The University of Adelaide

Code Cracking: Who Murdered the Somerton Man?

Group 142

Yifan Ma (a1658524)
2016/4/22

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. Corpse of the Somerton Man[1].

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.

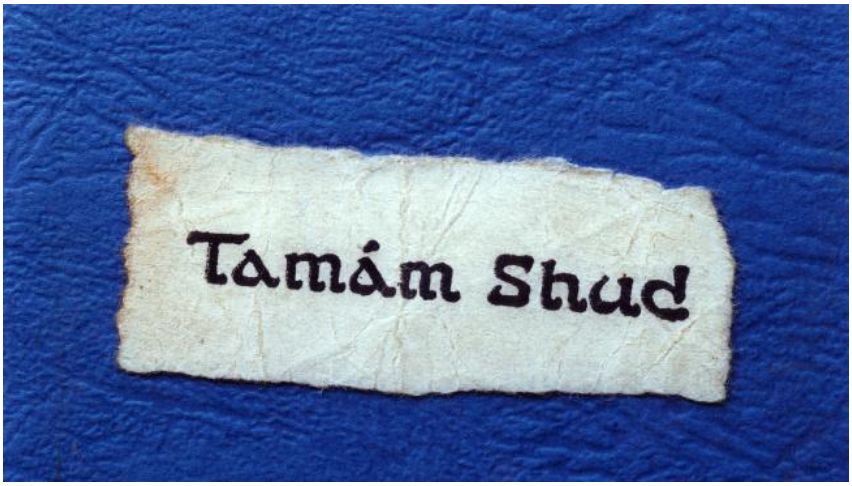

Figure 2. The paper scrap[2].

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 the figure below shows, it was a series of English letters. The code still remains uncracked to this today.

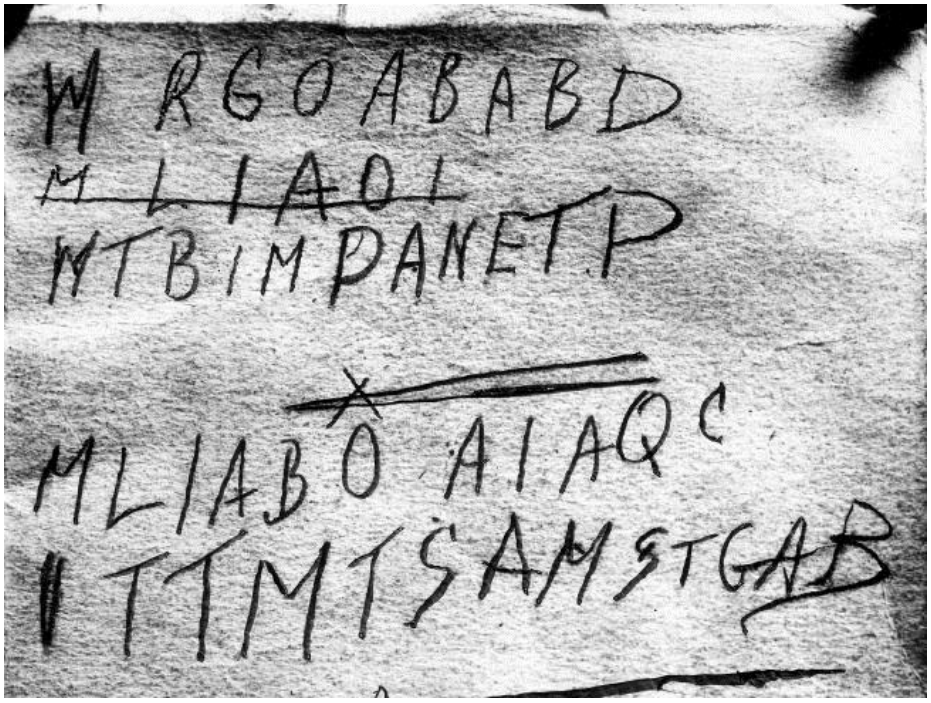

Figure 3. The hidden code[3]

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. For details please read in http://en.wikipedia.org/wiki/Taman_Shud_Case. 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.

The final destination of the case should be getting the code cracked and finding out

the murderer, yet it is not possible to crack the code directly. Hence the project group will work on the code to determine which language the secret code was intending to express, or whether the code was just a piece of letters generated arbitrarily. If the language could be determined, the further work is to seek for regular expressions under the condition of a specific language determined previously.

A library is required to be implemented. The library consists of letter arrays extracted from books in different languages with the same length of the original code. The code will be compared by those extracted letter segments in order to find out which language hold the biggest similarity. Hence the
language could be determined. The aforementioned work will be implemented by Java and Microsoft Excel.

Figure 3:

Another aspect of the project aim is to analysis the data generated by the victim's hair. By analyzing the trends of several chemical components it is possible to obtain more detailed clues and information 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 Tickle 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. [4].

5.2 n-grams searching

The n-grams represents n consecutive items in an string (array). In this report n = 2 (bigram) and n = 3 (trigram) modes will be applied. For example, the results of 2-grams expansion of the word “hello” would be “he”, “el”, “ll” and “lo”. Here is a good example of the text categorization using n-grams technique: N-Gram-Based Text Categorization [5].

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”.

6.2 Previous Project Groups in the University of Adelaide

There were exactly 6 final year student groups who had made contribution this project since 2009 and several aspects of the project had been investigated. The outcomes of the previous groups' work can be mainly concluded as below:

1. The letters appeared in the code are not randomly generated.
2. The letters are more likely to be initial letters of English word.
3. It may be taken from The Rubaiyat book.

