The University of Adelaide

Code Cracking: Who Murdered the Somerton Man?

2nd Draft Report

Yifan Ma (a1658524)

3rd June 2016

1. Introduction

1.1 Summary of the project

This project is related to an unsolved possible murder happened in Adelaide in 1948. The project group is suggested to draw the case closer to the truth by using engineering knowledge and skills. The team will work on two aspects of the project: analyzing the code and the analysis of mass spectrometry data of the victim's hair.

1.2 Background of the case

1.1.1 The Victim

The so called Somerton Man was found dead at 6.45 on 1st December, 1948. He was lying against a sea wall on Somerton Beach peacefully. The victim was a Caucasian male in his 40's, 180 centimeters high and with light sand to white colored hair. There were several evidences indicating that he used to be a ballet dancer. He was dressed decently with well-designed jacket and shirt. His trousers and shoes were also tidy; there was no sign of struggle and carrying of the corpus.

Figure 1 the Somerton Man

At the same time, some personal stuff in the victim's pockets was found. One of them which drew people's attentions the most was a wrinkled piece of paper printed

“Tamám Shud”. The case was hence named by the words on the paper scrap.

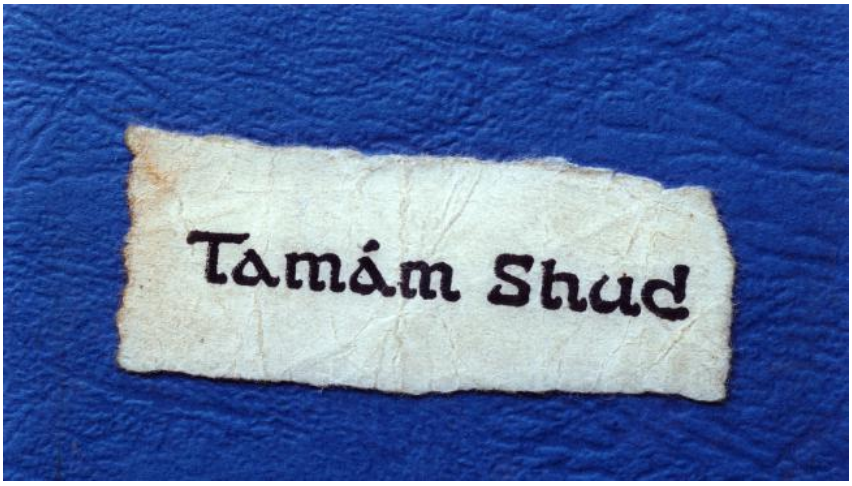

Figure 2 the paper scrap [2].

1.1.2 The paper scrap

The words on it were written in Persian (Cyrillic). It means *to the end* in English. The paper scrap was later confirmed to be torn off from a book: an uncommon edition of *The Rubáiyát*. Six month later the related book was found in a stranger’s car by the Police. The owner of the car had no idea about the victim and the dumped book.

1.1.3 Mysterious Code

The code was found under the irradiation of ultraviolet light when people inspecting the aforementioned book. As Figure3 shows, it was a series of English letters. The code still remains mysterious until today.

