Intelligent Tutoring System Chatbot (ITS-Chatbot)
Project Proposal

Group Members and Skills

- Samuel Toh (ST)
  - 3rd-year Computer Science Major
  - Programming Experience: Java, Python, C/C++, Django, SQL
  - Responsibility: Data processing, Frontend Development
  - Current Tasks: Research and Data pre-processing
- Andrew Chen (AC)
  - 2nd-year Computer Science Major
  - Programming Experience: Java, Python, C/C++
  - Responsibility: Natural Language Processing and Backend Development
  - Current Tasks: Research and Create Machine Learning model prototype
- Kunal Arora (KA)
  - 1st year Computer Science Graduate Student
  - Programming Experience: C/C++, Python, Java, Scala, C#, HTML, PHP, SQL, Cassandra
  - Responsibility: Machine Learning Algorithm Development
  - Current Tasks: Research and Data pre-processing

ITS-Chatbot Description

The main purpose of ITS is to help students succeed in the class through providing or directing them to the resources based on the strengths and weaknesses of the student in regards to the content of the course. While TAs serve as an integral resource for students, the team believes that a chatbot is needed to help share the workload of TAs and provide more data-driven responses to students' questions. Our goal is to develop an embedded chatbot that will help TAs to handle high volumes of questions before deadlines or exams as well as provide a more personalized experience.

In order to provide sufficient information to the proposed chatbot, data retrieval and preprocessing required. As a result, this semester, the team aims to understand, clean, and label the data available on the Piazza site of current and past courses, annotated textbook, and the syllabus. In addition, we will categorize the Piazza questions into three categories–Policy, Assignment, and Conceptual–to better understand the frequency and distribution of the types of questions students make throughout a semester. The end goal for this semester is to be able to complete a chatbot interface for Policy and Assignment questions, to which the responses can be easier to extract from the data.

Possible Solution/Things to Try

Pre-processing of dataset:

- Tokenization
- Stemming and Lemmatization
- Parts of Speech (POS) tagging
- Named entity recognition
- Stopword removal
- Word embeddings

Dataset Split:

Divide the dataset into training data, cross validation data and testing data in a ratio of 80:10:10.

Model Design:

- **Approach 1 - Question Similarity**: Using word embeddings and tf-idf, create document vectors. Use the cosine similarity score to find which question Q’ from the dataset is the new question Q* most similar to. Then either post the link to Q’ or post the instructor’s answer to Q’ as the answer to Q*.

- **Approach 2 - Transfer Learning:** Take a multi-layer deep learning model that is already trained on conversational data. Use our data and train the last few layers of the model. Tune the hyper-parameters until good accuracy is achieved.

- **Approach 3 - Deep Learning Model from scratch:** Take a publically available conversational dataset like Persona Chat (provided by Facebook) and combine it with our data. Develop a multi-layer deep learning model from scratch and train it on the combined dataset. Tweak the model and tune hyper-parameters until good accuracy is achieved.

- **Other Approaches/Things to try:**
    - Use Convolutional Neural Networks, Recurrent Neural Networks instead of just Deep Neural Networks
    - Rule-based methods
    - Intent classification
    - Generative Adversarial Networks
    - Use state of the art language models like BERT and GPT-2
    - Reinforcement Learning
    - IBM watson

Project Goals

- Retrieve data from the available data file such as contributions.csv, users.json, etc.
- Connect the endpoints between Piazza and the ITS database using Piazza API

- Pre-process data using tokenization, lemmatization, parts of speech (POS) tagging, dependency parsing, named entity recognition etc.
- Categorize data into Policy, Assignment and Conceptual questions
- Design and develop a machine learning model that generates an answer A given a question Q.
- Build a front-end interface for demo for responses to Policy and Assignment questions

Project Timeline

Week	Task
Week 4	Project Proposal Draft Due
Week 5	Final Project Proposal Due
Week 6-7 Phase 1: Resource Setup/Research/Data Preprocessing	Continue the research on building the bot and gathering data on policy-related questions from the class syllabus and Piazza. Start data pre-processing. **Milestone 1: Finalized at least 1 approach with fixed steps. All required data is available, cleaned and pre-processed. Data ready to be consumed.**
Week 8-12 Phase 2: Bot building	Start building the bot based on the data we have, only focus on answering policy-related questions this semester **Milestone 2: Chatbot works locally using command line**
Week 13-14 Phase 3: Robustness	Finalize everything, have a front-end to show how the bot functions **Milestone 3: Chatbot works with a basic frontend**
Week 15	Prepare for the final presentation and upload all of the code to GitHub
Week 16	Final presentation

Implementation Tools & Resources

- GitHub: https://github.gatech.edu/VIP-ITS
- Python
- Machine Learning Libraries: scikit-learn, Tensorflow/Keras/Pytorch

- NLP libraries: nltk, spacy
- Data processing libraries: numpy, pandas
- Communication: Slack
- Project Documentation Notebook
- Task Management: Trello