到了许多同学的喝彩。“从浩瀚的地球仪上，我认识了
       ↪  我的祖国……”突然，熙熙攘攘的人群中传来了诗歌朗诵声，这
       ↪  边朗诵声初歇，那边人群中又突...
    </div>

    <div class="others">
      <span class="date">2018-11-08</span>
    </div>

  </div>

</li>



<li>

  <div class="title">
    <a target="_blank" class="tit"
     ↪  href="/xydt/xyxw/content_39893" title="“红色匠心，青春向
     ↪  党”—我校十月“祖国颂”诗歌朗诵比赛圆满结束">“红色匠心，
     ↪  青春向党”—我校十月“祖国颂”诗歌朗诵比赛圆满结束</a>
  </div>

  <div class="pic">
    <a target="_blank" href="/xydt/xyxw/content_39893">
      <img alt="“红
       ↪  色匠心，青春向党”—我校十月“祖国颂”诗歌朗诵比赛圆满结束"
       ↪  src="/upload/sxszjzx/contentmanage/article/image/2018/11/07/1b861b6320
```

```
          </a>
        </div>

        <div class="con">

          <div class="intro">
            &nbsp; &nbsp; 为庆祝祖国 69 岁华诞，培养学生爱党爱国爱校情怀，配合省全民终身学习宣传周活动，提升学生诗歌朗诵水平及语文素养，营造朝气蓬勃、积极向上的校园文化氛围，11 月 1 日下午，绍兴技师学院（筹)、绍兴市职教中心在报告厅隆重举行十月“祖国颂”诗歌朗诵比赛决赛。...
          </div>

          <div class="others">
            <span class="date">2018-11-07</span>
          </div>

        </div>

      </li>



      <li>

        <div class="title">
          <a target="_blank" class="tit" href="/xydt/xyxw/content_39799" title=" 第三届校教职工气排球圆满结束 ">第三届校教职工气排球圆满结束 </a>
        </div>

```

```
    <div class="pic">
     <a target="_blank" href="/xydt/xyxw/content_39799">
      <img alt="第三届校教职工气排球圆满结束" src="/upload/sxszjzx/contentmanage/article/image/2018/11/02/bf87db2b58
     </a>
    </div>

    <div class="con">

     <div class="intro">
      &nbsp; &nbsp; 第三届校教职工气排球比赛，经过二轮26场激烈比拼于周二中午在校体育馆落下帷幕。最终比赛成绩如下：第一名：新疆学部，第二名：办公行政教科组，第三名：德育实习安监组，第四名：艺术设计组。本届气排球比赛，由校教职工10个工会小组分别组队参加，比赛共分...
     </div>

     <div class="others">
      <span class="date">2018-11-02</span>
     </div>

    </div>

   </li>



   <li>

    <div class="title">
```

```
            <a target="_blank" class="tit"
            ↪  href="/xydt/xyxw/content_39808" title=" 关注课堂, 携手育
            ↪  人——我校开展家长开放日观摩课活动报道">关注课堂, 携手育人
            ↪  ——我校开展家长开放日观摩课活动报道</a>
          </div>

          <div class="pic">
            <a target="_blank" href="/xydt/xyxw/content_39808">
              <img alt=" 关注课堂, 携手育人——我校开展家长开放日观摩课活
              ↪  动报道"
              ↪  src="/upload/sxszjzx/contentmanage/article/image/2018/11/04/6a62a56c68
            </a>
          </div>

          <div class="con">

            <div class="intro">
              为更好的架设学校、家庭、社会沟通的桥梁, 让家长们走进课堂, 走
              ↪  近孩子, 走进学校, 10 月 24 日上午我校举行了“家长开放日”
              ↪  活动。“请家长听一堂课”作为这次活动的重头戏, 学校和各系部
              ↪  都高度重视。除了语文、数学等文化课以外, 各系部还展示了包
              ↪  括声乐、建筑设计、...
            </div>

            <div class="others">
              <span class="date">2018-10-30</span>
            </div>

          </div>

        </li>
```

```



<li>

 <div class="title">
  <a target="_blank" class="tit" href="/xydt/xyxw/content_39589" title="“更高、更快、更强”——绍兴技师学院（筹）绍兴市职教中心成功举办第七十一届田径运动会">“更高、更快、更强”——绍兴技师学院（筹）绍兴市职教中心成功举办第七十一届田径运动会</a>
 </div>

 <div class="pic">
  <a target="_blank" href="/xydt/xyxw/content_39589">
   <img alt="“更高、更快、更强”——绍兴技师学院（筹）绍兴市职教中心成功举办第七十一届田径运动会" src="/upload/sxszjzx/contentmanage/article/image/2018/10/28/4ea0ecdfcc
  </a>
 </div>

 <div class="con">

  <div class="intro">
   &nbsp;10月17日至19日，我校隆重举办了第七十一届田径运动会。本届运动会共六个组别、十二个赛项，有89个班级、共1350名运动员参赛，参赛班级、参赛人数均创历史新高。17日下午，学校体育场上彩旗飘扬、音乐嘹亮，全校师生云集喜迎第七十一届田径运动会开幕。学校纪委书...
  </div>

  <div class="others">
   <span class="date">2018-10-28</span>
  </div>

 </div>

</li>



<li>

 <div class="title">
  <a target="_blank" class="tit"
  ↪  href="/xydt/xyxw/content_39560" title=" 弘扬垦荒精神 铸
  ↪  牢党性之魂——我校党员教师赴大陈岛开展专题党日活动">弘扬垦
  ↪  荒精神 铸牢党性之魂——我校党员教师赴大陈岛开展专题党日活
  ↪  动</a>
 </div>

 <div class="pic">
  <a target="_blank" href="/xydt/xyxw/content_39560">
   <img alt=" 弘扬垦荒精神 铸牢党性之魂——我校党员教师赴大陈岛
   ↪  开展专题党日活动"
   ↪  src="/upload/sxszjzx/contentmanage/article/image/2018/10/26/f65137cdca
  </a>
 </div>

 <div class="con">

  <div class="intro">
```

```
              为纪念中华人民共和国成立 69 周年，全面推进党的组织建设，激励
              ↪  全体党员教师继承和发扬党的优良传统和作风，绍兴市职教中心
              ↪  党委组织党员教师沿着习近平总书记的足迹奔赴浙江省直机关党
              ↪  员干部教育基地台州大陈岛开展现场学习教育，重温入党誓词，
              ↪  追溯红色记忆，学习“...
            </div>

            <div class="others">
              <span class="date">2018-10-26</span>
            </div>

          </div>

        </li>



        <li>

          <div class="title">
            <a target="_blank" class="tit"
            ↪  href="/xydt/xyxw/content_39499" title="加强师资建设，争
            ↪  创名师团队——我校召开《人性的追问与教师的职业成长》专题讲
            ↪  座">加强师资建设，争创名师团队——我校召开《人性的追问与教
            ↪  师的职业成长》专题讲座</a>
          </div>

          <div class="pic">
            <a target="_blank" href="/xydt/xyxw/content_39499">
```

```
   <img alt=" 加强师资建设，争创名师团队——我校召开《人性的追
   ↪  问与教师的职业成长》专题讲座"
   ↪  src="/upload/sxszjzx/contentmanage/article/image/2018/10/24/cfc07cf774
  </a>
 </div>

 <div class="con">

  <div class="intro">
   加强师资建设，争创名师团队——我校召开《人性的追问与教师的职
   ↪  业成长》专题讲座 10 月 12 日下午，学校邀请平湖职业中专校
   ↪  长贺陆军到我校报告厅召开《人性的追问与教师的职业成长》专
   ↪  题讲座。讲座由校长钱金星主持。全体教工认真聆听了讲座。贺
   ↪  校长从人的物质性与文化...
  </div>

  <div class="others">
   <span class="date">2018-10-24</span>
  </div>

 </div>

</li>



<li class="last">

 <div class="title">
```

```
        <a target="_blank" class="tit" href="/xydt/xyxw/content_39157" title=" 五星三名 · 用行动践行先锋力量——我校开展系列 “党员育人” 活动">五星三名 · 用行动践行先锋力量——我校开展系列 “党员育人” 活动</a>
      </div>

      <div class="pic">
        <a target="_blank" href="/xydt/xyxw/content_39157">
          <img alt=" 五星三名·用行动践行先锋力量——我校开展系列“党员育人”活动" src="/upload/sxszjzx/contentmanage/article/image/2018/10/10/0e90c18e22
        </a>
      </div>

