SPL-1 Project Report, 2019

Reference Mapping

SE 305: Software project Lab I

Submitted by

Hasnatul Jelan Pranto

BSSE Roll No. : 1026

BSSE Session:2017-18

Supervised By

Kishan Kumar Ganguly

Designation: Lecturer

Institute of Information Technology

Institute of Information Technology

University of Dhaka

29-05-2019

Table of Contents

1. Introduction

Referencing or Citation plays an important role in any kind of work where we have drawn ideas or researches or words from the work of other authors. Citation allows us to acknowledge their contribution in our work & give them their due credit. Also it makes the work more persuasive & authoritative as we are linking our work with the ones that can provide evidence to support particular assertions in our work.

As the name suggests, this software is such a tool that will allow us to map the relationship between authors based on the references in their work.

1.1 Background Study

PDF Formats

A PDF file is a page description format which has 4 parts:

Objects: Contains the data of the PDF.

File Structure: Determines how the objects are stored in a PDF file, how they are accessed, and how they are updated.

Document Structure: specifies how the basic object types are used to represent components of a PDF document: pages, fonts, annotations, and so forth.

Content Stream: Contains a sequence of instructions de-scribing the appearance of a page or other graphical entity.

Almost all PDF files are now being encoded by different filters (e.g ASCII base 85, ASCII hexadecimal, LZW, Flate, DCT etc.) either to compress it or to convert it to a portable ASCII representation. In this case, a corresponding decoding filter is required to convert the information back to its original form [1,2]

Citation Style

Different citation style (e.g. APA style, MLA style, IEEE style etc.) follows different pattern for citation. This citation style sometimes depends on the academic discipline involved. In the context of IEEE citation style, the references should be numbered and appear in the order they appear in the text. When referring to a reference in the text of the document, put the number of the reference in square brackets. E.g: [1]

The IEEE citation style has 3 main features:

• The author name is first name (or initial) and last. This differs from MLA style where author's last name is first.

• The title of an article (or chapter, conference paper, patent etc.) is in quotation marks.

• The title of the journal or book is in italics.

These conventions allow the reader to distinguish between types of references at a glance. The correct placement of periods, commas and colons and of date and page numbers depends on the type of reference cited. [3,4]

XML Format

One crucial aspect of this project is to convert a text file into an XML file. So, I had to learn the basic syntax style of the XML format.

1.2 Challenges

I had to deal with a number of challenges to implement this project. Some were easy to resolve, others were quite stubborn. Some of the worth mentioning challenges are:

1. The first challenge was to convert the pdf into simple text format. As this challenge itself is a different project & beyond the scope of my project, I had to use an API (PDFBox) to get the raw text from a PDF.

2. Implementing algorithm to identify paper title, author's information & to extract information from reference section was quite challenging.

3. One tiny & tricky challenge was to identify UTF-8 characters that are not included in ASCII character table.

4. Planning & managing a project of this scale for the first time itself was a challenge for me.

2. Project Overview

The entire project can be divided into 3 parts:

1. Converting the PDF into text file.
2. Converting the text file into XML.
3. Extracting necessary information from the XML file to make citation trees.

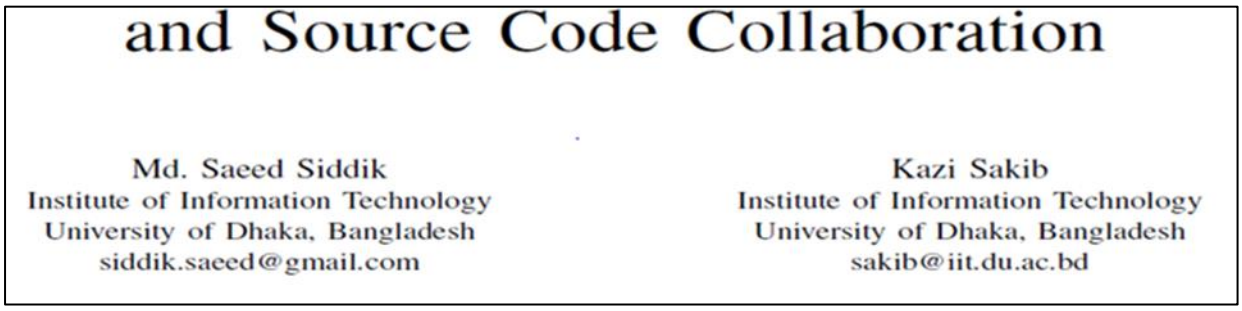
A brief explanation for each part is given below.

