PROJECT REPORT

Submitted in fulfillment of the requirement of
the course Cyber Security by

Harsimran Virk
(IT - 181080031)

Shubhankar Gupta
(IT - 181080030)

DEPARTMENT OF COMPUTER ENGINEERING AND INFORMATION TECHNOLOGY
VEERMATA JIJABAI TECHNOLOGICAL INSTITUTE
(An Autonomous Institute Affiliated to Mumbai University)
(Central Technological Institute, Maharashtra State)
Matunga, MUMBAI - 400019

TABLE OF CONTENTS

Sr. No.	Topic	Page No.
1	Problem Statement	3
2	Terminologies	4
3	Technology Stack	6
4	Algorithm	7
5	Demonstration	8
6	Conclusion	13

1. PROBLEM STATEMENT

Write a program that can perform a letter frequency attack on any monoalphabetic substitution cipher without human intervention. Your software should produce possible plaintexts in rough order of likelihood. It would be good if your user interface allowed the user to specify "give me the top 10 possible plaintexts".

2. TERMINOLOGIES

Substitution Cipher

In cryptography, a substitution cipher is a method of encrypting in which units of plaintext are replaced with the ciphertext, in a defined manner, with the help of a key; the "units" may be single letters (the most common), pairs of letters, triplets of letters, mixtures of the above, and so forth. The receiver deciphers the text by performing the inverse substitution process to extract the original message.

Monoalphabatic Cipher

A monoalphabetic substitution cipher, also known as a simple substitution cipher, relies on a fixed replacement structure. That is, the substitution is fixed for each letter of the alphabet. Thus, if "a" is encrypted to "R", then every time we see the letter "a" in the plaintext, we replace it with the letter "R" in the ciphertext.

A simple example is where each letter is encrypted as the next letter in the alphabet: "a simple message" becomes *"B TJNQMF NFTTBHF"*. In general, when performing a simple substitution manually, it is easiest to generate the ciphertext alphabet first, and encrypt by comparing this to the plaintext alphabet. The table below shows how one might choose to, and we will, lay them out for this example.

Cryptanalysis

Cryptanalysis (from the Greek kryptós, "hidden", and analýein, "to analyze") refers to the process of analyzing information systems in order to understand hidden aspects of the systems. Cryptanalysis is used to breach cryptographic security systems and gain access to the contents of encrypted messages, even if the cryptographic key is unknown.

In addition to mathematical analysis of cryptographic algorithms, cryptanalysis includes the study of side-channel attacks that do not target weaknesses in the cryptographic algorithms themselves, but instead exploit weaknesses in their implementation.

Frequency Analysis

In cryptanalysis, frequency analysis (also known as counting letters) is the study of the frequency of letters or groups of letters in a ciphertext. The method is used as an aid to breaking classical ciphers.

Frequency analysis is based on the fact that, in any given stretch of written language, certain letters and combinations of letters occur with varying frequencies. Moreover, there is a characteristic distribution of letters that is roughly the same for almost all samples of that language. For instance, given a section of English language, E, T, A and O are the most common, while Z, Q, X and J are rare. Likewise, TH, ER, ON, and AN are the most common pairs of letters (termed bigrams or digraphs), and SS, EE, TT, and FF are the most common repeats. The nonsense phrase "ETAOIN SHRDLU" represents the 12 most frequent letters in typical English language text.

In some ciphers, such properties of the natural language plaintext are preserved in the ciphertext, and these patterns have the potential to be exploited in a ciphertext-only attack.

3. TECHNOLOGY STACK

Frontend

React.js

React.js is a javascript library for building user interfaces. It allows us to develop UIs in a declarative manner and takes care of updating the application state whenever any data changes. It encourages component driven development which makes our codebase more modular.

Tailwind CSS

Tailwind CSS is basically a utility-first CSS framework for rapidly building custom user interfaces. It is a highly customizable, low-level CSS framework that gives you all of the building blocks you need to build bespoke designs without any annoying opinionated styles you have to fight to override.

Command Line Interface

Python

Python is a general-purpose interpreted, interactive, object-oriented, and high-level programming language. It was created by Guido van Rossum during 1985- 1990. Like Perl, Python source code is also available under the GNU General Public License (GPL). Python is named after a TV Show called ëMonty Pythonís Flying Circusí and not after Python-the snake.

Python 3.0 was released in 2008. Although this version is supposed to be backward incompatibles, later on many of its important features have been backported to be compatible with version 2.7.

4. ALGORITHM

Approach: The problem can be solved based on the following observations:

1. **Frequency analysis** is one of the known ciphertext attacks. It is based on the study of the frequency of letters or groups of letters in a ciphertext. In all languages, different letters are used with different frequencies.

2. The frequency array attack is based on the observation that in an English text, not all letters occur with the same frequency.

3. In the given problem, the string ***T = “ETAOINSHRDLCUMWFGYPBVKJXQZ”*** is used for deciphering.

4. Therefore, the idea is to find the difference between ith maximum occurring letter in the given string and the string T and then shift all the letters of the given string with that difference. The string obtained will be one of the possible decrypted strings.