      <div class="con">

        <div class="intro">
          为充分发挥党员教师的先锋模范作用，进一步拓宽党员育人的广度与深度，传播先锋力量，本学期伊始，我校党委组织六大支部开展了 “秉烛怀志，躬身明责” 的系列党员育人岗活动—— “党员育人示范岗”，用红正的党徽为学生的放学之路保驾护航；“党建带团建”，用党员力...
        </div>

        <div class="others">
          <span class="date">2018-09-29</span>
        </div>

      </div>

    </li>

    </ul>
    <div class="page">


      <a class="first disabled" href="javascript:void(0);">首页</a>
      <a class="prev disabled" href="javascript:void(0);">上一页</a>
      <a class="current">1</a> <a href="/xydt/xyxw_2">2</a> <a
       ↪  href="/xydt/xyxw_3">3</a> <a href="/xydt/xyxw_4">4</a> <a
       ↪  href="/xydt/xyxw_5">5</a> <a href="/xydt/xyxw_6">6</a> <a
       ↪  href="/xydt/xyxw_7">7</a> <a href="/xydt/xyxw_8">8</a> <a
       ↪  href="/xydt/xyxw_9">9</a>        <a class="next"
       ↪  href="/xydt/xyxw_2">下一页</a>
      <a class="last" href="/xydt/xyxw_166">尾页</a>
      <span class="total">共 1652 条信息/共 166 页</span>
      <span class="select">转到第<input title=" 按回车键跳转到指定页"
       ↪  onkeypress="javascript:return quickJumpPage(event,
       ↪  'quickJumpButton')" type="text" style="width:24px"
       ↪  id="quickJumpInput" class="quickJumpInput "
       ↪  value="1"><script type="text/javascript">var pageNameUrl =
       ↪  '/xydt/xyxw_{pageid}';function quickJumpPage(event){if
       ↪  (event.keyCode == 13 && !(event.srcElement &&
       ↪  (event.srcElement.tagName.toLowerCase() == 'textarea')))
       ↪  {var number =
       ↪  document.getElementById('quickJumpInput').value;if
       ↪  (/\d+/i.test(number)){if (number > 166) number =
       ↪  166;if(number < 1) number = 1;window.location.href =
       ↪  pageNameUrl.replace("{pageid}", number);}else{alert(' 输入的
       ↪  页数有误! ');}}}</script>页</span>
    </div>
    <!-- 正文内容 E -->
```

```
  </div>

</div>
```

```
'(defn get-mainContent-html
   "Get nav link page's mainContent."
   [nav-link]
   (html/select nav-link [:html :body :div.wrap :div#content
   ↪  :div.mainContent]))

(defn get-articles-list
  "Get page right articles list."
  [nav-link]
  (html/select (get-mainContent-html nav-link) [:div.mBd
  ↪  :ul.pageTPList
                                                :li :div.title
                                                 ↪  :a.tit]))

(pprint
 (map #(let [link  (first (html/attr-values % :href))
             title (html/text %)]
         {title link})
      (html/select
       (html/select
        (get-html "http://www.sxszjzx.com/xydt/xyxw")
        [:html :body :div.wrap :div#content :div.mainContent])
       [:div.mBd :ul.pageTPList :li :div.title :a.tit])))
```

```
({"六十载同心同德建名校　一甲子匠智匠力创品牌" "/xydt/xyxw/content_39935"}
 {"你的样子，我刚好喜欢　——我校开展“仪容仪表示范班”评比活动" "/xydt/xyxw/content_39927"}
 {"当快闪遇上诗歌——我们一起告白我的国" "/xydt/xyxw/content_39923"}
 {"“红色匠心，青春向党”—我校十月“祖国颂”诗歌朗诵比赛圆满结束" "/xydt/xyxw/content_39893"}
 {"第三届校教职工气排球圆满结束 " "/xydt/xyxw/content_39799"}
 {"关注课堂，携手育人——我校开展家长开放日观摩课活动报道" "/xydt/xyxw/content_39808"}
 {"“更高、更快、更强”——绍兴技师学院（筹）绍兴市职教中心成功举办第七十一届田径运动会"
  "/xydt/xyxw/content_39589"}
 {"弘扬垦荒精神 铸牢党性之魂——我校党员教师赴大陈岛开展专题党日活动" "/xydt/xyxw/content_39560"}
 {"加强师资建设，争创名师团队——我校召开《人性的追问与教师的职业成长》专题讲座" "/xydt/xyxw/content_39499"}
 {"五星三名·用行动践行先锋力量——我校开展系列“党员育人”活动" "/xydt/xyxw/content_39157"})
```

3. **DONE** iterate all result pages **CLOSED:** *[2018-11-20 Tue 16:49]*

```
<span class="total">共 1652 条信息/共 166 页</span>
```

```
(defn get-total-result-pages
  "How much result pages?"
  [nav-link]
  (Integer.
   (second
    (re-find #"/共 (.*) 页"
             (html/text
              (first (html/select (get-mainContent-html (get-html
              ↪  nav-link))
                                  [:div.mBd :div.page
                                  ↪  :span.total])))))))

(pprint (get-total-result-pages "http://www.sxszjzx.com/xydt/xyxw"))
```

```
166

http://www.sxszjzx.com/xydt/xyxw
OR:
http://www.sxszjzx.com/xydt/xyxw_1

http://www.sxszjzx.com/xydt/xyxw_2
....
http://www.sxszjzx.com/xydt/xyxw_166
```

```
(defn get-page-article-links
  "Get articles list's every article link and title."
  [nav-link]
  (map #(let [link  (first (html/attr-values % :href))
              title (html/text %)]
          {title link})
       (html/select (get-mainContent-html (get-html nav-link))
                    [:div.mBd :ul.pageTPList :li :div.title :a.tit])))

(defn get-all-article-links
  "Get a nav's all articles link and title with map as return."
  [nav-link]
  (for [n (range 1 (inc (get-total-result-pages nav-link)))]
    (let [url (str nav-link (format "_%d" n))]
      (get-page-article-links url))))

(def nav-articles-links (get-all-article-links
↪   "http://www.sxszjzx.com/xydt/xyxw"))
```

```
(pprint (take 3 nav-articles-links))
```

```
(({"六十载同心同德建名校 一甲子匠智匠力创品牌"
↪   "/xydt/xyxw/content_39935"}
  {"你的样子，我刚好喜欢 ——我校开展“仪容仪表示范班”评比活动"
  ↪   "/xydt/xyxw/content_39927"}
  {"当快闪遇上诗歌——我们一起告白我的国" "/xydt/xyxw/content_39923"}
  {"“红色匠心，青春向党”—我校十月“祖国颂”诗歌朗诵比赛圆满结束"
  ↪   "/xydt/xyxw/content_39893"}
  {"第三届校教职工气排球圆满结束 " "/xydt/xyxw/content_39799"}
  {"关注课堂，携手育人——我校开展家长开放日观摩课活动报道"
  ↪   "/xydt/xyxw/content_39808"}
  {"“更高、更快、更强”——绍兴技师学院（筹）绍兴市职教中心成功举办第七十
  ↪   一届田径运动会"
   "/xydt/xyxw/content_39589"}
  {"弘扬垦荒精神 铸牢党性之魂——我校党员教师赴大陈岛开展专题党日活动"
  ↪   "/xydt/xyxw/content_39560"}
  {"加强师资建设，争创名师团队——我校召开《人性的追问与教师的职业成长》
  ↪   专题讲座" "/xydt/xyxw/content_39499"}
  {"五星三名·用行动践行先锋力量——我校开展系列“党员育人”活动"
  ↪   "/xydt/xyxw/content_39157"})
```

