**Project: Que-Ans Machine with NLP**

1. Project Planning and Setup
   1. Project Overview
      1. Create a system that considers the entire document set or a small relevant subset to generate the correct answer.
      2. Scope of this project includes:
         1. The abstracts of research papers/articles which include the word “intelligence” published between 2013 – 2023 (both years included) and are available on pub-med(<https://pubmed.ncbi.nlm.nih.gov>)
         2. Forms of questions
            1. Yes/No questions.
            2. What, Which, When, Who, How questions
            3. List-type questions
            4. Casual Questions (why/how)
            5. Hypothetical questions
            6. Complex questions
   2. Stakeholder Identification
      1. Head of Project – Professor Dr. Michael Gertz [gertz@informatik.uni-heidelberg.de](mailto:gertz@informatik.uni-heidelberg.de)
      2. Project Mentor – Robbin Khanna [r.khanna@stud.uni-heidelberg.de](mailto:r.khanna@stud.uni-heidelberg.de)
      3. Team Members – Pranjal Sharma, Mateusz Stączek, Jan Tadeusz Smoleń, Agata Katarzyna Kaczmarek
   3. Technology Stack
      1. Python X.X
      2. NumPy X.X
      3. Pytorch X.X
      4. Frontend –
      5. Backend –
   4. Data Collection
      1. PubMed search criteria: abstract or title should contain word “intelligence.”
      2. Year from 2013 – 2023
      3. PubMed does not allow to process more than 10000 articles, so the years had to be divided in batches 2013-2017, 2018-2019, 2020, 2021, 2022, 2023.
      4. PubMed PMID extracted as .XLSX for the year as mentioned above in batches.
      5. These PMID had been inputted on the website pubmed2xl to download the excel file contained with abstracts of these PMID.
   5. **Project Timeline/Deliverables:**
      1. **Deadline 1:**
      2. **Deadline 2:**
      3. **Deadline 3:**
2. Data Preprocessing and Exploration
   1. Data Cleaning
      1. Steps of cleaning
   2. NLP Preprocessing
      1. Steps of preprocessing (tokenization, stop word removal, etc.) on the abstracts.
   3. Data Analysis
      1. Data analysis
3. Model Development
   1. Model Selection
      1. Hugging face model
   2. Model Training
      1. Steps to train the model.
   3. Evaluation Metrics
      1. Metrics defining
         1. Accuracy
         2. F1 score).
   4. Model Fine-Tuning
      1. Optimizing the model
4. Deployment and Documentation
   1. Deployment Strategy
      1. Strategy for deploying the question-answering machine.
   2. User Documentation
      1. Prepare readme file to understand installation and usage.
   3. Conclusion and Future Work
      1. Summarize the project outcomes and suggest possible future enhancements.
   4. License usage
   5. Project licensing/open sourcing
5. Web Application Development (UI)
   1. FRONTEND
   2. BACKEND
   3. Integration and Testing

**Version Control:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Version** | **Updated on (Date)** | **Author** | **Contact** | **Comment** |
| **0.1** | **21.11.2023** | **Pranjal Sharma** |  | **First Draft** |
|  |  |  |  |  |
|  |  |  |  |  |