

APRIL 28, 2017



N-GRAM CONDITIONAL ENTROPY IN SEQUENCE OF INDUS VALLEY SCRIPT

ARTIFICIAL INTELLIGENCE
Dr. Debasis Mitra

ALEESHA MISHRA, RAHUL DEV MISHRA

Contents

Problem Statement.....	2
Introduction	2
Corpus Details	3
Statistical Analysis.....	4
N-Gram Frequency.....	5
Cumulative Frequency Distribution	7
Analysis of Text Beginners	8
Analysis of Text Enders	9
Text Length Distribution	10
Information Theoretic Metrics	11
Entropy.....	11
Mutual Information	11
Conclusion.....	12
References	12

Problem Statement

The Indus river, which is currently in Pakistan had a civilization established around 2600 BCE called Indus Valley Civilization. It has the one of the most complex script which is yet to be deciphered. Its complex inscription patterns are a challenge to the linguists. Building on the research paper's analysis [1], we approach with the statistical analysis which uses n-gram model to analyze the syntax of the Indus script. Our focus is on the conditional probability of unigram, bigram, trigram, and quadrigram sequences. The model's performance is measured using the information theoretic metrics: entropy calculation and mutual information.

Introduction

The Indus Script is the ancient undeciphered script, which are collection of symbols written by Indus Valley civilization between 2600 and 1900 BCE [1]. The scripts are collected from Mohenjo-Daro, Harappa, Kalibangan and Lothal which are currently on the border of India, Pakistan, and Afghanistan. Most of the symbols are inscribed on different objects as seals, sealings and other copper or bronze artifacts as shown in Fig 1. The length of each inscription or text is the major limitation to deciphering the script. Due to its short length, identification of any syntax or pattern is intricate. In 1970s, Iravatham Mahadevan published a collection of those symbols in form of a corpus and concordance composed of 3573 texts and 417 distinguished symbols. He also substantiated the inscribed symbols as texts are mostly written from right to left as there were cramping towards the left of the texts. But there are texts written from right to left, up to down and down to up as well.



Fig 1: Example of Inscribed symbols on the seals

Corpus Details

The corpora are the collection of text and speech. For example, the famous English corpus is Brown Corpus. The corpus, authors have used for their statistical analysis is EBUDS which is the subset of Mahadevan's electronic concordance and corpus M77[1][3]. Our analysis used the corpus which we made mostly from the texts collected from Mohenjo-Daro, Kalibangan and Lothal. Our corpus is made up of 640 texts with 257 distinct symbols which otherwise known as signs. These signs are broadly classified into two types: basic signs which are composed of only one sign and composite signs composed of two or three basic signs as shown in figure 2.



Fig 2: Example of Basic and Composite Signs

The making of the corpus is composed of certain steps. Given a text of Indus script, we are required to find the identify each symbol with a unique number following a numbering system in which Mahadevan has allocated each symbol a unique number starting from 1[3] as shown in fig 3. Then we need to write the numbers separated by a hyphen from right to left. With the help of collection of these number series, which is now in a computer readable format is processed to get the statistical results. The sample corpus is shown in fig 4.

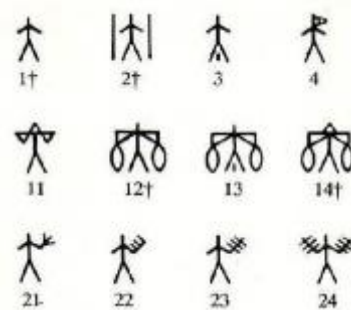


Fig 3: Sample of Mahadevan's unique numbering system

MOHENJODARO		
1194	100101	𑀩𑀲𑀸
1195	100101	𑀩𑀲𑀸
1196	103509	𑀩𑀲𑀸
1197	100101	𑀩𑀲𑀸
1198	100101	𑀩𑀲𑀸
1199	100101	𑀩𑀲𑀸
1200	100101	𑀩𑀲𑀸
1201	100101	𑀩𑀲𑀸
1202	100101	𑀩𑀲𑀸
1203	100101	𑀩𑀲𑀸
1204	100101	𑀩𑀲𑀸
1205	100101	𑀩𑀲𑀸
1206	100101	𑀩𑀲𑀸
1207	100101	𑀩𑀲𑀸
1208	100101	𑀩𑀲𑀸
1209	100111	𑀩𑀲𑀸