```
({" 柯桥区职教中心四位中层干部来我校交流学习"
↪  "/xydt/xyxw/content_39159"}
 {" 江苏泰州机电高等职业技术学校一行来校学习调研"
 ↪  "/xydt/xyxw/content_39552"}
 {" 倾听教师心声，助力学校发展——我校召开 2018 学年第一学期教师座谈会"
 ↪  "/xydt/xyxw/content_39158"}
 {" 校工会举行新学期首次工作会议" "/xydt/xyxw/content_38999"}
 {" “急救侠”集结，学校安全再添“保护伞”—我校教师积极参加“AED”培训"
 ↪  "/xydt/xyxw/content_29630"}
 {" “年轻的战士，我为你鼓掌”——2018 级新生军训会操纪实"
 ↪  "/xydt/xyxw/content_29629"}
 {" 以奖促学——让榜样的力量发光发热" "/xydt/xyxw/content_29628"}
 {" 未来工匠，领跑新征程——我校举行 2018 年第一学期开学典礼暨表彰大会
 ↪  (图文)" "/xydt/xyxw/content_29627"}
 {" 我校学子成功晋级第三届浙江省科学玩家青少年才能挑战赛复赛"
 ↪  "/xydt/xyxw/content_29626"}
 {" 告别“压力山大”  美好从“心”开始—我校第一堂“心理慕课”顺利开课"
 ↪  "/xydt/xyxw/content_29625"})
({" 树目标 重创新 团结协作创实效----市教育局副局长石鑫炯一行赴我校督查
↪  开学工作 (图文)"
  "/xydt/xyxw/content_29624"}
 {" 温情期许匠心传承-我校为 2018 级新生送上最走心入学礼"
 ↪  "/xydt/xyxw/content_29623"}
 {" 一个支部一个堡垒，2018 绍兴新疆班返校在路上 (图文)"
 ↪  "/xydt/xyxw/content_29622"}
 {" 记我校 2018 年文明单位、文明校园复评工作 (图文)"
 ↪  "/xydt/xyxw/content_29621"}
 {" 德育处、团委召开新学期第一次班主任会议" "/xydt/xyxw/content_34686"}
 {" 忠于职守，继往开来——2018 学年干部集体廉政谈话 (图文)"
 ↪  "/xydt/xyxw/content_34685"}
 {"2018 学年编外工作人员招聘公告" "/xydt/xyxw/content_34684"}
 {" 实践创新路上的中职名校建设——记我校 2018 学年第一期读书会（图文)" "/xydt/xyxw/content_34683"}
 {" 青春服务越青商，助力拥抱大湾区" "/xydt/xyxw/content_34682"}
 {"2018 年特招教师拟录用人员名单公示" "/xydt/xyxw/content_29640"}))
```

5.2.3 DONE benchmark new website two entry URL

CLOSED: *[2018-11-20 Tue 13:17]*

```
(require '[clj-http.client :as http])
(require '[net.cgrand.enlive-html :as html])
(use 'criterium.core)

(pr (quick-bench (http/get "http://www.sxszjzx.com")))
```

```
(pr (quick-bench (http/get "http://zjzx.sxsedu.net")))
```

5.3 TODO 从旧网站遍历所有文章，然后以标题在新网站中搜索，对比搜索结果中的第一个 [0/1]

- □ 有一个统一而且简单的办法是在旧网站上遍历所有文章，然后每条文章在新网站中搜索。获得新网站的搜索结果。然后比对新旧网站文章页面的内容。

`jyj.sxsedu.net/s?sid=22&wd="QUERY"`

```
;;; Search entry
;; http://jyj.sxsedu.net/s?sid=22&wd="QUERY"
(defn website-new-search
  "Search TITLE in new website http://jyj.sxsedu.net/."
  [title]
  )

;; div.s-result > div.result-list > ul > li > h4 > a[href="url"]
(defn website-root-search-result-select
  "Select the first result from `website-new-search` results."
  [results]
  (-> results
      (html/select [:.s-result :.result-list :ul :li :h4 :a])
      (html/attr-values :href))
  )

;;; TODO: encode Chinese into URL
```

5.3.1 TODO 如何遍历所有文章 [0/1]

- □ search

1. **DONE** 遍历爬取旧网站的所有文章 **CLOSED:** *[2018-09-27 Thu 14:33]*

   Old website 类目: `http://www.sxszjzx.com/e/action/ListInfo/?classid=390`
   文章页面: `http://www.sxszjzx.com/e/action/ShowInfo.php?classid=390&id=12767`

   文章页面的内容 tag

   ```
   :div.page_right > :.pright_t1 > :.pright_t4
        article          title           content
   ```

   Single entry test:

   ```
   (:status (http/get "http://www.sxszjzx.com/e/action/ShowInfo.php?classid=6&id=6"
                      {:as "GB2312"}));; => 200
   ```

```

(if (= " 信息提示" (first
                   (:content
                    (first
                     (html/select
                      (html/html-snippet
                       (:body (http/get
                       ↪   "http://www.sxszjzx.com/e/action/ShowInfo.php?classid=6&id=
                                        {:as "GB2312"})))
                      [:title])))))
  false
  true)
```

A huge double deep iteration:

```
(require 'clojure.java.io)

(for [classid (range 0 1000)]
  (for [id (range 0 1000)]
    (let [url     (format
    ↪   "http://www.sxszjzx.com/e/action/ShowInfo.php?classid=%s&id=%s"
                          classid id)
          respond (http/get url {:as "GB2312"})]
      (if-let [status (not (= " 信息提示" (first
                                          (:content (first
                                                     (html/select

                                                      ↪   (html/html-snippet
                                                      ↪   (:body
                                                      ↪   respond))
                                                      [:title])))))))]
```

```
        (let [*out* (clojure.java.io/writer "/home/stardiviner/titles.lst")]
          (println (str "OK! " url))
          (println (-> respond
                       html/html-snippet
                       (html/select [:pright_t1]) ; title
                       html/texts)))
        (println (str "FAILED! " url))
        ))))
```

After long time running and test, seems this does not work.

5.3.2 DONE 新网站搜索入口 API

CLOSED: *[2018-09-27 Thu 14:07]* **DEADLINE:** *<2018-09-27 Thu>*

```
http://jyj.sxsedu.net/s?sid=22&wd="QUERY"
```

```
(doseq [title (....)] ; TODO:
  (-> (http/get "http://jyj.sxsedu.net/s?sid=22&wd=%s" title)
      ))
```

5.3.3 DONE 获取第一个搜索结果是否和旧网站的标题相同？

CLOSED: *[2018-09-27 Thu 14:14]* **DEADLINE:** *<2018-09-27 Thu>*

```
div.s-result > div.result-list > ul > li > h4 > a[href="url"]
```

For testing:

`http://www.sxszjzx.com/e/action/ShowInfo.php?classid=115&id=15463` “致匠心”—-艺术设计系开学第一课

`http://www.sxszjzx.com/e/action/ShowInfo.php?classid=115&id=15453` 17 学前教育 2 班开学第一课—-君子以自强不息

`http://www.sxszjzx.com/e/action/ShowInfo.php?classid=115&id=15421` 童心颂师恩

```clojure
(require '[clojure.string :as str])

(defn website-new-search-result-same? [title]
  "Search TITLE in new website search API."
  (let [first-result       (first
                            (-> (:body
                                 (http/get
                                  (format
                                  ↪   "http://jyj.sxsedu.net/s?sid=22&wd=%s"
                                  ↪   title)))
                                html/html-snippet
                                (html/select [:.s-result :.result-list :ul
                                ↪   :h4 :a])))
        first-result-title (apply str (html/texts (html/select
        ↪   first-result [:span])))
        first-result-link  (first (html/attr-values first-result :href))]
    (= first-result-title (convert-old-title title))))

(defn- convert-old-title [title]
  "Remove some special characters in old website title."
  (str/replace title #"[ “” -]" ""))

(prn (website-new-search-result-same? " 童心颂师恩"))
(prn (website-new-search-result-same? " “致匠心” ----艺术设计系开学第一课"))
```

```
true
```

```
true
```

5.4 FAILED 按照 URL 的格式遍历所有类目和文章

CLOSED: *[2018-09-25 Tue 10:45]*

Old website 类目: `http://www.sxszjzx.com/e/action/ListInfo/?classid=390` 文章页面: `http://www.sxszjzx.com/e/action/ShowInfo.php?classid=390&id=12767`

`http://www.sxszjzx.com/e/action/ListInfo/index.php?page=109&classid=33&totalnum=1645`

New website 类目: `http://zjzx.sxsedu.net/xydt/xyxw_2` 文章页面: `http://zjzx.sxsedu.net/xydt/xyxw/content_34682`

6 DONE crawler.clj 爬虫存储提取连接的数据库 [5/5]

CLOSED: *[2018-11-20 Tue 21:58]*

6.1 DONE Redis

CLOSED: *[2018-11-20 Tue 21:58]*

Clojure Redis Packages

Redis

6.1.1 DONE how to access the Redis?

CLOSED: *[2018-11-21 Wed 19:59]*