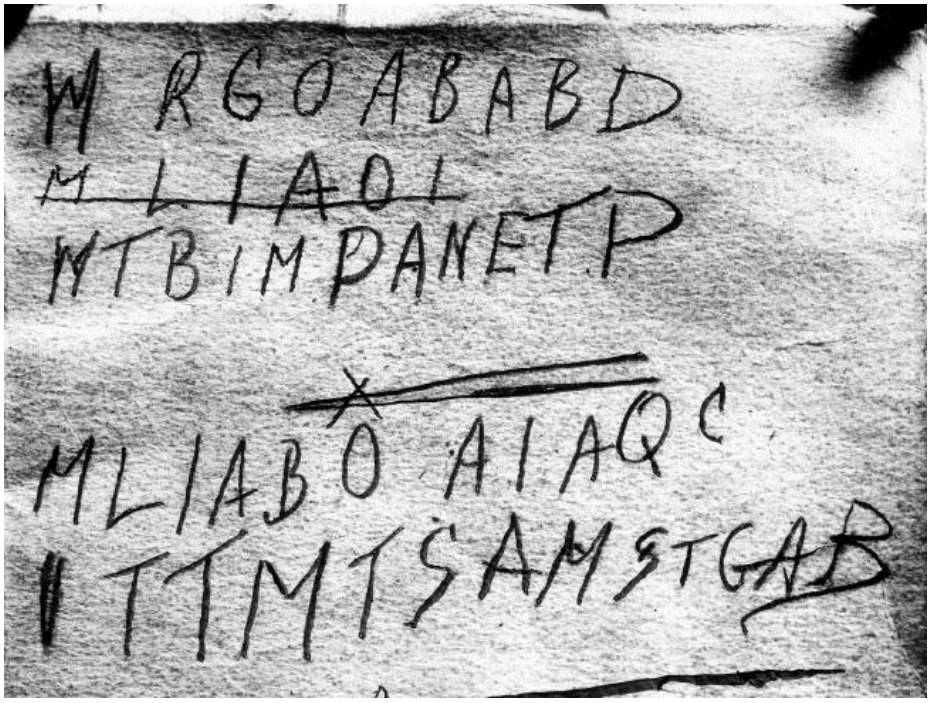

Figure 3. The hidden code [3]

1.1.4 Victim's hair

The hair was found in the plaster mode people made according to the victim's body. Chemical and biological analysis had been done by the University of Adelaide's research team. By sorting and analyzing the hair's data it is possible to draw out some clues about what environments the victim had been in before he met his demise.

2. Aims

The project is related to an unsolved murder case happened in Adelaide in 1948[4]. Based on the secret code and the hair data of the victim, the project group is supposed to pull the knowledge closer to the killer.

Despite the fact that the code is waiting to be cracked, previous groups had made outstanding outcomes which have reduced the gap enormously, while it is always

necessary to view outcomes worked out be other people critically.

The first aim is to review the statistical analysis of the letters as the previous group had not involved all possibilities of ambiguous letters, nor had they implemented the analysis with sufficient text samplings. Given the assumption that the code is a set of initializes, it is expected to generate more convincing conclusion to show whether the code was extracted with origin English words, or words in other languages.

The second aim is getting the decrypts of the code using Omar Khayyan as a one-time pad. Then filter the results to get top-20 most common words. As people have no idea where the one-time pad starts, it is vital to loop through the whole book (10,500 letters in total). The output decrypts will be as many as 10,000 and the expected outcome will be 20 words exactly.

Another aspect of the project aim is to analysis the mass data generated by scanning the victim's hair. By analyzing the trends of several chemical components it is possible to obtain more detailed clues about the victim's living conditions before he met his demise.

3. Motivation

As the previous chapter stated, this project is relating to an unsolved possible murder case which has been remaining in the public's interest for almost 70 years. The motivation of this project could be divided into the following three aspects:

3.1 For the victim and his family

This is the most important reason why the project is undertaken. The victim has been resting in West Terrance Cemetery without a name for decades. It will be meaningful

if the identity of the victim could be determined. This is also for the victim's family whom lost their relative and probably had no idea about it.

3.2 Satisfy people's curiosities

Since the happening of the Taman Shud Case, the general public has been giving utmost attention to it. There are so many questions in this mysterious case. People want to know whether it was a murder case or a suicide, what the Somerton Man came for and why he died without a name for such a long time, what the code stood for, etc.

3.3 Spy Hypothesis

The Somerton Man was suspected of being a spy from the former Soviet Union in part because of the hidden code found in the book. Another reason of the generation of the spy hypothesis was that the case happened during the cold war period. Hence, it comes to one of the motivations for this project.

4. Significance

The project is aiming to pull the case closer to the truth, make contribution to reveal the identity of the victim and the reason he came to Australia.

5. Technical Background

There are several concepts that needed to be illustrated in order to avoid audiences' confusions when reading this report.

