2.1 HTML 文档结构与文档树

2.1.1 HTML 文档结构

HTML 文档实际上类似一个 XML 文档，完整的 HTML 文档包含根元素<html>，然后在<html>中包含<head>,<body>等元素，下面是一个典型的 HTML 文档：

```
<html><head><title>The Dormouse's story</title></head>
<body>
<p class="title"><b>The Dormouse's story</b></p>
<p class="story">
Once upon a time there were three little sisters; and their names were
<a href="http://example.com/elsie" class="sister" id="link1">Elsie</a>,
<a href="http://example.com/lacie" class="sister" id="link2">Lacie</a> and
<a href="http://example.com/tillie" class="sister" id="link3">Tillie</a>;
and they lived at the bottom of a well.
</p>
<p class="story">...</p>
</body>
</html>
```

这个文档在浏览器中显示的效果如图 2-2-1 所示。

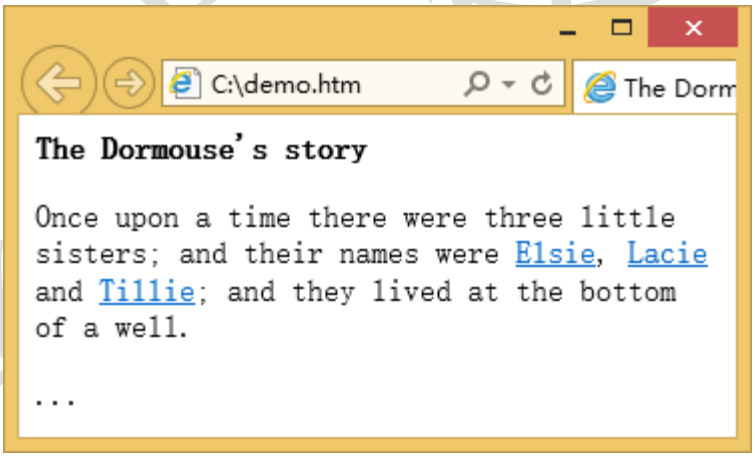

图 2-1-1 HTML 文档

HTML 文档中的<...>的元素称为一个 tag 元素或者 element 元素，例如<html>、<body>、<title>、 、<a>等都是这类的元素，每个 tag 元素都有对应的一个结束元素</...>，例如</html>、</body>、</title>、 、等。

注意 HTML 中的 tag 元素的名称是不区分大小写的，因此<html>、<HTML>、<Html>是一样的，这一点与 XML 不同。

一个 tag 元素可以有很多属性，例如 中的 元素有属性 class，属性值为 title。

特别注意的是 HTML 中除了 tag 元素外，穿插于 tag 元素之间的那些文本也是元素，称为 text 元素，例如<title>The Dormouse's story</title>中的文本 The Dormouse's story 也是一个元素，它是一个 text 文本元素，它的父节点是<title>。

2.1.2 HTML 文档树

HTML 的结构是一个树状结构，在内存中形成一棵树，例如 HTML 结构：

```
<html>
<head><title>Demo</title></head>
<body>
<div>A<p>B</p>C</div>
<span>D</span>
</body>
</html>
```

那么对应的文档树如图 2-1-2 所示。

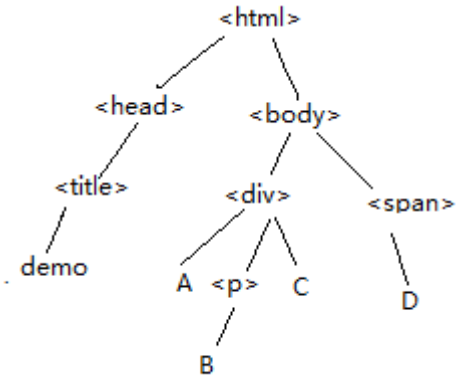

图 2-1-2 HTML 文档树

HTML 文档树结构的概念是十分重要的，它是我们今后查找 tag 节点的重要依据。