```clojure
(require '[taoensso.carmine :as redis])

;;; Redis store crawler links
(defonce redis-conn-pool {:pool {}
                          :spec {:host "127.0.0.1" :port 6379}})

(defmacro wcar*
  [& body]
  `(redis/wcar redis-conn-pool ~@body))

(comment
  (wcar*
   (redis/ping)))

(defn save-link-to-redis
  "Save crawled link to Redis.
  Usage: (save-link-to-redis :old title link)"
  [website-key title link]
  (wcar*
   (redis/hset website-key title link)))

(defn get-link-from-redis
  "Get link to crawl from Redis.
  Usage: (get-link-from-redis :old title"
  [website-key title]
  (wcar*
   (redis/hget website-key title)))
```

6.1.2 DONE extract all links in nav bar

CLOSED: *[2018-11-22 Thu 08:44]*

```
(defn crawl-website-old-links []
  "Crawl all links of website old."
  (for [nav-link (map #(-> % vals first) old/nav-bar-links-map)]
    (for [n (range 1 (inc (count old/nav-bar-links-map)))]
      (wcar* (redis/hset :old/links n nav-link)))))

(defn crawl-website-new-links []
  "Crawl all links of website new."
  (for [nav-link (map #(-> % vals first) new/nav-bar-links-map)]
    (for [n (range 1 (inc (count new/nav-bar-links-map)))]
      (wcar* (redis/hset :new/links n nav-link)))))
```

6.1.3 DONE save all links to Redis

CLOSED: *[2018-11-21 Wed 21:58]*

```
(defn crawl-website-old-links []
  "Crawl all links of website old."
  (for [nav-link (map #(-> % vals first) old/nav-bar-links-map)]
    (for [n (range 1 (inc (count old/nav-bar-links-map)))]
      (redis/hset :old/links n nav-link))))

(defn crawl-website-new-links []
  "Crawl all links of website new."
  (for [nav-link (map #(-> % vals first) new/nav-bar-links-map)]
    (for [n (range 1 (inc (count new/nav-bar-links-map)))]
      (redis/hset :new/links n nav-link))))
```

Check out the stored links:

```
http://www.sxti.zj.cn/html/recruit/plan.html
http://www.sxti.zj.cn/html/recruit/plan.html
http://www.sxti.zj.cn/html/recruit/plan.html
```

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

http://www.sxti.zj.cn/html/recruit/plan.html

```
http://www.sxti.zj.cn/html/recruit/plan.html
```

6.1.4 DONE literal over all redis Hash key-values

CLOSED: *[2018-11-21 Wed 19:59]*

```clojure
(defn all-links-in-redis
  "Get all links in Redis."
  [website-key]
  (wcar*
   ;; (redis/hgetall website-key)
   (redis/hvals website-key)))
```

```clojure
(wcar*
 (redis/hvals :old/links))
```

7 TODO crawler.clj 爬取页面的内容 [3/6]

7.1 DONE scrape the article

CLOSED: *[2018-11-21 Wed 19:27]*

```clojure
(defn crawl-website-old-articles []
  "Crawl all articles of website old."
  (for [link (all-links-in-redis :old/links)]
    (old/get-html link)))

(defn crawl-website-new-articles []
  "Crawl all articles of website new."
  (for [link (all-links-in-redis :new/links)]
    (new/get-html link)))
```

7.1.1 DONE old

CLOSED: *[2018-11-21 Wed 19:14]*

```
(defn get-page-article-content-html
  "Get page's article content html."
  [url]
  (html/select (get-html url) [:div.page_right]))

(defn parse-article-title-and-content
  "Parse and extract the title and content text."
  [url]
  (let [content-html (get-page-article-content-html url)
        title        (html/text (first (html/select content-html
         ↪  [:div.news_Title1])))
        content      (html/text (first (html/select content-html
         ↪  [:div.news_Content1]))))]
    {:title title, :content content}))

(comment
  (html/text
   (first
    (html/select
     (get-page-article-content
      ↪  "http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=61&id=15533")
     [:div.news_Content1]))))
```

7.1.2 DONE new

CLOSED: *[2018-11-21 Wed 19:27]*

```clojure
(defn get-page-article-content-html
  "Get page's article content html."
  [url]
  (html/select (get-html url) [:div.mainContent]))

(defn parse-article-title-and-content
  "Parse and extract the title and content text."
  [url]
  (let [content-html (get-page-article-content-html url)
        title        (html/text
                      (first
                       (html/select
                        content-html
                        [:div.mBd :article.articleCon :div.printArea
                         ↪  :h2.title])))
        content      (html/text
                      (first
                       (html/select
                        content-html
                        [:div.mBd :article.articleCon :div.printArea
                         ↪  :div.conTxt])))]
    {:title title, :content content}))

(comment
  (html/text
   (first
    (html/select
     (get-page-article-content-html
      ↪  "http://www.sxszjzx.com/xydt/xyxw/content_39935")
     [:h2.title]))))
```

7.2 TODO process the crawled article content text [0/1]

7.2.1 TODO strip some special characters and spaces

- □
- □ _

7.3 TODO auto find content element

How to auto find the max scope content element's text. So that crawler can auto extract the text. No need the selector rules.

8 TODO store.clj 爬虫内容存储到数据库 [3/4]

8.1 DONE SQLite

CLOSED: *[2018-11-22 Thu 10:28]*

```
(defn save-to-sqlite
  "Save data into SQLite DB."
  [title article]
  (jdbc/with-db-connection [db sqlite]
    (jdbc/insert! db :articles {:title   title ; TODO: do I need insert
     ↪  the `:id`?
                                :article article})))


(defn save-all-articles-to-sqlite
  []
  (for [article (crawler/crawl-website-old-articles)]
    (save-to-sqlite (key article) (val article))))
```

8.1.1 DONE detect table exist?

CLOSED: *[2018-11-22 Thu 10:08]*

Use `{:conditional? true}` for `jdbc/create-table-ddl`.

8.1.2 DONE save title and content to SQLite

CLOSED: *[2018-11-22 Thu 10:28]*

```
(defn all-links-in-redis
  "Get all links in Redis."
  [website-key]
  (wcar*
   ;; (redis/hgetall website-key)
   (redis/hvals website-key)))

(defn crawl-website-old-articles []
  "Crawl all articles of website old."
  (crawl-website-old-articles)
  (for [link (all-links-in-redis :old/links)]
    (let [page (old/parse-article-title-and-content link)]
      (store/save-to-sqlite :old (:title page) (:content page)))))

(comment
  (let [kk (old/parse-article-title-and-content (first (all-links-in-redis
  ↪  :old/links)))]
    (store/save-to-sqlite :old (:title kk) (:content kk))))
```

```

(defn crawl-website-new-articles []
  "Crawl all articles of website new."
  (crawl-website-new-links)
  (for [link (all-links-in-redis :new/links)]
    (let [page (new/parse-article-title-and-content link)]
      (store/save-to-sqlite :new (:title page) (:content page)))))

(comment
  (let [kk (new/parse-article-title-and-content (first (all-links-in-redis
   ↪   :new/links)))]
    (store/save-to-sqlite :new (:title kk) (:content kk))))
```

8.2 TODO MongoDB

9 TODO validate.clj (检查页面的可用性)

9.1 Validate Links

9.2 Validate Images

9.3 Validate Videos

9.4 Validate Docs

10 TODO compare.clj 对比文章

☒ 对比文章标题

☐ 对比 html 页面中的 img，video 之类的 tags 的数量和类型。

10.1 TODO compare.clj 比较页面的相似性 [0/1]

☐ 对比文字部分 tag 的内容，用 MD5 之类的 digest 算法

- SimHash

- MinHash

□ search Simhash Java libraries Java MinHash

11 DONE 比较上传视频

CLOSED: *[2018-09-26 Wed 19:08]*

11.1 DONE 旧网站的视频

CLOSED: *[2018-09-26 Wed 19:08]*

```
(def website-old-url "http://www.sxti.zj.cn")
```

```
nil#'user/website-new-url
```

get a video link in list example:

```
(pprint
 (first
  (:content
   (first
    (html/select
     (-> (:body
          (http/get
           (format (str website-old-url
           ↪  "/e/action/ListInfo/index.php?classid=10&page=%s") 0)
           {:as "GB2312"}))
         html/html-snippet
         (html/select [:.imgList1]))
     [:ul])))))
```