Converting the PDF into text file

This part was done by the help of an API, in this case, I used Apache PDFBox which is a java tool to work with PDFs. Using PDFBox, I was able to get the simple text format from the PDF.

Converting the text file into XML

In one sense, the main work of this project starts from here. This part can be further divided into smaller parts. First thing to make the XML file was to identify the title of the paper along with the authors name, mail, institution, address & other information. I analyzed several papers & found two patterns for this. One is-


and Source Code Collaboration

Md. Saeed Siddik
Institute of Information Technology
University of Dhaka, Bangladesh
siddik.saeed@gmail.com

Kazi Sakib
Institute of Information Technology
University of Dhaka, Bangladesh
sakib@iit.du.ac.bd

Figure 01- A pattern where author's information are separated individually (Source: From the research paper entitled "**RDCC: An Effective Test Case Prioritization Framework using Software Requirements, Design and Source Code Collaboration**")

In the upper pattern, the authors informations have been described individually. Another pattern is-


A Reusable Adaptation Component Design for Learning-Based Self-Adaptive Systems

Kishan Kumar Ganguly, Kazi Sakib
Institute of Information Technology
University of Dhaka, Dhaka, Bangladesh
Emails: bsse0505@iit.du.ac.bd, sakib@iit.du.ac.bd

Figure 2- Another pattern where author's information are merged together (Source: From the research paper entitled **"A Reusable Adaptation Component Design for Learning-Based Self-Adaptive Systems**")

Here, as we can see, the informations are in conglomerate form. The software is designed in a way to detect these two styles.

Next comes, this software is able to identify the citation numbers inside the article.

Last of all, We have to extract information from the reference section. As there are various citation style, Following all of them can be self-colliding. So this project strictly follows the IEEE citation style for this purpose. The software is able to identify & extract information from 7 distinct IEEE citation style both for print & electronic sources-

1. Book
2. Book Chapters
3. Article In a Journal
4. Articles From Conference Proceedings (published)
5. Papers Presented at Conferences (unpublished)
6. Books (electronic source)
7. Journal (electronic source) [2]

Extracting necessary information from the XML file to make citation trees

After having the XML files, the next thing is to build the citation trees. In this project, the trees are built in such a way that the parent-node paper is referring to the child-node papers. A graphical demonstration here might be useful to simplify the structure-

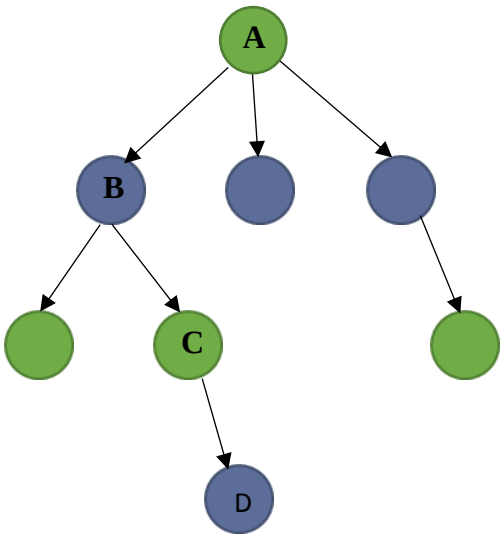


Figure 3- A Citation Tree

Figure 3 shows- Paper “A” is referencing to paper “B”, “B” referencing to “C”, “C” referencing to “D”. Here a node will contain the paper name & the authors name. Having this tree helps to find authors who is referencing to whom & find the information about a paper.

