第二章 万维网网页信息的表达及解析作业

提交日期：3 月 31 日
提交形式：作业以 word 或 pdf 形式，提交到助教邮箱，命名方式为“姓名”+“学号”
提交邮箱：邮箱地址 cwt@bupt.edu.cn ，邮件主题需包含 "信息表达"这四个字作为关键（以方便自动归档），提交成功后会有自动回复。

2.1 万维网架构及网页表达

1、【填空题】
- ✧ URL 标志分布在整个因特网上的万维网文档，URI 和 URL 是两个相近的概念，但 URL 只是 URI 的一种，那么 URI 和 URL 的**中文全称**分别是_____和_____。
- ✧ 实现万维网上各种超链的链接，采用的协议为_____
- ✧ _____标识语言使各种万维网文档都能在因特网上的各种计算机上显示出来，同时使用户清楚地知道在什么地方存在着超链接。

2、【选择题】下面关于“投票算法模型”的描述错误的是（ ）
A、通过定义一些固定不变的规则，对每一个文本块打分。
B、得分最高的文本被认为是正文的可能性足够大，可以接受。
C、可以通过足够多的网页对规则进行反馈。
D、一些网页的正文段抽取错误可能是某个规则器打分不合理导致的。

3、【简答题】网页结构化的意义？

2.2 网页信息抽取

1、【填空题】
- ✧ 万维网的静态结构式一个互相联通的_____，网页可以看作_____，网页之间的链接可以看作_____。
- ✧ 网页具有三大特征，我们将网页的诞生到消亡称为_____，由于 HTML 语言描述的网页是一种半结构化的数据，所以网页具有_____，而对于隐藏的网页则体现了网页的_____。

2、【简答题】网页抓取策略采用深度优先策略还是宽度优先策略？为什么？

4、【简答题】MD5 是什么算法？在网页抓取过程中用它来做什么？

5、【实践题】网上查找一个 I-Match、Shingle 或者 Jaccard 系数应用简例（任意一个都可以），按照自己的理解简述应用过程（用自己的语言描述，不要完全 copy）（给出参照网页的网址）