```
{:tag :li,
 :attrs nil,
 :content
 ({:tag :a,
   :attrs
   {:href "/e/action/ShowInfo.php?classid=173&id=15401",
    :target "_blank"},
   :content
   ({:tag :img,
     :attrs
     {:src

      ↪   "/d/file/html/news/4/4/2018-09-07/small29350dd5474f452348699de611ad62f8.jpg",
      :alt "心理课堂六",
      :width "220",
      :height "180"},
     :content nil})}
  {:tag :div, :attrs {:class "movie1"}, :content nil}
  {:tag :h1,
   :attrs nil,
   :content
   ({:tag :a,
     :attrs
     {:href "/e/action/ShowInfo.php?classid=173&id=15401",
      :title "心理课堂六",
      :target "_blank"},
     :content ("心理课堂六")})})}
```

```
{:tag     :li,
 :attrs   nil,
 :content ({:tag     :a,
            :attrs   {:href "/e/action/ShowInfo.php?classid=378&id=12761",
            ↪   :target "_blank"},
            :content ({:tag :img, :attrs {:src
            ↪   "/d/file/html/news/4/5/2017-05-24/small9bfeabf676860034d33a510fcd583c35.jpg
            ↪   :alt " 重拾自我---16 旅服高工", :width "220", :height
            ↪   "180"}, :content nil})
            }
           {:tag     :div,
            :attrs   {:class "movie1"},
            :content nil}
           {:tag     :h1,
            :attrs   nil,
            :content ({:tag     :a,
                       :attrs   {:href
                       ↪   "/e/action/ShowInfo.php?classid=378&id=12761",
                       ↪   :title " 重拾自我---16 旅服高工", :target
                       ↪   "_blank"}
                       :content (" 重拾自我---16 旅服高工")})
            })}
```

```
;;; get first list page for testing
(comment
  (def page-list
    (-> (:body
          (http/get
           (format (str website-old-url
           ↪   "/e/action/ListInfo/index.php?classid=10&page=%s") 0)
          {:as "GB2312"}))
        html/html-snippet
        (html/select [:.imgList1]))))

(defn- extract-video-page-content [page]
  (-> (:body
       (http/get
        (format (str website-old-url "/e/action/ListInfo/index.php?classid=10&page=%s") page)
        {:as "GB2312"}))
      html/html-snippet
      (html/select [:.imgList1])))

;;; get all title and URL
(defn- extract-video-link [node]
  (let [lists ((juxt #(first (html/attr-values % :href))
                     #(first (html/attr-values % :title)))
               (-> node
                   (html/select [:h1 :a])
                   first))]
    {(second lists)
     ;; construct url
     (str website-old-url (first lists))}))

(comment
  (map
   extract-video-link
   (html/select-nodes* (html/select page-list [:ul]) [:li])))
```

#+RESULTS[*<2018-10-16 Tue 14:51>* c93a85f8a06485728b2f1196ae9883d418693abb]: extract-video-page-info-old

```
(defn extract-page-videos-maxpage-old
  "Get the max page number of
   ↪  http://www.sxti.zj.cn/e/action/ListInfo/?classid=10"
  [url]
  (let [last_page (-> (:body (http/get url {:as "GB2312"}))
                      html/html-snippet
                      (html/select [:div.yema1])
                      first
                      :content
                      last)]
    (if (= (first (:content last_page)) " 尾页")
      ;; (print (format "http://www.sxti.zj.cn/%s" (first
          (html/attr-values last_page : href))))
      (Integer. (first (re-seq #"\d+"
                               (re-find #"page=\d+&" (first
                               ↪  (html/attr-values last_page
                               ↪  :href)))))))
    ))

(print (extract-page-videos-maxpage-old (str website-old-url
↪  "/e/action/ListInfo/?classid=10")))
```

```
13
```

#+RESULTS[*<2018-10-16 Tue 14:51>* bdecab31eb8bad0dfc2c2d281094d39fe174346c]: extract-video-pages-max-old

```
<<define-old-website-url>>
<<extract-video-page-info-old>>
<<extract-video-pages-max-old>>

```

```
(def videos-links-old
  (reduce
   merge
   (reduce
    concat
    (for [n (range 0 (extract-page-videos-maxpage-old (str website-old-url "/e/action/ListInfo/?classid=10")))]
      (map
       #(-> %
            extract-video-link)
       (html/select-nodes*
        (html/select (extract-video-page-content n) [:ul])
        [:li]))))))

(pprint videos-links-old)
```

11.2 DONE 新网站的视频

CLOSED: *[2018-09-26 Wed 19:08]*

```
(def website-new-url "http://www.sxszjzx.com/")
;; (def website-new-url "http://zjzx.sxsedu.net")
```

```
nil#'user/website-new-url
```

get a video link in list example:

```
(pprint
 (first
  (html/select
   (-> (:body
        (http/get
         (format (str website-new-url "/ztlm/xysp_%s") 1)))
       html/html-snippet
       (html/select [:ul.vedioPageList]))
   [:li :h3 :a])))
```

```
{:tag :a,
 :attrs {:href "/ztlm/xysp/zycz/content_129", :title "《青春 梦想》 14 大专
 ↪  动漫"},
 :content ("《青春 梦想》 14 大专动漫")}
```

```
(defn extract-page-videos-maxpage-new
  "Get the max page number of http://www.sxszjzx.com/ztlm/xysp."
  [url]
  (Integer.
   (clojure.string/replace
    (re-find #"/*\d+ 页"
             (-> (:body (http/get url))
                 html/html-snippet
                 (html/select [:div.page :span.total])
                 first
                 :content
                 first))
    " 页" "")))

(print (extract-page-videos-maxpage-new (str website-new-url
↪   "/ztlm/xysp")))
```

```
14
```

#+RESULTS[*<2018-10-16 Tue 14:53>* 92e10ba363a1a208d764be2f96efcc2f1e4a5be4]: extract-video-pages-max-new

```
<<define-new-website-url>>
<<extract-video-pages-max-new>>

(def videos-links-new
  (reduce
   merge
   (reduce
    concat
    (for [n (range 1 (extract-page-videos-maxpage-new (str website-new-url "/ztlm/xysp")))]
      (map
       (fn [a]
         {(first (html/attr-values a :title))
          (str "http://zjzx.sxsedu.net" (first (html/attr-values a :href)))})
       (html/select
        (-> (:body
             (http/get
              (format "http://zjzx.sxsedu.net/ztlm/xysp_%s" n)))
            html/html-snippet
            (html/select [:ul.vedioPageList]))
        [:li :h3 :a]))))))

