UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE ESTUDIOS SUPERIORES ARAGÓN

Ingeniería en Computación

DISEÑO Y ANALISIS DE ALGORITMOS

Tarea 7 Descubrir en que idiomas están escritos los textos del proyecto que les corresponde

Profesor: Marcelo Pérez Medel

Manzano Ponce Josué

Olvera Martínez Leonardo

Ortega Batún Luis Fernando

Grupo: 1507

Fecha: jueves 16 de noviembre de 2023

Descubrir en que idiomas están escritos los textos del proyecto que les corresponde

En clase hicimos un programa para hacer el histograma de un texto, haga el histograma de cada texto y ordénelo de mayor a menor, investigue las frecuencias de los idiomas:

Inglés

Francés

Alemán

Portugués

Italiano

haga sus histogramas ordenados de mayor a menor y compárelo con sus textos, encuentre la mejor suposición de en qué idiomas se encuentran.

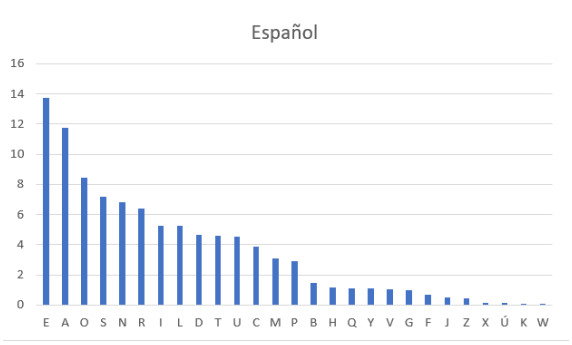
Nosotros pertenecemos al equipo numero 17, los textos que se nos fueron asignados para descifrar fueron los siguientes:

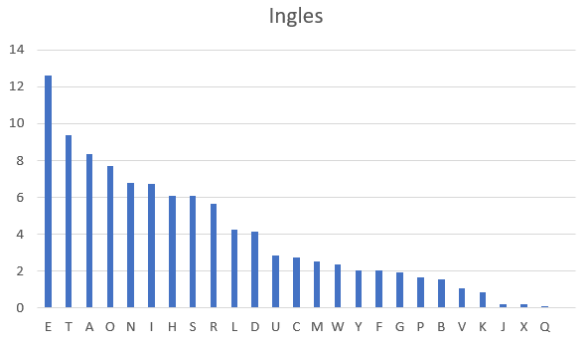
Texto 1: Oitlt xlt onb ovjtc bd jtbjqt nib ngqq otqq vbr oixo vbr zxmmbo yxkt x egddtltmzt gm oigc nblqe: oibct nib xlt xdlxge ob olv xme oibct nib xlt xdlxge vbr ngqq crzztte Crzztcc gc qgkgmu vbrlctqd, qgkgmu nixo vbr eb, xme qgkgmu ibn vbr eb go Nitmtftl vbr dgme vbrlctqd bm oit cget bd oit yxwblgov, go gc ogyt ob jxrct xme ltdqtzo Kttj vbrl dxzt xqnxvc obnxle oit crmcigmt xme cixebnc ngqq dxqq htigme vbr Gd vbr ixft tftlvoigmu rmetl zbmolbq, vbr xlt mbo ybfgmu dxco tmbrui Go gc erlgmu brl exlktco ybytmoc oixo nt yrco dbzrc ob ctt oit qguio

Texto 2: Jz bzz qwjjd ped nevvd le vkzhd eo mod hdlvd kelwkhd odvmkdjw d neknd hwove naejzpwvke qe qelvdobd. Vmvvd jd bzod w qwjepevdvd odvmkdjpwovw qd gdkwve kznnezlw z kegeqe gwoqee, umeoqe gwk jd pdffezk gdkvw qwfje doepdje ozo lzoz ownwlldkew fdccew, ow lgwnedje kwneolezoe. Lzjz rwjeoe dcevdoz j'oend gdkvw naemld, naw ad qe rdjvz jd lvwlld vkelvwbd qwe jzkz znnae kdllwfodve. Lzoz dggwod dkkehdvz eo Evdjed wq w mo jmzfz qdhhwkz eondovwhzjw. Ewke az helevdvz ej nwovkz qe Kzpd w qwhz qekw naw w qdhhwkz mod nevvd nzo pzjve pzompwove dovenae, eordvve lzoz keldjwove djj'wgznd Kzpdod. Az helvz ej Nzjzllwz w qwhz dppwvvwkw naw gwoldhz rzllw gem gennzjz, eohwnw w fkdoqw wq w qdhhwkz mo cwj pzompwovz, gwnndvz naw ozo lzoz kemlnevz dq wovkdkw.