329-89-171-8-342

Fig 4: Corpus in electronic format

Statistical Analysis

We have used N-grams of text for the statistical analysis of the Indus Script. N-grams are clearly the combinations of text, it could be letters or words, of length n that is available in the given text. The n-gram of size 1 denotes unigram, 2 denotes bigram and 3 denotes trigram and so on. Claude Shannon postulated the idea of n-gram when he raised the question: "For a given sequence of letter, what is the likelihood of the next letter?" [10]

A n-gram model uses $x_{i-(n-1)}, \dots, x_{i-1}$ to predict x_i . N-gram is the probability of a word given its history. This is also defined using the conditional probability: $P(x_i | x_{i-(n-1)}, \dots, x_{i-1})$. In other words given a history h and a word w , the conditional probability that w will follow h is given by $P(w | h)$. Formally conditional probability can be defined as:

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

$$P(B/A) = \frac{P(A \cap B)}{P(A)}$$

Fig 5: Conditional Probability Equation

For example, consider the text: “The statistical analysis of”. We would like to find the probability of the word “the” following the text. This is given by answering the question that how many times in the history h has the word w followed, which is given by:

$$P(\text{the} \mid \text{The statistical analysis of}) = \frac{\text{count}(\text{The statistical analysis of the})}{\text{count}(\text{The statistical analysis of})}$$

With this concept of n-gram, we have observed the following statistical details:

N-Gram Frequency

Frequency is the ordered ranking of the occurrence of each text in the corpus. We subjected the corpus for unigram, bigram, trigram and quadrigram frequency analysis.

Unigram Frequency:

The frequency analysis using 1-gram is nothing but the count of occurrence of each symbol in the corpus. The frequency distribution for unigram is shown in Fig 6.

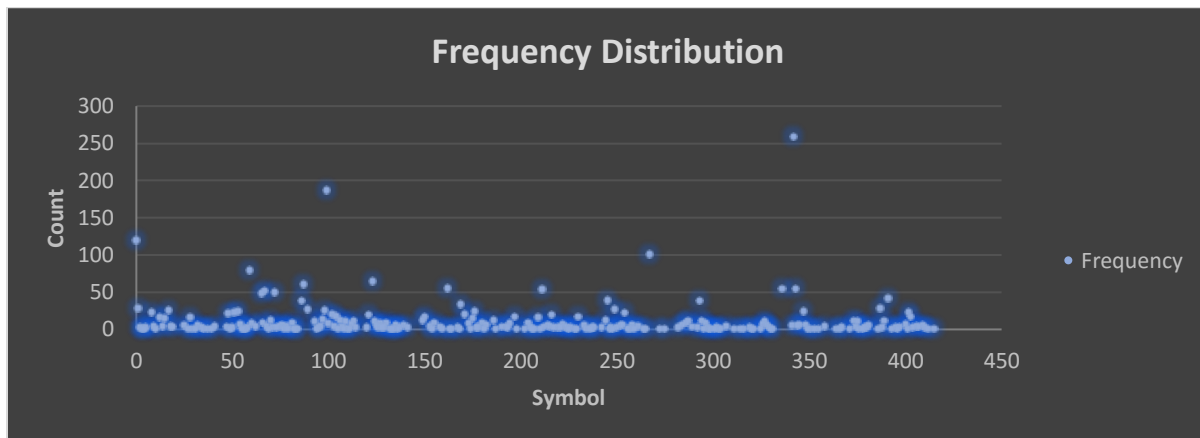


Fig 6: Frequency Distribution of Unigram

It can be seen that the symbol 342 occurs maximum number of times followed by symbols 99 and 267. The symbol 0 refers to illegible symbol i.e. the historians couldn't decipher the exact symbol and hence it has been left empty in the sequence.

Bigram Frequency:

Next using 2-gram, we observed the frequency of occurrences of pair of symbols.

