PROJECT PLAN

Extraction of Structured Information from Dictionary PDFs

Caleb Smith
Josephine Paculio
Joshua Dib
Roger Bernardo

Contents

Executive Summary

This project plan sets out the various tasks involved in the Professional Experience project "Extraction of Structured Information from Dictionary PDFs" for Dr. Rachel Hendery from the School of Humanities and Communication Arts. The main goal of the project is to automate searching dictionary PDF files to extract the headword and translation and produce an output preferably of XML file type.

The project is to be undertaken by students studying Computer Science and Information and Communications Technology from the School of Computing, Engineering, and Mathematics. The project team will be working under the supervision of Jonathan Weightman.

Included in this document are all the tasks to be completed for the project and the working timelines for each deliverable. These are outlined in the work breakdown structure, Gantt chart, and milestone plan.

Detailed information on the project team as well as individual tasks have been included. Planned meetings with the client, supervisor, and for the project team itself are also included.

Upon completion of the project plan, a detailed proposal will be submitted to the client, with a working protype to be demonstrated in week 6, and the final system to be completed and available for testing in week 10, with the final system handover in week 13.

Introduction

As technology becomes more widespread and readily available, applications and processes have been put in place to improve quality of life and provide society with increased ease and efficiency for tasks. Linguistic analysis is one field to see the benefits technology can have through use of automation and algorithm application.

This document will begin to lay the foundation of an OCR based solution to aid with the extraction of information to be used in further research. Contained will detail a designated, well-rounded team to deliver the project, an analysis of the tasks involved with the associated risks, as well as a proposed timeline.

Following on from this document a better insight will be gained into the expected plan going forward and will lead into a more in-depth discussion regarding real solutions to be developed and implemented.

Group Details

The project group is composed of students from the School of Computing, Engineering, and Mathematics, studying Computer Science and Information and Communications Technology respectively.

Caleb Smith, an ICT student, knowledgeable in the following:

- HTML and CSS
- JavaScript
- SQL
- Systems Analysis and Design
- Project Management

Josephine Paculio, an ICT student, knowledgeable in the following:

- HTML and CSS
- JavaScript and PHP
- Java
- Systems Analysis and Design

Joshua Dib, Computer Science student, proficient in:

- Logic algorithms - strong understanding of data flow and tailoring a solution towards maximum performance
- Object oriented programming - useful skillset to provide a more efficient and intuitive product
- SQL queries - able to manipulate data into designated categories and further extract parameters from the dataset
- Web Development - undertaken multiple course exploring different technologies and frameworks

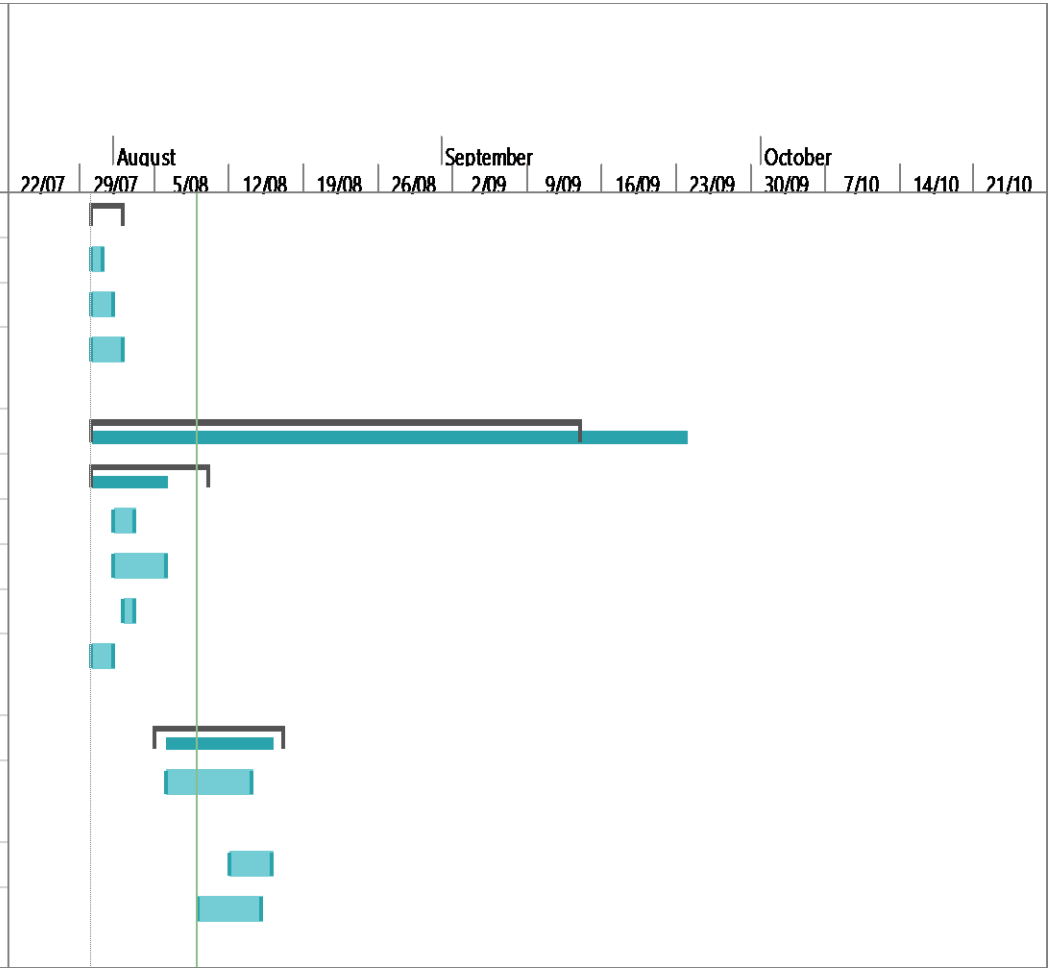
Roger Bernardo, a Computer Science student, knowledgeable in the following:

- C, C++
- HTML and CSS
- JavaScript and PHP

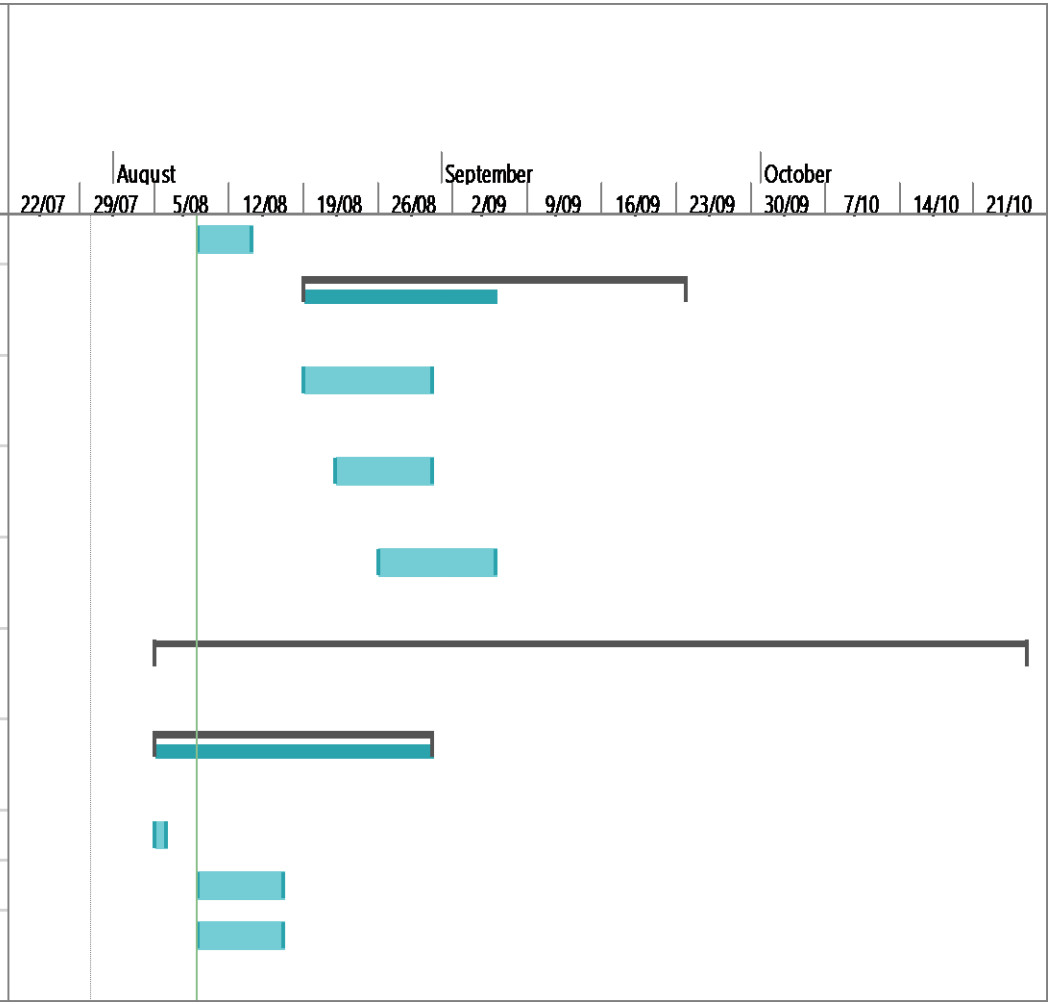
Work Breakdown Structure

Gantt Chart

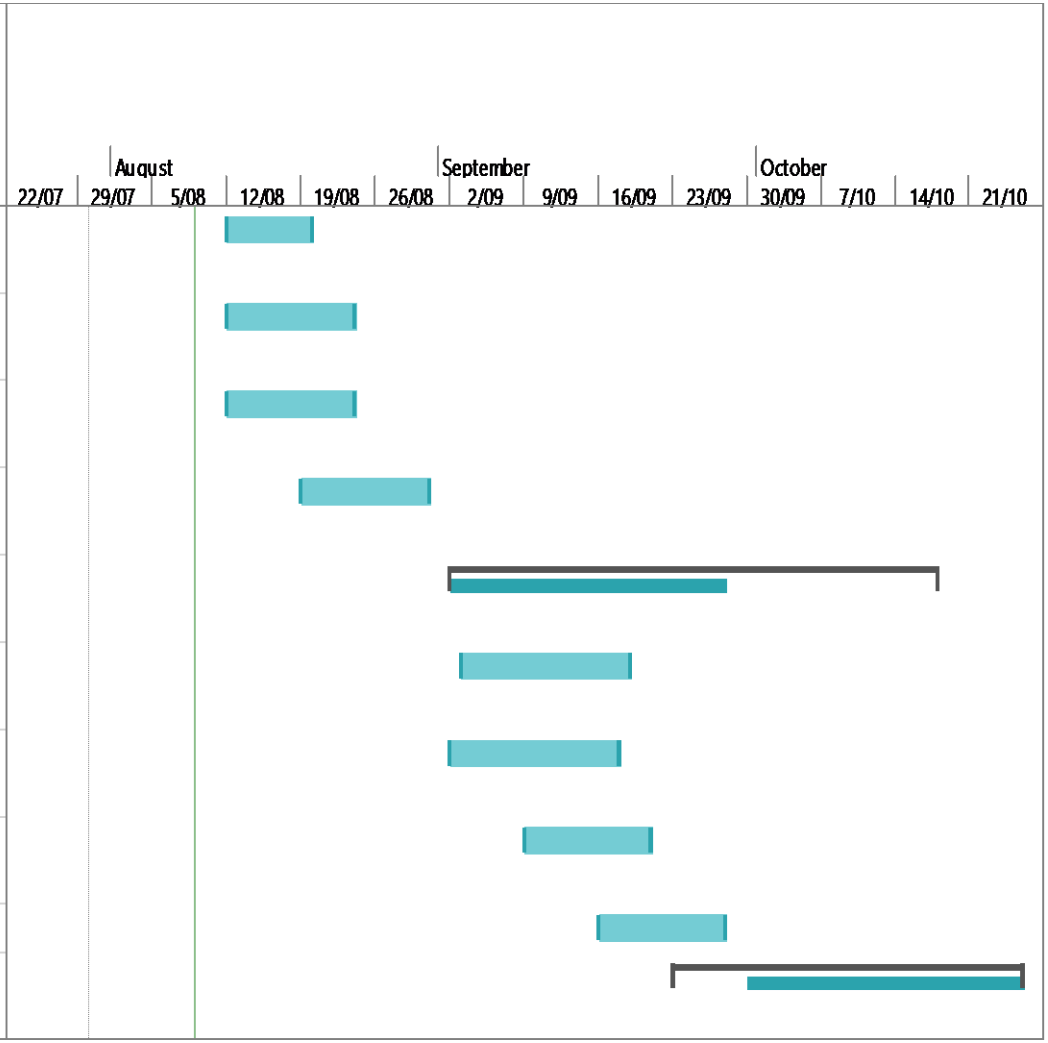
ID	Task Name	Duration	Start	Finish
1	**Pre-planning**	**3 days**	**Tue 30/07/19**	**Thu 1/08/19**
2	Assign a project manager	1 day	Tue 30/07/19	Tue 30/07/19
3	Allocate primary roles	2 days	Tue 30/07/19	Wed 31/07/19
4	Set up development resources	3 days	Tue 30/07/19	Thu 1/08/19
5	**Project Plan**	**34 days?**	**Tue 30/07/19**	**Fri 13/09/19**
6	**Project Plan**	**9 days?**	**Tue 30/07/19**	**Fri 9/08/19**
7	List group details	2 days	Thu 1/08/19	Fri 2/08/19
8	Conduct a risk analysis	3 days	Thu 1/08/19	Mon 5/08/19
9	Construct a Gantt Chart	1 day	Fri 2/08/19	Fri 2/08/19
10	Establish a meeting structure	2 days	Tue 30/07/19	Wed 31/07/19
11	**Project Proposal**	**10 days?**	**Mon 5/08/19**	**Fri 16/08/19**
12	Research alternate solutions	6 days	Tue 6/08/19	Tue 13/08/19
13	Identify any constraints	4 days	Mon 12/08/19	Thu 15/08/19
14	List functions and use cases	4 days	Fri 9/08/19	Wed 14/08/19



