

网页爬取

赵耀

大纲

- ▶ Html标签简要介绍
- ▶ Jsoup的使用



```
<html>
```

```
  <head>  
  </head>
```

```
  <body>
```

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

```
    <span></span>
```

```
    <table></table>
```

```
  </body>
```

```
</html>
```

- <html> 元素是 HTML 页面的根元素
- <head> 元素包含了文档的元（meta）数据
- <title> 元素描述了文档的标题
- <body> 元素包含了可见的页面内容
- <h1> 元素定义一个大标题
- <p> 元素定义一个段落
- <a 这是一个链接 元素定义一个链接
- 元素定义一个图像
- <div>元素是用于分组 HTML 元素的块级元素。用于布局
- 用来组合文档中的行内元素。标签提供了一种将文本的一部分或者文档的一部分独立出来的方式。
- <table>呈现表格化数据

属性	描述
class	为html元素定义一个或多个类名 (classname) (类名从样式文件引入)
id	定义元素的唯一id
style	规定元素的行内样式 (inline style)
title	描述了元素的额外信息 (作为工具条使用)

更多标签和属性

Html 标签大全

<http://www.runoob.com/tags/html-reference.html>

Html属性大全

<http://www.runoob.com/tags/ref-standardattributes.html>

```
<table border="1">
  <tr>
    <th>Header 1</th>
    <th>Header 2</th>
  </tr>
  <tr>
    <td>row 1, cell 1</td>
    <td>row 1, cell 2</td>
  </tr>
  <tr>
    <td>row 2, cell 1</td>
    <td>row 2, cell 2</td>
  </tr>
</table>
```

- **<table>** 表格
- **<th>** 表格的表头
- **<tr>** 表格中的一行
- **<td>** 表格中的单元格

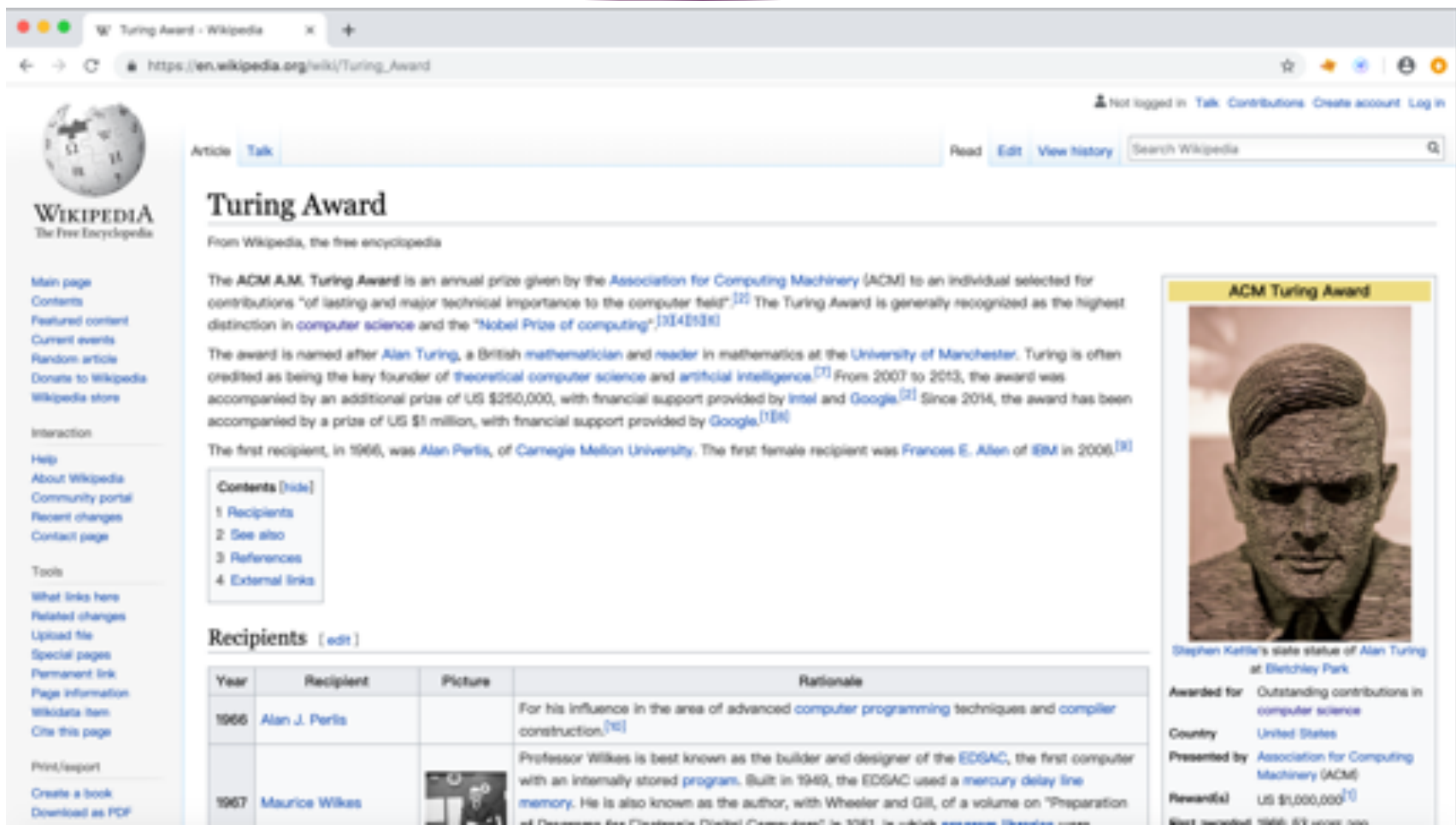
Header 1	Header 2
row 1, cell 1	row 1, cell 2
row 2, cell 1	row 2, cell 2