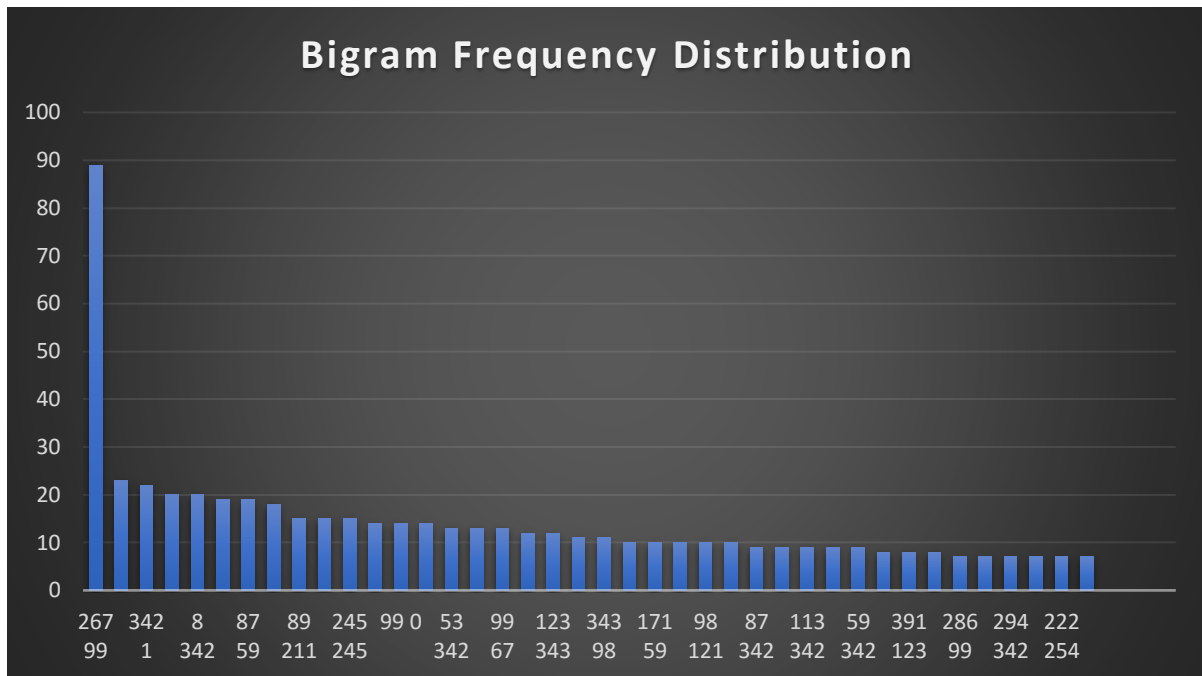


Fig 7: Frequency Distribution of Bigram

In this case, the pair 267 and 99 are the most frequently occurring pair, which is then followed by the pairs 342 and 1 & 342 and 8.

Trigram Frequency:

Next, we were interested in finding the occurrence of 3-gram i.e. taking three symbols at a time and finding their presence in the corpus.

From the graph in Fig 8 it can be seen that the triple with the symbols 336, 89 and 211 occurs the most.

Quadrigram Frequency:

When we analyzed by taking 4 symbols at a time we found out that the 4 symbols 209, 343, 98 and 121 appears maximum number of times.