3. User Manual

1. Running the executable file, the user will see a interface like this-

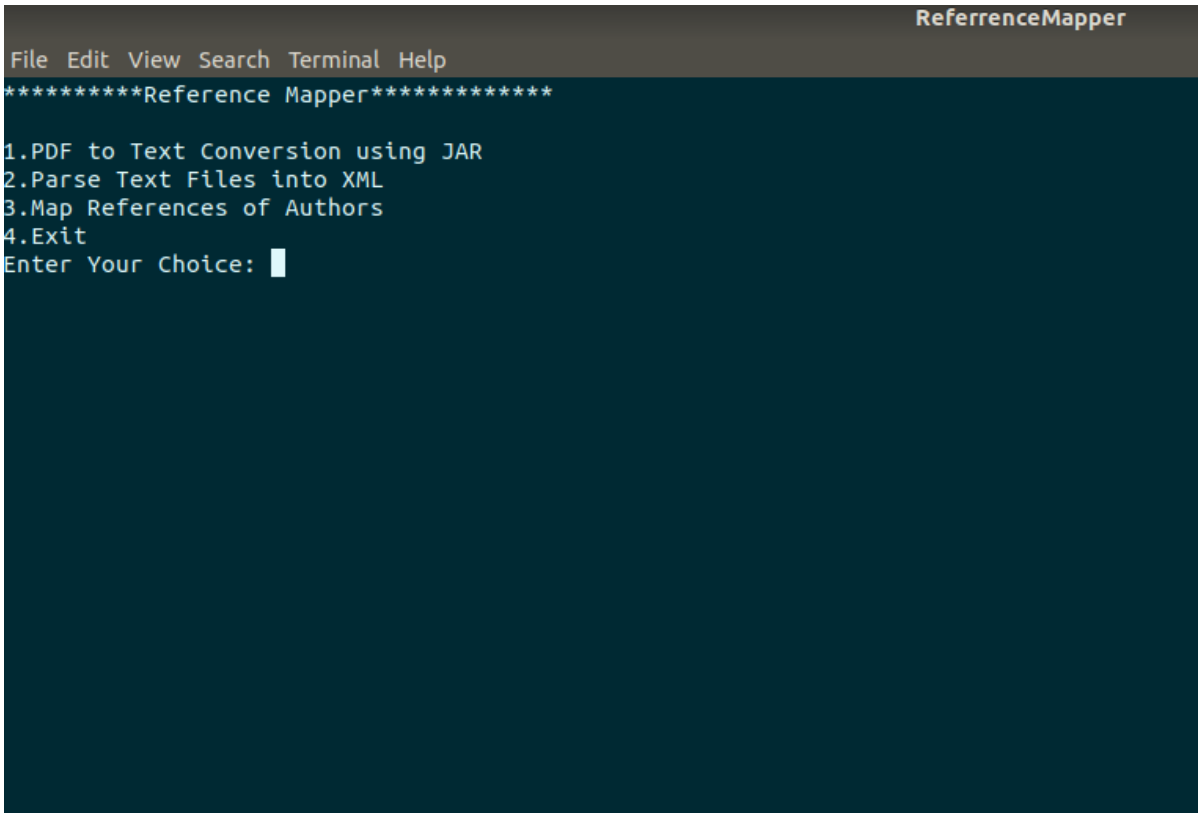


Figure 4- The Main Window

2. If the user has a JAR file that contains the program to convert PDF into text file, he can choose option 1. This will make text files from pdf & store them in the papers repository. But if no JAR file is available, the user have to supply the text files.