Follow the steps below to solve the problem:

1. Initialize a string say T as “ETAOINSHRDLCUMWFGYPBVKJXQZ”.

2. Find the frequency of each character of the string S, and store it in a variable, say freq[].

3. Iterate over the range [0, 5] using the variable i and perform the following steps:

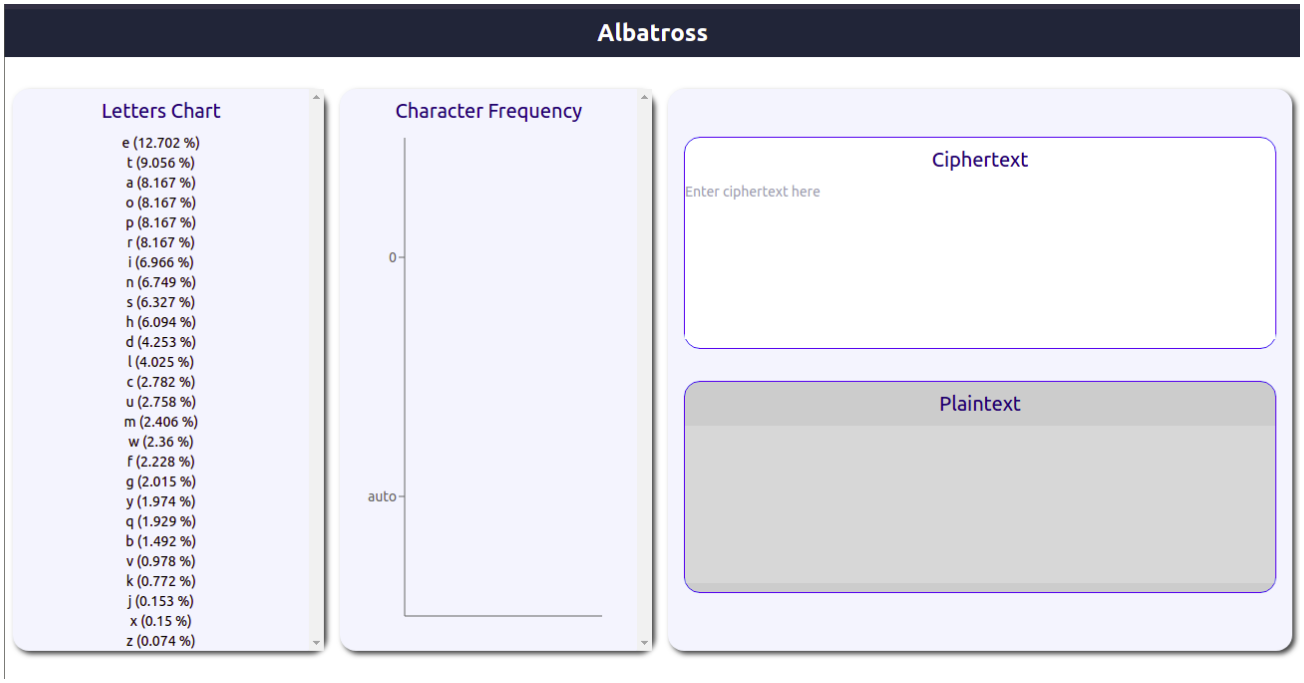
    a. Find the ith most occurring element in the string S and store it in a variable, say ch.
    b. Find the difference between the ch and ith character of the string T and store it in a variable, say x.
    c. Iterate over the characters of string S, and shift all characters by x and then push the obtained string into an array plaintext[].

4. Finally, after the above steps, print the strings obtained in the array plaintext[].

5. DEMONSTRATION

We have deployed a website that performs the frequency analysis attack. The website can be found at https://cocky-swanson-d482fb.netlify.app/ .

