Project 2

CS 421: Natural Language Processing

Due: April 18, 2025 (11:59 p.m. CST)

1 Introduction

Welcome to the Project 2 for CS 421! In this project, you will explore the fascinating world of conversational agents by building and comparing different types of chatbots. You will start by implementing the classic ELIZA chatbot, a **rule-based chatbot** that uses pattern matching to respond to user inputs. Next, you will build a **corpus-based chatbot**, leveraging data-driven techniques to generate responses. Finally, you will interact with a **large language model (LLM)** to experience the state-of-the-art in conversational AI.

For all three chatbots, you will store the chat history to analyze and compare the interactions. At the end of the project, you will conduct a human evaluation to evaluate the performance and user experience of each chatbot.

Don't hesitate to reach out on Piazza or during office hours with any questions you have as you complete it! Happy Coding!

2 Instructions

Each part of this deliverable is labeled as Code or Written. The guidelines for these two types of components are provided below.

2.1 Code

The questions need to be completed using Python (version 3.10+). There are **no external packages** required to complete this project. If you want to use an external package for any reason, you are required to get approval from the course staff on Piazza prior to submission. Starter code is provided for the project in the form of a Jupyter notebook in the supplementary material. Do not rename/delete any functions or global variables provided in this python notebook, and write your solution in the specified sections. The python notebook may also contain important information and/or examples in comments, so please read them carefully. This part of the project will be graded manually after submission on Gradescope.

To implement your models, please utilize Google Colab[1]. Please take note of the usage limits on Google Colab and plan your work accordingly. You may also implement the assignment locally, however, please be prepared for a longer wait time.

To submit your solution for Code questions, you need to download the `.ipynb` file[2]. You need to submit the following file as a part of your submission:

- ☐ `chatbots.ipynb`
- ☐ `*.txt` files for all conversations

Submit this `.ipynb` file on Gradescope under Project 2. All specified files need to be submitted to receive full credit.

2.2 Written

You are required to submit all Written questions in a single PDF file. You may create this PDF using Microsoft Word, scans of your handwritten solution, LaTeXor any other method or design tool you prefer. To submit your solution for Written questions, you need to provide answers to the following questions in a single PDF.

- ☐ Q4
- ☐ Q5

Before submission, ensure that all pages of your solution are present and in order. Submit this PDF on Gradescope under Project 2. All questions need to be completed to receive full credit.

3 Questions

3.1 Code (50 points)

Q1 (15): Rule-Based Chatbot
In this part, you will be implementing the rule-based chatbot ELIZA using the rules given in `eliza.py`. The given `.py` file provides an interface to use the ELIZA model by defining an instance using the `.Eliza()` method and generating responses using the `.respond()` method. Use these methods to generate responses from the ELIZA model, while storing the chat history to a `.txt` file. Follow the file `chatbots.ipynb` for step-by-step instructions. Generate 3 chat conversations with at least 10 utterances in each conversation. A sample conversation is provided in `sample_conversation.txt`.
Supplementary material: `chatbots.ipynb, eliza.py`

[1] http://colab.research.google.com/

[2] To download the `.ipynb`, select `Download > Download .ipynb` from the `File` menu on Google Colab

Q2 (30): Corpus-Based Chatbot
In this part, you will be implementing a corpus-based chatbot using the corpus provided in `'dialogues.csv` (derived from the Empathetic Dialogues dataset[3]). You will perform the following steps:

1. Load the data using Pandas Dataframe[4].
2. Calculate the contextual embeddings of the sentences in the corpus using the SentenceTransformer[5] model.
3. Find the most similar sentence in the training corpus using cosine similarity[6]
4. Output the response as the corresponding value of 'Agent' for the most similar sentence.

Generate 3 chat conversations with at least 10 utterances in each conversation for performing human analysis.
Supplementary material: `chatbots.ipynb, dialogues.csv`

Q3 (5): LLM Chatbot
Utilize a Large Language Model (LLM) such as ChatGPT[7], Llama 2[8], Mistral[9] or any other LLM of your choice to converse with it to generate 3 chat conversations with at least 10 utterances in each conversation. Save these conversation histories `.txt` files to be utilized for human analysis.

3.2 Written (50 points)

Q4 (30): Human Analysis Survey
Conduct a human-analysis of the responses generated from your chatbots. As a part of this analysis, you must score each chatbot (given all the conversations associated with that chatbot) on the following questions to be answer on a scale of 1-5 (Likert Scale - where 1 means 'Strongly Disagree', 2 means 'Disagree', 3 means 'Neither agree or disagree', 4 means 'Agree' and 5 means 'Strongly Agree') for each conversation:

1. Were the responses generated by the chatbot **fluent**? That is, were they grammatically well formed and fluent in the English language?
2. Were the responses generated by the chatbot **relevant**? That is, in the context of the user input, did the chatbot provide a relevant response?

[3] `https://github.com/facebookresearch/EmpatheticDialogues`
[4] `https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.html`
[5] `https://sbert.net/docs/package_reference/sentence_transformer/SentenceTransformer.html`
[6] `https://sbert.net/docs/package_reference/sentence_transformer/SentenceTransformer.html`
[7] `https://chatgpt.com`
[8] `https://huggingface.co/spaces/huggingface-projects/llama-2-13b-chat`
[9] `https://huggingface.co/spaces/hysts/mistral-7b`

3. Were the responses generated by the chatbot **coherent**? That is, do the chatbot provide consistent responses to user input given the previous utterances in the conversation?

4. Were the responses generated by the chatbot **empathetic**? That is, is the model empathetic towards the user and provides responses that exhibit empathy?

5. How was the **tone** of the chatbot responses (1: Extremely Negative to 5: Extremely Positive)?

At the end of this survey for all conversations, provide the average scores achieved by each chatbot model for this survey results.

Q5 (20): Comparing chatbots
Based on your interaction with the three kinds of chatbot models, provide a short description of the strengths and weaknesses of each model.

4 Rubric

This assignment will be graded according to the rubric below. Partial points may be awarded for rubric items at the discretion of the course staff.

Q1 (15 points possible)	
Correct Implementation	+10
Generating conversation logs	+5
Q2 (30 points possible)	
Correct Implementation	+25
Generating conversation logs	+5
Q3 (5 points possible)	
Generating conversation logs	+5
Q4 (30 points possible)	
Successfully conducting human analysis	+30
Q5 (20 points possible)	
Appropriate responses to all questions	+20