5.1 Levenshtein Distance

This concept exists in Information Theory and Computer Science areas. It represents how much difference there is when comparing a pair of sequences. The edit distance (or Levenshtein distance) between two words is the smallest number of **substitutions, insertions, and deletions** of symbols that can be used to transform one of the words into the other. [5] In other words, the Levenshtein Distance is the minimum edit distance between two strings.

Here is an example demonstrating the calculation of the Levenshtein distance, substitution will be marked as **s** and **d** for deletion, **i** for insertion. String1: INTENSION, string2: EXECUSION.

I	N	T	E	#	N	S	I	O N
|	|	|	|	|	|	|	|	|
#	E	X	E	C	U	S	I	O N
|	|	|	|	|	|	|	|	|
d	s	s	—	i	s	———		

Table1. Levenshtein distance

According to Table1, the minimum cost to turn string1 into string2 is 5: 3 substitutions, 1 deletion and 1 insertion.

5.2 Vector Space Model and TF-IDF method

Vector Space Model (VSM) is a widely applied method in information filtering, text comparison and relevancy calculating. It turns text into algebraic form. A vector space consists of a set of elements (like x, y and z). Vectors can be operated (adding and multiplying will be used in this case), while text could be treated as a bag of words or a bag of letters.

A vector v can be expressed as follow:

$$v = a_1 v_{i1} + a_2 v_{i2} + \ldots + a_n v_{in,}$$

Where the coefficients a_n are the **weights** and v_n are the **elements** in the vector. The figure below shows two document vectors $V(d_1)$, $V(d_2)$ and a query vector V(Q). The number of terms being considered (t_1 and t_2) was set to 2 for simplicity, as 2-d figure is easy to present on a paper. More axes will be required as the terms increase. The document d_1 includes elements $\{t_1\ldots\}$ while d_2 contains $\{t_2\ldots\}$. Suggest that t_1 represents the word '**chocolate'** and t_2 represents '**glazzy'**. Document d_1 includes the two words '**chocolate doughnut $3.5 …**' and d_2 includes the two words '**glazzy doughnut $3.0 …**'. Document Q may be '**go and buy some milk chocolate …**'

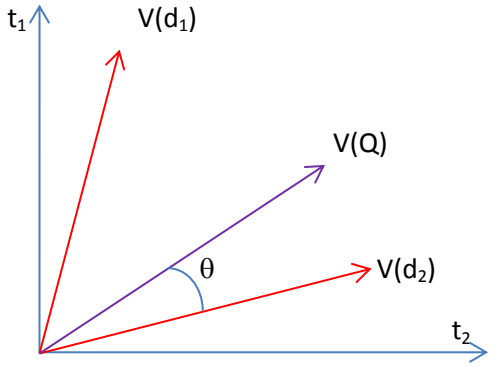

Figure4. Vector Space Model

Hence as the figure presented $V(d_1)$ stands closer to t_1 axis and $V(d_2)$ stands closer to t_2 axis. The cosine of θ represents the closeness of a vector to the query vector.

The method to calculate the weights is TF-IDF algorithm which represents the weights by a combination of Term Frequency (TF) and Inverse Document Frequency (IDF). How many times the term t occurred in the document D is recorded as Term Frequency and the Inverse Document Frequency is defined as $\log(\frac{N}{d_f})$, where N means the total number of documents in the space and d_f is the amount of documents where term t occurred. Hence the TF-IDF weight could be expressed as:

$$\textbf{\textit{Weight (TF-IDF}}_{t,d}\textbf{\textit{) = TF}}_{t,d}\textbf{\textit{ * IDF}}_{t}$$

Finally the two documents' similarity can be found by calculating the Cosine Similarity of their corresponding vectors (dot product of the two vectors divided by the product of the two vectors' Euclidean lengths):

$$\textbf{\textit{Similarity <d}}_1\textbf{\textit{, d}}_2\textbf{\textit{ > = cos(}}\theta_{d1,d2}\textbf{\textit{) = }}\frac{V(d_1)\cdot V(d_2)}{|V(d_1)||V(d_2)|}$$

5.3 One-time Pad

One-time pad (OTP) is an encryption technique in cryptography. One-time pad encryption system generates a set of random keys and each key will be used only once to encrypt its corresponding block. By using the proper key sequence (which should be exactly the same as the one used when encrypting) the receiver could decrypt the blocks. If the key sequence is generated in a truly random process and is longer than the original text blocks, by keeping the key sequence secret people will have no chance to decipher. In a word, one-time pad encryption makes it almost impossible to decipher the encrypted text without the proper key sequence.