1. This is the landing page of our website.

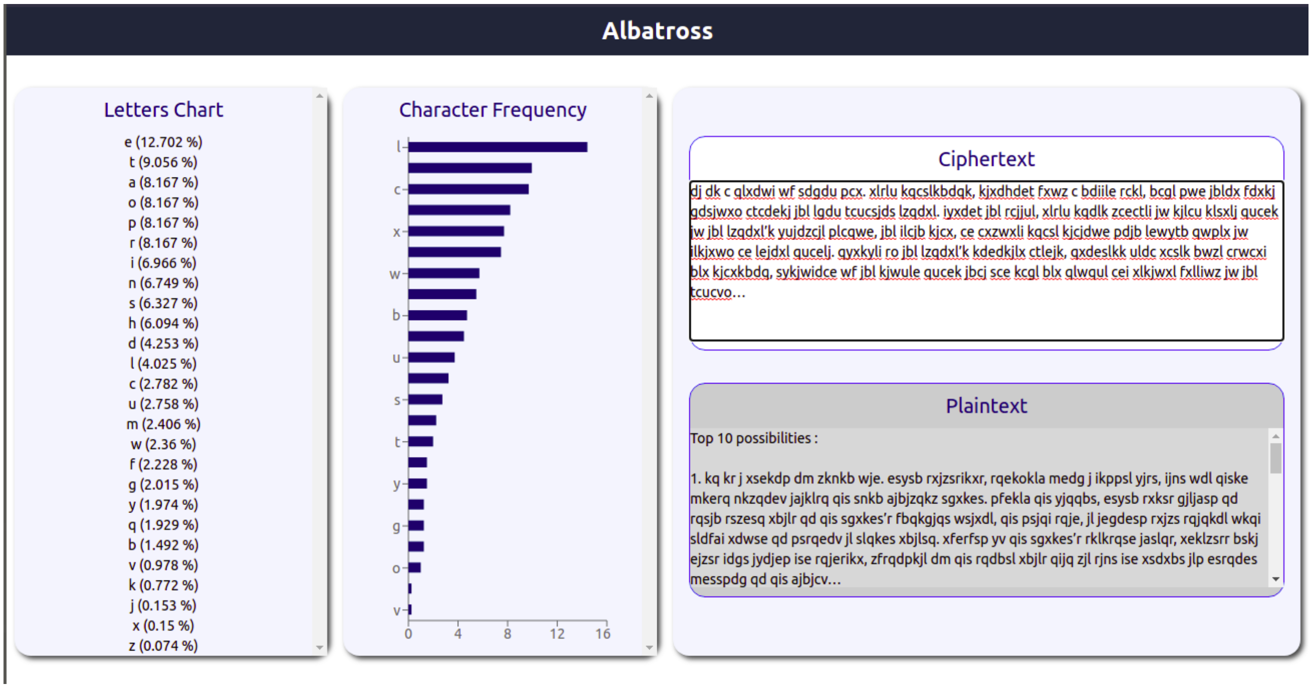


Letters Chart

e (12.702 %)
t (9.056 %)
a (8.167 %)
o (8.167 %)
p (8.167 %)
r (8.167 %)
i (6.966 %)
n (6.749 %)
s (6.327 %)
h (6.094 %)
d (4.253 %)
l (4.025 %)
c (2.782 %)
u (2.758 %)
m (2.406 %)
w (2.36 %)
f (2.228 %)
g (2.015 %)
y (1.974 %)
q (1.929 %)
b (1.492 %)
v (0.978 %)
k (0.772 %)
j (0.153 %)
x (0.15 %)
z (0.074 %)

On the left, we can see a letter frequency chart. The chart shows the sorted list of most commonly used letters in the English language. The next column will be filled in a later step. We can enter our text in ciphertext textarea, and we will see the possibilities in the Plaintext area.

2. We will enter a ciphertext and see the page visualizations.



Letters Chart

e (12.702 %)
t (9.056 %)
a (8.167 %)
o (8.167 %)
p (8.167 %)
r (8.167 %)
i (6.966 %)
n (6.749 %)
s (6.327 %)
h (6.094 %)
d (4.253 %)
l (4.025 %)
c (2.782 %)
u (2.758 %)
m (2.406 %)
w (2.36 %)
f (2.228 %)
g (2.015 %)
y (1.974 %)
q (1.929 %)
b (1.492 %)
v (0.978 %)
k (0.772 %)
j (0.153 %)
x (0.15 %)
z (0.074 %)

Ciphertext

dj dk c qlxdwi wf sdgdu pcx. xlrlu kqcslkbdqk, kjxdhdet fxwz c bdiile rckl, bcgl pwe jbldx fdxkj gdsjwxo ctcdekj jbl lgdu tcucsjds lzqdxl. iyxdet jbl rcjjul, xlrlu kqdlk zcectli jw kjlcu klsxlj qucek jw jbl lzqdxl'k yujdzcjl plcqwe, jbl ilcjb kjcx, ce cxzwxli kqcsl kjcjdwe pdjb lewytb qwplx jw ilkjxwo ce lejdxl qucelj. qyxkyli ro jbl lzqdxl'k kdedkjlx ctlejk, qxdeslkk uldc xcslk bwzl crwcxi blx kjcxkbdq, sykjwidce wf jbl kjwule qucek jbcj sce kcgl blx qlwqul cei xlkjwxl fxlliwz jw jbl tcucvo...

Plaintext

Top 10 possibilities :

1. kq kr j xsekdp dm zknkb wje. esysb rxjzsrikxr, rqekokla medg j ikppsl yjrs, ijns wdl qiske mkerq nkzqdev jajklrq qis snkb ajbjzqkz sgxkes. pfekla qis yjqqbs, esysb rxksr gjljasp qd rqsjb rszesq xbjlr qd qis sgxkes'r fbqkgjqs wsjxdl, qis psjqi rqje, jl jegdesp rxjzs rqjqkdl wkqi sldfai xdwse qd psrqedv jl slqkes xbjlsq. xferfsp yv qis sgxkes'r rklkrqse jaslqr, xeklzsrr bskj ejzsr idgs jydjep ise rqjerikx, zfrqdpkjl dm qis rqdbsl xbjlr qijq zjl rjns ise xsdxbs jlp esrqdes messpdg qd qis ajbjcv...

