## 结构分析

在csdn.sql和yahoo.sql中分别有6428632条、453491条记录。其中可用的记录分别有和442837条 使用python（/结构/structure.py）得到下表

|  |  |  |
| --- | --- | --- |
|  | Csdn | Yahoo |
| 总数 | 6428620 | 442837 |
| 纯数字 | 2894247（45.0%） | 26080（5.9%） |
| 纯字母 | 794274 （12.4%） | 153410 （34.6%） |
| 数字+字母 | 6194963 （96.4%） | 430223 （97.2%） |
| 包含特殊符号 | 233657 （3.6%） | 12614 （2.8%） |
| AANN型（A为字母，N为数字，例MAGA2020，纯数字和纯字母不算） | 1830676 （28.5%） | 185310（41.8%） |
| NNAA型（A为字母，N为数字，例321stop，纯数字和纯字母不算）） | 415405（6.5%） | 24962（5.6%） |

从上表可以看出，国内用户（csdn）相比于国际用户（yahoo）更喜欢用纯数字做密码，国际用户（yahoo）比国内用户（csdn）更喜欢用字母做密码。且绝大部分的用户都不会使用特殊字符作为密码。

使用mysql（/结构/mysql.doc）对csdn.sql和yahoo.sql查寻出现次数最多的密码，结果存在/结构（csdn\_max.txt,yahoo\_max.txt）

使用AC自动机（/结构/ac自动机.py）统计csdn.sql和yahoo.sql出现次数最多的字母和数字，结果存在/结构，其中字母（csdn\_alpha.txt,yahoo\_alpha.txt），数字（csdn\_number.txt,yahoo\_number.txt）

## 英文和拼音分析

对密码中英文字母所组成的单词和拼音进行分析。为了进一步关注问题本身，对所有密码仅保留字母部分（使用python中"".join(filter(str.isalpha,s))命令），除去不包含任何字母的密码，在csdn.sql和yahoo.sql中分别还有3303221，413309条记录，并分别使用AC自动机和动态规划的方法对它们进行英文单词和拼音的分析。

#### 英文

单词表（/英文/word2.txt）中包含常用单词，使用频率从高到低排列

AC自动机（/英文/ac自动机.py）:

利用单词表统计所有单词出现的次数，只要出现就计数。

该方法的统计会造成一种“重复统计”，例如单词mango会被使得man，go，mango三个单词的出现次数都增加1

对csdn和yahoo.的分析结果分别存放在（/英文/csdn\_output\_AC.txt）和（/英文/yahoo\_output\_AC.txt）

动态规划（/英文/动态规划.py）:

利用常用单词的使用频率和zipf定律对每个单词给一个权值，然后使用动态规划的思想将一个密码进行权值最大的分割，密码会被看成“一句话”，被算法切割为若干个互不包含的部分，如果切割后的每个部分都是单词表中的单词，则这些切割后单词的出现次数增加1。

该统计方法会造成一种“遗漏统计”,例如IlikePKQ，虽然I和like都是单词，但是因为PKQ不是单词，所以I和like也不会计入出现次数。

介于密码中有大量IloveXXX这种类型，其中XXX很可能是人名，所以加入1000个常用英文名（/英文/name2.txt）进入单词表（/英文/word2.txt）中，同时考虑大小写，一个名字保存三种形式（如ALICE，Alice，alice）。

对csdn和yahoo的分析结果分别存放在（/英文/csdn\_output\_DP.txt）和（/英文/yahoo\_output\_DP.txt）

#### 拼音

与英文分析类似，将单词表更换为（/拼音/pinyin2.txt）

AC自动机（/拼音/ac自动机.py）：

对csdn和yahoo.的分析结果分别存放在（/拼音/csdn\_output\_AC.txt）和（/拼音/yahoo\_output\_AC.txt）

动态规划（/拼音/动态规划.py）：

对csdn和yahoo的分析结果分别存放在（/拼音/csdn\_output\_DP.txt）和（/拼音/yahoo\_output\_DP.txt）

#### 总结

对于单词表中的长串（长单词或长拼音）来说，AC自动机的统计要更精准一些，一方面动态规划会“遗漏统计”，另一方面长串很难成为其他串的子串，其在语义中出现大概率会作为一个单独的串（单词或拼音）出现，所以此时不太容易造成“重复统计”；而单词表中的短串，例如a，很容易成为其他串（如man）的子串,而从语义上来分析应该要把man看成一个整体而不是m,a,n三个独立的单词，AC自动机会把a出现的所有地方都统计一遍造成“重复统计”，而动态规划算法进行的条件是密码分割的各个部分都要属于单词表，因此a只有被认定为一个独立的单词或拼音时才计入统计，所以对于短串（短单词或短拼音）来说，动态规划的统计要更具备参考意义。