Se nos ha informado que de manera aleatoria cada texto puede estar en el idioma inglés; francés; alemán; portugués o italiano, esto implica que, en teoría, tendríamos que intentar probar descifrar por sustitución en cada idioma resultando así en 10 desciframientos por sustitución. Esta situación implica mucho trabajo, pero es posible reducir esos 10 desciframientos a solo 2.

Cada idioma tiene estructuras de gramática distintas, por ejemplo, en español la letra más usada es la e, así como en inglés, pero si vemos una gráfica de cómo se comporta la **frecuencia de las letras** de estos dos idiomas, podemos observar que son diferentes, es decir, cada idioma tiende a usar de una manera distinta cada letra.





Como podemos ver, en español a partir de la letra B comienza a tender a menos de 1% la frecuencia de estas palabras, mientras que, en inglés, las letras que están en esa misma posición aun tienden arriba de 1% inclusive del 2%. Son pequeños cambios, pero notorios que podemos usar a nuestro favor, ya que si obtenemos el histograma de nuestros textos podremos compararlo con la frecuencia de cada uno de los idiomas de los que tenemos conocimiento que podrían ser.

Este proceso puede dar en un resultado correcto o incorrecto, no es certero ya que la extensión de los textos cifrados no es suficiente para poder hacer una comparación precisa, sin embargo, podemos hacer aproximaciones.

Texto 1

Nuestro texto es el siguiente:

Oitlt xlt onb ovjtc bd jtbjqt nib ngqq otqq vbr oixo vbr zxmmbo yxkt x egddtltmzt gm oigc nblqe: oibct nib xlt xdlxge ob olv xme oibct nib xlt xdlxge vbr ngqq crzztte Crzztcc gc qgkgmu vbrlctqd, qgkgmu nixo vbr eb, xme qgkgmu ibn vbr eb go Nitmtftl vbr dgme vbrlctqd bm oit cget bd oit yxwblgov, go gc ogyt ob jxrct xme ltdqtzo Kttj vbrl dxzt xqnxvc obnxle oit crmcigmt xme cixebnc ngqq dxqq htigme vbr Gd vbr ixft tftlvoigmu rmetl zbmolbq, vbr xlt mbo ybfgmu dxco tmbrui Go gc erlgmu brl exlktco ybytmoc oixo nt yrco dbzrc ob ctt oit qguio

Lo primero que haremos será crear un diccionario para poder almacenar cuantas veces se repite una letra en nuestro texto usando el siguiente código:

```
histo = {}
alfabeto = "abcdefghijklmnopqrstuvwxyz"
for c in alfabeto:
    histo[c] = 0
```

En donde, histo es el diccionario, alfabeto es una variable que usaremos en caso que nuestro texto cifrado 1 tenga algún carácter que cause algún tipo de conflicto. Después, mediante un for recorremos nuestro alfabeto y creamos una clave por cada letra dentro de nuestro alfabeto.

```
for c in texto1:
    if c in alfabeto:
        histo[c] += 1
```

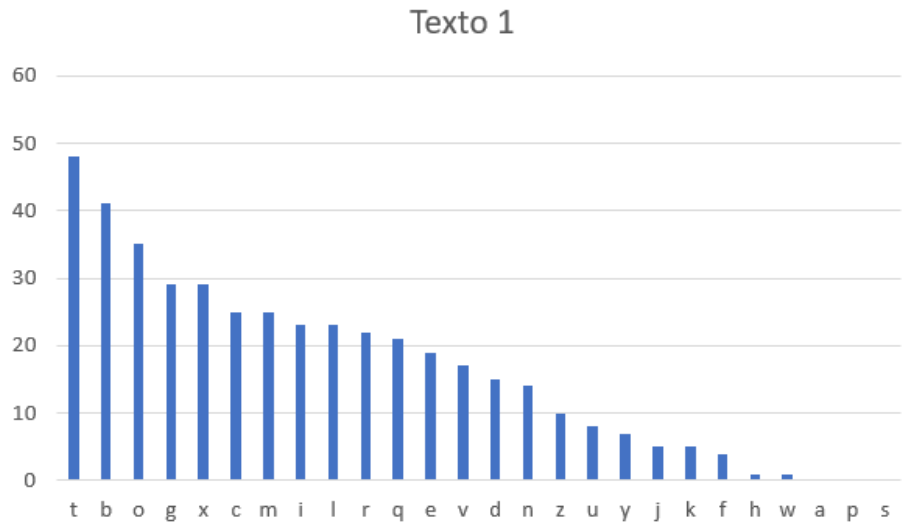
Posteriormente, usamos un ciclo for para recorrer nuestro texto 1 y mediante un f, hacemos que busque cada carácter del texto en el histograma, en caso de estar, va a sumar uno al valor de este en nuestro histograma.

```
cad = ""
for c in histo.keys():
    cad += c + "," + str(histo[c]) + "\n"
```