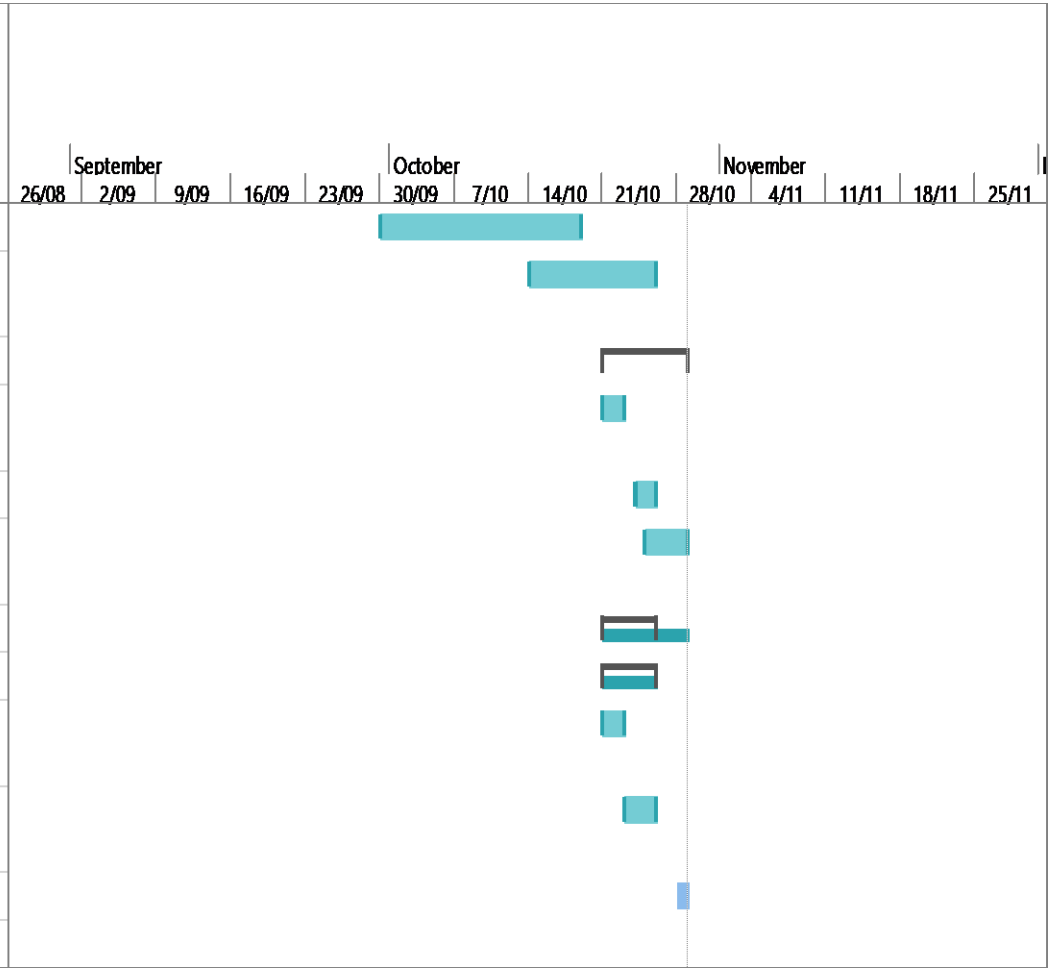
ID		T N	Task Name	Duration	Start	Finish
15			Relsease schedule	3 days	Fri 9/08/19	Tue 13/08/19
16			**System Analysis Design**	**26 days?**	**Mon 19/08/19**	**Mon 23/09/19**
17			Construct variety of System Diagrams	10 days	Mon 19/08/19	Fri 30/08/19
18			Define project requirements	7 days	Thu 22/08/19	Fri 30/08/19
19			Demostrate test cases	9 days	Mon 26/08/19	Thu 5/09/19
20			**Product Implemetation**	**60 days?**	**Mon 5/08/19**	**Fri 25/10/19**
21			**Working prototype**	**20 days?**	**Mon 5/08/19**	**Fri 30/08/19**
22			Resource set up	1 day	Mon 5/08/19	Mon 5/08/19
23			Create initial interface	6 days	Fri 9/08/19	Fri 16/08/19
24			Extract text from an already OCRed PDF file	6 days	Fri 9/08/19	Fri 16/08/19



