题目 102

要求：

1、用 python 将所给数据集中的样本数据，针对不同的网页特征进行特征提取，分别统计出特征数据集；

2、特征数据预处理：对所有样本进行特征提取后，对每个特征的特征值进行归一化处理（这里可采用线性归一化处理）和向量化（文本特征可采用 TF/TF-IDF 等向量化）处理，将特征值缩放至区间[0,1]。

3、经过归一化处理后，将不同类的数据集进行特征融合，所有网页样本均表示为一个 N 维的特征向量 $f=[u_1,...,u_7,h_1,...,h_{10},j_1,...,j_{18},...]^T$。

一、URL 特征提取规则：

U1：URL 的长度（含 http://）

U2：URL 中点的个数

U3：URL 中数字的个数

U4：URL 中特殊字符的个数（特殊字符：“#、@、_、-、&、/、=”）

U5：URL 中子域名的个数

U6：URL 路径深度（层数）

U7：URL 中是否包含 IP 地址（为二值特征，包含 IP 地址时为 1，否则为 0。）

Flag 为：n，d，p。（可将其表示：n 为 0，d 为 1，p 为 2）。

表 1：URL 特征：

Id	U1	U2	U3	U4	U5	U6	U7	Flag
1								
2								
3								
...								
n								

二、HTML 特征提取规则：

H1：HTML 长度

H2：网页文本内容占整个 HTML 文件长度的比重

H3：URL 出现次数

H4：隐藏标签的数量（主要考察标签的 size 属性(包括 width 属性和 height 属性)、hidden 属性、display 属性以及 visible 属性）

H5：<iframe>标签数量

H6：<meta>标签个数

H7：<image>标签数量

H8：<script>标签数量

H9：<object>标签数量

H10：<embed>标签数量

Flag 为：n，d，p。（可将其表示：n 为 0，d 为 1，p 为 2）。

表 2:：HTML 特征：

Id	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	Flag
1											
2											
3											
...											
n											

三、JavaScript 特征提取规则：

J1：eval 函数调用次数

J2：长字符串数量(长度大于 30)

J3：JavaScript 代码长度

J4：可疑文件出现次数。(可疑文件主要指.exe、.ini、.dll、.tmp 等后缀文件)

J5：setTimeout 函数数量

J6：setInterval 函数数量

J7：window.location 函数数量

J8：Window.open 函数数量

J9：scriptObject.src 数量

J10：scriptObject.setAttribute 数量

J11：scriptObject.innerHTML 数量

J12：document.location 函数数量

J13：document.cookie 函数数量

J14：document. Write()函数个数

J15：unescape()和 escape()函数个数

J16：split()和 replaoe()函数个数

J17：JavaScript 代码占 HTML 文档的比率

J18：Navigator 字符串出现次数

Flag 为：n，d，p。（可将其表示：n 为 0，d 为 1，p 为 2）。

表 3：JavaScript 特征：

Id	J1	J2	J3	...	J16	J17	J18	Flag
1								
2								
3								
...								
n								

四、文本特征：

提取页面文本特征 W1，W2，...，Wn。

(1) 文本抽取:首先对网页进行 HTML 解析，并利用 Xpath 分别提取 Web 网页中<title>标签里面的网页标题、<meta>标签里面的 keyword 和 description 以及<script>标签中 alert 方法中的指定消息，并将抓取出来的文本信息按行存入 txt 文本中。

如：抽取的部分文本如下所示。

中国工商银行
山西建行
Dropbox
Google Docs
中国工商
淘宝网-魔兽世界300元=3w发货
QQ
本店支持信用卡直接付款
工商银琛
工商手机网
掌上门户-中国移动
工商
账户充值 - 支付宝 - 网上支付 安全快速!
工商银行
Google Docs Drive
Apple - Apple ID
Alibaba Manufacturer Directory - Suppliers, Manufacturers, Exporters &
支付宝收银台

(2) **分词**:对于抓取出来的网页文本信息，需要对其进行中文分词来构建特征词袋，从而获得钓鱼网页的关键词信息。可采用 Jieba 分词的精确模式对文本进行分词，精确模式可以将中文句子以最精确的方式进行分割，非常适用于文本分析。

(3) **去重及停用词并排序**:首先对分词后的结果利用停用词词表对其进行去停用词操作。由于分词之后得到的词语太多，所以对分词后的文本关键词进行去重，并基于 TF 的策略对词经行词频排序。

例如，统计出的特征如下：

特征词汇	词频统计	归属类型
银行	378	银行
淘宝	231	电商
工商	209	银行
Apple	153	电商
信用卡	106	支付

特征词汇	词频统计	归属类型
网上投注	1863	博彩信息
娱乐平台	1252	娱乐信息
代孕	776	非法医疗
中国友链网	226	非法链接
毕业证	80	非法办证

表 4：文本特征，例如：

根据上面统计，W1：网上投注（对应 Flag 为 p），W2：代孕（对应 Flag 为 p），W3：银行（对应 Flag 为 d）等。**Flag 为：n，d，p。（可将其表示：n 为 0，d 为 1，p 为 2）。**

Id	W1	W2	...	Wn	Flag
1					
2					
...					
n					

五、将以上所有特征数据特征融合

表 5：所有网页特征

Id	U1	...	U7	H1	...	H10	J1	...	J18	W...	Flag
1											
2											
3											
...											
n											