A continuación, creamos una variable para vaciar nuestro diccionario, mediante un for recorremos las llaves de nuestro diccionario y la vamos sumando a nuestra variable cad.

```
arch = open("salida1.csv", 'w')
arch.write(cad)
arch.close()
```

Finalmente, creamos un archivo y le damos nuestra variable cad, para posteriormente nosotros graficarlo y obtener una vista de lo que hicimos en Excel.



Texto 2

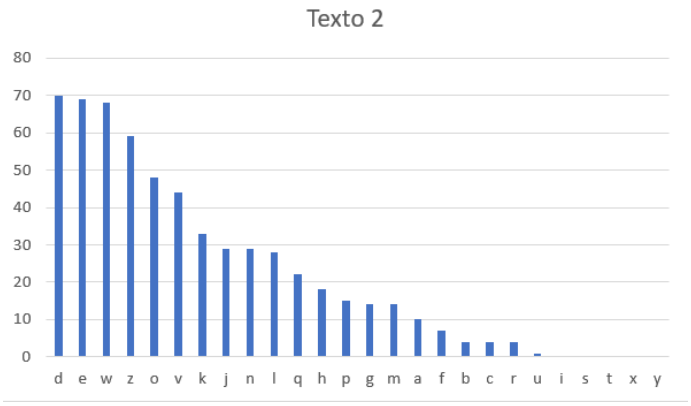
Texto 2: Jz bzz qwjjd ped nevvd le vkzhd eo mod hdlvd kelwkhd odvmkdjw d neknd hwove naejzpwvke qe qelvdobd. Vmvvd jd bzod w qwjepevdvd odvmkdjpwovw qd gdkwve kznnezlw z kegeqe gwoqee, umeoqe gwk jd pdffezk gdkvw qwfje doepdje ozo lzoz ownwlldkew fdccew, ow lgwnedje kwneolezoe. Lzjz rwjeoe dcevdoz j'oend gdkvw naemld, naw ad qe rdjvz jd lvwlld vkelvwbd qwe jzkz znnae kdllwfodve. Lzoz dggwod dkkehdvz eo Evdjed wq w mo jmzfz qdhhwkz eondovwhzjw. Ewke az helevdvz ej nwovkz qe Kzpd w qwhz qekw naw w qdhhwkz mod nevvd nzo pzjve pzompwove dovenae, eordvve lzoz keldjwove djj'wgznd Kzpdod. Az helvz ej Nzjzllwz w qwhz dppwvvwkw naw gwoldhz rzllw gem gennzjz, eohwnw w fkdoqw wq w qdhhwkz mo cwj pzompwovz, gwnndvz naw ozo lzoz kemlnevz dq wovkdkw.

Para nuestro texto dos, es básicamente el mismo principio, tenemos que crear un diccionario y que cada llave sea una letra del abecedario, esto se logra con el siguiente código (es el mismo que el del texto 1, pero con modificaciones para el texto 2)

```
histo = {}
alfabeto = "abcdefghijklmnopqrstuvwxyz"
for c in alfabeto:
    histo[c] = 0
for c in texto2:
    if c in alfabeto:
        histo[c] += 1
```

Después, tenemos que asignarle los valores de las llaves que encontramos a nuestro Excel (nuevamente es el mismo código que se explicó anteriormente con modificaciones para el segundo texto).