On entering the ciphertext, we can see that we have a histogram of letter frequencies. From the frequency chart, we can see that the letter **l** (Small L) has the most frequency in the ciphertext. So, our first possibility maps **l** to **e**, and so on.

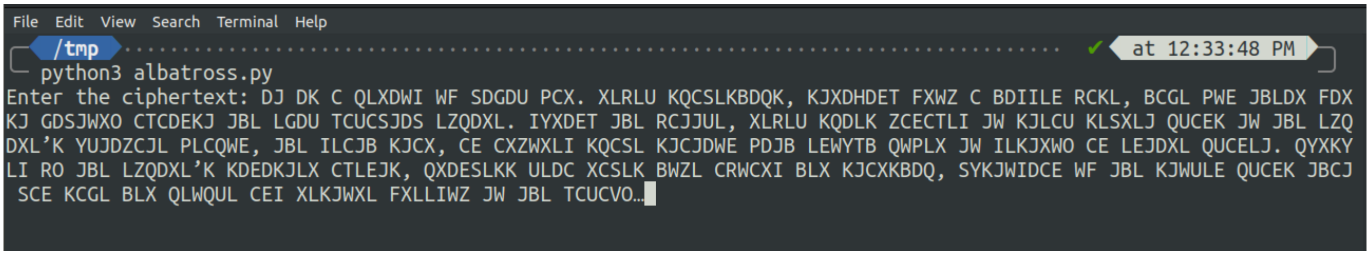
Our website shows 10 possible plaintexts, in decreasing order of probabilities.

We also have prepared a Command Line Interface that performs the same task. The file is named albatross.py. It is a CLI written in Python programming language. The steps to use that as follows:

1. On a terminal, type:

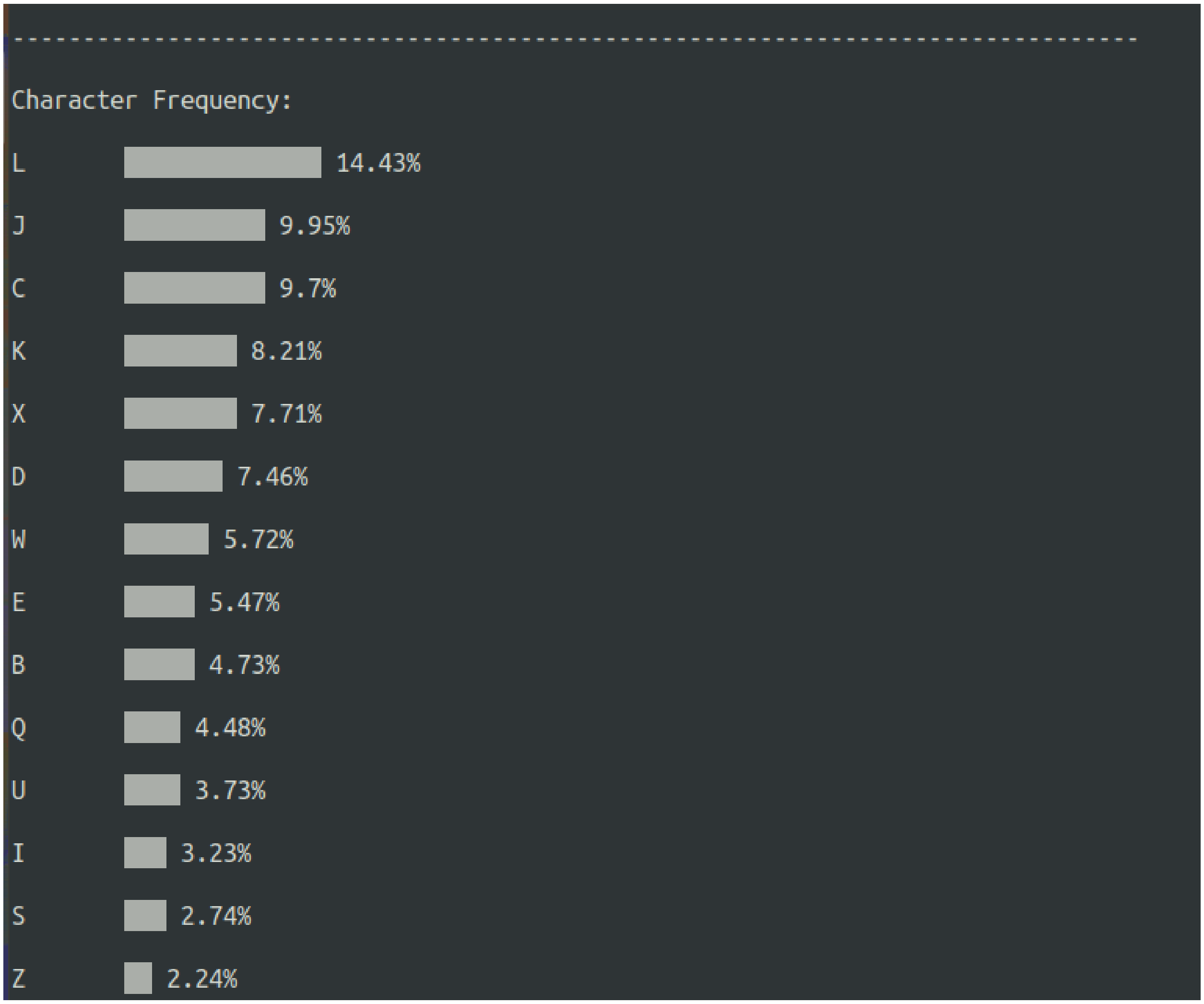
```
python3 albatross.py
```

2. The input prompt asks for ciphertext, so we add the ciphertext:

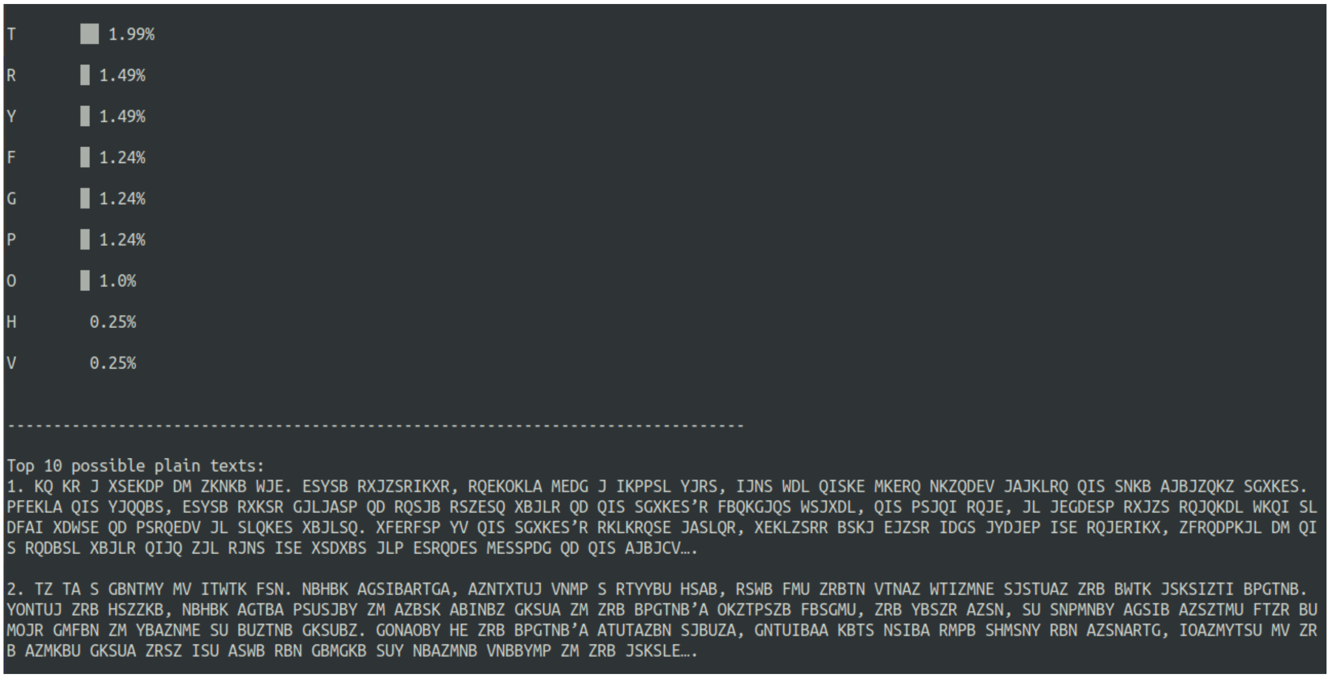


```
 python3 albatross.py
Enter the ciphertext: DJ DK C QLXDWI WF SDGDU PCX. XLRLU KQCSLKBDQK, KJXDHDET FXWZ C BDIILE RCKL, BCGL PWE JBLDX FDX
KJ GDSJWXO CTCDEKJ JBL LGDU TCUCSJDS LZQDXL. IYXDET JBL RCJJUL, XLRLU KQDLK ZCECTLI JW KJLCU KLSXLJ QUCEK JW JBL LZQ
DXL’K YUJDZCJL PLCQWE, JBL ILCJB KJCX, CE CXZWXLI KQCSL KJCJDWE PDJB LEWYTB QWPLX JW ILKJXWO CE LEJDXL QUCELJ. QYXKY
LI RO JBL LZQDXL’K KDEDKJLX CTLEJK, QXDESLKK ULDC XCSLK BWZL CRWCXI BLX KJCXKBDQ, SYKJWIDCE WF JBL KJWULE QUCEK JBCJ
 SCE KCGL BLX QLWQUL CEI XLKJWXL FXLLIWZ JW JBL TCUCVO…
```