(pprint videos-links-new)
```

```
14{"纪念刘和珍君" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_106",
 "走近绍兴黄酒1" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_76",
 "《青春" "http://zjzx.sxsedu.net/ztlm/xysp/zycz/content_129",
 "人生因设计而美丽2" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_90",
 "3" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_94",
 "基本站姿组合手位1" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_84",
 "计数原理" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_97",
 "《展望未来我们的明天会更加精彩》14大专城建"
 "http://zjzx.sxsedu.net/ztlm/xysp/zycz/content_132",
 "Art" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_93",
 "动画动作修饰2" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_86",
 "绞孔攻螺纹3" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_83",
 "方言与普通话2" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_107",
 "07-课堂礼仪" "http://zjzx.sxsedu.net/ztlm/xysp/jysp/content_119",
 "色彩空间" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_101",
 "局域网配置综合实训3" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_75",
 "——有一种爱叫做亲情" "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_37",
 "我被十三所学校开除(2)" "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_28",
 "中央电视台-开学第一课" "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_141",
 "人生因设计而美丽1" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_89",
 "王金云" "http://zjzx.sxsedu.net/ztlm/xysp/zycz/content_128",
 "前厅模拟接待" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_66",
 "人生因设计而美丽3" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_91",
 "重拾自我---16旅服高工" "http://zjzx.sxsedu.net/ztlm/xysp/zycz/content_136",
 "Body" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_63",
 "2017年学第二学期开学第一课" "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_61",
 "The" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_104",
 "《拾梦者》" "http://zjzx.sxsedu.net/ztlm/xysp/zycz/content_138",
 "室内操" "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_21",
 "懂得体贴孝敬父母 - 下" "http://zjzx.sxsedu.net/ztlm/xysp/jysp/content_112",
 "第二届三字经PK赛" "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_41",
 "08-校园礼仪" "http://zjzx.sxsedu.net/ztlm/xysp/jysp/content_120",
 "12-社会交往（下）" "http://zjzx.sxsedu.net/ztlm/xysp/jysp/content_124",
 "绞孔攻螺纹1" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_81",
 "我喜欢我的学校（校园生活纪录）" "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_51",
 "葛豪焙" "http://zjzx.sxsedu.net/ztlm/xysp/zycz/content_135",
 "孝亲尊师电视特别节目（三）" "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_39",
 "礼仪操" "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_46",
 "05-教师的装饰" "http://zjzx.sxsedu.net/ztlm/xysp/jysp/content_117",
 "永恒的舞台" "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_53",
 "全国人口普查中小学生一堂课动画短剧"
 "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_31",
 "纺织贸易实务" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_68",
 "心理课堂六" "http://zjzx.sxsedu.net/ztlm/xysp/jysp/content_126",
 "我和我的学校（浙江省影视大赛第一名作品）"
 "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_52",
 "社团成果展示片" "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_49",
 "2014预科生活——预习青春" "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_59",
 "01-教师礼仪概述" "http://zjzx.sxsedu.net/ztlm/xysp/jysp/content_113",
 "孝亲尊师电视特别节目(二)" "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_38",
 "绍兴市职教中心建校50周年校庆" "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_139",
 "第三套中学生广播体操-舞动青春分解示范"
 "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_23",
 "商务学区班主任论坛2" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_96",
 "一个人---15大专动漫" "http://zjzx.sxsedu.net/ztlm/xysp/zycz/content_137",
 "合情推理3" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_80",
 "商务学区班主任论坛1" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_95",
 "直系电机系统故障排除" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_69",
 "礼仪教育视频" "http://zjzx.sxsedu.net/ztlm/xysp/jysp/content_125",
 "健康消费\"茶文化知识讲座(2)" "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_33",
 "开学第一课（2011年9月1日）"
 "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_140",
 "14计算机1班" "http://zjzx.sxsedu.net/ztlm/xysp/zycz/content_134",
 "健康消费\"茶文化知识讲座(3)" "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_34",
 "《青春誓言》13美术高考2班" "http://zjzx.sxsedu.net/ztlm/xysp/zycz/content_133",
 "绞孔攻螺纹2" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_82",
 "《技能成就梦想》13大专机电徐俊亮、傅卓楷"
 "http://zjzx.sxsedu.net/ztlm/xysp/zycz/content_131",
 "第五十八届田径运动会" "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_43",
 "职教三字经PK赛1" "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_26",
 "10-师生关系" "http://zjzx.sxsedu.net/ztlm/xysp/jysp/content_122",
 "2017新生军训纪录片" "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_60",
 "健康消费\"茶文化知识讲座(4)" "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_35",
 "《正步人生》绍兴市职教中心2010军训特别节目"
 "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_142",
 "邹越：让生命充满爱" "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_29",
 "基本站姿组合手位3" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_88",
 "懂得体贴孝敬父母－中" "http://zjzx.sxsedu.net/ztlm/xysp/jysp/content_111",
 "合情推理1" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_72",
 "绍兴市职教中心2011军训特别节目"
 "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_40",
 "艺术教育特色宣传片" "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_25",
 "04-教师的语言" "http://zjzx.sxsedu.net/ztlm/xysp/jysp/content_116",
 "“国家改革发展示范校”建设纪实" "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_55",
 "徐涵宗" "http://zjzx.sxsedu.net/ztlm/xysp/zycz/content_130",
 "\"走进\"世贸" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_100",
 "房志成" "http://zjzx.sxsedu.net/ztlm/xysp/zycz/content_127",
 "11-社会交往（上）" "http://zjzx.sxsedu.net/ztlm/xysp/jysp/content_123",
 "我被十三所学校开除(1)" "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_27",
 "方言与普通话3" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_70",
 "茶为国饮" "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_36",
 "职教之歌" "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_56",
 "学会节俭，不做奢侈的孩子－上" "http://zjzx.sxsedu.net/ztlm/xysp/jysp/content_108",
 "绍兴市职教中心" "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_18",
 "第三套中学生广播体操-舞动青春完整音乐"
 "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_24",
 "方言与普通话1" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_105",
 "06-教师的仪表" "http://zjzx.sxsedu.net/ztlm/xysp/jysp/content_118",
 "局域网配置综合实训2" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_74",
 "合情推理2" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_78",
 "走近绍兴黄酒2" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_77",
 "学会节俭，不做奢侈的孩子 - 下" "http://zjzx.sxsedu.net/ztlm/xysp/jysp/content_109",
 "神神慢" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_103",
 "走近绍兴黄酒3" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_79",
 "2007校园文化节" "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_19",
 "09-办公室礼仪" "http://zjzx.sxsedu.net/ztlm/xysp/jysp/content_121",
 "石膏像的打形" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_67",
 "新兴力量的崛起" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_98",
 "多彩的消费" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_64",
 ...}
```

11.3 DONE 两个视频数据集合比较

CLOSED: *[2018-09-26 Wed 19:08]*

```clojure
(require '[clojure.set :as set])

<<website-old-videos>>
<<website-new-videos>>

(println "--------------------------------------")

(map
 println
 (set/difference
  (set (keys videos-links-old))
  (set (keys videos-links-new))))
```

```
13{nil "http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=172&id=5601",
 "十月主题演讲比赛"
 "http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=171&id=13200",
 "纪念刘和珍君"
 "http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=172&id=6427",
 "走近绍兴黄酒1"
 "http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=172&id=5658",
 "《青春"
 "http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=378&id=11369",
 "人生因设计而美丽2"
 "http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=172&id=5638",
 "3" "http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=172&id=5687",
 "基本站姿组合手位1"
 "http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=172&id=5634",
 "计数原理"
 "http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=172&id=5599",
 "《展望未来我们的明天会更加精彩》14大专城建"
 "http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=378&id=11373",
 "Art"
 "http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=172&id=5685",
 "动画动作修饰2"
 "http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=172&id=5665",
 "绞孔攻螺纹3"
 "http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=172&id=5663",
 "方言与普通话2"
 "http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=172&id=5669",
 "07-课堂礼仪"
 "http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=173&id=5531",
 "色彩空间"
 "http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=172&id=5603",
 "局域网配置综合实训3"
 "http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=172&id=5655",
 "中央电视台-开学第一课"
 "http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=171&id=5564",
 "人生因设计而美丽1"
 "http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=172&id=5637",
 "王金云"
 "http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=378&id=11368",
 "我的学校（专业选报指导）"
 "http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=171&id=6434",
 "人生因设计而美丽3"
 "http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=172&id=5639",
 "重拾自我---16旅服高工"
 "http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=378&id=12761",
 "2017年学第二学期开学第一课"
 "http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=171&id=14247",
 "The"
 "http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=172&id=5598",
 "《拾梦者》"
 "http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=378&id=11375",
 "室内操"
 "http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=171&id=5628",
 "懂得体贴孝敬父母 - 下"
 "http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=173&id=5572",
 "08-校园礼仪"
 "http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=173&id=5532",
 "预防电信网络诈骗专题讲座"
 "http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=171&id=13151",
 "12-社会交往（下）"
 "http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=173&id=5536",
 "绞孔攻螺纹1"
```

```
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=172&id=5661",
"我喜欢我的学校（校园生活纪录）"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=171&id=6438",
"葛豪焙"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=378&id=11376",
"我被十三所学校开除1"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=171&id=5558",
"母亲节"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=171&id=9315",
"礼仪操"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=171&id=6428",
"05-教师的装饰"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=173&id=5529",
"永恒的舞台"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=171&id=6440",
"全国人口普查中小学生一堂课动画短剧"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=171&id=5556",
"纺织贸易实务"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=172&id=5751",
"心理课堂六"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=173&id=15401",
"我和我的学校（浙江省影视大赛第一名作品）"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=171&id=6439",
"社团成果展示片"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=171&id=6436",
"01-教师礼仪概述"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=173&id=5521",
"孝亲尊师电视特别节目(二)"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=171&id=5496",
"绍兴市职教中心建校50周年校庆"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=171&id=5630",
"第三套中学生广播体操-舞动青春分解示范"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=171&id=5612",
"共创美好未来”大合唱比赛决赛"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=171&id=15167",
"商务学区班主任论坛2"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=172&id=5641",
"一个人---15大专动漫"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=378&id=12762",
"合情推理3"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=172&id=5633",
"商务学区班主任论坛1"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=172&id=5640",
"直系电机系统故障排除"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=172&id=5755",
"礼仪教育视频"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=173&id=6429",
"开学第一课（2011年9月1日）"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=171&id=5491",
"14计算机1班"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=378&id=11374",
"2017校园新歌声总决赛"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=171&id=13711",
"《青春誓言》13美术高考2班"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=378&id=11372",
"绞孔攻螺纹2"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=172&id=5662",
"《技能成就梦想》13大专机电徐俊亮、傅卓楷"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=378&id=11371",
"第五十八届田径运动会"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=171&id=5831",
"职教三字经PK赛1"
```