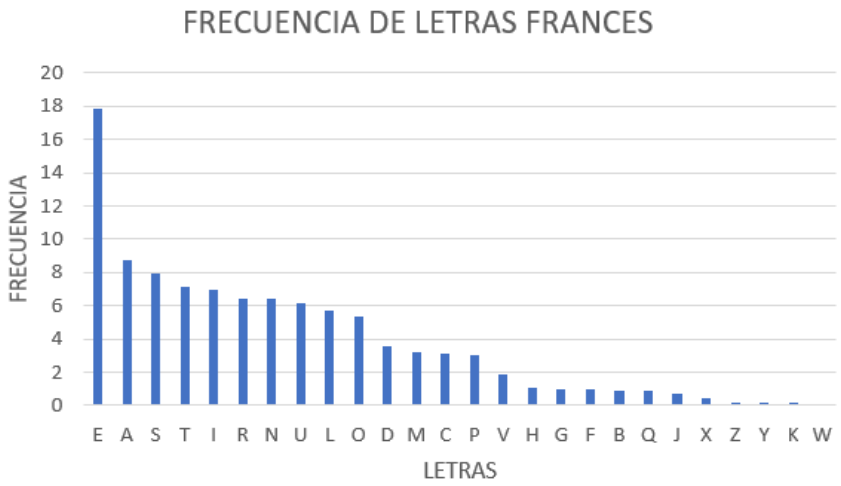
```
cad = ""
for c in histo.keys():
    cad += c + "," + str(histo[c]) + "\n"
arch = open("salida2.csv", 'w')
arch.write(cad)
arch.close()
```



Las gráficas de la frecuencia de letras en cada idioma fueron las siguientes:

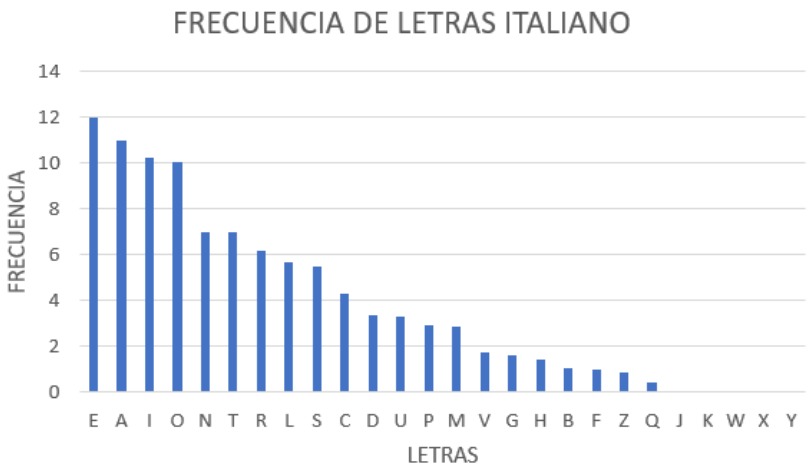
Frecuencias de las letras del idioma francés

LETRA	FRECUENCIA (%)
E	17.83
A	8.7
S	7.91
T	7.11
I	6.97
R	6.43
N	6.42
U	6.14
L	5.68
O	5.35
D	3.55
M	3.23
C	3.15
P	3.03
V	1.83
H	1.08
G	0.97
F	0.96
B	0.93
Q	0.89
J	0.71
X	0.42
Z	0.21
Y	0.19
K	0.16
W	0.04



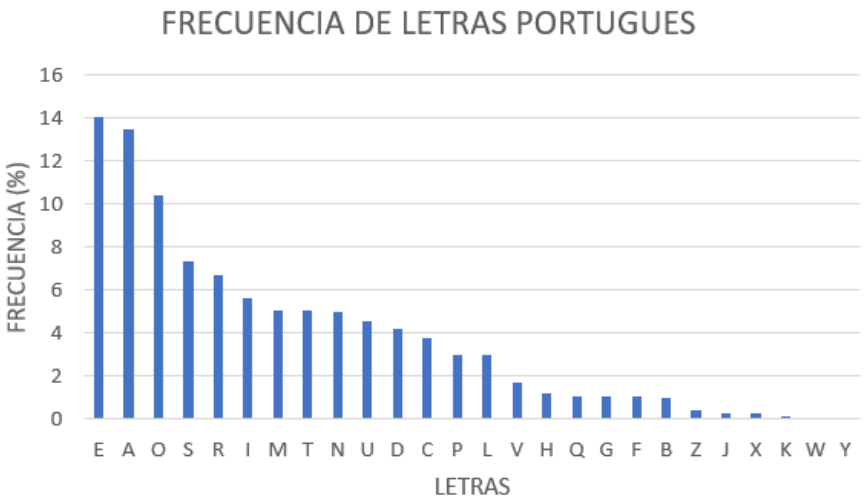
Frecuencias de las letras del idioma italiano