ID	Task Name	Duration	Start	Finish
25	Display extracted text on HTML page	6 days	Mon 12/08/19	Mon 19/08/19
26	Organise extracted text into structured format	10 days	Mon 12/08/19	Fri 23/08/19
27	Place extracted text into Datasets	10 days	Mon 12/08/19	Fri 23/08/19
28	Turn extracted text into JSON file	10 days	Mon 19/08/19	Fri 30/08/19
29	**Feature Development**	**34 days?**	**Mon 2/09/19**	**Thu 17/10/19**
30	Dynamically attach a PDF file to the system	12 days	Tue 3/09/19	Wed 18/09/19
31	Save system's output to user's system	12 days	Mon 2/09/19	Tue 17/09/19
32	Extract text using Optical Characted Recognition	10 days	Mon 9/09/19	Fri 20/09/19
33	Result in an output file	10 days	Mon 16/09/19	Fri 27/09/19
34	**Final Product**	**25 days?**	**Mon 23/09/19**	**Fri 25/10/19**



ID		T N	Task Name	Duration	Start	Finish
35			Conduct error checking	15 days	Mon 30/09/19	Fri 18/10/19
36			Make quality of life improvements	10 days	Mon 14/10/19	Fri 25/10/19
37			**Product handover**	**6 days?**	**Mon 21/10/19**	**Mon 28/10/19**
38			Describe the implementation schema	2 days	Mon 21/10/19	Tue 22/10/19
39			List any product variations	2 days	Thu 24/10/19	Fri 25/10/19
40			List any outstanding issues	2 days	Fri 25/10/19	Mon 28/10/19
41			**Final Presentation**	**5 days?**	**Mon 21/10/19**	**Fri 25/10/19**
42			**Promotional media**	**5 days?**	**Mon 21/10/19**	**Fri 25/10/19**
43			Define the project abstract	2 days	Mon 21/10/19	Tue 22/10/19
44			Create video demonstration	3 days	Wed 23/10/19	Fri 25/10/19
45			**Panel presentation**	1 day	Mon 28/10/19	Mon 28/10/19
46						



Milestone Plan

Milestone	Main Responsibility	Planned Date	Actual Date
Project Plan	Project Group Collaboration	09/08/2019	
Project Proposal	Project Group Collaboration	16/08/2019	
Working Prototype	Project Group Collaboration	30/08/2019	
Systems Analysis and Design Report	Project Group Collaboration	13/09/2019	
OCR Error Checking	Project Group Collaboration	13/09/2019	
Alpha Release	Project Group Collaboration	27/09/19	
Beta Release	Project Group Collaboration	11/10/2019	
Final System Handover	Project Group Collaboration	18/10/2019	
Project Abstract and Video	Project Group Collaboration	25/10/2019	
Final Presentation	Project Group Collaboration	28/10/2019	

Project Management Related Issues and Risks

Risk/Issue	Resolution Expected to be Actioned
The PDF dictionaries vary in quality. Some dictionaries are scanned PDFs which we may have trouble using OCR technologies on them.	Extensive error checking once working prototype is finalised.
Project team has never worked together before – inexperienced team dynamic.	Constant communication between team members throughout development.
OCR technology may not be 100% accurate going into the development process.	Substantial testing phase to be implemented.

