THE UNIVERSITY OF ADELAIDE

SCHOOL OF ELECTRICAL & ELECTRONIC ENGINEERING

ADELAIDE, SOUTH AUSTRALIA, 5000

Cracking the Voynich manuscript code

Prof. Derek Abbott, Yaxin Hu

ELEC ENG MASTER PROJECT NO. 141

Date submitted: 22 April 2016

Supervisor: Prof. Derek Abbott and Dr. Brain Ng.

Signature of Supervisor:

Contents

Abstract

This project is to crack the Voynich manuscript which is an unknown hand-written book. This book is considered to be an unknown language, cipher code or hoax. This thesis proposal is aimed to provide methods in determining possible features of the Voynich manuscript. All the methods are related to data mining, computer coding and statistical methods. There will be specific explanation of the methods that will be carried out in the whole project. Furthermore, this document provides the management of this projects.

1. Introduction

1.1 Project Background

The Voynich the manuscript was created in the first half of the fifteenth century (probably between 1404 and 1438) [1]. No one today knows what it says or who wrote it. The book is in a strange alphabet. At 1912, a book collector named Wilfried Voynich found it in an Italian Jesuit college [1].

Since this book cannot be read, it is divided into six different sections by illustrations with different styles and images:

a) **Herbal:**

   There are one or more plants on each page, which is a format of European herbals [2].

b) **Astronomical**

   There are circular diagrams such as suns, moons, and stars which suggest this part as something about astronomy or astrology [2].

c) **Biological**

   Mostly naked women shows that this part should be biological section [2].

d) **Cosmological**

   Circular diagrams of obscure nature make this section as cosmological section [2].

e) **Pharmaceutical**

   Drawings of isolated plants parts and objects resembling apothecary jars show that this section should be something about pharmaceutical [2].

f) **Recipes**

   This part are full pages of text in short paragraphs [2].

1.2 Aim

The aim of this project is to search the text and determine whether there are any possible features that can be used to decode the Voynich manuscript using statistical methods. The investigation of languages and linguistics is required to be processed with the unknown text. But, it is not necessary to fully decode the Voynich manuscript since it is not possible to be done in a one-year project.

1.3 Significance and Motivation

With statistical methods, trying to carry out a project that is used to investigate the language and linguistics of an unknown book is an attempt that may beyond excellent. Trying to find any features of relationships and patterns of the Voynich manuscript could be used to decode the unknown text with unknown languages. It may contribute significant progress in attempting decode a part of the book. The outcomes can be used to further linguistic or language decryption. Such as information decoding, search engines and data mining. Specific applications such as Google, Turn-it-in, Google translate, Yahoo, and Grammarly.

1.4 Technical Background and Challenges

Data mining as an important part in this project, it is the foundation of analysing the Voynich manuscript. It is an interdisciplinary subfield of computer science that is used to process the discovering patterns in large data sets involving methods such as artificial intelligence, machine learning, statistics and database systems [3]. In this project, data mining should be used to test and analyse the specific linguistic and language features.

As the Voynich manuscript has been transcript into English alphabet version with several kinds of method such as European Voynich Alphabet (EVA). There is an example of a part of text of the Voynich manuscript and EVA in appendix 1.

Since it is an unknown hand-written book for more than five century, there is no useful material that can be used to determine the symbols of the manuscript. The way that can be used to determine word allocation is the spacing between different sets of symbols. Also, it is believed that this manuscript should have several pages missing. Also, there is strong evidence that many of the book's bifolios were reordered at various points in its history, and that the page order may be different from what it is today [4].

Due to the pre-study, no useful technical can be used to translate or determine the manuscript [5]. Therefore, what we can use is basic linguistics and languages.

1.5 Knowledge Gaps

This project is a decoding project, therefore, it will require a lot of software work with variety of statistical methods to access the aim. None of us have master so much kinds of particular knowledge in different subjects. Therefore, each of us will be required to develop software programming skill and statistics skill. Beside, since we have no evidence that any kind of particular skill can be used to solve this project, several different kinds of skill will be needed to grasp in processing the Voynich manuscript.

