**Proposal for HistoriBot – AI-Powered History Q&A System**

**1. Introduction and Motivation**

The goal of this project is to develop **HistoriBot**, an AI-powered **Question Answering (Q&A) system** specializing in **Canadian History**. Historical information is often scattered across multiple sources, making it difficult to retrieve precise answers efficiently. By leveraging **Natural Language Processing (NLP) and Machine Learning**, this project aims to create an intelligent system that provides **accurate and relevant responses** to historical queries about Canada.

**2. Team Details**

**Team Member:** *Drayton,*  *Stephen*

**3. Dataset Selection and Preprocessing**

The dataset for this project will be sourced from **The Canadian Encyclopedia**, which contains a detailed timeline of significant events in Canadian history.

**Steps to Prepare the Dataset:**

* **Web Scraping**: Using **BeautifulSoup4** and **Requests** to extract historical events from [The Canadian Encyclopedia](https://www.thecanadianencyclopedia.ca/en/timeline/100-great-events-in-canadian-history).
* **Data Cleaning & Preprocessing**:
  + Removing **HTML tags, special characters, and irrelevant content**.
  + Tokenizing and **normalizing text** (lowercasing, removing stopwords).
  + Structuring the dataset into **context-question-answer** pairs for model training.
* **Training & Fine-tuning**:
  + Using a **pre-trained transformer model** (e.g., **BERT, RoBERTa, or DistilBERT**) and fine-tuning it on the **Canadian history dataset** for better accuracy.

**4. Software and Tools**

To develop and deploy the Q&A system, we will use the following technologies:

* **Web Scraping & Data Processing**: BeautifulSoup4, Requests, Pandas, NLTK
* **Machine Learning & NLP**: Hugging Face Transformers, PyTorch, TensorFlow, Scikit-learn
* **Model Fine-tuning**: BERT, RoBERTa, or DistilBERT trained on the historical dataset
* **Evaluation Metrics**: F1-score, Exact Match, and human-annotated accuracy assessment
* **Web Deployment**: Flask for building an interactive web interface

**5. Plan of Work (3-Week Timeline)**

| **Week** | **Milestone** | **Task Description** |
| --- | --- | --- |
| **Week 1** | **Data Collection & Preprocessing** | - Scrape data from The Canadian Encyclopedia  - Clean and preprocess the dataset  - Structure data into context-question-answer pairs |
| **Week 2** | **Model Development & Fine-Tuning** | - Select a **pre-trained transformer model** (e.g., BERT)  - Fine-tune the model on Canadian history dataset  - Evaluate accuracy using test questions |
| **Week 3** | **Web App Development & Final Evaluation** | - Develop a **Flask web app** for user interaction  - Integrate the trained model into the web app  - Conduct **final performance testing** and document results |

**6. Expected Outcome**

* A functional **AI-powered Q&A system** capable of answering Canadian history-related questions.
* A **Flask web app** where users can input questions and receive AI-generated responses.
* An evaluation report detailing the system’s **accuracy, strengths, and limitations**.

This project will contribute to **historical education and research**, making **historical knowledge more accessible and interactive** through AI.