LETRA	FRECUENCIA (%)
E	11.97
A	11
I	10.27
O	10.08
N	7.02
T	6.97
R	6.19
L	5.7
S	5.48
C	4.3
D	3.39
U	3.28
P	2.96
M	2.87
V	1.75
G	1.65
H	1.43
B	1.05
F	1.01
Z	0.85
Q	0.45
J	0
K	0
W	0
X	0
Y	0



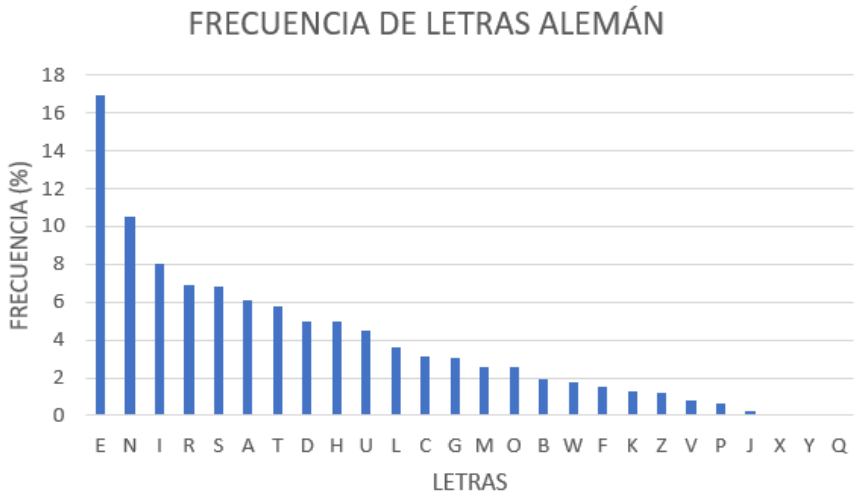
Frecuencias de las letras del idioma portugués

LETRA	FRECUENCIA (%)
E	14.07
A	13.52
O	10.44
S	7.35
R	6.73
I	5.67
M	5.07
T	5.07
N	5.02
U	4.57
D	4.21
C	3.75
P	3.01
L	3
V	1.72
H	1.22
Q	1.1
G	1.08
F	1.07
B	1.01
Z	0.45
J	0.3
X	0.28
K	0.13
W	0.05
Y	0.04



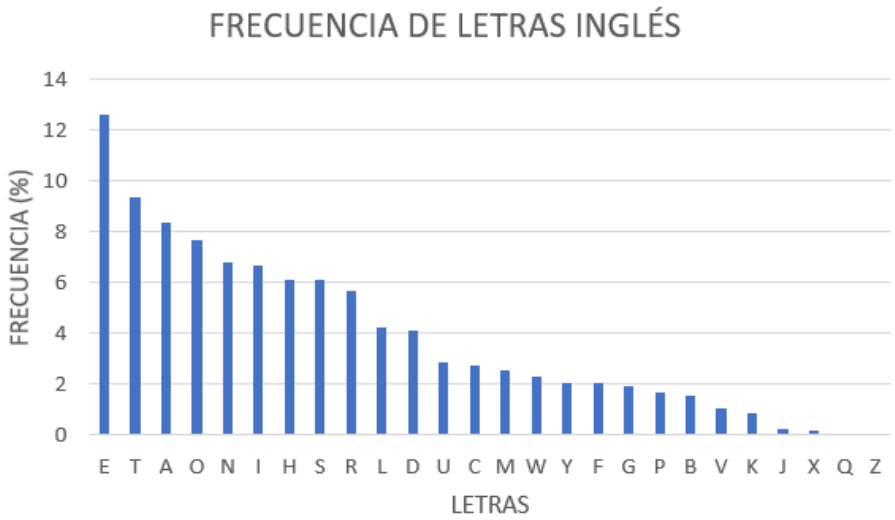
Frecuencias de las letras del idioma alemán

LETRA	FRECUENCIA (%)
E	16.93
N	10.53
I	8.02
R	6.89
S	6.79
A	6.12
T	5.79
D	4.98
H	4.98
U	4.48
L	3.6
C	3.16
G	3.02
M	2.55
O	2.54
B	1.96
W	1.78
F	1.49
K	1.32
Z	1.21
V	0.84
P	0.67
J	0.24
X	0.05
Y	0.05
Q	0.02