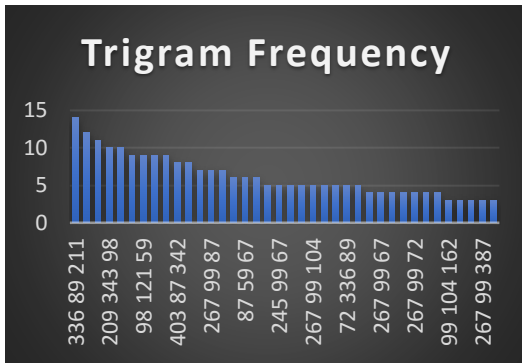


Fig 8: Frequency Distribution of Trigram

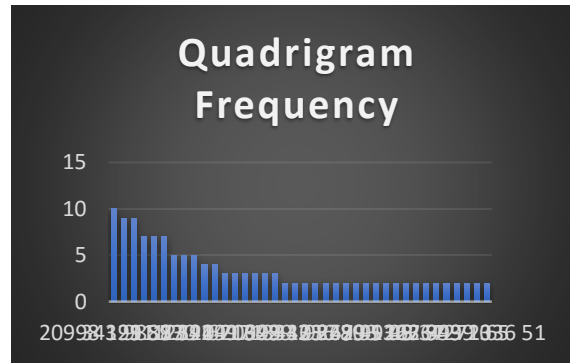


Fig 9: Frequency Distribution of Quadrigram

To summarise the analysis, the table in Fig 10 shows the most frequently occurring symbols as per the type of analysis





Type of Frequency Analysis	Most Frequent Symbol(s)
Unigram	 342†
Bigram	 267† 99
Trigram	 336† 89† 211
Quadrigram	 209 343† 98† 121†

Fig 10: Most frequently occurring sign in n-gram models

Cumulative Frequency Distribution

The cumulative frequency distribution is the sum of the class and all the classes below it.[13] The graph in Fig 11 depicts the cumulative frequency of all the unigram symbols, only text enders and only text beginners. From the graph one can deduce that almost 65 – 70 symbols in the corpus account for close to 80% of all the text in the corpus, which in turn means that the remaining number of symbol occur highly infrequently in the corpus. Hence, this also agrees with the Zipf-Mandelbrot law.

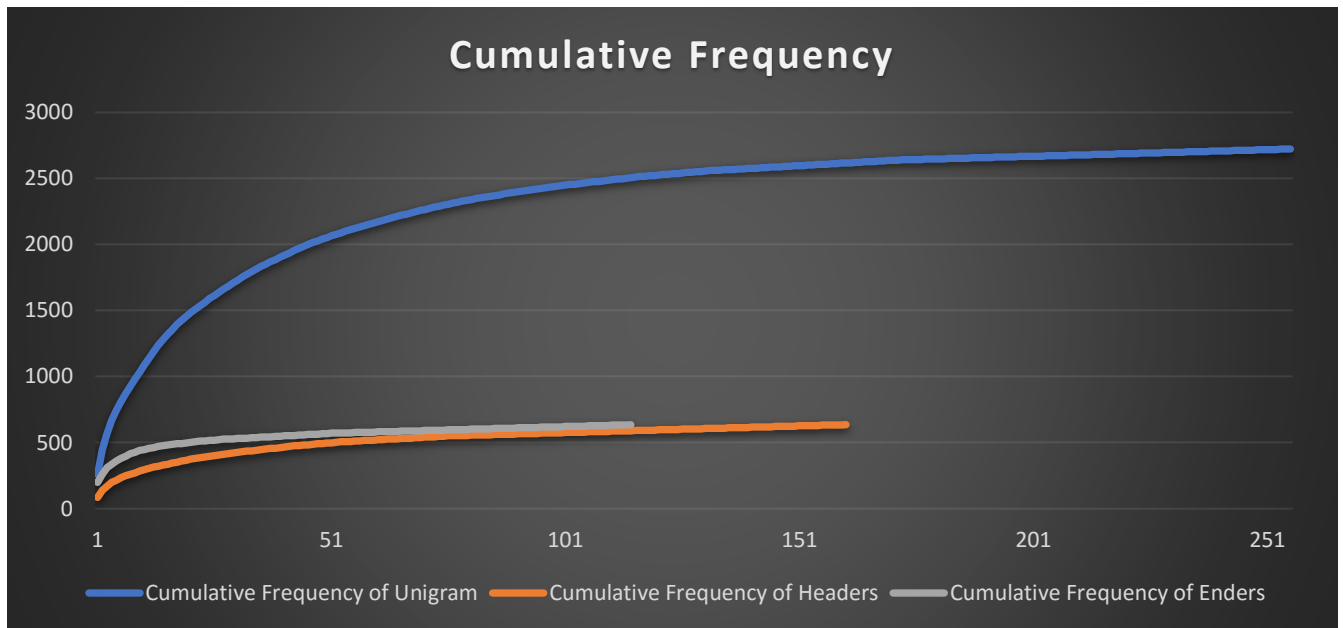


Fig 11: Cumulative Frequency Distribution of Unigram, Headers and Enders

Analysis of Text Beginners

We now present the results of bigram analysis on Text Beginners. Text Beginners are those symbols which occur as the very first text or symbol of the line. Let's say the token # indicates the beginning of a line. By convention, the unigram probability for the start token is unity, $P(\#) = 1$, since every text must begin with #. The probability of sign a being a beginner is then $P(\#a) = P(a|\#)$, since $P(\#) = 1$.