Individual Task List

Caleb

Task	Deliverable	Start Date	End Date	% Complete
Create agenda and take minutes for second meeting	Project Plan	6/08/2019	6/08/2019	100%
Project management related risks and issues	Project Plan	6/08/2019	9/08/2019	100%
Research an alternative solution	Project Proposal	5/8/19		40%
Create an interface for product development	System Implementation			0%
Draft system design diagrams	System Analysis and Design			0%
List non-functional requirements	System Analysis and Design			0%
Perform quality of life improvements	System Implementation			0%

Josephine

Task	Deliverable	Start Date	End Date	% Complete
Create agenda and take minutes for first meeting	Project plan	30/7/19	30/7/19	100%
Request group server from IT		30/7/19	30/7/19	100%
Create meeting plan	Project plan	31/7/19	4/8/19	100%
Write executive summary	Project plan	31/7/19	4/8/19	100%
Further research for recommended solution	Project Proposal	05/08/2019		
Create an interface for product development	System Implementation			0%
Draft system design diagrams	System Analysis and Design			0%
List non-functional requirements	System Analysis and Design			0%

Joshua

Task	Deliverable	Start Date	End Date	% Complete
Construct the work breakdown structure	Project plan	30/7/19	2/7/19	100%
Research an alternative solution	Project Proposal	5/8/19		40%
Draft system design diagrams	System Analysis and Design			0%
List non-functional requirements	System Analysis and Design			0%
Output extracted information	System Implementation			0%
Perform quality of life improvements	System Implementation			0%
Construct test plans	System Implementation			0%

Roger

Task	Deliverable	Start Date	End Date	% Complete
Gantt Chart	Project Plan	04/08/2019	09/08/2019	100%
Research about JavaScript and some of its libraries such as tesseract.js and pdf.js	Project System	01/08/2019	30/08/2019 18/10/2019	10%
Write a program that extract text from an already OCRed PDF file and print it to the HTML page.	Project System	06/08/2019	08/08/2019	100%
Write a function that turns the extracted text to a structured format	Project System	10/08/2019	13/08/2019	0%
Construct test plans	System Implementation			0%

Planned Meetings

Type	Frequency	First Meeting	Number of Meetings Planned	Duration	Location and Mode
Client	As needed	30/7/2019	As needed	1 hour	Face to face at the client's office – PS.ED. G.81 or Zoom
Supervisor	Weekly, Tuesday 6PM	30/7/2019	13	1 hour	Face to face – Kingswood campus library
Team	Weekly, Tuesday, 5PM	30/7/2019	13	1 hour	Face to face – Kingswood campus library

Meeting Agendas and Minutes

Client Meeting 1

Meeting Agenda

Project Number	**KS1912**
Meeting Type	Client - Meeting 1
Meeting Date	30 July 2019 (Week 2)
Location	PS - ED.G.81
Start Time	11:00 AM
Duration	1 hour
Chairperson	Dr. Rachel Hendery
Invitees	Caleb Smith Josephine Paculio Joshua Dib Roger Bernardo

A. Purpose of Meeting

Client introduction. Project discussion and clarification.

B. Agenda Items

Item	Description	Who	Time
1	Introduction	Dr. Hendery	11:00
2	Project discussion • Overview • Scope • Current system demonstration	Dr. Hendery	11:05
3	Questions and clarification • Questions from the project team	PX group	11:35

Meeting Minutes

Project Number	**KS1912**
Meeting Type	Client - Meeting 1
Meeting Date	30 July 2019 (Week 2)
Location	PS - ED.G.81
Start Time	11:00AM
Duration	1 hour
Chairperson	Dr. Rachel Hendery
Invitees	Caleb Smith Josephine Paculio Joshua Dib Roger Bernardo
Apologies	N/A
Minute Recorded by	Josephine Paculio

A. Discussions and Outcomes

Item 1: Project Discussion

The main goal of the project is to be able to automate searching for words in dictionaries that have been read through OCR, instead of the current

process which is to literally go through a pdf file or multiple pdf files to search for a word.