Here demonstrated a brief example of OTP encryption:

1. Firstly establish the rule for converting 26 letters into numbers:

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26

2. Then generate a letter sequence randomly and shape it:

Random Sequence: K P N E X V Q M D U T M

Shaped Sequence: K P|N E|X V|Q M|D U|T M

3. Generate one-time pad by turning sequence into numbers according to the rule in step1:

Shaped Sequence: K P|N E|X V|Q M|D U|T M

Number Sequence: 1116|1405|2422|1713|0421|2013

The number sequence above is the one time pad. Now the one time pad has been created. Assume the text for encryption is 'Come at eleven'.

4. Format text need to be encrypted and turn it to numbers:

Shaped Text: C O|M E|A T|E L|E V|E N

Number Sequence: 0315|1305|0120|0512|0522|0514

5. Encrypting with one-time pad generated in step3:

Now add the above number sequence in step4 with the one-time pad together using Fibonicci addidition(no carrying addition, for example 5+9 = 4, not 14):

Sequence: 0315|1305|0120|0512|0522|0514

One-time pad: 1116|1405|2422|1713|0421|2013

Encrypted Sequence: 4421|2700|2562|1225|0943|2527

6. Related Work

Myriad of individuals and research groups have been devoted to this project since the happening of the Tamám Shud case 70 years ago.

6.1 Australian Department of Defence

In response to the request from journalist Stuart Littlemore the Australian Department of Defence had worked on cracking the code left in the Tamám Shud case. Unfortunately after a time of working the cryptographers defined the code as unable to crack. The code was said to either "have insufficient symbols" or it was just a meaningless product generated under a "disturbed mind" [6].

6.2 Previous Groups in the University of Adelaide

The conclusions drawn out by previous groups are: the code is unlikely to be generated randomly, the code is unlikely to be initial letters from words, the book *Rubaiyat of Omar Khayyam* was unlikely to be used as a one-time pad for encryption, the original language of the code is likely to be English, the code is unlikely to be initialisms extracted from poems, the book *Rubaiyat of Omar Khayyam* was not used as a straight substitution one-time pad for encryption and the code was not created using the *Rubaiyat of Omar Khayyam* as a one-time pad.[7][8][9][10][11][12]

This project will critically review the aforementioned conclusions while aim to draw out more convincing conclusions.

7. Deliverables.

As the Gantt chart shows, there are six key milestones in this project: 1st Draft and

2nd Draft Thesis, conclusion of code cracking task, project wiki page, video, slides and presentation for exhibition and the final report. Inside those milestones there are three key deliverables: the project wiki page, video, slides and presentation for exhibition and the final report (The Gantt chart has been modified yet has not been updated due to software errors. It will be updated in the next report).