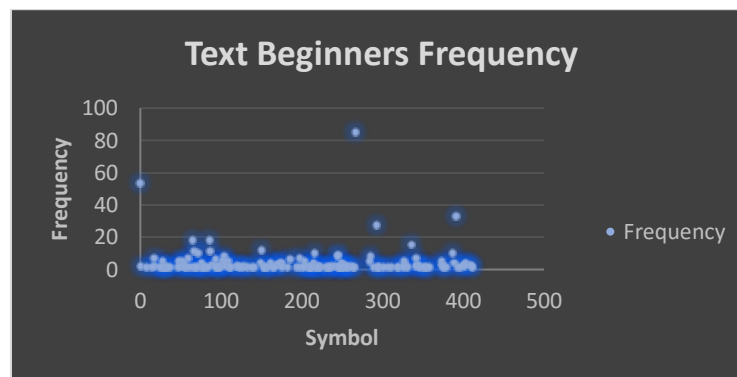


Fig 12: Frequency Distribution of Text Beginners

From our analysis, we found that there are 167 distinct symbols which occur as text beginners. Out of these symbols, symbol 267 appear highly frequently as text beginner.

Now after finding the text beginners, we were interested in finding the symbols following the text beginners. We chose the top three text beginners i.e. symbol 267, 391 and 293 for this analysis. It can be inferred from the plots of conditional probabilities in Fig 13, i.e. $P(b|267)$, $P(b|391)$ and $P(b|293)$ that the

text beginners 267, 391 and 293 are more selective in terms of the number of signs which can follow them.