2. Requirements

Although it is not necessary to fully decode the Voynich manuscript, this project should present several outcomes:

a) A clear investigation of language and linguistics of the Voynich manuscript

b) Any critical attempts.

c) Any possible results within the attempts.

d) Any hypotheses within the results.

e) Any decoded text if possible.

3. Related work

The Voynich manuscript has been investigated for almost a century by a large number of professors and specialists. They have contributed several possible hypotheses that can be used in this project through their analysis.

Stephen Bax (2014) states that the Voynich manuscript is not a hoax, and it is probably an explanatory treatise which appears to act as a type of manual for interpreting and transmitting information across cultures [6]. If it is possible, it may lead to a new direction of analysing the Voynich manuscript. The work centrality may should move to this specific section.

Another work that may contribute possible impact is Gbariel Landin (2001)'s "Evidence of Linguistic Structure in the Voynich Manuscript Using Spectral Analysis". He used statistical method to character the Voynich manuscript with nature languages. Zipf's law that he used to analyse on entropy in this book shows that there may exist some linguistic form in Voynich manuscript because the long range correlation, length modal and periodic structures in the Voynich manuscript [7].

A multiple tests of the Voynich manuscript carried out by Roush (2014) shows that there may needs several kinds of attempting such as:

a) Word length distribution
b) Word and image association
c) Word recurrence intervals
d) Zipf's law
e) N-Grams [8]

They made a brief conclusion of these attempting, however, none significant result is approached by them, which may indicated that further attempting should be taken.

Another statistical investigation token by Costa (2013) on the Voynich manuscript in related to vocabulary statistics shows that the Voynich manuscript is similar to natural languages [9].

4. Proposed Methods

4.1 Characterisation of the Voynich manuscript

Mainly, there are several task in characterisation of the manuscript.

a) Total words in the whole manuscript
b) Total characters in the whole manuscript
c) Unique words
d) Unique character
e) Frequency of words

f) Frequency of character

g) Character that only appear at the start or the end of words

Compare these statistical results with known languages may contribute significant progress in determining the features of the Voynich manuscript.

4.2 Text investigation: Digits

Digits investigation will be our first breakpoint in decoding the Voynich manuscript. This part will be taken following by several steps.

a) Find patterns in known language digits such as Roman digits and Greek digits.

b) Trying to search any words in the Voynich manuscript that may related to any patterns in known language digits and locate all of them.

c) Translate all the possible words and check whether these words conform to the images that may nearby the words.

d) Use statistical methods analyse any possible digital patterns that may conform to the Voynich manuscript.

e) Decode all the digits if step d is success.

Digital investigation may contribute significant influence in the whole investigation. If not, the follow investigation will become more important.

4.3 Illustration investigation

Illustrations investigations is associated to the digitals investigation. It will follows several steps:

a) Locate all the images that contains one thing that appears more than once in an image.

b) Number the time that things appears in each image.

c) Trying to search words nearby the image that may conform any digital patterns in known language digits.

d) Decode all the digits if step c is success.

Illustration investigation is a different way that used to investigate digits. The difference is that there may contains different kinds of encryption in the Voynich manuscript if it is encrypted, therefore, it is a way to ensure that digital investigation can solute this possibility.

4.4 Marginal symbol investigation

Marginal symbol investigation is a method that is used to investigate the last section of the Voynich manuscript. In the recipes section, there are many solid stars or hollow stars in front of each paragraph. This method will to goes in the following steps:

a) Locate all the stars in recipes section.

b) Count the number of solid stars and hollow stars separately.

c) Search the texts nearby all the stars that may contain any possible numbers.

d) Compare all the recipes sections and try to find any pattern for any possible numbers.

Marginal symbol investigation is a way that if both digital investigation and Illustration investigation cannot get significant result. It may provide another breakpoint in the whole digital analysis.

5. Project Management

5.1 Time Management

As shown in figure1, time management is divided into five parts. Background research should be taken from week 1 to week 5 in semester 1. After that, text analysis will be taken between week6 and week 7. Then illustration investigation should begin from week 8. Following should be marginal symbol research which will begin from week 10. The last task will be translation of any possible test from week 5 to week 9 in semester 2.

