Ngram Analysis on Classical Ciphers for ease of Understanding of Cryptography

Author: José E. Jiménez Correa
Advisor: Dr. Alfredo Cruz
Electric & Computer Engineering and Computer Science

Abstract

An Ngram is defined as a group of characters in a text. They are used to determine the validity of texts in the decryption results and show how they can be affected. Using classical ciphers, we show how changes in their encryption and decryption processes can affect de data being processed. By analyzing the results and creating Ngram Stamps we determine the validity of the text. This paper will provide examples of how the Ngram detection can help validate texts and show how cryptography works in securing and validating your data using the processes of encryption and decryption.

Introduction

A cipher is an algorithm used to encrypt and decrypt data, it is one of the methods used to secure and encrypt data on the net. It does not only have to encrypt data it can also be used to maintain integrity of the data being transmitted. Any kind of interruption on the cipher or method of decryption can prove fatal for the message or data. To illustrate this, we will use the classical ciphers since they give very clear and concise examples of these properties. These ciphers are the ones that can be resolved in very short times and where at the time of their creation top choices for encryption and decryption of data and messages. These classical ciphers are known as symmetrical ciphers[4] since they use the same key for encryption and decryption. They can also be broken by brute force which will be the method that we will use for the Ngram analysis. Thanks to this they are perfect for demonstration purposes.

Background

On this project we will be working with various limitations. This is due to the quantity of variations that can exist on the encryption and decryption ciphers. The first limitation is that we are going to use classical ciphers due to the ease of decryption, encryption, and analysis. A second limitation is one of language. The language that we will be using for all source text and analysis is the English language we will be limiting the language to a subset of it. This subset is defined as all the letters of the alphabet only using the capitalizations of these letters (A,B,C, …, X,Y,Z) a total of 26 recognized characters or letter which would be our symbols. Another limitation would be that of the analysis, it would be based on the Ngrams for the English alphabet. The Ngrams used for the analysis[1] will be the ones that are repeated at least 100,000 times in English texts and all other iterations of less mentions are discarded. This is based in a study by Mayzner[2] which catalogs the positions and locations of different Ngrams in English texts

Problem

Thanks to the advancement of technology the knowledge of how data is secured, managed, and encrypted is almost non-existent. This is a factor for people to make mistakes on their security thinking that they are protected[5]. The common users just listen to all the features of an encryption service and decide that the one that has more options is the best. This is done without taking in consideration the security or integrity of the data since they lack the knowledge on how the processes of encryption and decryption works.

Methodology

To validate the results of the analysis for the various ciphers one must first get the Ngram stamp of the data. A Ngram stamp is defined as the result of the Ngram analysis on a given text, it is the composition of Ngram hits from 2Ngrams to 9Ngrams. The source text for analysis is a selection of text from the Wikipedia web page of the Rigel Star [3] this text is composed of 21,344 characters composed of capital letters, small caps, numbers, and special characters. This Text is then saved on a text file(.txt) for ease of access exactly as it is copied from the source without any modification. For an analysis to be completed the source text must be modified in a way that our analysis data can recognized the texts. For this what we will do is take the source data text and modify it so that all the letters of the alphabet are in their capitalized forms. This is done without modifying any of the format on the text file. All other characters that are not part of our pre-defined alphabet are left as they are and not modified in anyway as shown in Table 1.

Table 1 Original Text Versus Modified Text

Original
… Rigel as β Orionis (Latinized to Beta Orionis) was made by Johann Bayer in 1603. The "beta" designation is commonly given to the second-brightest star in each constellation, but Rigel is almost always brighter than α Orionis (Betelgeuse)…
Modified
… RIGEL AS B ORIONIS (LATINIZED TO BETA ORIONIS) WAS MADE BY JOHANN BAYER IN 1603. THE "BETA" DESIGNATION IS COMMONLY GIVEN TO THE SECOND-BRIGHTEST STAR IN EACH CONSTELLATION, BUT RIGEL IS ALMOST ALWAYS BRIGHTER THAN A ORIONIS (BETELGEUSE)…

As one can see the modification do not affect the format of the text just the letters. There exists **363** 2Ngram, **1569** 3Ngrams, **2252** 4Ngrams, **2013** 5Ngrams, **1544** 6Ngrams, **1099** 7Ngrams, **715** 8Ngrams, and **427** 9Ngrams. This result is obtained by counting how many times the Ngrams are mentioned in the text at least once creating the Ngram Stamp. Some examples of the Ngrams are shown in Table 2.