3. User can choose option 2 once he has the text files. If the user does so, the following stage will appear-

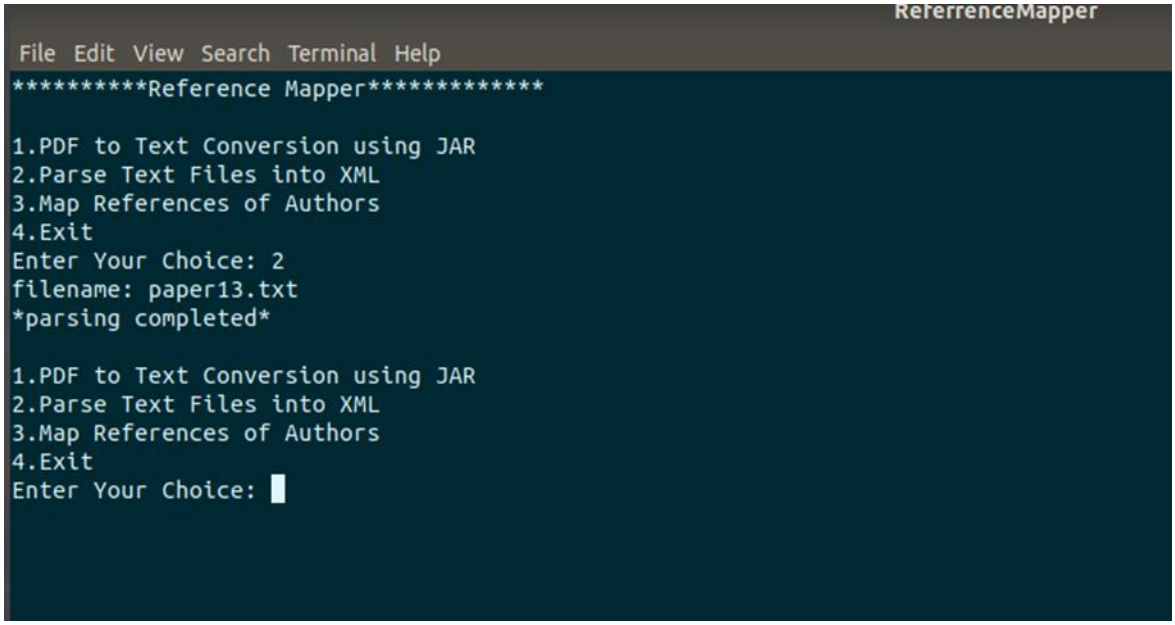


Figure 5.choosing option 2

Here, suppose the user has a text file named “paper13.txt” like below-

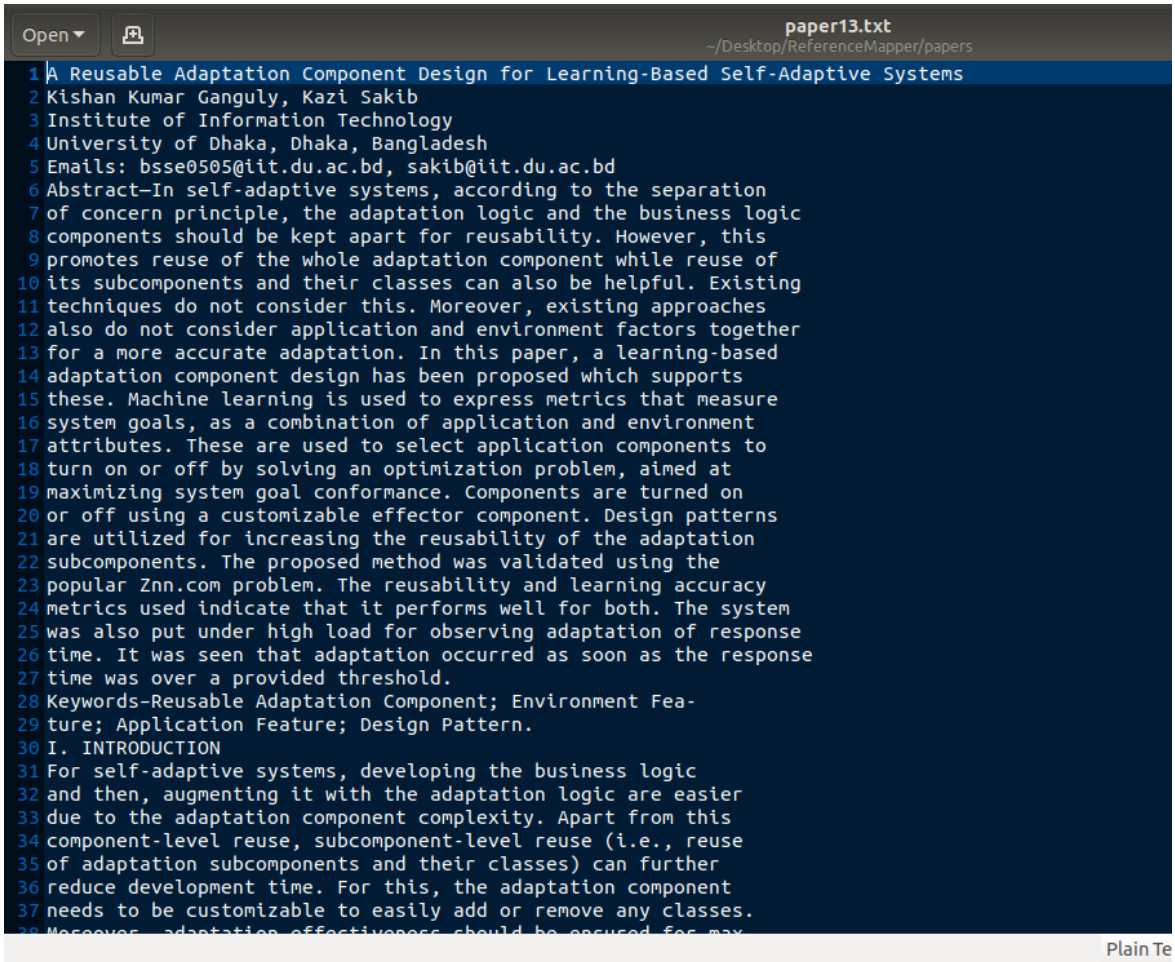


Figure 6- A Plain Text File

Parsing it into a XML, the user will get a new file like this-

```xml
<xml>
<document>
<title> A Reusable Adaptation Component Design for Learning-Based Self-Adaptive Systems </title>
<authors>
<author>
<name> <firstName> Kishan Kumar </firstname> <lastname> Ganguly </lastname> </name>
<name> <firstName> Kazi </firstname> <lastname> Sakib </lastname> </name>
<Department> Institute of Information Technology </Department>
<University> University of Dhaka, Dhaka, Bangladesh </University>
<mail> Emails: </mail>
<mail> bsse0505@iit.du.ac.bd, </mail>
<mail> sakib@iit.du.ac.bd </mail>
</author>
</authors>
<article>
Abstract—In self-adaptive systems, according to the separation
components should be kept apart for reusability. However, this
promotes reuse of the whole adaptation component while reuse of
its subcomponents and their classes can also be helpful. Existing
techniques do not consider this. Moreover, existing approaches
also do not consider application and environment factors together
for a more accurate adaptation. In this paper, a learning-based
adaptation component design has been proposed which supports
these. Machine learning is used to express metrics that measure
system goals, as a combination of application and environment
attributes. These are used to select application components to
turn on or off by solving an optimization problem, aimed at
maximizing system goal conformance. Components are turned on
or off using a customizable effector component. Design patterns
are utilized for increasing the reusability of the adaptation
subcomponents. The proposed method was validated using the
popular Znn.com problem. The reusability and learning accuracy
metrics used indicate that it performs well for both. The system
was also put under high load for observing adaptation of response
time. It was seen that adaptation occurred as soon as the response
time was over a provided threshold.
Keywords-Reusable Adaptation Component; Environment Fea-
```