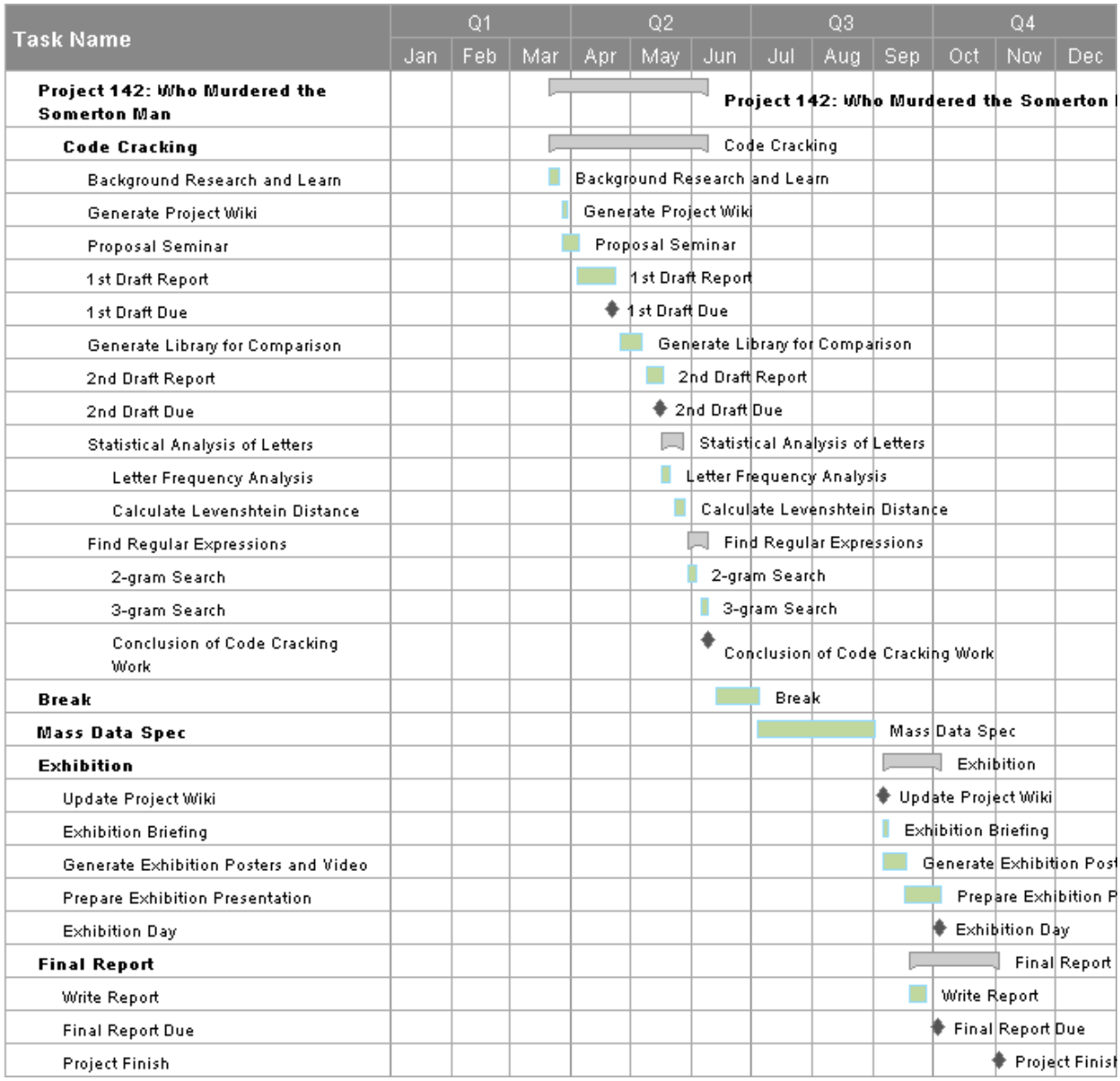

Figure 5 Project Gantt chart

8. Knowledge Gaps

The text comparison procedure requires a library. Hence text processing skill will be required. It is vital to get familiar with the text processing toolkit in Java library. Algorithms design and implement skills in Java are also required when dealing with the calculation of Levenshtein distance and the Vector Space Modelling. Data mining related skills will be used when doing the filtering of one-time pad decryption results. Besides, knowledge related to statistics will be required when analyzing the letters.

9. Technical Challenges

Building library is undoubtedly the first challenge in this project. It requires a wide category of texts in multiple languages. Also they should be sorted in a way that the comparison can be carried out easily.

Generating algorithms for text comparison comes to the second technical challenge. As the size of the library will be relatively huge, a good algorithm will help to get the comparison work done effectively.

Programming and data processing skills are also required in this project. It is necessary to get familiar with Java's text processing libraries and data processing software (R, Matlab and MS Excel).

10. Method

10.1 Language Similarity Analysis.

Prior to determine the code it is necessary to firstly check whether the analyzing method applied is effective or not. This will be done by check the similarity of strings

extracted in one language and then check the similarity of the strings extracted from two different languages (For example: compare the similarities of English-English versus English-French). All the combinations of inner language similarity and cross-language similarity will be compared. Only if the difference of two similarities is exceptional obvious, the analyzing method will be defined as effective.

