单表代换统计分析工具

PB21000004　吴越

1　实验原理与算法

单表代换密码（Monoalphabet Cipher）的定义为：每个明文字母按照代换表（密钥）替换为一个新的字母。对其的分析主要基于单表代换以下几个性质：

1. 单表替换不改变相应字母的出现频率

   由于自然语言中存在高冗余，因此每个字母出现频率不同，存在统计性质。以英语为例，在英语中

   - e的频率最高
   - 其次是t,a,o,i,n,s,h,r
   - z,q,x,j的频率近似为0

   因此当密文文本量较大时，对代换后的密文做字母出现频率的分析有可能破解单表代换。

2. 单表代换不改变前后字母存在的关联性

   由于文本具有连接特征，前后字母存在一定关联性，因此根据字母的常用组合也能分析单表代换，同样以英语为例，在英语中常见的连接特征包括：

   - q后几乎百分之百连接着u
   - x前几乎总是i和e，只在极个别情况下是o和a
   - e和e之间，r的出现频率很高

   因此在已知某些字母的代换规则，再结合文本的连接特征，就能推测出其他字母的代换规则。

综上分析，单表代换统计分析具体的算法如下：

① 统计密文中各个字母的出现频率
② 将统计结果与自然语言频率表对比，确定部分密钥
③ 再结合多字母组合和连接特征，确定部分密钥
④ 从语义上，猜测其它密钥，验证破译结果

2　代码实现

2.1　统计密文中各个字母的出现频率

为了统计密文中字母的出现频率并将其可视化展示出来，这里利用了Python中的ntlk库统计频率以及matplotlib库将结果可视化展示出来，具体代码实现如下：

```
# 分析密文中每个字母的出现频率，并将结果可视化
freq = nltk.FreqDist(text)
top_five = freq.most_common(5)
letters = [x[0] for x in top_five]
counts = [x[1] for x in top_five]
fig, ax = plt.subplots()
ax.bar(letters, counts)
ax.set_title('Top 5 Most Frequent')
ax.set_xlabel('Letters')
ax.set_ylabel('Counts')
canvas = FigureCanvas(fig)
layout = QVBoxLayout()
layout.addWidget(canvas)
```

首先调用了ntlk库中的nltk.FreqDist方法来分析密文，可以自动获得一个统计有各个字符出现频率的列表，为方便展示，将列表中频率前5位的数据利用matplotlib库中的方法绘制成柱状图展示出来，有助于用户确定出现最为频繁的字母的代换规则。

2.2 结合自然语言频率统计给出破译建议

由于单表替换不改变相应字母的出现频率，因此统计密文字母频率后与已知的英语中字母出现的频率相匹配就很有可能是代换规则，具体代码如下：

```
freq = nltk.FreqDist(text)
letter_freq = ['e', 't', 'a', 'o', 'i', 'n', 's', 'r', 'h', 'l', 'd', 'c', 'u', 'm', 'f', 'p', 'g', 'w', 'y', 'b', 'v', 'k', 'x', 'j', 'q', 'z']

suggestion+="I.词频破译建议：\n根据词频统计，有可能的单表代换对应如下：\n"
for item in freq:
    # 在英文字母的出现频率列表中查找对应的字母
    letter = letter_freq.pop(0)
    suggestion+=(item+ ':' + letter + ';')
```

即根据统计的freq与已知的letter_freq按大小关系匹配，作为破译建议输出。

2.3 结合多字母组合和连接特征给出破译建议

这里选取了最为常见的字母组合：the（最常用三字组合）和th（最常用二字组合）以及常见的连接关系：1. q后几乎百分之百连接着u，2. x前几乎总是i和e，只在极个别情况下是o和a，3. e和e之间，r的出现频率很高来给出破译建议。具体代码实现如下：

```
suggestion+="\n\nII.连接特征建议："
suggestionlist=[[],[],[],[],[]]
suggestionword=["\n1.the（最常用三字组合）：","\n2.th（最常用二字组合）：","\n3.q后几乎百分之百连接着u：","\n4.x前几乎总是i和e：","\n5.e和e之间，r的出现频率很高："]
for i in range(len(decodelist)):
    word=decodelist[i]
    for j in range(len(word)):
        if(word[j]=='t'and j+2<=len(word)-1 and word[j+2]=='e'):
            suggestionlist[0].append(wordlist[i][j:j+3]+'->the;')
        if(word[j]=='t' and j!=len(word)-1):
            suggestionlist[1].append(wordlist[i][j:j+2]+'->th;')
        if(word[j]=='q' and j!=len(word)-1):
            suggestionlist[2].append(wordlist[i][j:j+2]+'->qu;')
```

```
        if(word[j]=='x' and j!=0):
            suggestionlist[3].append(wordlist[i][j-1:j+1]+'->ix/ex;')
        if(word[j]=='e'and j+2<=len(word)-1 and word[j+2]=='e'):
            suggestionlist[4].append(wordlist[i][j:j+3]+'->ere;')
for i in range(len(suggestionlist)):
    suggestionlist[i]=list(set(suggestionlist[i]))
    suggestion+=suggestionword[i]+''.join(suggestionlist[i])
```

