腾讯微博抓取分析

1、入口 URL：

http://t.qq.com/yang-shangchuan/

2、抓取动态网页：

#-[?*!@=]

3、URL 范围限制：

-^http://t.qq.com/login.php
+^http://t.qq.com/yang-shangchuan/
-.

4、自定义解析插件 parse-html：

4.1 增加依赖 ivy.xml：

```
<dependencies>
   <dependency  org="org.jsoup"  name="jsoup"  rev="1.7.3"/>
</dependencies>
```

4.2 导出依赖 plugin.xml：

```
<library  name="jsoup-1.7.3.jar"/>
```

4.3 修改代码 HtmlParser.java：

在代码 text = sb.toString(); 后面增加：

```
String parseText = getText(content,
metadata.get(Metadata.ORIGINAL_CHAR_ENCODING));
if(parseText != null){
  System.out.println("忽略原来的文本:"+text);
  text = parseText;
  System.out.println("使用自定义提取的文本:"+text);
}
```

增加方法：

```
private String getText(Content content, String encoding) {
  //提取微博内容的CSS PATH
  String cssPath = "html body.apS div.apSwrap div#topWrap
div#mainWrapper.clear div.main div.AL div#listWrapper
ul#talkList.LC li div.msgBox div.msgCnt";

  try {
    byte[] contentInOctets = content.getContent();
    InputStream in = new ByteArrayInputStream(contentInOctets);
```

```
    Document doc = Jsoup.parse(in, encoding,
content.getBaseUrl());
    Elements elements = doc.select(cssPath);
    StringBuilder str = new StringBuilder();
    for(Element element : elements){
       String text = element.text();
       if(StringUtils.isNotBlank(text)){
          str.append(text).append("\n\n");
       }
    }
    return str.toString();
  } catch (Exception e) {
    LOG.error("Error: ", e);
  }
  return null;
}
```

增加方法：

```
  public static byte[] readAll(InputStream in){
      ByteArrayOutputStream out = new ByteArrayOutputStream();
      try {
          byte[] buffer = new byte[1024];
          for (int n ; (n = in.read(buffer))>0 ; ) {
              out.write(buffer, 0, n);
          }
      } catch (IOException ex) {
          LOG.error("读取文件失败",ex);
      }
      return out.toByteArray();
  }
```

替换 main 方法：

```
  public static void main(String[] args) throws Exception {
    String url =
"http://t.qq.com/yang-shangchuan/?mode=0&id=255108772573&pi=33
&time=1334435957";

    InputStream in = new URL(url).openStream();
    byte[] bytes = readAll(in);

    Configuration conf = NutchConfiguration.create();
    HtmlParser parser = new HtmlParser();
    parser.setConf(conf);
    Parse parse = parser.getParse(
            new Content(url, url, bytes, "text/html", new
Metadata(), conf)).get(url);
```

```
    System.out.println("data: "+parse.getData());

    System.out.println("text: "+parse.getText());

  }
```

5、抓取参数：

urls -solr http://localhost:8983/solr/collection1 -dir data -depth 100 -topN 10000