The pre-defined methods for language similarity analysis are calculating and comparing the distance of two arrays and distance will be calculated using Levenshtein distance algorithm and Space Vector Model distance algorithm. The flow chart below (figure6) demonstrates the procedures involved in this task (English-English versus English-French); same rule applies for the comparison of in and between other languages.

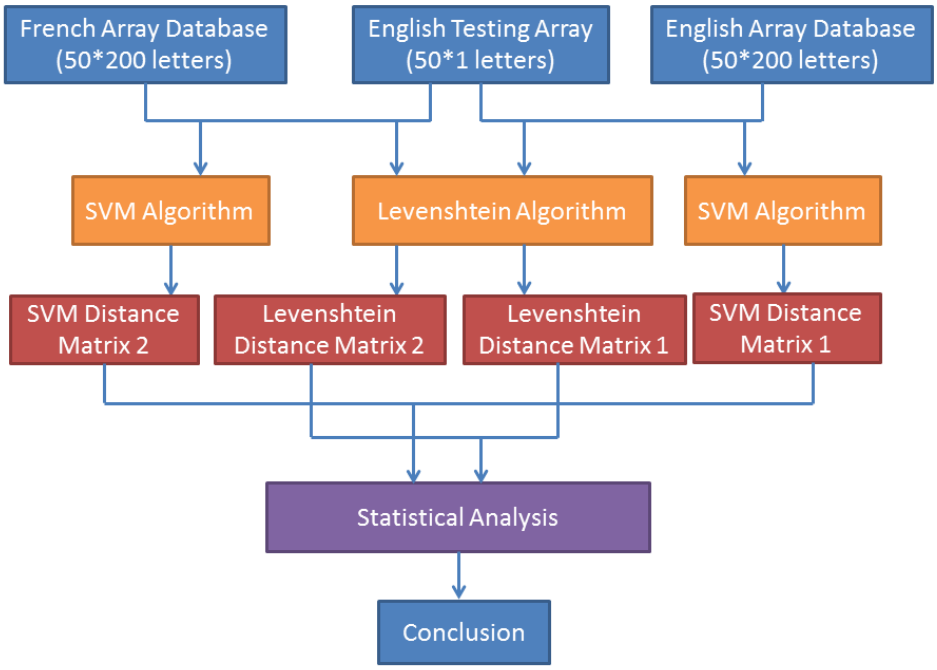

Figure 6

10.2 Analysis with the Somerton Man's Codes.

After the similarity analysis methods are proved to be effective it is reasonable to start the analysis with the Somerton man's codes. As the hand-written code (please

refer to the Figure7 below) was ambiguous all the possible versions will be taken into consideration.

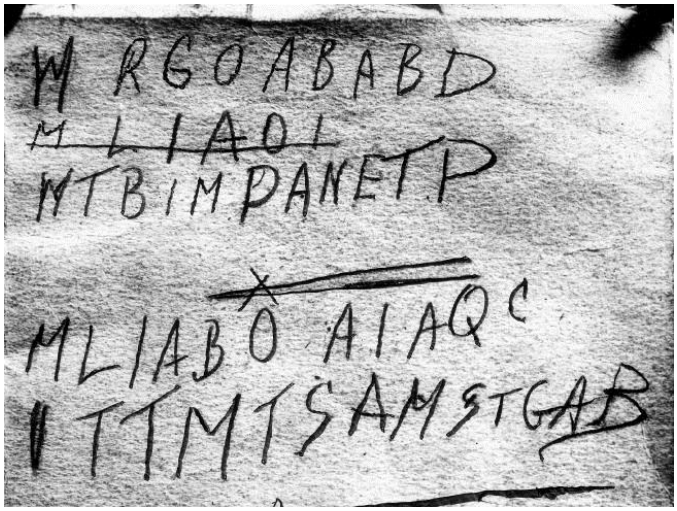

Figure 7

By comparing the Somerton Man's code with every database in different languages there will be a set of matrixes indicating the similarities. The matrix with greatest mean value will be marked and the code will be proved to be written from the matrix's corresponding language. The figure8 below shows English and French similarity calculation procedures; same rule applies for other languages.