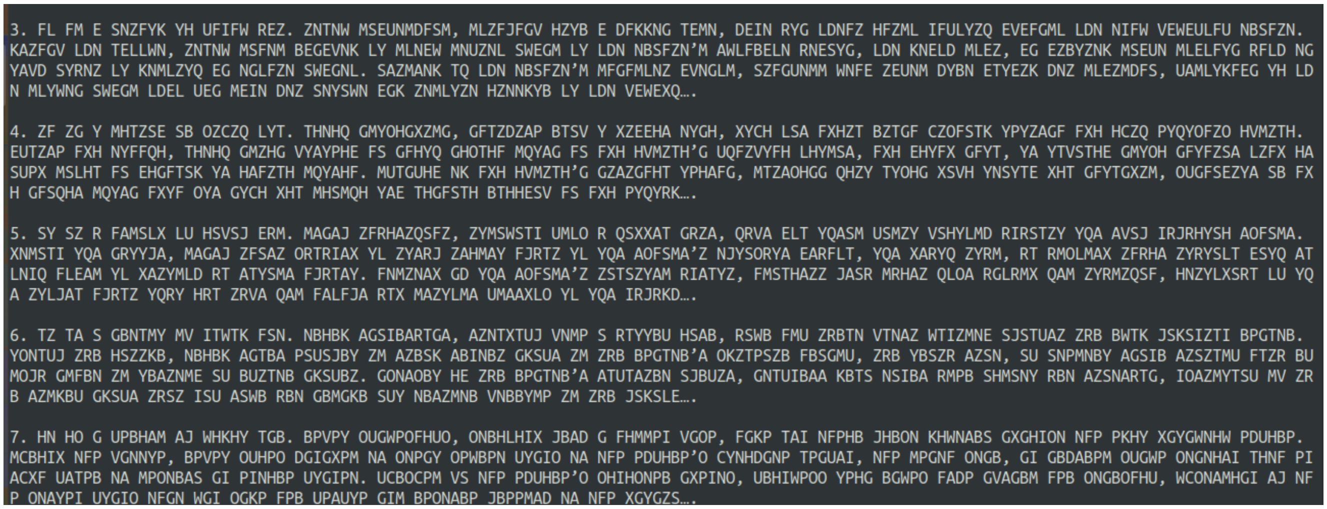
3. We see the outputs : A character frequency chart, and the top 10 probabilities of plaintext:



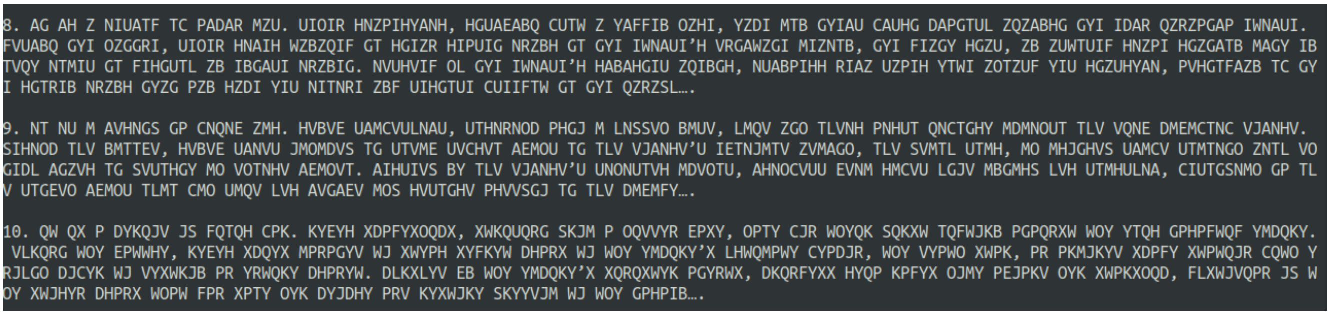
```
--------------------------------------------------------------------------------

Character Frequency:

L                     14.43%

J                 9.95%

C                 9.7%

K               8.21%

X               7.71%

D              7.46%

W             5.72%

E            5.47%

B            4.73%

Q           4.48%

U           3.73%

I          3.23%

S          2.74%

Z         2.24%
```



```
T       █ 1.99%

R       ▌ 1.49%

Y       ▌ 1.49%

F       ▌ 1.24%

G       ▌ 1.24%

P       ▌ 1.24%

O       ▌ 1.0%

H         0.25%

V         0.25%


--------------------------------------------------------------------------------

Top 10 possible plain texts:
1. KQ KR J XSEKDP DM ZKNKB WJE. ESYSB RXJZSRIKXR, RQEKOKLA MEDG J IKPPSL YJRS, IJNS WDL QISKE MKERQ NKZQDEV JAJKLRQ QIS SNKB AJBJZQKZ SGXKES.
PFEKLA QIS YJQQBS, ESYSB RXKSR GJLJASP QD RQSJB RSZESQ XBJLR QD QIS SGXKES'R FBQKGJQS WSJXDL, QIS PSJQI RQJE, JL JEGDESP RXJZS RQJQKDL WKQI SL
DFAI XDWSE QD PSRQEDV JL SLQKES XBJLSQ. XFERFSP YV QIS SGXKES'R RKLKRQSE JASLQR, XEKLZSRR BSKJ EJZSR IDGS JYDJEP ISE RQJERIKX, ZFRQDPKJL DM QI
S RQDBSL XBJLR QIJQ ZJL RJNS ISE XSDXBS JLP ESRQDES MESSPDG QD QIS AJBJCV….

2. TZ TA S GBNTMY MV ITWTK FSN. NBHBK AGSIBARTGA, AZNTXTUJ VNMP S RTYYBU HSAB, RSWB FMU ZRBTN VTNAZ WTIZMNE SJSTUAZ ZRB BWTK JSKSIZTI BPGTNB.
YONTUJ ZRB HSZZKB, NBHBK AGTBA PSUSJBY ZM AZBSK ABINBZ GKSUA ZM ZRB BPGTNB'A OKZTPSZB FBSGMU, ZRB YBSZR AZSN, SU SNPMNBY AGSIB AZSZTMU FTZR BU
MOJR GMFBN ZM YBAZNME SU BUZTNB GKSUBZ. GONAOBY HE ZRB BPGTNB'A ATUTAZBN SJBUZA, GNTUIBAA KBTS NSIBA RMPB SHMSNY RBN AZSNARTG, IOAZMYTSU MV ZR
B AZMKBU GKSUA ZRSZ ISU ASWB RBN GBMGKB SUY NBAZMNB VNBBYMP ZM ZRB JSKSLE….
```



```
3. FL FM E SNZFYK YH UFIFW REZ. ZNTNW MSEUNMDFSM, MLZFJFGV HZYB E DFKKNG TEMN, DEIN RYG LDNFZ HFZML IFULYZQ EVEFGML LDN NIFW VEWEULFU NBSFZN.
KAZFGV LDN TELLWN, ZNTNW MSFNM BEGEVNK LY MLNEW MNUZNL SWEGM LY LDN NBSFZN'M AWLFBELN RNESYG, LDN KNELD MLEZ, EG EZBYZNK MSEUN MLELFYG RFLD NG
YAVD SYRNZ LY KNMLZYQ EG NGLFZN SWEGNL. SAZMANK TQ LDN NBSFZN'M MFGFMLNZ EVNGLM, SZFGUNMM WNFE ZEUNM DYBN ETYEZK DNZ MLEZMDFS, UAMLYKFEG YH LD
N MLYWNG SWEGM LDEL UEG MEIN DNZ SNYSWN EGK ZNMLYZN HZNNKYB LY LDN VEWEXQ….

4. ZF ZG Y MHTZSE SB OZCZQ LYT. THNHQ GMYOHGXZMG, GFTZDZAP BTSV Y XZEEHA NYGH, XYCH LSA FXHZT BZTGF CZOFSTK YPYZAGF FXH HCZQ PYQYOFZO HVMZTH.
EUTZAP FXH NYFFQH, THNHQ GMZHG VYAYPHE FS GFHYQ GHOTHF MQYAG FS FXH HVMZTH'G UQFZVYFH LHYMSA, FXH EHYFX GFYT, YA YTVSTHE GMYOH GFYFZSA LZFX HA
SUPX MSLHT FS EHGFTSK YA HAFZTH MQYAHF. MUTGUHE NK FXH HVMZTH'G GZAZGFHT YPHAFG, MTZAOHGG QHZY TYOHG XSVH YNSYTE XHT GFYTGXZM, OUGFSEZYA SB FX
H GFSQHA MQYAG FXYF OYA GYCH XHT MHSMQH YAE THGFSTH BTHHESV FS FXH PYQYRK….

5. SY SZ R FAMSLX LU HSVSJ ERM. MAGAJ ZFRHAZQSFZ, ZYMSWSTI UMLO R QSXXAT GRZA, QRVA ELT YQASM USMZY VSHYLMD RIRSTZY YQA AVSJ IRJRHYSH AOFSMA.
XNMSTI YQA GRYYJA, MAGAJ ZFSAZ ORTRIAX YL ZYARJ ZAHMAY FJRTZ YL YQA AOFSMA'Z NJYSORYA EARFLT, YQA XARYQ ZYRM, RT RMOLMAX ZFRHA ZYRYSLT ESYQ AT
LNIQ FLEAM YL XAZYMLD RT ATYSMA FJRTAY. FNMZNAX GD YQA AOFSMA'Z ZSTSZYAM RIATYZ, FMSTHAZZ JASR MRHAZ QLOA RGLRMX QAM ZYRMZQSF, HNZYLXSRT LU YQ