Figure 7- The Upper Part of the XML

```xml
dates adaptation effectiveness and utilization of these features.
In future, the technique will be enhanced to take adaptation
decision by foreseeing future effects of the decision on the
system. It will also be extended to automate threshold selection
for metrics and failed adaptations.
</article>
<reference>
<citationid> [1] </citationid>
<authors>
<name> <firstName> D. </firstname> <lastname> Garlan </lastname> </name>
<name> <firstName> S.-W. </firstname> <lastname> Cheng </lastname> </name>
<name> <firstName> A.-C. </firstname> <lastname> Huang </lastname> </name>
<name> <firstName> B. </firstname> <lastname> Schmerl </lastname> </name>
<name> <firstName> P. </firstname> <lastname> Steenkiste </lastname> </name>
</authors>
<article> “Rainbow: Architecture-based self-adaptation with reusable infrastruc- ture,” </article>
<journal> IEEE Computer, vol. 37, no. 10, pp. 46-54, 2004. </journal>
<citationid> [2] </citationid>
<authors>
<name> <firstName> N. </firstname> <lastname> Esfahani </lastname> </name>
<name> <firstName> A. </firstname> <lastname> Elkhodary </lastname> </name>
<name> <firstName> S. </firstname> <lastname> Malek </lastname> </name>
</authors>
<article> “A learning-based frame- work for engineering feature-oriented self-adaptive software systems,” </article>
<journal> Software Engineering, IEEE Transactions on, vol. 39, no. 11, pp. 1467– 1493, 2013. </journal>
<citationid> [3] </citationid>
<authors>
<name> <firstName> A. J. </firstname> <lastname> Ramirez  </lastname> </name>
<name> <firstName> B. H. </firstname> <lastname> Cheng </lastname> </name>
</authors>
<article> “Design patterns for developing dynam- ically adaptive systems,” </article>
<journal> in Proceedings of the 2010 ICSE Workshop on Software Engineering for Adaptive and Self-Managing Systems. ACM, 2010, pp.
49-58. </journal>
<citationid> [4] </citationid>
<authors>
<name> <firstName> M. L. </firstname> <lastname> Berkane </lastname> </name>
<name> <firstName> L. </firstname> <lastname> Seinturier </lastname> </name>
<name> <firstName> M. </firstname> <lastname> Boufaida </lastname> </name>
```

Figure 8- The Reference part of the XML

4. Now, if the user chooses option 3 “Map References of Authors”, he will get this stage-

```
1.PDF to Text Conversion using JAR
2.Parse Text Files into XML
3.Map References of Authors
4.Exit
Enter Your Choice: 3
How many files: 7
filename: testxml0.xml
filename: testxml1.xml
filename: testxml2.xml
filename: testxml4.xml
filename: testxml5.xml
filename: testxml6.xml
filename: testxml7.xml
Reference Trees Have been created Successfully

1.See The Overall Mapping
2.See References of An Author
3.See Reference Thread For An Author
4.Return
```

Figure 9- Choosing option 3

Here, the user have to give the XML file names which are in the “papers” repository. After giving the file names, the user has three options to choose from. If he chooses option 1, he will find all of the citation trees like the below image-

```
Reference Trees Have been created Successfully

1.See The Overall Mapping
2.See References of An Author
3.See Reference Thread For An Author
4.Return
1
Rahim Khan
        Jonas Mockus
        Vytautas Tiesis
                Asad Khan
                Jahir Khan
        Karim Khan
Asad Khan
        Karim Khan
        Atikur Rahman
Jonas Mockus
        Asad Khan
Atikur Rahman
        Kamal Khan
        Monir Khan
        Helal Khan
        Belal Khan
Monir Khan
        Atikur Rahman
To See Papers Name, press 1
Go back 0
```