以the（最常用三字组合）为例，当用户确定尝试了t字母与e字母的代换规则后，如果代换后的密文中出现形如“t#e”的组合，则将建议：这三个字母可能被代换为the组合加入到建议列表中。其它的建议产生同理，最后拼接字符串得到结合多字母组合和连接特征给出的破译建议。

2.4 结合单词语义给出破译建议

这里选取了英文中常用的几千个单词，用户根据需要也可以自行添加单词到wordlist.txt中。思路是如果代换后的单词与单词库中的单词相差较小，就提示可能的代换关系（例如：eithej有可能对应为either）。具体代码实现如下：

```
suggestion+="\n\nIII.字典建议：\n根据词义判断，有可能的对应词对应如下："
suggestionlist=[]
for i in range(len(decodelist)):
    word=decodelist[i]
    for dic in dictionary:
        if(len(dic)==len(word) and self.count(dic,word)>=len(word)-1):
            suggestionlist.append(wordlist[i]+'->'+dic+';')
suggestionlist=list(set(suggestionlist))
suggestion+=''.join(suggestionlist)
```

即如果代换后的单词与单词库中的单词相差较小，只有一个及以下字母的不同差异，就给出提示有可能这个单词对应的就是字典中的单词，将该建议加入到建议列表中。最后拼接字符串得到结合单词语义给出的破译建议。

2.5 根据用户所给密钥字设置代换表

要根据用户需求，依照用户所给密钥字设置代换表并给出代换后的密文，具体代码如下：

```
# 依照用户所给密钥字设置代换表
try:
    key=key.split('-')
    assert key[0] in string.ascii_lowercase
    assert key[1] in string.ascii_lowercase
    self.match_dict[key[0]]=key[1]
except:
    pass
# 根据代换表密文代换
def replace_string(self,s, dct):
    word=''
    for i in range(len(s)):
        if s[i] in string.ascii_lowercase:
            word+=dct[s[i]]
        else:
            word+=s[i]
    return word
```

即首先根据用户的输入确定代换表match_list，利用assert保证用户输入数据的有效性。代换密文时对密文中每个字母根据match_list中的对应关系做字母替换得到最终结果。

