SJMMA 2021 C 题

网页标记问题

过去的三十年是互联网蓬勃发展的三十年。如今，互联网上已有超过十亿个网页。本问题中，一个“网页”指一个 URL 字符串，如“www.baidu.com”。我们的问题是，对于一个网页（URL），在互联网上抓取对应的 HTML 文件，通过分析这个文件的内容，来给网页标记一个公司名字。

比如，将网页“www.shanghai.gov.cn”的 HTML 文件进行抓取，我们会看到如下内容（忽略无关内容）：

```
<html>
  <head>
    <meta name="SiteName" content="上海市人民政府">
    <meta name="SiteDomain" content="www.shanghai.gov.cn">
    <meta name="SiteIDCode" content="3100000044">
    <meta name="ColumnName" content="上海市人民政府">
    <meta name="ColumnType" content="首页">
    <title>上海市人民政府</title>
    <meta name="keywords" content="上海,上海市政府">
    <meta name="description" content="中国上海">
  </head>
/html>
```

通过分析这些数据，我们可以简单的使用<title>里的内容，也就是“上海市人民政府”来作为公司名字。这个方法对于很多网站都是有效的。

但是，这个方法在分析一些网页时效果不太好。例如，“www.baidu.com”的 HTML 文件有如下内容

```
<html>
  <head>
    <meta name="description" content="全球最大的中文搜索引擎、致力于让网民更便捷地获取信息，找到所求。百度超过千亿的中文网页数据库，可以瞬间找到相关的搜索结果。">
    <title>百度一下，你就知道</title>t>
  </head>
/html>
```

如果使用<title>，我们就会把公司名字标记为“百度一下，你就知道”。显然，

“百度”或者“百度公司”是一个更好的名字，但却不容易从网页里获得了。

我们希望你的团队完成以下几个任务：

任务一：给定一个 URL，请选择一门计算机语言来编程爬取该 URL 对应的 HTML 文件，通过返回 HTML 文件的<head>里面的<title>的内容，来获得公司名字的第一个版本。我们期待每个团队都能得到相同的结果。

任务二：分析任务一方法的局限性。研究在什么情况下<title>不是很好的公司名字。想办法把你的计算机程序升级到第二版，输出一个更好的公司名字。例如，你可以分析<head>里面的其他内容来互相验证，也可以直接对内容进行提取。请注意这个问题非常的困难，你们的第二版只需要为一小部分网页找到更好的公司名字即可。

任务三：用你们的第二版程序对各种网页进行实验，详细分析并列出你们的算法的效果与不足。这个问题不存在很完美的解法，因此你们的方法肯定也有一定的适用范围和局限性。请指出你们的方法对哪些网页效果较好，哪些网页效果较差。请在这个复杂问题里找到一个相对容易的网页子集来提升效果。

你们的提交需要包括：

1 论文

你的团队所提交的论文应包含 1 页“总结摘要”，其正文不可超过 20 页（总页数限于 21 页）。附录和参考文献应置于正文之后，不计入 20 页之限。

2 源代码

为任务一和任务二提交一个可运行的计算机程序。输入为一个 URL，输出为一个公司名字。提交格式为源代码。

SJMMA2021 Problem C

Problem on Labeling Webpages

The Internet has been growing at an incredible rate in the past 30 years. Today, there are over 1 billion webpages. In this problem, a webpage is defined by a URL string like "www.baidu.com". For a given webpage (URL), our problem is to fetch its HTML file and label it with a company name.

For example, this is the HTML file of webpage "www.www.shanghai.gov.cn" (omitting unrelated):

```
<html>
 <head>
  <meta name="SiteName" content="上海市人民政府">
  <meta name="SiteDomain" content="www.shanghai.gov.cn">
  <meta name="SiteIDCode" content="3100000044">
  <meta name="ColumnName" content="上海市人民政府">
  <meta name="ColumnType" content="首页">
  <title>上海市人民政府</title>
  <meta name="keywords" content="上海,上海市政府">
  <meta name="description" content="中国上海">
 </head>
/html>
```

The simplest way is to use "上海市人民政府" within the tag <title> as a company name. It works for many webpages.

However, this method doesn't work well for many other webpages. For webpage "www.baidu.com", its HTML looks like this:

```
<html>
 <head>
  <meta name="description" content="全球最大的中文搜索引擎、致力于让网民更便捷地获取信息，找到所求。百度超过千亿的中文网页数据库，可以瞬间找到相关的搜索结果。">
  <title>百度一下，你就知道</title>t>
 </head>
/html>
```

The company name would be "百度一下，你就知道". We believe "百度" or "百度公司" is a better company name.

We would like your team to solve the following problems.

Task 1: For a given URL, use any programming language to fetch the HTML file from the Internet and return its <title> tag inside <head>. This is version 1. We expect all teams to get the same result.

Task 2: Please research the limitations of using <title> and implement version 2 with a better company name. For example, you could use the metadata in <head>, or the text in <body>. Since this is a difficult problem, you can focus on a subset of all webpages.

Task 3: Apply your version 2 to various webpages and analyze its pros and cons. There won't be a perfect solution for this problem. So just find a subset of this problem to solve.

Submission

1 paper

Your solution paper should include a 1-page Summary Sheet. The body cannot exceed 20 pages for a maximum of 21 pages with the Summary Sheet inclusive. The appendices and references should appear at the end of the paper and do not count towards the 21 pages limit.

2 source code

Submit a runnable computer program for **task one** and **task two**. Input as a URL, output as a company name. The submission format is source code.