6.3 Extension of the previous work

This project will be based on the conclusions delivered by previous groups that the letters are initial letters. But it is necessary to confirm whether the language was English or not. As the amount of samples for comparison used in previous work was not large enough to draw out the conclusion that the language was exactly English.

Further work will be performed after the target language of the code is determined. N-grams searching and mass data comparison techniques will be applied to find out the regular expressions inside the code.

Another aspect of the project is the analysis of mass spectrometer data of the Somerton man's hair.

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.

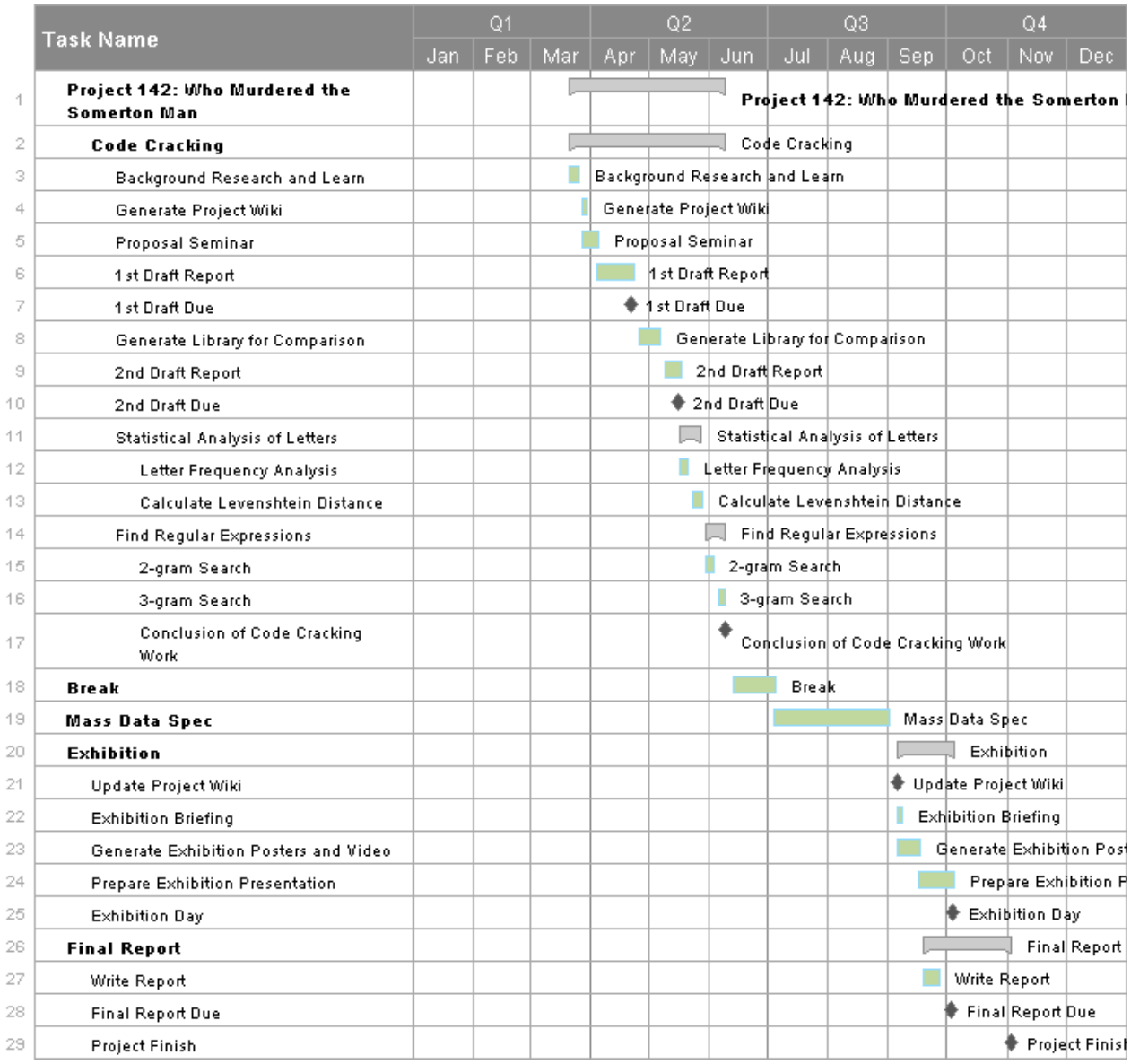

Figure 4 Project Gantt chart

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 calculating the Levenshtein distance and n-grams searching. Besides, information theory and statistics knowledge will be used in analyzing the comparison outcomes.

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 softwares (R, Matlab and MS Excel).

10. Method

There are myriad examples of text processing using Java on the Internet (e.g. Github, Khan Academy). The best way to get fast progress in text processing is learning the code written by others. This method is also suitable for the algorithm design knowledge gap aforementioned.

The information theory and statistics knowledge gap can be fulfilled by reviewing the statistics course learned in the previous semester and learning through the online education website (e.g. Khan Academy, Youtube).

11. Planning and Feasibility

11.1 Work breakdown

Mainly there are two tasks inside the project: Code cracking 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. 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]. 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.

[5]. William B. Cavnar and John M. Trenkle, "*N-Gram-Based Text Categorization*", Environmental Research Institute of Michigan, Ann Arbor Michigan.