Jsoup

```
import org.jsoup.Jsoup;
Document doc = Jsoup
    .connect("https://en.wikipedia.org/wiki/Turing_Award")
    .get();
```


如何开始爬取网页？

先用chrome浏览器打开想要爬取的页面



The screenshot shows a web browser window displaying the Wikipedia article for the Turing Award. The browser's address bar shows the URL https://en.wikipedia.org/wiki/Turing_Award. The page features the Wikipedia logo, a sidebar with navigation links, and the main article content. The article title is "Turing Award", and it includes a summary, a detailed description, and a table of recipients.

Turing Award

From Wikipedia, the free encyclopedia

The ACM A.M. Turing Award is an annual prize given by the [Association for Computing Machinery](#) (ACM) to an individual selected for contributions "of lasting and major technical importance to the computer field".^[2] The Turing Award is generally recognized as the highest distinction in computer science and the "Nobel Prize of computing".^{[3][4][5][6]}


The award is named after [Alan Turing](#), a British mathematician and reader in mathematics at the University of Manchester. Turing is often credited as being the key founder of [theoretical computer science](#) and [artificial intelligence](#).^[7] From 2007 to 2013, the award was accompanied by an additional prize of US \$250,000, with financial support provided by [Intel](#) and [Google](#).^[8] Since 2014, the award has been accompanied by a prize of US \$1 million, with financial support provided by [Google](#).^{[1][9]}

The first recipient, in 1966, was [Alan Perlis](#), of [Carnegie Mellon University](#). The first female recipient was [Frances E. Allen](#) of [IBM](#) in 2006.^[3]


Contents [hide]

- [Recipients](#)
- [See also](#)
- [References](#)
- [External links](#)

Recipients [edit]

Year	Recipient	Picture	Rationale
1966	Alan J. Perlis		For his influence in the area of advanced computer programming techniques and compiler construction. ^[10]
1967	Maurice Wilkes		Professor Wilkes is best known as the builder and designer of the EDSAC , the first computer with an internally stored program . Built in 1949, the EDSAC used a mercury delay line memory . He is also known as the author, with Wheeler and Gill , of a volume on "Preparation of Documents for Electronic Digital Computers" in 1961, in which assembly language was

ACM Turing Award



Stephen Kellie's state statue of Alan Turing at Bletchley Park

Awarded for Outstanding contributions in [computer science](#)

Country [United States](#)

Presented by [Association for Computing Machinery](#) (ACM)

Reward(s) US \$1,000,000^[10]

First awarded 1966; 61 years ago

打开开发者工具


视图->开发者->
开发者工具



分析待爬取内容

要爬取的内容为
table.wikitable

The screenshot shows a web browser displaying the Wikipedia page for the Turing Award. The page features a table with the following data:

Year	Recipient	Picture	Rationale
1966	Alan J. Perlis		For his influence in the area of advanced computer programming techniques and compiler construction. ^[9]
1967	Maurice Wilkes		Professor Wilkes is best known as the builder and designer of the EDSAC, the first computer with an internally stored program. Built in 1949, the EDSAC used a mercury delay line memory. He is also known as the author, with Wheeler and Gill, of a volume on "Preparation of Programs for Electronic Digital Computers" in 1951, in which program libraries were effectively introduced. ^[7]

The developer tools are open, showing the HTML structure of the table. The selected element is `table.wikitable`, which has the following structure:

```
<table class="wikitable">
  <tbody>
    <tr bgcolor="#cccc">
      <th style="width:10px">Year
      </th>
      <th style="width:15px">Recipient
      </th>
      <th>Picture
      </th>
      <th>Rationale
      </th>
    </tr>
    <tr>
      <td>1966
      </td>
      <td>Alan J. Perlis
      </td>
      <td>
      </td>
      <td>For his influence in the area of advanced computer programming techniques and compiler construction.[9]
      </td>
    </tr>
    <tr>
      <td>1967
      </td>
      <td>Maurice Wilkes
      </td>
      <td>
      </td>
      <td>Professor Wilkes is best known as the builder and designer of the EDSAC, the first computer with an internally stored program. Built in 1949, the EDSAC used a mercury delay line memory. He is also known as the author, with Wheeler and Gill, of a volume on "Preparation of Programs for Electronic Digital Computers" in 1951, in which program libraries were effectively introduced.[7]
      </td>
    </tr>
  </tbody>
</table>
```

The Styles pane shows the computed styles for the selected element, including `width: 300px` and `background-color: #cccc`.

提取信息

- ▶ 逐级提取图灵奖获得者的信息，见ScrapingExample.java（课堂讲解代码）