Dictionaries are in native Australian languages and pacific languages. Dictionaries tend to be structured in the same way with the head word and translation being the results to be derived. Most dictionaries have not been read through OCR, and if possible, automating the system might include the ability to read those through OCR.

Each search output to be a separate file (for the client to put into a database in the future). Search results could be saved as csv or xml file, preferably xml to follow the standard mark-up used in xml: example to be sent.

Item 2: Questions and clarification

At a minimum the system could be run using command line, but a simple GUI is better, to take PDF files that have been read through OCR. Ideally the system should run in Linux and Windows. No programming language preferred, could be a web app if easier.

Team meeting 1

Meeting Agenda

Project Number	**KS1912**
Meeting Type	Team - meeting 1
Meeting Date	30 July 2019 (Week 2)
Location	Parramatta South campus library
Start Time	12:00PM
Duration	Josephine Paculio
Chairperson	Josephine Paculio
Invitees	Caleb Smith Joshua Dib Roger Bernardo

B. Purpose of Meeting

Project plan discussion

B. Agenda Items

Item	Description	Who	Time
1	Project plan discussion	PX group	12:00

Meeting Minutes

Project Number	**KS1912**
Meeting Type	Team – meeting 1
Meeting Date	30 July 2019 (Week 2)
Location	Parramatta South campus library
Start Time	12:00
Duration	1 hour
Chairperson	Josephine Paculio
Invitees	Caleb Smith Joshua Dib Roger Bernardo
Apologies	N/A
Minute Recorded by	Josephine Paculio

B. Discussions and Outcomes

Josephine will put up a draft project plan and will do executive summary and project management related issues and risks. Caleb will do the Gantt chart. Group details will be an individual write up.

Other project members will divide up parts after the meeting with the supervisor is done.

No need to meet the client next week, the team will work on the project plan and will email it to the client for approval before submitting on vUWS.

Roger will be the one to submit on vUWS on or before the due date.

Agenda and minutes for subsequent meetings will be done on a rotation basis.

The team will wait for the supervisor meeting and see what he recommends in terms of technology to be used for the project.

Supervisor meeting 1

Meeting Agenda

Project Number	**KS1912**
Meeting Type	Supervisor - Meeting 1
Meeting Date	30 July 2019 (Week 2)
Location	Kingswood campus library
Start Time	6:15 PM
Duration	1 hour
Chairperson	Jonathan Weightman
Invitees	Caleb Smith Josephine Paculio Joshua Dib Roger Bernardo

C. Purpose of Meeting

Supervisor – team introduction. Project discussion and clarification based on information gathered from client meeting earlier in the day.

B. Agenda Items

Item	Description	Who	Time
1	Introduction	Jonathan Weightman, PX group	6:15
2	Project discussion • Overview • Scope	Jonathan Weightman, PX group	6:20
3	Questions and clarification • Questions from the project team	PX group	6:50

Meeting Minutes

Project Number	**KS1912**
Meeting Type	Supervisor- Meeting 1
Meeting Date	30 July 2019 (Week 2)
Location	Kingswood campus library
Start Time	6:15PM
Duration	1 hour
Chairperson	Jonathan Weightman
Invitees	Caleb Smith Josephine Paculio Joshua Dib Roger Bernardo
Apologies	N/A
Minute Recorded by	Josephine Paculio

C. Discussions and Outcomes

Communicating with Jonathan is easiest through Google Hangouts, University email for more official communication

For the project, a multi-platform language like JavaScript would be suitable and making a web application would be easier. There are libraries available for OCR, or from Adobe.

Error detection will be important because of the OCR-ed document. The project team needs to plan the data structure output to make sure it's consistent across all pdf files so cleaning the dataset (dictionary PDF files) may be necessary.

A JSON file output would be good because it's an object and can be converted into xml or csv as well. It can also be exported to a table using a plug-in called datatables which has filtering options and options to export to and has sorting as well.

By week six a working prototype with search and extraction working and the output file. Error detection maybe not be fully working by then but will have to be by week 10. For deployment to the client, the project team will just transfer the files over, no need for server.

Server should be requested from IT for presentation if not development.