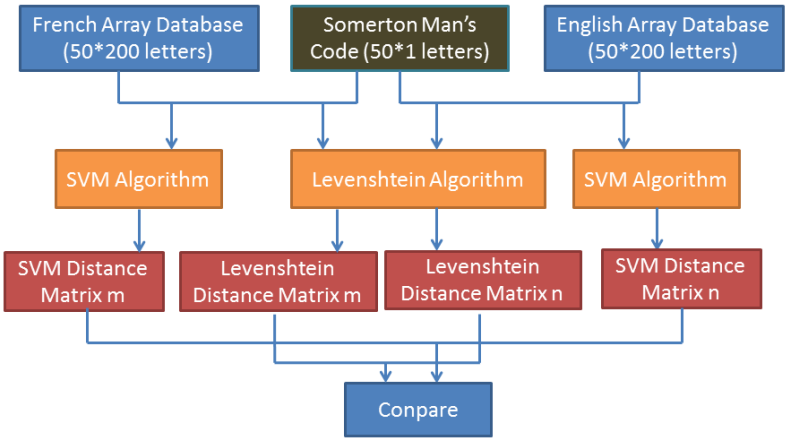

Figure 8

10.3 Letter Frequency Analysis.

Here counts the letter frequency of the Somerton Man's Code, and compare it with the average values of letter frequencies calculated from each language's database.

According to the result it is easy to observe which language has the letter frequency most similar to the frequency calculated from the Somerton Man's code.

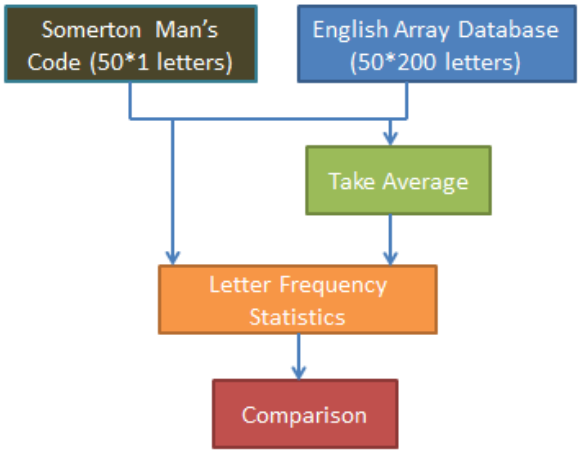

Figure 9

10.4 One-Time Pad Decryption

Needs to be fulfilled.

10.5 Finding Most Common Phrases

Needs to be fulfilled.

11. Planning and Feasibility

11.1 Work breakdown

Mainly there are two tasks inside the project: Code analysing and mass spectrometry data analysis. The two tasks are totally irrelevant.

The code cracking task can be divided into two sub tasks: Determine the language and find out the regular expression. The second one will depend on the outcome of the first one.

11.2 Timeline

Please refer to the Gantt chart (Figure 4) pasted in Chapter7 and the attached MS Excel file for more detailed information.

11.3 Budget

This project will not require any kinds of hardware equipment except USB flash drive for data storage use. Some software, online book resources and online storage services may be needed. The cost of the project will not exceed the budget.

11.4 Task Allocation

Yami Li will be in charge of the mass spectrometry data analysis task while Yifan Ma will be in charge of the code cracking tasks. This allocation scheme is not fixed. If anyone finishes their corresponding task in advance, they have the obligation to help their team mate finish the rest of the tasks.

11.4 Risk Management

Risks	Likelihood	Impact	Mitigation
Going out of Budget	Low	Low	Almost Impossible to happen
Supervisor Unavailable	Low	Moderate	Communicate with Supervisors in advance.
Lack of Communication	Moderate	Moderate	Meet supervisors regularly. Keep everyone informed by

			email and Facebook.
Team Member Quits	Low	High	Inform Supervisor ASAP
Team member Sickness	Moderate	Moderate	Keep exercising regularly. Inform other remembers in advance.
Computer Failure	Moderate	High	Use Google drive to get everything saved.