```
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=171&id=5589",
"10-师生关系"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=173&id=5534",
"2017新生军训纪录片"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=171&id=13117",
"《正步人生》绍兴市职教中心2010军训特别节目"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=171&id=5557",
"邹越：让生命充满爱"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=171&id=5560",
"2014预科生活 预习青春"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=171&id=9374",
"基本站姿组合手位3"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=172&id=5636",
"懂得体贴孝敬父母 - 中"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=173&id=5571",
"我被十三所学校开除2"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=171&id=5559",
"合情推理1"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=172&id=5631",
"绍兴市职教中心2011军训特别节目"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=171&id=5440",
"艺术教育特色宣传片"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=171&id=5607",
"04-教师的语言"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=173&id=5528",
"“国家改革发展示范校”建设纪实"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=171&id=8017",
"徐涵宗"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=378&id=11370",
"房志成"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=378&id=11367",
"11-社会交往（上）"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=173&id=5535",
"方言与普通话3"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=172&id=5670",
"我的学校 绍兴市职教中心"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=171&id=6435",
"茶为国饮"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=171&id=5550",
"职教之歌"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=171&id=8288",
"学会节俭，不做奢侈的孩子－上"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=173&id=5566",
"第三套中学生广播体操-舞动青春完整音乐"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=171&id=5613",
"方言与普通话1"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=172&id=5668",
"06-教师的仪表"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=173&id=5530",
"职教三字经PK赛2"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=171&id=5590",
"局域网配置综合实训2"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=172&id=5654",
"合情推理2"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=172&id=5632",
"走近绍兴黄酒2"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=172&id=5659",
"学会节俭，不做奢侈的孩子－下"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=173&id=5569",
"神神慢"
"http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=172&id=5604",
"走近绍兴黄酒3"
```

```
 "http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=172&id=5660",
 "09-办公室礼仪"
 "http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=173&id=5533",
 "新兴力量的崛起"
 "http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=172&id=5600",
 "藏族民族风情"
 "http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=172&id=5605",
 "07级新生军训"
 "http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=171&id=5828",
 "动画动作修饰3"
 "http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=172&id=5666",
 "孝亲尊师电视特别节目(二)成长足迹"
 "http://www.sxti.zj.cn/e/action/ShowInfo.php?classid=171&id=5495",
 ...}
14{"纪念刘和珍君" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_106",
 "走近绍兴黄酒1" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_76",
 "《青春" "http://zjzx.sxsedu.net/ztlm/xysp/zycz/content_129",
 "人生因设计而美丽2" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_90",
 "3" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_94",
 "基本站姿组合手位1" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_84",
 "计数原理" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_97",
 "《展望未来我们的明天会更加精彩》14大专城建"
 "http://zjzx.sxsedu.net/ztlm/xysp/zycz/content_132",
 "Art" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_93",
 "动画动作修饰2" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_86",
 "绞孔攻螺纹3" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_83",
 "方言与普通话2" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_107",
 "07-课堂礼仪" "http://zjzx.sxsedu.net/ztlm/xysp/jysp/content_119",
 "色彩空间" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_101",
 "局域网配置综合实训3" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_75",
 "———有一种爱叫做亲情" "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_37",
 "我被十三所学校开除(2)" "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_28",
 "中央电视台-开学第一课" "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_141",
 "人生因设计而美丽1" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_89",
 "王金云" "http://zjzx.sxsedu.net/ztlm/xysp/zycz/content_128",
 "前厅模拟接待" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_66",
 "人生因设计而美丽3" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_91",
 "重拾自我---16旅服高工" "http://zjzx.sxsedu.net/ztlm/xysp/zycz/content_136",
 "Body" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_63",
 "2017年学第二学期开学第一课" "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_61",
 "The" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_104",
 "《拾梦者》" "http://zjzx.sxsedu.net/ztlm/xysp/zycz/content_138",
 "室内操" "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_21",
 "懂得体贴孝敬父母-下" "http://zjzx.sxsedu.net/ztlm/xysp/jysp/content_112",
 "第二届三字经PK赛" "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_41",
 "08-校园礼仪" "http://zjzx.sxsedu.net/ztlm/xysp/jysp/content_120",
 "12-社会交往（下）" "http://zjzx.sxsedu.net/ztlm/xysp/jysp/content_124",
 "绞孔攻螺纹1" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_81",
 "我喜欢我的学校（校园生活纪录）" "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_51",
 "葛豪焙" "http://zjzx.sxsedu.net/ztlm/xysp/zycz/content_135",
 "孝亲尊师电视特别节目（三）" "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_39",
 "礼仪操" "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_46",
 "05-教师的装饰" "http://zjzx.sxsedu.net/ztlm/xysp/jysp/content_117",
 "永恒的舞台" "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_53",
 "全国人口普查中小学生一堂课动画短剧"
 "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_31",
 "纺织贸易实务" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_68",
 "心理课堂六" "http://zjzx.sxsedu.net/ztlm/xysp/jysp/content_126",
 "我和我的学校（浙江省影视大赛第一名作品）"
 "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_52",
 "社团成果展示片" "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_49",
 "2014预科生活——预习青春" "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_59",
 "01-教师礼仪概述" "http://zjzx.sxsedu.net/ztlm/xysp/jysp/content_113",
 "孝亲尊师电视特别节目(二)" "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_38",
 "绍兴市职教中心建校50周年校庆" "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_139",
 "第三套中学生广播体操-舞动青春分解示范"
 "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_23",
 "商务学区班主任论坛2" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_96",
 "一个人---15大专动漫" "http://zjzx.sxsedu.net/ztlm/xysp/zycz/content_137",
 "合情推理3" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_80",
 "商务学区班主任论坛1" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_95",
 "直系电机系统故障排除" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_69",
 "礼仪教育视频" "http://zjzx.sxsedu.net/ztlm/xysp/jysp/content_125",
 "健康消费\"茶文化知识讲座(2)" "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_33",
 "开学第一课（2011年9月1日）"
 "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_140",
 "14计算机1班" "http://zjzx.sxsedu.net/ztlm/xysp/zycz/content_134",
 "健康消费\"茶文化知识讲座(3)" "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_34",
 "《青春誓言》13美术高考2班" "http://zjzx.sxsedu.net/ztlm/xysp/zycz/content_133",
 "绞孔攻螺纹2" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_82",
 "《技能成就梦想》13大专机电徐俊亮、傅卓楷"
 "http://zjzx.sxsedu.net/ztlm/xysp/zycz/content_131",
 "第五十八届田径运动会" "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_43",
 "职教三字经PK赛1" "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_26",
 "10-师生关系" "http://zjzx.sxsedu.net/ztlm/xysp/jysp/content_122",
 "2017新生军训纪录片" "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_60",
 "健康消费\"茶文化知识讲座(4)" "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_35",
 "《正步人生》绍兴市职教中心2010军训特别节目"
 "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_142",
 "邹越：让生命充满爱" "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_29",
 "基本站姿组合手位3" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_88",
 "懂得体贴孝敬父母－中" "http://zjzx.sxsedu.net/ztlm/xysp/jysp/content_111",
 "合情推理1" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_72",
 "绍兴市职教中心2011军训特别节目"
 "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_40",
 "艺术教育特色宣传片" "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_25",
 "04-教师的语言" "http://zjzx.sxsedu.net/ztlm/xysp/jysp/content_116",
 "“国家改革发展示范校”建设纪实" "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_55",
 "徐涵宗" "http://zjzx.sxsedu.net/ztlm/xysp/zycz/content_130",
 "\"走进\"世贸" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_100",
 "房志成" "http://zjzx.sxsedu.net/ztlm/xysp/zycz/content_127",
 "11-社会交往（上）" "http://zjzx.sxsedu.net/ztlm/xysp/jysp/content_123",
 "我被十三所学校开除(1)" "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_27",
 "方言与普通话3" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_70",
 "茶为国饮" "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_36",
 "职教之歌" "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_56",
 "学会节俭，不做奢侈的孩子－上" "http://zjzx.sxsedu.net/ztlm/xysp/jysp/content_108",
 "绍兴市职教中心" "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_18",
 "第三套中学生广播体操-舞动青春完整音乐"
 "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_24",
 "方言与普通话1" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_105",
 "06-教师的仪表" "http://zjzx.sxsedu.net/ztlm/xysp/jysp/content_118",
 "局域网配置综合实训2" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_74",
 "合情推理2" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_78",
 "走近绍兴黄酒2" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_77",
 "学会节俭，不做奢侈的孩子－下" "http://zjzx.sxsedu.net/ztlm/xysp/jysp/content_109",
 "神神慢" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_103",
 "走近绍兴黄酒3" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_79",
 "2007校园文化节" "http://zjzx.sxsedu.net/ztlm/xysp/xnhd/content_19",
 "09-办公室礼仪" "http://zjzx.sxsedu.net/ztlm/xysp/jysp/content_121",
 "石膏像的打形" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_67",
 "新兴力量的崛起" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_98",
 "多彩的消费" "http://zjzx.sxsedu.net/ztlm/xysp/jxfc/content_64",
 ...}
```

