C#程序设计及应用

唐大仕

dstang2000@263.net

北京大学

课堂作业：网络爬虫

课堂作业：网络爬虫

- 网络信息获取　　DownloadString、DownloadStringAndGuessEncoding
- 正则表达式解析　　DownloadImages、Crawler
- 集合的使用　　Crawler
- 事件的使用　　EventWhenDownload
- 线程更新界面　　ThreadAndWinform
- 写日志文件　　Log

- 也可以做其他网络信息获取的题目（见“text-net.rar”文件）

正则表达式

- 三要素： 字符、数量、位置
- [0-9]{8,11}\b
- 详细：
  - 字符: [0-9] \d [a-zA-Z_] .即任意 [^0-9]非数字 \s空白 \w常规(含汉字)
  - 数量: {0,1}即? {1,} 即 + {0,}即*
  - 位置: ^首 $尾 \b \b
  - 或者 | 分组( ) 如 \.txt|\.docx?
  - 非贪婪（数量后面用个?）如 .*?
- 在线测试 http://tool.chinaz.com/regex/
  - 注意 java/js中 \w与C#中的\w含义不同，前者不包含汉字，后者包含

网络爬虫

- 网页右击，查看源代码
- 在浏览器中用F12 (或Ctrl+Shift+I)
  - 查看network
  - 审查元素