12. Outcomes and Following Works

Up to now the first stage (analyzing letters from code) of the project is almost finished. The database for comparison has been established which consists of plain text retrieved from several English books and their corresponding translations in French, German, Spanish, Irish, etc. First letter arrays from text have been established by extracting first letter of every word from texts, and convert all the accent letters into Standard English letters. Algorithm for the calculation of Levenshtein distance has been realized. The Space Vector Modelling algorithm is relatively complex yet it will be established soon.

The next move will focus on the analysis of results generated from the previous calculation. After the analysis the first task of the project will be finished. The second task will be started immediately once the first task is finished.

13. Project Status

The project plan has been changed as the project aims and expected outcomes had

been modified after consulting the tutor. Currently the progress is a little slower than the plan while there will be no problem catching up with the plan. There will be no lectures and assignments due in the next few weeks hence all the times could be allocated to the project.

14. References

[1]. Renato Castello, "New twist in Somerton Man mystery as fresh claims emerge," Sunday Mail SA, November 23th, 2013. Access via Internet: http://www.adelaidenow.com.au/news/south-australia/new-twist-in-somerton-man-mystery-as-fresh-claims-emerge/story-fni6uo1m-1226766905157

[2]. Lynton Grace, "Somerton Man mystery: New details revealed of Jo Thomson, nurse in the case", The Advertiser, 29th May 2015. Access via Internet: http://www.adelaidenow.com.au/news/south-australia/somerton-man-mystery-new-details-revealed-of-jo-thomson-nurse-in-the-case/news-story/4c6bccbd2318584ad0cc6daaf3d8abd4

[3]. Lynton Grace, "Somerton Man mystery: New details revealed of Jo Thomson, nurse in the case", The Advertiser, 29th May 2015. Access via Internet: http://www.adelaidenow.com.au/news/south-australia/somerton-man-mystery-new-details-revealed-of-jo-thomson-nurse-in-the-case/news-story/4c6bccbd2318584ad0cc6daaf3d8abd4

[4]. From Wikipedia, the Taman Shud Case. Access via Internet: https://en.wikipedia.org/wiki/Taman_Shud_Case

[5]. Stavros Konstantinidis, "*Computing the Levenshtein Distance of a Regular Language*", Dept. Math. and Computing Sci., Saint Mary's University, Canada, IEEE Information Theory Workshop, 2005.

[6]. Inside Story, presented by Stuart Littlemore, ABC TV, screened at 8 pm, Thursday, August 24th, 1978.

[7]. A. Turnbull and D. Bihari. (2009). Final Report 2009: Who killed the Somerton man? [online]. Available: https://www.eleceng.adelaide.edu.au/personal/dabbott/wiki/index.php/Final_report_2009:_Who_killed_the_Somerton_man%3F

[8]. K. Ramirez and L-V. Michael. (2010). Final Report 2010 [online]. Available: https://www.eleceng.adelaide.edu.au/personal/dabbott/wiki/index.php/Final_Report_2010

[9]. S. Maxwell and P. Johnson. (2011). Final Report 2011 [online]. Available: https://www.eleceng.adelaide.edu.au/personal/dabbott/wiki/index.php/Final_Report_2011

[10]. A. Duffy and T. Stratfold. (2012). Final Report 2012 [online]. Available: https://www.eleceng.adelaide.edu.au/personal/dabbott/wiki/index.php/Final_Report_2012

[11]. L. Griffith and P. Varsos. (2013). Semester B Final Report 2013 – Cipher Cracking [online]. Available: https://www.eleceng.adelaide.edu.au/personal/dabbott/wiki/index.php/Semester_B_Final_Repo

rt_2013_-_Cipher_cracking
[12]. N. Gencarelli and J-K. Yang. (2015). Semester B Final Report 2015 – Cipher Cracking [online]. Available:
https://www.eleceng.adelaide.edu.au/personal/dabbott/wiki/index.php/Final_Report/Thesis_2015#Conclusions