```
----------------------------------------
nil
十月主题演讲比赛
我的学校（专业选报指导）
预防电信网络诈骗专题讲座
我被十三所学校开除1
母亲节
共创美好未来”大合唱比赛决赛
2017校园新歌声总决赛
2014预科生活 预习青春
我被十三所学校开除2
我的学校 绍兴市职教中心
职教三字经PK赛2
孝亲尊师电视特别节目(二)成长足迹
孝亲尊师电视特别节目(一)---有一种爱叫做亲情
```

11.4 DONE 检查新网站视频的缩略图是否存在

CLOSED: *[2018-10-11 Thu 15:00]* **SCHEDULED:** *<2018-10-11 Thu>*

failed thumbnail example: `http://www.sxszjzx.com/ztlm/xysp_4`

\\

(Common\\\-translation240160.gif

```
<li>
    <div class="pic">
        <a href="/ztlm/xysp/xnhd/content_43">
            <img alt=" 第五十八届田径运动会" src="\Content\_Common\Base\img\error-translation_240_160.gif" width="240" height="160">
        </a>
    </div>
    <h3 class="title">
```

```
        <a href="/ztlm/xysp/xnhd/content_43" title="第五十八届田径运动会">第五十八届田径运动会</a>
    </h3>
    <span class="titleBg"></span>
</li>
```

```
(defn extract-page-video-links
  "Extract all video links from web page."
  [url]
  ;; div>a
  (html/select
   (-> (:body
        (http/get url))
       html/html-snippet
       (html/select [:ul.vedioPageList]))
   [:li :div.pic :a]))

(defn extract-page-video-filename
  "Extract video filename from web page."
  [url]
  (last
   (clojure.string/split
    (first
     (html/attr-values
      (first
       (-> (:body (http/get url))
           html/html-snippet
           (html/select [:div.vedioPlayer :div])))
      :data-url))
    #"video/")))
```

```

(defn extract-page-videos-maxpage
  "Get the max page number of http://www.sxszjzx.com/ztlm/xysp."
  [url]
  (Integer.
   (clojure.string/replace
    (re-find #"/*\d+ 页"
             (-> (:body (http/get url))
                 html/html-snippet
                 (html/select [:div.page :span.total])
                 first
                 :content
                 first))
    " 页" "")))

;; (extract-page-videos-maxpage "http://www.sxszjzx.com/ztlm/xysp")
```

```
<<extract page videos>>

(for [n (range 1 (extract-page-videos-maxpage-new
↪   "http://www.sxszjzx.com/ztlm/xysp"))]
  (map
   (fn [a]
     (let [img            (first (html/select a [:img]))
           img-src        (first (html/attr-values img :src))
           title          (first (html/attr-values img :alt))
           video-page-url (str "http://www.sxszjzx.com" (first
           ↪   (html/attr-values a :href)))]
       (when (and (not (clojure.string/blank? img-src))
                  (re-find #"\\Content*" img-src))
```

```
         ;; get video info link
         (print {title video-page-url})
         ;; get video filename
         (println
          (extract-page-video-filename
           (str "http://www.sxszjzx.com" (first (html/attr-values a
           ↪   :href))))))))
   (extract-page-video-links (format "http://www.sxszjzx.com/ztlm/xysp_%s"
   ↪   n))))
```

#+RESULTS[*<2018-10-16 Tue 15:00>* 4c0092ec2c05b70d98b0ba30aae6b8f2ce3c5347]:

11.5 DONE 获取新网站视频页面的视频源信息

CLOSED: *[2018-10-21 Sun 11:02]*

`http://www.sxszjzx.com/ztlm/xysp/zycz/content_129`

```
<<define-new-website-url>>

(defn website-new-video-page-extract-info
  [url]
  (let [video-object (-> (:body (http/get url))
                         html/html-snippet
                         (html/select [:div.vedioPlayer :div])
                         first)]
    {(apply str (html/attr-values video-object :data-title))
     (str website-new-url (first (html/attr-values video-object
     ↪   :data-url)))}))

(website-new-video-page-extract-info
↪   "http://www.sxszjzx.com/ztlm/xysp/zycz/content_129")
```

```
{"《青春梦想》14大专动漫"
 "http://www.sxszjzx.com//upload/sxszjzx/contentmanage/video/2018/09/27/qcmx[自定义].mp4"}
```

11.6 TODO 检查视频是否可以播放 [4/5]

- ☒ 检测视频源文件是否存在?
    - ☒ 像 这个视频连接源文件地址 就是可以打开播放的，但是用新网站的播放器就是不行。
    - ☒ 应该是需要 flash player，但是我检测我安装了 flash plugin 了。

GET `http://www.sxszjzx.com/Content/_Common/Assets/Scripts/swfobject.js?_=1540091532521` GET `http://www.sxszjzx.com/Content/_Common/Base/swf/PowerPlayback.swf`

When I click play button on video frame:

There is a request return 404 result status:

```
http://www.sxszjzx.com/upload/sxszjzx/contentmanage/video/2018/09/27/qcmx[
```

- □ 这里应该给他们报告个 Bug，在 Windows 下请求的 URL 是正确的，在 Linux 下却是缺失部分文件名的，看来是无法正确处理一些文件名中的特殊字符。
- ☒ 在 Linux Firefox/Firefox Developer Edition 两个浏览器下测试，因为其他的浏览器都不支持 Flash。

request headers:

```
Host: www.sxszjzx.com
User-Agent: Mozilla/5.0 (X11; Linux x86_64; rv:62.0) Gecko/20100101 Firefox/62.0
Accept: text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8
Accept-Language: en-US,en;q=0.7,zh-CN;q=0.3
Accept-Encoding: gzip, deflate
Referer: http://www.sxszjzx.com/Content/_Common/Base/swf/PowerPlayback.swf
Cookie: PowerUniqueVisitor=c3f5685d-dfca-46fb-86d9-b58efc5eef6b_2018%2F10%2F21%200%3A00%3A0
```

```
DNT: 1
Connection: keep-alive
```

response headers:

```
HTTP/1.1 404 Not Found
Content-Type: text/html
ServerResponseDuration: 15.6279ms
X-PowerEasy-Version: 1.9.1.2
X-PowerEasy-Product: SiteAzure
X-Frame-Options: SAMEORIGIN
X-XSS-Protection: 1; mode=block
X-Content-Type-Options: nosniff
Date: Sun, 21 Oct 2018 03:15:39 GMT
Content-Length: 1163
```

## Usage

FIXME

## License

Copyright © 2018 FIXME

Distributed under the Eclipse Public License either version 1.0 or (at your option) any later version.

12 DONE Use Git as Version Control System

CLOSED: *[2018-11-23 Fri 08:28]*

13 Generate Project Report

13.1 update project clocktable dynamic block

13.2 **TODO** Org Mode export