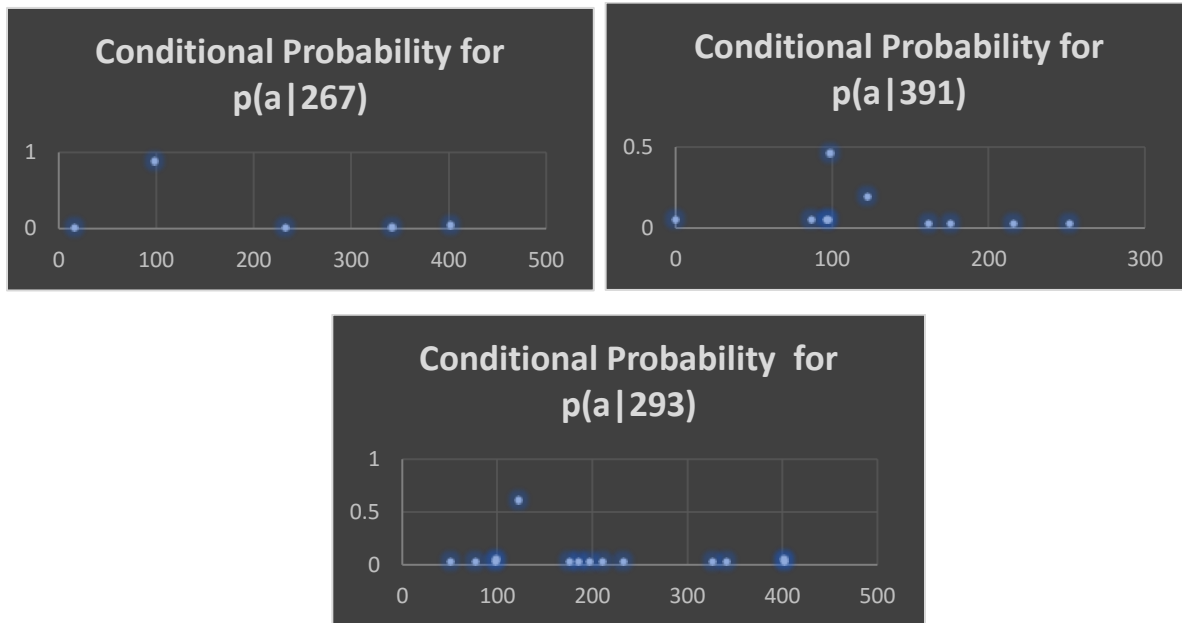


Fig 13: Conditional Probability of Text Beginners

Analysis of Text Enders

Similarly, if we consider \$ as the end of a text, we are interested in observing the symbols that occur as a text ender. As above, using bigrams of the symbols, we had to find the frequency with the given condition that the symbol must precede the end of a text \$.

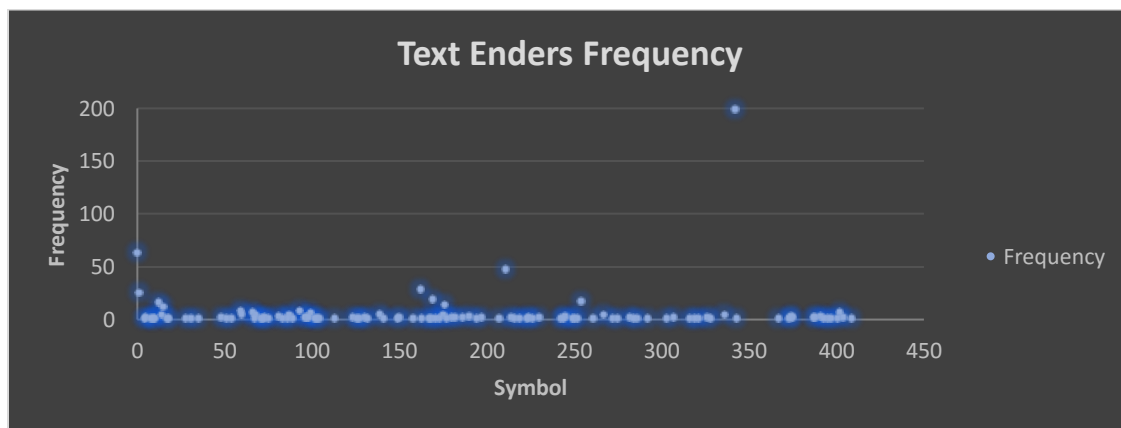


Fig 14: Frequency Distribution of Text Enders

Unlike text beginners, the population of symbols as text enders is relatively small. There are only 116 symbols occurring as text ender which symbolizes that not much flexibility of symbols is present while choosing a text ender. Also, it draws a correlation of a sign with

other signs preceding or following it using the results of bigram analysis. From the graph, it is clear that the symbol 342 occur the most as text ender.

Similarly, we chose the top three text ends i.e. symbol 342, 176 and 211 to understand the occurrence of symbols that precede the text ends. This is given by conditional probability $P(342|a)$, $P(176|a)$ and $P(211|a)$ for the text ends as shown in Fig 15.

