PROJECT INITIATION DOCUMENT:

Text Summarization using Open-Source Python Tools

PROJECT INFORMATION

- **Project Name:** Text Summarization using Open-Source Python Tools
- **Date of Document Submission:** February 3, 2023
- **Version:** 1.0
- **Client:** IMPaCT. 

PROJECT TEAM

The members working in project are distributed in three categories:

A **Product Owner**, will be working as Liaison between the client(sponsor) and the rest of team members. The development team and Scrum Master will be reporting to the Product Owner about how the team is doing and the Product Owner will be answerable to the client about the reports of project completion.

A **Scrum Master** keeps track of the team performance, makes sure everyone knows what to do and also makes sure the problems are mitigated and managed. The Scrum Master is also responsible for keeping the track of backlog and reporting the functioning of the team to the product owner.

And the **Development Team** consists of **developers** who will receive tasks from the Product Owner and will be looked after by the Scrum Master. The role of the development team is to essentially understand the requirements of the client,

researching on how to build the product, making sure to get the product completed according to the defined scope and get it delivered.

Team Member and their Responsibilities:

Team Member	Role
Akshat Baranwal	Product Owner
Steven Warner	Scrum Master
Cameron Caruso	Developer in Development Team
Parth Dalwadi	Developer in Development Team
Michael Provenzano	Developer in Development Team
Mark O'Connor	Developer in Development Team

TABLE OF CONTENT

1. PROJECT SCOPE AND DEFINITION

Project Definition:

This project will use AI tools to generate summaries of text conversations. These conversation summaries will be generated as timestamped pdf reports, which can be shared or stored as needed.

Scope of this project:

This project will be deployed on servers run by Impact and users will be able to use this service on demand via the main Impact application. They will also be able to define time ranges for the messages that need to be summarized. This project will also deliver its summaries with a response time between 100ms and 150ms.

Problem to be solved:

Criminals and their organizations have the ability to evade law enforcement from one jurisdiction to another by utilizing information silos. To mitigate this problem, departments from different jurisdictions need to be able to share information with each other in a timely manner.

Outcomes of the project and how are they measured:

By generating summaries of text conversations between two or more officers, it allows information to be quickly dispersed between departments, higher-ups, or other offices within the department. The summaries themselves will be evaluated based on the ability to pick the most relevant information.

The Deliverables:

The deliverables for this project will involve a Python script to process text, and a machine learning algorithm that will generate the summaries. The Python script is used to gather the messages from the original text channels and turn them into a readable format for the machine learning algorithm. The machine learning algorithm will process this text into summaries and return them as pdf files with appropriate timestamps.

Projected Task for all 5 Sprints(Tentative):

Sprint 1: The team will be familiar with the provided server and database environments. Team will also be familiar with basic language processing techniques, processes, and tools.

Sprint 2: The Python text processing script will be completed. It will be able to properly select messages based on a time range and process them so that it will be readable by the machine learning algorithm.

Sprint 3: The machine learning algorithm will be completed. It will be able to process the inputted text into readable format. It will prioritize the important details of a text conversation.

Sprint 4: The algorithm and script will be tested within a sandbox server environment. Any discovered bugs will be fixed. The deliverables will be made to work within the given server environment without interfering with other processes.

Sprint 5: The clients will have all deliverables in their hands. A user manual will be provided. A final presentation showing project usage and outputs will be given to the client.

Constraints:

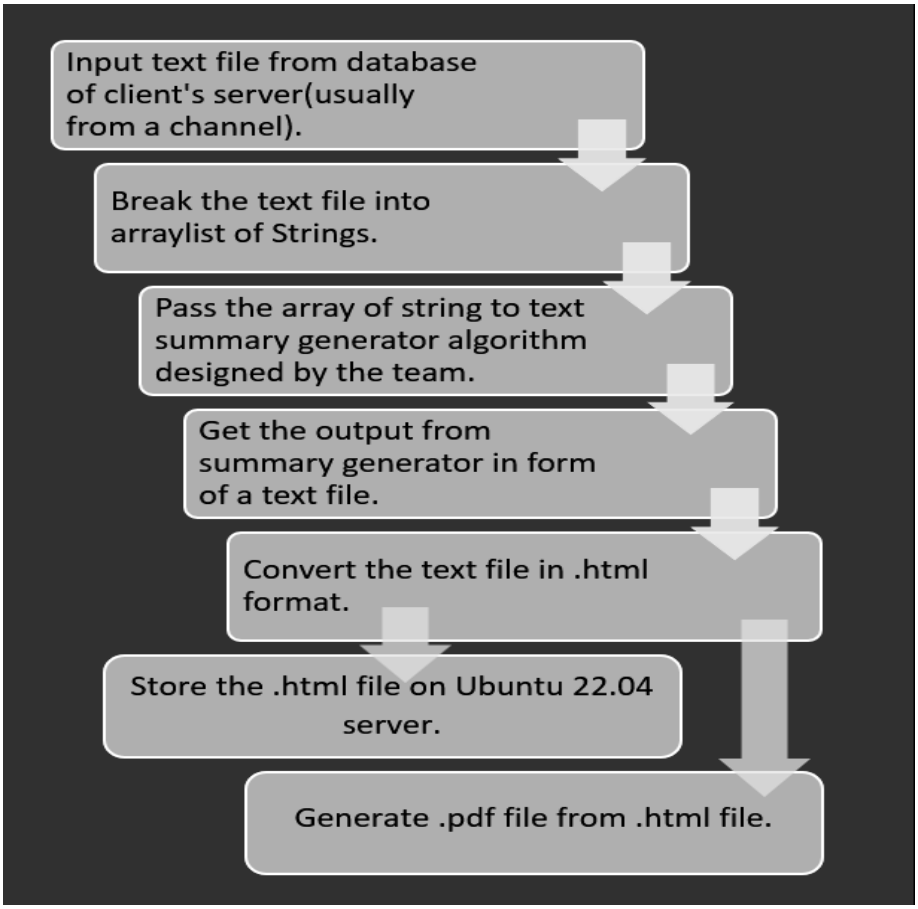
Privacy concerns and server environment will constrain our deliverables. The server environment will dictate that all deliverables will have to be written in the Python programming language. Due to privacy concerns about data, no data being accessed can be sent to any third parties. All Python tools used must be open sourced to avoid this concern.

2. COMMON TERMS AND DEFINITIONS

Text Channel: A text channel includes text conversations between two individuals.

Chat Rooms: Chat rooms are basically places where two or more than two individuals can text and share information.

3. PROJECT DIAGRAM



Flow chart showing the working of the Product.

4. PROJECT METHODOLOGY

The methodology involved in completion of this project.

1. To complete this project, we will be using "**The Scrum**" agile methodology. Scrum defines goals that must be completed by the end of a project rather than a typical thorough plan of how each step of the project must be completed. The major advantage of this methodology is how it is easy to change the direction of where the project is going based on both team meetings and meetings with the sponsor.

2. Throughout the whole project, the team members will share their ideas to the product owner and the product owner will be responsible to convey it to the client. Zoom meetings can be arranged between the client and the project members according to their convenience.

3. The team members will have meetings scheduled at least thrice a week, in order to understand and discuss the project. In the meetings itself, under the presence of all the team members, the crucial decisions will be made on how to lead the project forward. The only decisions made individually are going to be about the responsibilities each team member has been assigned to them and how they are going to execute it. Every decision involving the whole team is going to be counted as milestones in the process of completion of the project.

4. As mentioned in the above point, the decisions made on how to further proceed the project are going to be counted as milestones in the process of completion of the project. Achieving the milestones will let the team keep track of the progress. A "Trello Board" will be used as the team's product backlog, or roadmap, to better prioritize a list of project requirements and tasks. All the changes and setbacks will be reflected in the Trello Board, later they are going to be involved after discussion of the whole team member to the "Sprint-To Do list," where it gets executed.