No.	Task	Week
	Semester 1	
1	Background research	1
2	Text analysis	6
3	Illustration investigation	8
4	Marginal symbol research	10
	Semester 2	
5	Translation	5-9

Figure 1 Time management

5.2 Risk Management

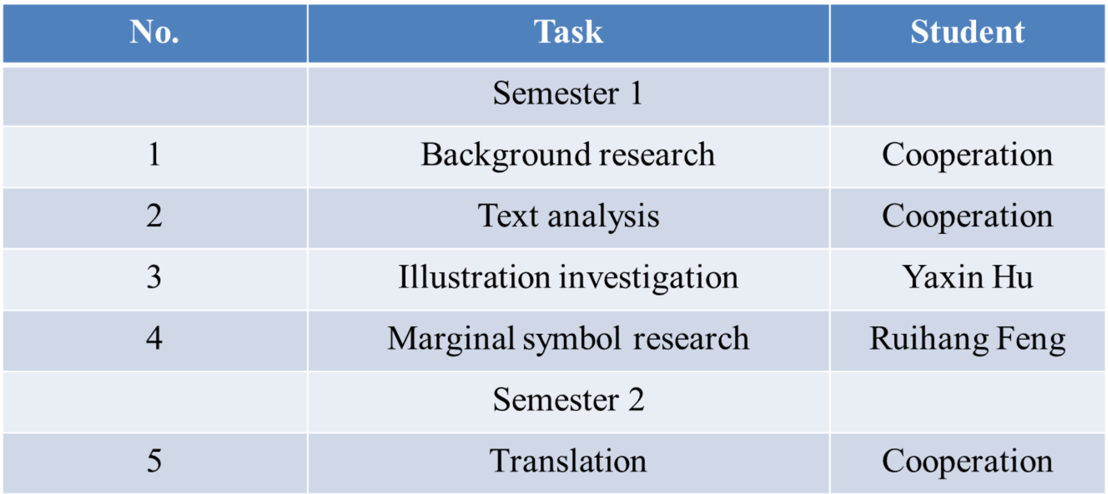
As shown in figure 2, there should be six risk. There important risk that may cause important impact also may occur in a significant probability should be mismanagement of time, lack of references and health issues. Each of them may cause significant impact to the final result.

NO.	Risk	Probability	Impact
1	Mismanagement of time	Moderate	High
2	Loss of data or files	Low	High
3	Team member's quit	Low	High
4	Lack of references	High	High
5	Health issues	Moderate	Moderate
6	Supervisor unavailable	Low	Moderate

Figure 2 Risk management

5.3 Task Allocation

As shown in figure 3, the task management is associated to the time management. Except the illustration investigation will be carried out by Yaxin Hu, marginal symbol research will be carried out by Ruihang Feng, other tasks will be done by both of us.



No.	Task	Student
	Semester 1	
1	Background research	Cooperation
2	Text analysis	Cooperation
3	Illustration investigation	Yaxin Hu
4	Marginal symbol research	Ruihang Feng
	Semester 2	
5	Translation	Cooperation

Figure 3 Task allocation

5.4 Budget

a) 500 AUS dollars for each member.

b) Research need to be carried out further research.

c) All program that need to be used are available on University system.

d) Major work are based on computer.

5.5 Management Strategy

Meetings will be the main way to exchange project progress between project members and supervisors. At least one meeting should be held between project members each week and a minimum of one meeting should be held between project members and supervisors each three weeks.

These meetings should involve phase progress, issues occurs, further ideas that could help processing the project and any possible results.

6. Conclusion

With a literature review of the Voynich manuscript, the background and proposal methods have been settled down. As discussed in section 4, digital investigation will be the main breakpoint in the whole project. All possible methods will be carried out to determine any possible features of the Voynich manuscript.

For now, the steps taken by project members is correspond to the time management. All the thing that have been done would contribute significant influence in the whole project.