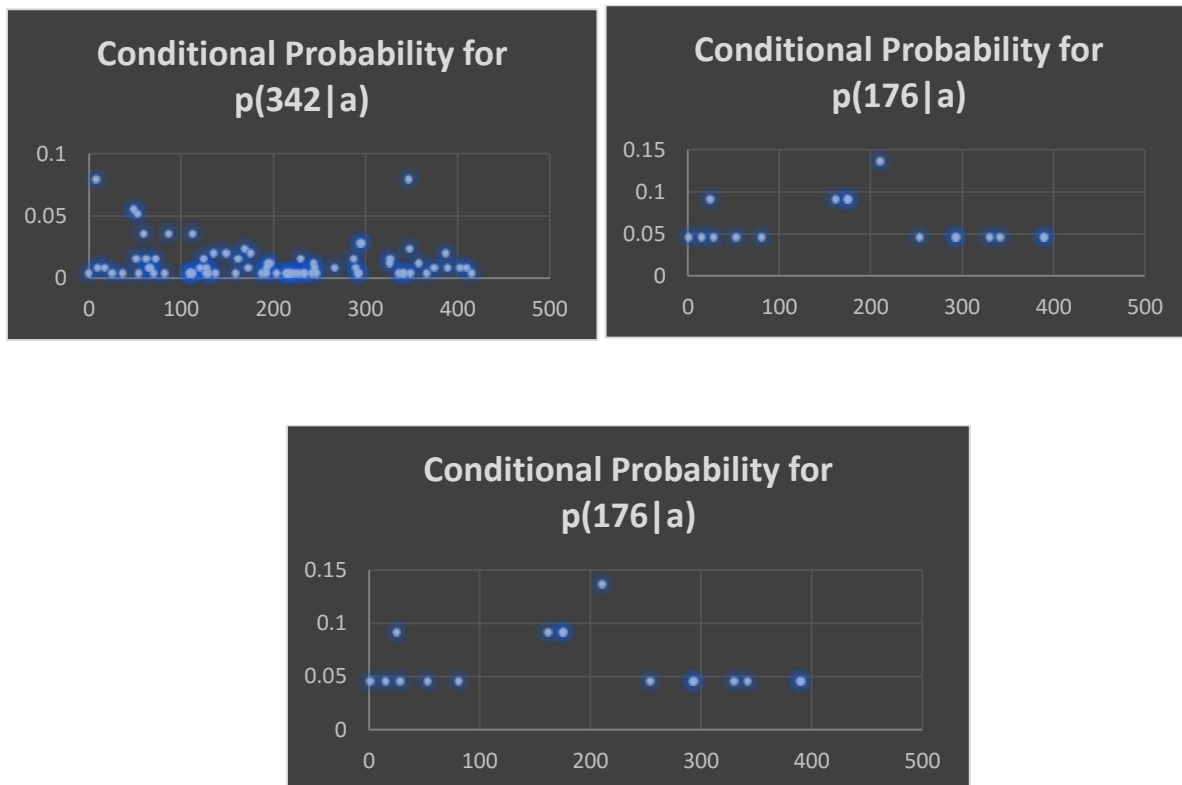


Fig 15: Conditional Probability of Text Enders

Text Length Distribution

We have analyzed the text length distribution of the corpus. The text length distribution is shown in Fig 16. From the distribution, it can be observed that text with length 3 and 4 occurs predominantly in the corpus. The corpus that we used has around 650 lines of text out of which lines of text length 3 and 4 account for almost 40% of the corpus population.