Team meeting 2

Meeting Agenda

Project Number	**KS1912**
Meeting Type	*Group meeting*
Meeting Date	6/08/2019
Location	Kingswood Library
Start Time	5pm
Duration	1 Hour
Chairperson	Caleb Smith
Invitees	*Everyone in the group.*

D. Purpose of Meeting

Double checking plan requirements, discussing risks and ironing out errors in the project plan for submission on Friday.

B. Agenda Items

Item	Description	Who	Time
1	Going over plan requirements and checking off what has been done so far.	All	5pm
2	Brain storming risks	All	5pm
3	Checking each other work for errors	All	5pm
4	General discussion or project	All	5pm
5			

Meeting Minutes

Project Number	**KS1912**
Meeting Type	*(Client/Supervisor/Team)*
Meeting Date	6/08/2019
Location	Kingswood Library
Start Time	5pm
Duration	1 hour
Chairperson	Caleb Smith
Invitees	*Everyone in the group*
Apologies	N/A
Minute Recorded by	Caleb Smith

A. Discussions and Outcomes

Item 1:

Details of the Discussion: First point of discussion is an introduction on how everyone is going with the project so far and any concerns anyone may have.

Actions Items, Responsibility and Due Date:

1. Discussing a plan to meet plan deadline Friday week 3.
2. All responsible
3. Friday 9/08/2019

Item 2:

Details of the Discussion: Discussion of project risks and possible solutions to those risks.

Actions Items, Responsibility and Due Date:

1. Project plan
2.
3.

Item 3:

Details of the Discussion: Communication with our client and our future meetings with her.

Actions Items, Responsibility and Due Date:

1. Create new meeting date and time
2. Caleb will communicate with client.
3. 6/08/2019

Supervisor meeting 2

Meeting Agenda

Project Number	**KS1219**
Meeting Type	*Supervisor*
Meeting Date	6/08/2019
Location	Kingswood Library
Start Time	6:00pm
Duration	45mins
Chairperson	Caleb Smith
Invitees	*All*

E. Purpose of Meeting

Weekly meeting. We set a purpose to gather more planning details from Jonathon and more insights into development technologies.

B. Agenda Items

Item	Description	Who	Time
1	Discuss planning process.	Supervisor + team	6pm
2	Discuss planning work completed.	Supervisor + Project team	6pm
3	Discuss development technologies most efficient for project requirements.	Supervisor + Project team	6pm
4			
5			

Meeting Minutes

Project Number	**KS1912**
Meeting Type	*Supervisor*
Meeting Date	6/08/2019
Location	Kingswood Campus
Start Time	6:04pm – 6:49
Duration	45MINS
Chairperson	Caleb Smith
Invitees	*(List all personal invited including client/supervisor/team member/other external parties)*
Apologies	None
Minute Recorded by	Caleb Smith

A. Discussions and Outcomes

Item 1:

Details of the Discussion:

How is project planning going and what still needs implementing?

Actions Items, Responsibility and Due Date:

1. Gant chart needs to be modified with more details in the developmental phase.

2. Removal of "presentation phase" in the gant chart as it does not have relevance to the developed system.

3. Due date: Friday.

Item 2:

Details of the Discussion:

Half of the given dictionaries have been OCR so we will need to find out a way to OCR the dictionary PDF's that are not.

Actions Items, Responsibility and Due Date:

1. Problem space, find a solution.

2. Whole team is responsible for research and solution.

3. Development deadline.

Item 3:

Details of the Discussion:

Simple spell checks will be needed for error correction.

Actions Items, Responsibility and Due Date:

1. Create a method for spell checks to negate errors in the dictionaries.

2. Due date, developmental process deadline.

3.

Item 4:

Details of the Discussion:

Design an interface – draw it on paper or design it through HTML.

Actions Items, Responsibility and Due Date:

1. Due date next week.

2. Caleb's responsibility

3.

Item 5:

Details of the Discussion:

Capitalise on screen size. EG: Bootstrap etc. Pick a framework and add in navigation functionality.

Actions Items, Responsibility and Due Date:

1. Research.

2. Team contribution.

3. Development process.

Conclusion

After concluding this document there are several key takeaways including, the group allocated for the task, an insight into the work structure and the key milestones during development, risks that must be kept in mind, as well as a timeline such a project may follow.

A more substantial report building upon what has been mentioned is to follow and will detail the proposed technologies and features for the final product.