A ZYLJAT FJRTZ YQRY HRT ZRVA QAM FALFJA RTX MAZYLMA UMAAXLO YL YQA IRJRKD….

6. TZ TA S GBNTMY MV ITWTK FSN. NBHBK AGSIBARTGA, AZNTXTUJ VNMP S RTYYBU HSAB, RSWB FMU ZRBTN VTNAZ WTIZMNE SJSTUAZ ZRB BWTK JSKSIZTI BPGTNB.
YONTUJ ZRB HSZZKB, NBHBK AGTBA PSUSJBY ZM AZBSK ABINBZ GKSUA ZM ZRB BPGTNB'A OKZTPSZB FBSGMU, ZRB YBSZR AZSN, SU SNPMNBY AGSIB AZSZTMU FTZR BU
MOJR GMFBN ZM YBAZNME SU BUZTNB GKSUBZ. GONAOBY HE ZRB BPGTNB'A ATUTAZBN SJBUZA, GNTUIBAA KBTS NSIBA RMPB SHMSNY RBN AZSNARTG, IOAZMYTSU MV ZR
B AZMKBU GKSUA ZRSZ ISU ASWB RBN GBMGKB SUY NBAZMNB VNBBYMP ZM ZRB JSKSLE….

7. HN HO G UPBHAM AJ WHKHY TGB. BPVPY OUGWPOFHUO, ONBHLHIX JBAD G FHMMPI VGOP, FGKP TAI NFPHB JHBON KHWNABS GXGHION NFP PKHY XGYGWNHW PDUHBP.
MCBHIX NFP VGNNYP, BPVPY OUHPO DGIGXPM NA ONPGY OPWBPN UYGIO NA NFP PDUHBP'O CYNHDGNP TPGUAI, NFP MPGNF ONGB, GI GBDABPM OUGWP ONGNHAI THNF PI
ACXF UATPB NA MPONBAS GI PINHBP UYGIPN. UCBOCPM VS NFP PDUHBP'O OHIHONPB GXPINO, UBHIWPOO YPHG BGWPO FADP GVAGBM FPB ONGBOFHU, WCONAMHGI AJ NF
P ONAYPI UYGIO NFGN WGI OGKP FPB UPAUYP GIM BPONABP JBPPMAD NA NFP XGYGZS….
```



```
8. AG AH Z NIUATF TC PADAR MZU. UIOIR HNZPIHYANH, HGUAEABQ CUTW Z YAFFIB OZHI, YZDI MTB GYIAU CAUHG DAPGTUL ZQZABHG GYI IDAR QZRZPGAP IWNAUI.
FVUABQ GYI OZGGRI, UIOIR HNAIH WZBZQIF GT HGIZR HIPUIG NRZBH GT GYI IWNAUI'H VRGAWZGI MIZNTB, GYI FIZGY HGZU, ZB ZUWTUIF HNZPI HGZGATB MAGY IB
TVQY NTMIU GT FIHGUTL ZB IBGAUI NRZBIG. NVUHVIF OL GYI IWNAUI'H HABAHGIU ZQIBGH, NUABPIHH RIAZ UZPIH YTWI ZOTZUF YIU HGZUHYAN, PVHGTFAZB TC GY
I HGTRIB NRZBH GYZG PZB HZDI YIU NITNRI ZBF UIHGTUI CUIIFTW GT GYI QZRZSL….

9. NT NU M AVHNGS GP CNQNE ZMH. HVBVE UAMCVULNAU, UTHNRNOD PHGJ M LNSSVO BMUV, LMQV ZGO TLVNH PNHUT QNCTGHY MDMNOUT TLV VQNE DMEMCTNC VJANHV.
SIHNOD TLV BMTTEV, HVBVE UANVU JMOMDVS TG UTVME UVCHVT AEMOU TG TLV VJANHV'U IETNJMTV ZVMAGO, TLV SVMTL UTMH, MO MHJGHVS UAMCV UTMTNGO ZNTL VO
GIDL AGZVH TG SVUTHGY MO VOTNHV AEMOVT. AIHUIVS BY TLV VJANHV'U UNONUTVH MDVOTU, AHNOCVUU EVNM HMCVU LGJV MBGMHS LVH UTMHULNA, CIUTGSNMO GP TL
V UTGEVO AEMOU TLMT CMO UMQV LVH AVGAEV MOS HVUTGHV PHVVSGJ TG TLV DMEMFY….

10. QW QX P DYKQJV JS FQTQH CPK. KYEYH XDPFYXOQDX, XWKQUQRG SKJM P OQVVYR EPXY, OPTY CJR WOYQK SQKXW TQFWJKB PGPQRXW WOY YTQH GPHPFWQF YMDQKY.
 VLKQRG WOY EPWWHY, KYEYH XDQYX MPRPGYV WJ XWYPH XYFKYW DHPRX WJ WOY YMDQKY'X LHWQMPWY CYPDJR, WOY VYPWO XWPK, PR PKMJKYV XDPFY XWPWQJR CQWO Y
RJLGO DJCYK WJ VYXWKJB PR YRWQKY DHPRYW. DLKXLYV EB WOY YMDQKY'X XQRQXWYK PGYRWX, DKQRFYXX HYQP KPFYX OJMY PEJPKV OYK XWPKXOQD, FLXWJVQPR JS W
OY XWJHYR DHPRX WOPW FPR XPTY OYK DYJDHY PRV KYXWJKY SKYYVJM WJ WOY GPHPIB….
```

6. CONCLUSION

Through this mini project, we have learnt and understood how frequency attacks work on Monoalphabetic Ciphers. We have designed and implemented a solution that allows us to view the top 10 possibilities for deciphering text using a frequency analysis attack.

While the technique itself is immune to language, it needs to be noted that completely automating the process is bound to be error prone. After an automatic check, some characters will need manual inspection for the same.