Table 2 Ngram Example List

2Ngram	3Ngram	4Ngram	5Ngram
UR, CA, EL, TA, LA	STA, CTI, ICA, IST, EAR	FORM, NING, ECTI, SOME, PORT	PLACE, ETWEE, BETWE, RIGHT, AGAIN
6Ngram	**7Ngram**	**8Ngram**	**9Ngram**
ENERAL, SYSTEM, RELATI, CTIONS, ECAUSE	ORMATIO, CERTAIN, INCREAS, RELATIO, SPECIAL	ALTHOUGH, RODUCTIO, ODUCTION, ASSOCIAT, MATERIAL	HARACTERI, ERNATIONA, NTERNATIO, RNATIONAL, INTERNATI

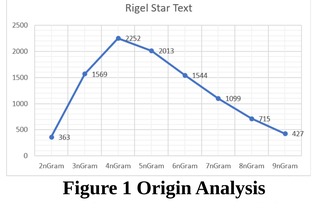

Figure 1 Origin Analysis

In Figure 1 we can see the Ngram Stamp of the source text. This Origin Ngram Stamp will give us a base for comparison on the other results of the cipher analysis. This will help us know which results are the correct answers in subsequent tests and analysis. This Ngram stamp results in a chart outlining the peaks of the mentions of the Ngrams in the origin text showing the expected max values and correct values of the Ngram stamp analysis.

Results and Discussion

The result of the analysis gave predictable results and some surprising ones. Of the four ciphers that we analyzed the Caesar and Affine ciphers where the ones that behaved as predicted. Their Ngram stamps where under the predictable range. All the analysis were performed using a brute force approach utilizing all possible values of the alphabet or until the limitations were met according to their ciphers. This analysis is composed of searching for matches for of **669** Ngrams of two characters, **8653** Ngrams of three characters, **42171** Ngrams of four characters, **93713** Ngrams of five characters, **114565** Ngrams of six characters, **104610** Ngrams of seven characters, **82347** Ngrams of eight characters, and **59030** Ngrams of nine characters and are the ones that will be searched for one by one on the subsequent analysis. For the Caesar cipher since we used the English alphabet only **26** result are available including the correct one. In Figure 2 we see the analysis result of the Caesar cipher Ngram analysis. For the Affine analysis we use **312** possible values in the English alphabet as shown in Figure 3.

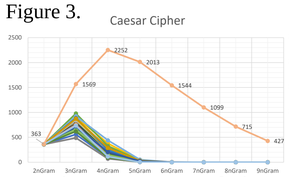

Figure 2 Caesar Analysis

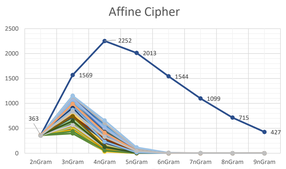

Figure 3 Affine Analysis

In the Vigenère and Playfair cipher a limit to the words to be available for possible keys to be used for the analysis was enforced at 100. In the Vigenère cipher there where a total of 100 results to be analyzed based on the top mentioned letters of the English alphabet they can be seen in Figure 4. In this cipher we can see how the keys if not different or close enough can affect the data being analyzed. In these previous cipher the data was affected character by character thus not changing the structure of the data and the analysis behaved predictably[6]. Meanwhile in the analysis of the Playfair cipher gave a different result as shown in Figure 5.