7. Reference

[1] Schmeh, Klaus (January–February 2011). "The Voynich Manuscript: The Book Nobody Can Read". *Skeptical Inquirer.* Retrieved 2013-09-05.

[2] Shailor, Barbara A.,Beinecke MS 408, Yale University, Beinecke Rare Book and Manuscript Library, General Collection of Rare Books and Manuscripts, Medieval and Renaissance Manuscripts, accessed 24 June 2013.

[3] "*Data Mining Curriculum*". ACM SIGKDD. 2006-04-30. Retrieved 2014-01-27.

[4] Barabe, Joseph G. (McCrone Associates) (April 1, 2009). "Materials analysis of the Voynich Manuscript". Beinecke Library.

[5] G. Landini, "Evidence of Linguistic Structure in the Voynich Manuscript Using Spectral Analysis," *Cryptologia*, pp. 275-295, 2001.

[6] B. Stephen, *"A proposed partial decoding of the Voynich script",* www.stephenbax.net, Version 1, January 2014.

[7] G. Landini, "Evidence of Linguistic Structure in the Voynich Manuscript Using Spectral Analysis," *Cryptologia*, pp. 275-295, 2001.

[8] B. Shi and P. Roush, "Semester B Final Report 2014 - Cracking the Voynich code,"
University of Adelaide, Adelaide, 2014.

[9] D. R. Amancio, E. G. Altmann, D. Rybski, O. N. Oliveira Jr. and L. d. F. Costa, "Probing the Statistical Properties of Unknown Texts: Application to the Voynich Manuscript," PLoS ONE 8(7), vol. 8, no. 7, pp. 1-10, 2013.

Appendix

Appendix 1

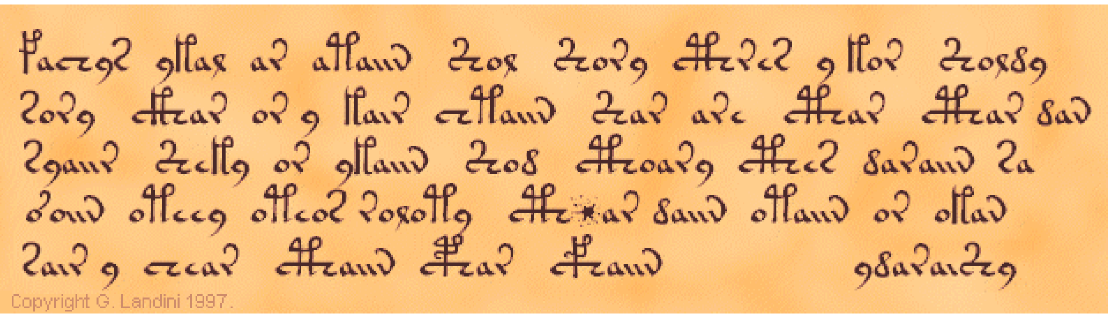
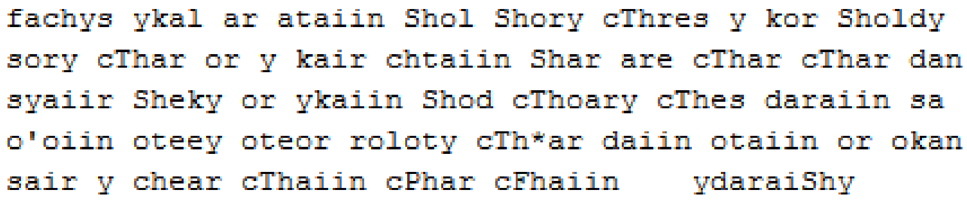

Copyright G. Landini 1997.

Figure 4 An original part of the Voynich manuscript



```
fachys ykal ar ataiin Shol Shory cThres y kor Sholdy
sory cThar or y kair chtaiin Shar are cThar cThar dan
syaiir Sheky or ykaiin Shod cThoary cThes daraiin sa
o'oiin oteey oteor roloty cTh*ar daiin otaiin or okan
sair y chear cThaiin cPhar cFhaiin    ydaraiShy
```

Figure 5 The transcription of this part of the Voynich manuscript