3　实验测试

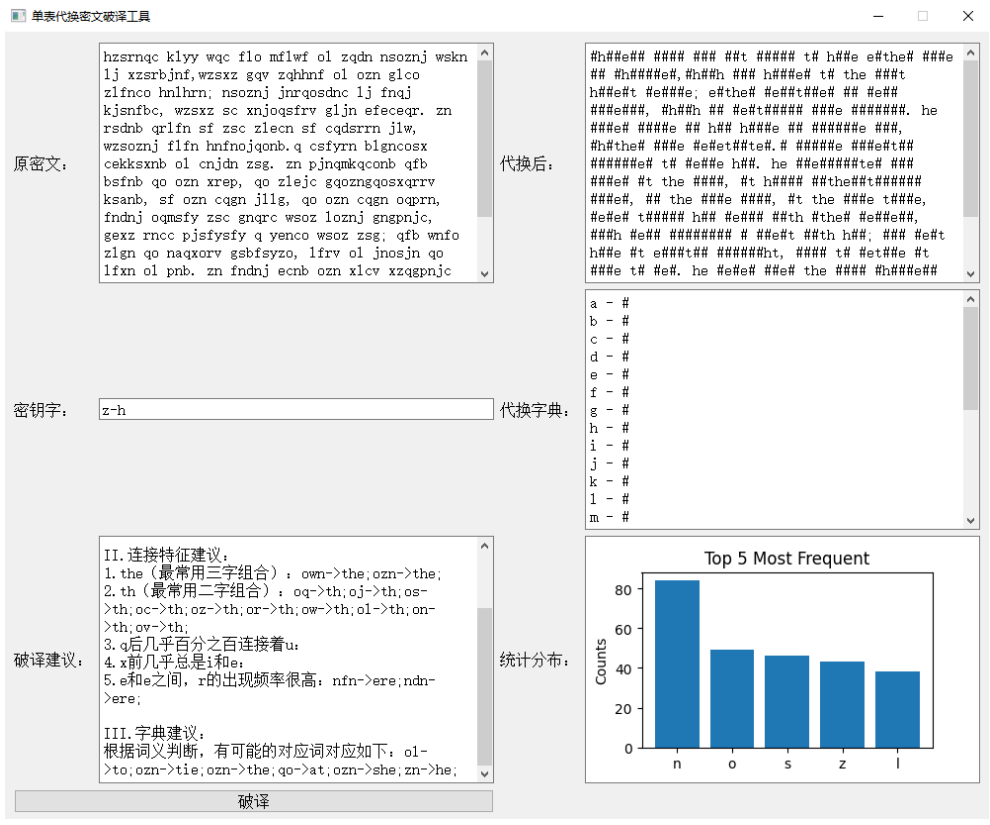
以PPT中给出的密文解密为测试例：

hzsrnqc klyy wqc flo mflwf ol zqdn nsoznj wskn lj xzsrbjnf,wzsxz gqv zqhhnf ol ozn glco zlfnco hnlhrn; nsoznj jnrqosdnc lj fnqj kjsnfbc, wzsxz sc xnjoqsfrv gljn efeceqr. zn rsdnb qrlfn sf zsc zlecn sf cqdsrrn jlw, wzsoznj flfn hnfnojqonb.q csfyrn blgncosx cekksxnb ol cnjdn zsg. zn pjnqmkqconb qfb bsfnb qo ozn xrep, qo zlejc gqozngqosxqrrv ksanb, sf ozn cqgn jllg, qo ozn cqgn oqprn, fndnj oqmsfy zsc gnqrc wsoz loznj gngpnjc, gexz rncc pjsfysfy q yenco wsoz zsg; qfb wnfo zlgn qo naqxorv gsbfsyzo, lfrv ol jnosjn qo lfxn ol pnb. zn fndnj ecnb ozn xlcv xzqgpnjc wzsxz ozn jnkljg hjldsbnc klj soc kqdlejnb gngpnjc. zn hqccnb onf zlejc leo lk ozn ownfov-klej sf cqdsrrn jlw, nsoznj sf crnnhsfy lj gqmsfy zsc olsrno.

将其输入原密文框，点击破译，得到词频统计以及破译建议如下：

根据通过词频统计给出的破译建议，我们猜测最有可能的代换 n - e ， o - t 。将其输入密钥字框再次破译得到新的破译建议以及代换后的密文如下：

根据更新给出的连接特征建议 ozn - the ，我们猜测 z - h ，将其输入密钥字框再破译，结果如下：



这里根据更新的字典建议 ol - to 猜测 l - o，根据qo - at 猜测 q-a 同理输入到密钥字框中。建议再次更新。

同理重复上述过程，根据更新的破译建议猜测代换关系，更新密钥字后参考更新后的破译建议，最终的结果如下：

即完成了最终的破译，代换字典显示于右边界面。# 表示密文中未出现的字母。

4　结果分析

最终的破译结果如下：

代换字典：

a - x　b - d　c - s　d - v　e - u　f - n　g - m　h - p　i - #　j - r　k - f　l - o　m - k

n - e　o - t　p - b　q - a　r - l　s - i　t - #　u - #　v - y　w - w　x - c　y - g　z - h

表示密文中未出现的字母，代换后的密文如下：

phileas fogg was not known to have either wife or children,which may happen to the most honest people; either relatives or near friends, which is certainly more unusual. he lived alone in his house in saville row, whither none penetrated.a single domestic sufficed to serve him. he breakfasted and dined at the club, at hours mathematically fixed, in the same room, at the same table, never taking his meals with other members, much less bringing a guest with him; and went home at exactly midnight, only to retire at once to bed. he never used the cosy chambers which the reform provides for its favoured members. he passed ten hours out of the twenty-four in saville row, either in sleeping or making his toilet.

语义上连贯清晰，文本可读性较高，可以认为完成了破译任务。

5　实验总结

本次单表代换统计分析工具主要基于两点性质：1.单表替换不改变相应字母的出现频率，2. 单表代换不改变前后字母存在的关联性。具体实现上通过统计密文中各个字母的出现频率与自然语言频率对比，再结合多字母组合和连接特征，最后从英文单词语义上，猜测代换的密钥，为用户提供三个方面的破译建议。最后根据用户提供的密钥字，由用户验证破译结果，完成对单表代换的分析。