Figure 10- The Citation Trees

```
To See Papers Name, press 1
Go back 0
1
Rahim Khan (Practical Bayesian Optimization of Machine Learning Algorithms)
        Jonas Mockus (The application of Bayesian methods for seeking the extremum)
        Vytautas Tiesis (The application of Bayesian methods for seeking the extremum)
                Asad Khan (The application of Bayesian methods for seeking the infimum)
                Jahir Khan (The application of Bayesian methods for seeking the infimum)
        Karim Khan (The application of Bayesian methods for seeking the extremum)
Asad Khan (Practical Bayesian Optimization of Machine Learning Algorithms)
        Karim Khan (The application of Bayesian methods for seeking the extremum)
        Atikur Rahman (The application of Bayesian methods for seeking the extremum)
Jonas Mockus (Practical Bayesian Optimization of Machine Learning Algorithms)
        Asad Khan (The application of Bayesian methods for seeking the extremum)
Atikur Rahman (Practical Bayesian Optimization of Machine Learning Algorithms)
        Kamal Khan (The application of Bayesian methods for seeking the extremum)
        Monir Khan (The application of Bayesian methods for seeking the extremum)
        Helal Khan (The application of Bayesian methods for seeking the extremum)
        Belal Khan (The application of Bayesian methods for seeking the extremum)
Monir Khan (Practical Bayesian Optimization of Machine Learning Algorithms)
        Atikur Rahman (The application of Bayesian methods for seeking the extremum)

1.See The Overall Mapping
2.See References of An Author
3.See Reference Thread For An Author
4.Return
```

Figure 11- Citation tree with paper names

If the user chooses option 2, he will have to give a authors name to see the bi-directional reference relation & this screen will look like this-

```
        Belal Khan (The application of Bayesian methods for seeking the extremum)
Monir Khan (Practical Bayesian Optimization of Machine Learning Algorithms)
        Atikur Rahman (The application of Bayesian methods for seeking the extremum)

1.See The Overall Mapping
2.See References of An Author
3.See Reference Thread For An Author
4.Return
2
Enter Author Name: Asad Khan
Referred By                              Referred To
Vytautas Tiesis ------>> Asad Khan ------>> Karim Khan
Jonas Mockus                             Atikur Rahman

1.See The Overall Mapping
2.See References of An Author
3.See Reference Thread For An Author
4.Return
2
Enter Author Name: Rahim Khan
Referred By                                 Referred To
           ------>> Rahim Khan ------>> Vytautas Tiesis
                                        Karim Khan
                                        Jonas Mockus

1.See The Overall Mapping
2.See References of An Author
3.See Reference Thread For An Author
4.Return
```

Figure 12- The Reference Relation of an Author

If he chooses option 3, he have to give a authors name & he can see how this author is connected in the citation tree, This stage will be like this-

```
1.See The Overall Mapping
2.See References of An Author
3.See Reference Thread For An Author
4.Return
3
Enter Author Name: Asad Khan
Enter Paper Name: The application of Bayesian methods for seeking the infimum

A Chain Reference From The Root Author To Author Asad Khan:

Rahim Khan ---> Vytautas Tiesis ---> Asad Khan

1.See The Overall Mapping
2.See References of An Author
3.See Reference Thread For An Author
4.Return
```

Figure 13- Reference Thread In a Citation Tree

4. Conclusion

Working on this project proved very beneficial to improve my coding skills & to get a more clear perception of what goes on through the memory of the computer when I make such calls. It also helped me to learn how

to handle large codes in a project. I believe the experience that I gathered from working on this project will anyhow help me in my future works. I cordially thank my supervisor for all the help, support & guide I got from him to finish this project.

5.Appendix

Some scopes that this project does not cover are- other citation styles(i.e. MLA, APA etc), more robustness, including authors personal information in the tree etc. I'll like to work on these scopes in future.

6.References

1. https://www.glyphandcog.com/textext.html, Glyph & Cog.

2. PDF Reference, 3rd ed.

3. https://pitt.libguides.com/citationhelp, Course & Subject Guides.

4. IEEE-Citation-StyleGuide