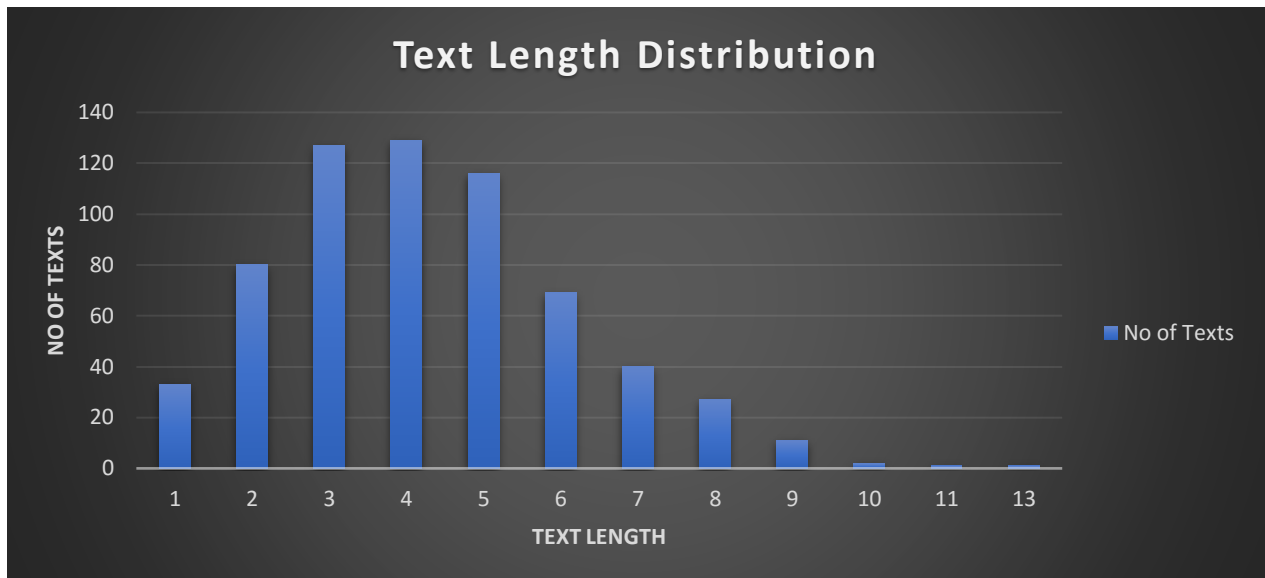


Figure 16 Text length distribution in the corpus used in the analysis

Information Theoretic Metrics

These are the metrics used to determine the measurement of how well a language model models a natural language or corpus. We have used entropy and mutual information metrics on our corpus to get a statistical result of the models.

Entropy

An entropy is the calculation of expected value of the information in each text. Given a set of probabilities (a probability distribution) $P = \{p_1, p_2, \dots, p_n\}$, we define entropy as :

$$H(P) = \sum_{i=1}^n p_i * \log(1/p_i)$$

where n is the number of outcomes

The range of the entropy is based on the number of outcome. The range of entropy is defined as:

$$0 \leq \text{Entropy} \leq \log(n).$$

We have calculated the entropy based on our corpus shown in fig 5.

Models	Entropy (H)
Unigram	6.44
Bigram	9.33
Trigram	9.94
Quadrigram	9.60

Fig 17: Entropy Calculation of n-gram models

Mutual Information

The mutual information is the measurement of mutual dependence of two given models. This concept is intricately linked with entropy of a given probability P. The information contained in the probability P is defined by mutual information. Given two model X and Y, the mutual information is :

$$I(X; Y) = \sum_{y \in Y} \sum_{x \in X} p(x, y) \log \left(\frac{p(x, y)}{p(x)p(y)} \right)$$

We calculated the mutual information of Bigram based on unigram as 2.89.

Conclusion

The immense challenge that we faced while working on this project was to get an electronic form of corpus. While it took us plenty of time in converting the corpus to an electronic format, the size of the corpus is somewhat limited. Using this corpus, we have performed the statistical analysis of the following measures:

1. Zipf-Mandelbrot Law
2. Cumulative Frequency Distribution
3. Bigram Probability
4. Conditional Probabilities of Text Enders and Text Beginners
5. Entropy
6. Mutual Information
7. Text Length Distribution

We have keenly observed that even with the limited size of the corpus, the statistical output of our tests or measurements is in-line with the results of the authors.

References

- [1] Statistical Analysis of the Indus Script Using n-Grams. Nisha Yadav1*, Hrishikesh Joglekar2, Rajesh P. N. Rao3, Mayank N. Vahia1, Ronojoy Adhikari4*, Iravatham Mahadevan5
- [2] Foundations of Statistical Natural Language Processing. Manning C, Hinrich Schutze (1999) Cambridge: MIT Press.
- [3] The Indus Script: Texts, Concordance and Tables, Memoirs of the Archaeological Survey of India. Mahadevan I (1977)
- [4] A mathematical theory of communication. The Bell System, Technical Journal 27: 379–423. Shannon CE (1948)
- [5] A statistical approach for pattern search in Indus writing. International Journal of Dravidian Linguistics37: 39–52. Yadav N, Vahia MN, Mahadevan I, Joglekar H (2008)
- [6] Cracking the Indus Script: A Potential Breakthrough ([Link](#))
- [7] www.harappa.com
- [8] Indus Script ([Link](#))
- [9] Entropy ([Link](#))
- [10] N-Grams ([link](#))
- [11] Text Mining, Analytics & More ([link](#))
- [12] N-gram Wikipedia ([link](#))
- [13] What is cumulative frequency distribution ([link](#))