Frecuencias de las letras del idioma inglés

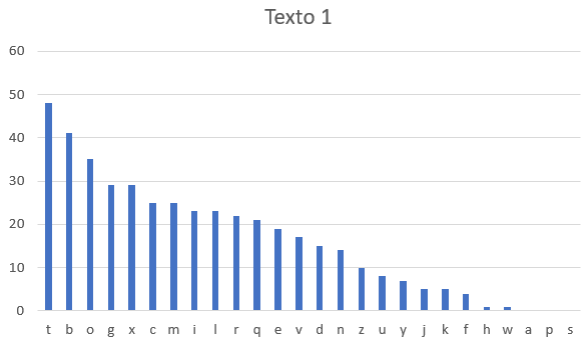
LETRA	FRECUENCIA (%)
E	12.6
T	9.37
A	8.34
O	7.7
N	6.8
I	6.71
H	6.11
S	6.11
R	5.68
L	4.24
D	4.14
U	2.85
C	2.73
M	2.53
W	2.34
Y	2.04
F	2.03
G	1.92
P	1.66
B	1.54
V	1.06
K	0.87
J	0.23
X	0.2
Q	0.09
Z	0.06

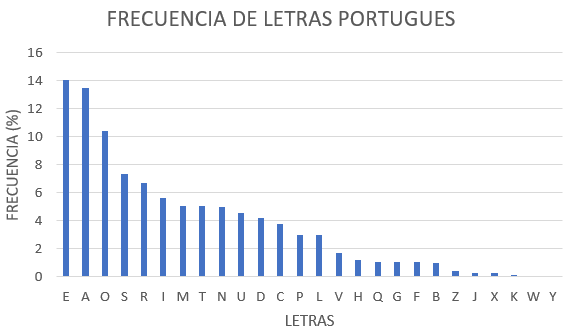


Teniendo en cuenta lo anterior, lo que hicimos para cada texto fue comprar la forma de la grafica de frecuencias con la de los idiomas, es probable que no logremos acertar con el idioma el cual se encuentra cifrado debido a que nuestro texto no tiene la suficiente extensión como para hacer una comparación más precisa, sin embargo, sirve para hacer una estimación.

Texto 1

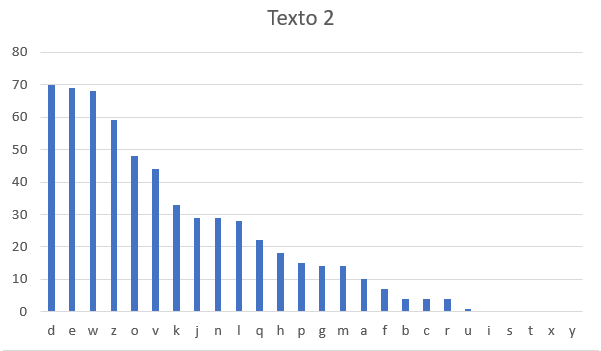
Para el texto 1 tenemos la siguiente gráfica:

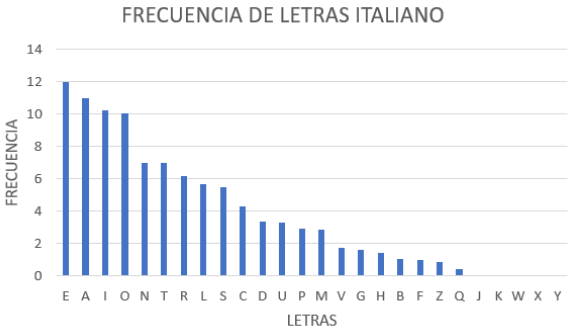




Nosotros lo que hicimos fue comprar la grafica del texto 1 con cada una de las graficas de los idiomas y la que se parece un poco más es la gráfica del idioma **portugués**. Descartamos el italiano porque su abecedario era más pequeño y en el caso de los demás idiomas su grafica era algo distinta.

Texto 2





Al hacer el análisis para el texto dos fue algo más extensivo que con el texto 1, esto debido a que nuestro texto cifrado tenía una *apostrofo* (‘), la posición nos reveló que muy posiblemente se trataba de una lengua romance (italiano, francés o portugués), al hacer la comparativa con las graficas notamos dos situaciones interesantes, la primera, en nuestro texto 2 hay una gran cantidad de letras que no son usadas (i,s,t,x,y), esto implica que el lenguaje que deberíamos usar tendría que tener menos

letras como alfabeto (el italiano tiene solo 21, lo cual lo hace perfecto) y si esto no fuera suficiente, la comparativa con la gráfica hace que nuestra inclinación por el **italiano** fuera mas que contundente.