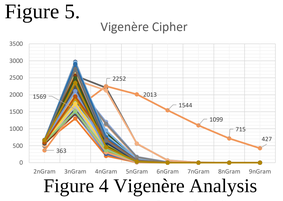

Figure 4 Vigenère Analysis

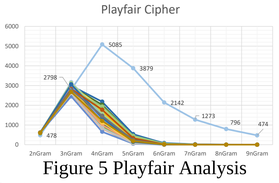

Figure 5 Playfair Analysis

As we can see in the Playfair analysis we can see a mayor increase of the false positives in the analysis and this is due to the method of the cipher to encrypt the data. The Playfair encrypts the data in a bi-gram method thus eliminating all non-alphabetic mentions in the source text and this affect the analysis. It took an average of 126.17s for each analysis per file of the Caesar cipher, 120.26s for the Affine cipher, 134.35s for the Vigenère cipher, and 88.51s for the Playfair cipher. In Table 3 we can see an Average of mentions per Ngram in each cipher.

Table 3 Total Ngrams Vs. Average Vs. Source

Source:	**2nGram**	**3nGram**	**4nGram**	**5nGram**
Rigel Text	***363***	***1569***	***2252***	***2013***
Total	*669*	*8653*	*42171*	*93713*
Caesar	359.32	762.60	201.68	15.40
Affine	359.24	785.68	234.23	22.54
Vigenère	625.01	2131.70	559.99	55.15
Playfair	610.32	2851.95	1318.16	190.12
Source:	**6nGram**	**7nGram**	**8nGram**	**9nGram**
Rigel Text	***1544***	***1099***	***715***	***427***
Total	*114565*	*104610*	*82347*	*59030*
Caesar	0.44	0.00	0.00	0.00
Affine	0.96	0.03	0.00	0.00
Vigenère	3.29	0.18	0.01	0.00
Playfair	15.17	1.40	0.26	0.04

Conclusions

By watching the result of the cipher algorithm and their analysis we can determine how secure they are and how they can ensure it. The ciphers show how when and erroneous method is used for a decryption of the text it gives us garbage data. This can tell us how secure the data is and how its integrity and availability is. At this age, these classical ciphers are not very secure with the advent of the computer era due to the processing power of the machines, but it gives us an idea of how the process works. We can see how the source text is processed into cipher text with the given ciphers and how changes in the method can give us different results. Depending on the cipher a change of letter or change of number greatly affects the consistency of the message. By not following the correct steps of encryption or decryption the data that is to be processed changes greatly in response to wrong input. By using an incorrect key in the decryption of the analysis the analysis shows a garbled text and becomes and erroneous solution. With the classical cipher we can see how an encryption and decryption works and we can have an idea of how difficult it is to break the encryption if you do not have a correct starting point. This can show us where the security is weak and can be broken and show how a better understanding of cryptography methods can protect our data in this age of information technology[7].

Future Work

In the future what can be done is use different methods of analysis and see how the result vary and add a variety of different ciphers to analyze.

Acknowledgements

I give thanks to my advisor Dr. Alfredo Cruz for his guidance, advice, and patience in the process of this project.
To my family for being there and giving all the support needed for me to finish this work.

References

[1] Norvig, P. (2012). English Letter Frequency Counts: Mayzner Revisited or ETAOIN SRHLDCU. Retrieved September 24, 2019, from https://norvig.com/mayzner.html
[2] Mayzner, M. S., & Tresselt, M. E. (1965). Tables of single-letter and diagram frequency counts for various word-length and letter-position combinations [by] M.S. Mayzner and M.E. Tresselt. Goleta, Calif.] Psychonomic Press.
[3] Rigel. (2020, November 01). Retrieved January 1, 2021, from https://en.wikipedia.org/wiki/Rigel
[4] Stallings, W. (2011). Cryptography and network security: Principles and practice. Boston: Prentice Hall.
[5] Pfleeger, C. P., & Pfleeger, S. L. (2012). Analyzing computer security: A threat/vulnerability/countermeasure approach. Upper Saddle River, NJ: Prentice Hall.
[6] Whitman, M. E., & Mattord, H. J. (2005). Principles of information security 2nd. ed. Boston, MA: Thomson Course Technology.
[7] Quade, P. (2019). The digital big bang: The hard stuff, the soft stuff, and the future of cybersecurity. Indianapolis, IN: Wiley.