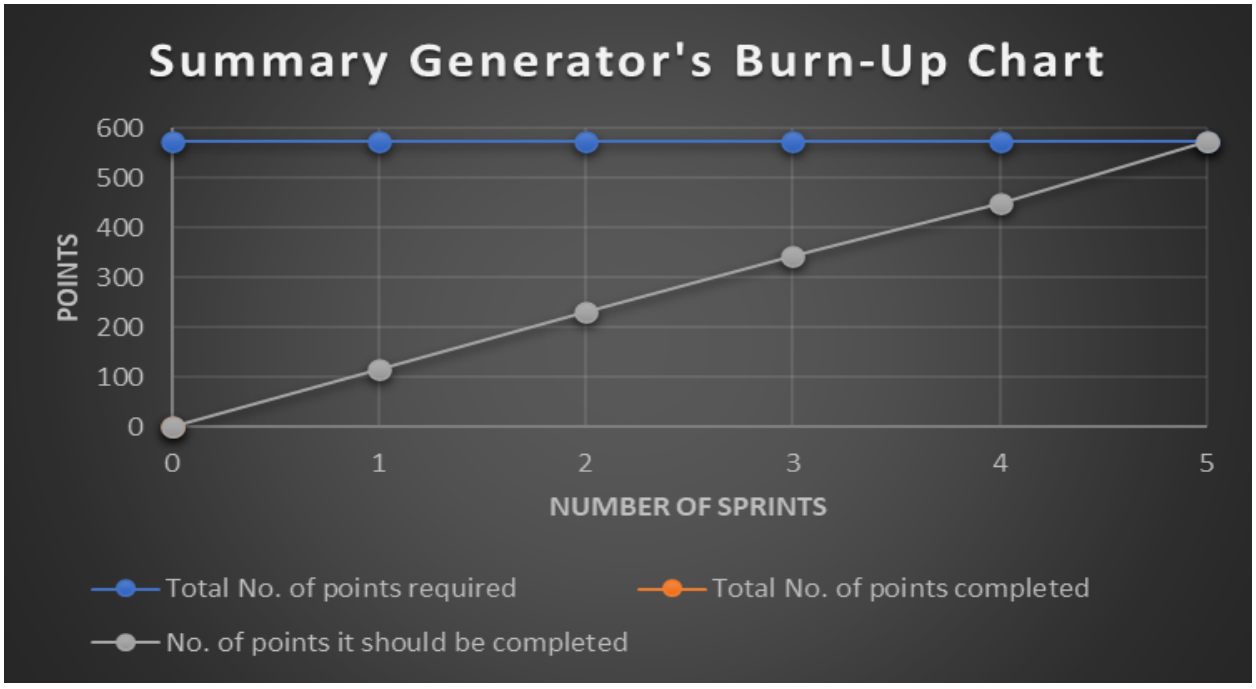
5. PROJECT REQUIREMENTS

The project is mainly maintained using a "**Trello Board**" to keep track of the progress. All the required tasks are going to be converted into "User Stories Backlog" which later on will be distributed in "**Sprint's Task Backlog**" and "**Sprint's To-Do List**."

Our backlog on Trello: https://trello.com/b/nm3NqUPr/african-elephant-team

Apart from this, we will be needing the client's Git Repository, some external tools (like the one to convert HTML to pdf), branch in client's server and maybe some other tools or access which for now is unforeseen for the Project Team.

6. BURN UP CHART



The total project is divided into five sprints, where for each sprint, some random but evenly distributed points have been assigned by the team. The points add to 575 in number (An arbitrary Number), giving a "linear" line representation to the process of completion of Project.

7. PROJECT LIMITATIONS, CONSTRAINTS, AND FEASIBILITY

Project Feasibility and its limitations:

According to the planning of the development team and other project members, the team believes that the project will be completed within our five sprints and on the 5th sprint, the client(sponsor) will be having the product in their hand.

Project integration with other existing systems:

The project will have to be integrated with our sponsor's servers and the input will come from the client's databases and the output will be sent to their server(s).

Limitations to the success of the project:

Since the nature of our project is linguistic based, the project may make summaries that feel robotic, poor grammar, or probably include irrelevant details in the summary due to limited time and data for machine learning training.

Constraints:

Privacy concerns and server environment will constrain our deliverables. The server environment will dictate that all deliverables will have to be written in the Python programming language. Due to privacy concerns about data, no data being accessed can be sent to any third parties. All Python tools used must be open sourced to avoid this concern.

Technical Unknowns:

- Currently undefined project tools.
- Database Formatting.
- Use of Celery (mentioned in a previous meeting with the client).

Foreseen points of Failure:

The team does not foresee any point of failure up to the date. The requirements from the client end are pretty clear to the team and the process of building the product is clear between the development team.

8. DEFINITION OF SUCCESS

Definition of Success:

All deliverables provided to the clients work well within the given server environment. The deliverables produce the required summaries in HTML format (later to pdf), with all pertinent details.

Metrics defining Success:

Metrics that measure success can be defined include a text-summarization tool that works functionally and the ability to have it work on a Ubuntu 22.04 server.

To Elaborate, it should take input of text conversation between two people (A text Channel), and generate an output in the form of a summary report of it in the form of HTML (later on can be converted into a pdf file). The output should be a working format on Ubuntu 22.04 Server.

Ensuring Product Quality:

Product quality will be ensured by constantly testing the application in a private branch of the client's server environment. The team will provide these test results to the sponsor and implement any feedback. In addition to this, the team will also have scheduled meetings with the clients to discuss any additional requirements that may arise.