题目二：DGA 域名识别和聚类分析

设计说明

DGA（Domain generation algorithm，域名生成算法）是一种利用特定种子字符，结合加密算法，进而生成一系列伪随机恶意域名的方法。恶意软件可使用 DGA 逃避域名黑名单的检测。

本题中，选手需从给定的数据包中通过数据分析发现其中存在的 DGA 域名，并自行设计算法完成 DGA 域名的家族聚类。

数据说明

提供的流量为 pcap 格式数据包，大小为 2.1GB。

持续时间

本题的答题时间为 2019 年 4 月 9 日至 2019 年 4 月 28 日。

提交形式和文件格式

选手提交 1 个 csv 文件、Writeup、程序代码，统一使用 zip 格式打包提交。注意，压缩文件请勿加密，并确保可以使用 unzip 命令无参数解压。请勿压缩文件夹。

提交文件的命名规则和内容如下：

- dga.csv：DGA 域名信息，**列名：domain,family**
  - 每一行包含两列：域名 **domain**、该域名的家族代码 **family**。
  - 在书写域名时，请省略末尾代表 DNS 根的点号。
  - 数据包中包含的 DGA 家族种类数量不确定，域名的家族代码请使用从 1 开始递增的数字（1、2、3、4、…、10、…），使用其他家族代码将不得分。
  - 同一个域名不会同时属于两个及以上家族；若提交的答案中某个域名同时出现在多个家族中，将以文件中第一条出现的记录为准。
  - 请将你认为的所有 DGA 域名都列在本文件中，并使用家族代码进行区分；请不要将非 DGA 域名列在本文件中。
- Writeup 和代码
  - 所有的队伍都需要提交解题 Writeup 和代码，详细说明解题步骤和思路。Writeup 和代码将由组委会审核并作为成绩评定依据，应详细包含解题思路、数据分析步骤、聚类算法、代码功能、代码运行环境和说明、每一类别的特征等内容。
  - Writeup 请使用 PDF 格式。代码请打包为 zip 格式文件。
  - 每支队伍在 2019 年 4 月 28 日答题截止前，最后一次的提交包含 Writeup 和代码即可。未提交 Writeup 和解题代码将影响题目最终得分。

- 示例（见 example_dns2.zip）

在如下的数据包中，我认为 www.example.com、www.example1.com、www.example2.com 是 DGA 产生的域名，www.example3.com 是正常域名。并且，www.example.com 和 www.example2.com 属于同一个家族，www.example1.com 属于另外一个家族。

```
123 131.641335    192.168.56.1    192.168.56.128    75 DNS    Standard query 0x8a7a A www.example.com
125 136.065661    192.168.56.1    192.168.56.128    76 DNS    Standard query 0xba7d A www.example1.com
129 138.780224    192.168.56.1    192.168.56.128    76 DNS    Standard query 0xc306 A www.example2.com
131 141.450533    192.168.56.1    192.168.56.128    76 DNS    Standard query 0x5463 A www.example3.com
```

我提交的 dga.csv 内容应为：

domain	family
www.example.com	1
www.example2.com	1
www.example1.com	2

我最后一次提交的 zip 文件应该包含上面这个 csv 文件、一个 Writeup PDF 文件，和一个代码 zip 文件。之前用于阶段评分的提交，可以只包含上面这个答案 csv 文件。

提交规则

选手在本题的持续时间内，提交次数不限。

主办方将不定期进行阶段评分，将队伍本题的得分以及当前得分排名范围反馈给选手。在答题截止前，选手可以根据阶段反馈情况调整答案。

本题最终得分为选手在 2019 年 4 月 28 日 24 时答题截止前**最后一次提交结果**的得分。

评分规则

本题的得分由两部分组成，每一部分各占 50%，累加得到本题得分。

第一部分：DGA 域名识别准确性

计算公式为：

$$Score_1 = max\left(\frac{1}{N} \cdot w \cdot \sum_{k=1}^{m}([D_k \in FTrue]), 0\right)$$

各项符号的含义：

$FTrue$：标准答案集

N：标准答案集（$FTrue$）的长度

m：选手提交的域名列表长度

D_k：选手提交的域名列表中，第 k 个域名

[]：其中为判断语句。语句成立时，值为 1；不成立时，值为-1

$max()$：取最大值函数

w：积分权重

第二部分：DGA 家族聚类准确性

计算公式为：

$$Score_2 = \frac{1}{N} \cdot \sum_{k=1}^{m} \frac{count\left(F_{k \in F} \cap FTrue_{k \in FTrue}\right)}{count\left(F_{k \in F}\right) + count\left(FTrue_{k \in FTrue}\right) - count\left(F_{k \in F} \cap FTrue_{k \in FTrue}\right)}$$

各项符号的含义：

$count()$：计算集合长度的函数

N：标准答案集的长度

$FTrue_{k \in FTrue}$：标准答案中，第k个域名对应的家族集合

m：选手提交的域名列表长度

$F_{k \in F}$：选手提交的答案中，第k个域名对应的家族集合

其他

主办方拥有对题目内容和评分规则的解释权，并保留在极端情形下更正题目内容和评分规则的权利。