Chatbot Report

**System design.** The chatbot was developed using several NLP techniques and integrated with ChatGPT. The NLP techniques used include document vectorization, cosine similarity, sentence tokenization, Named Entity Recognition (NER), sentiment analysis, word lemmatization, and part-of-speech (POS) tagging.

**Description of NLP techniques.** Text normalization is foundational to any NLP task. It prepares text so that it can be used in an insightful way. Tokenization is the dividing of text into smaller units. The chosen units were sentences and/or words, depending on the context of the process. A second preprocessing measure taken was word lemmatization. This is the conversion of words to their root form. For example, additionally and additional will both be interpreted as addition. Both of these were facilitated through the Natural Language Toolkit (NLTK) library.

The following three techniques are used through pretrained models provided by spaCy, which is a library for industrial-strength natural language processing. POS is the assignment of grammatical classifications (verbs, adjectives, nouns, etc.) to words of a text. NER refers to information extraction from unstructured text that highlights and labels different types of identities such as person, date, money, place, and organization. Sentiment analysis is about inferring the emotion behind a text, whether it is positive or negative.

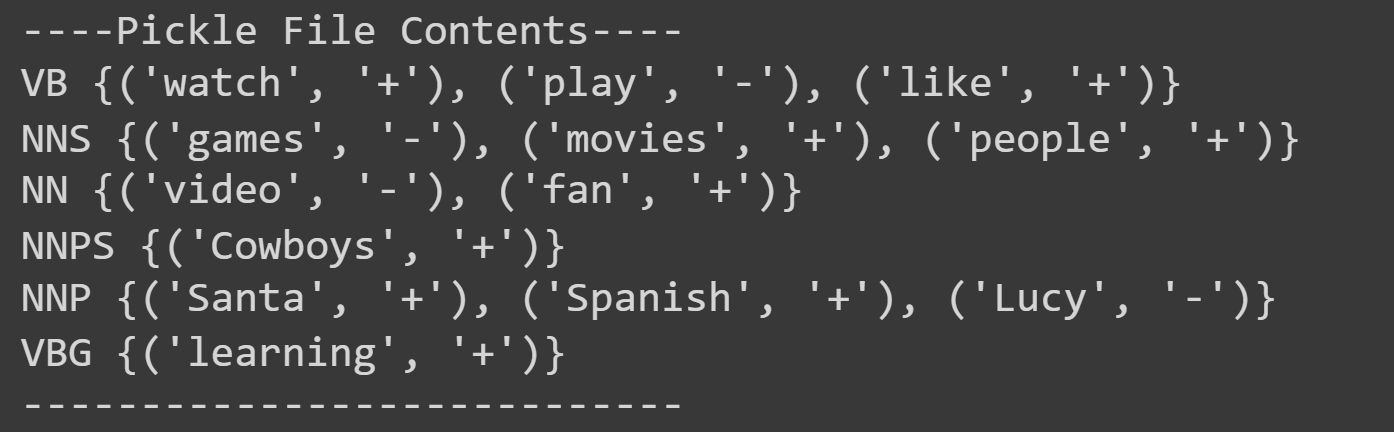
Document vectorization is about representing documents as a simple mathematical structure based on the vocabulary present in that document compared to the entire corpus. Thus, by comparing two vectors of a corpus, their similarity can be assessed. Cosine similarity is a formularized comparison of two vectors. It also normalizes the number so that if the cosine is closer to one then the vectors are more similar. These two techniques were trivially implemented from scratch based on theory learned in class.

|  |
| --- |
| Figure: Dialogue tree |

**Usage of NLP techniques.** The first main component of the chatbot is support for user models. This means that the chatbot is able to remember the user’s basic information, likes, and dislikes. Two different approaches were tested for this functionality. The first is one that is rules-based. The second one depends on the more elaborate GPT-3.5 model provided by the OpenAI API. Both approaches use NER to extract the user’s name and date of birth (age). This had to be consistent because these were used to identify individuals. Each person gets an individual pickle file identifiable by <name><dob>.p which contains a Python dictionary of their additional information.

The rules-based approach collects that “additional information” by a broad prompt such as, “Tell me more about yourself.” From there, sentence segmentation and sentiment analysis is used to assess the emotional tone of the topics of that sentence (Is X and Y a like or dislike?). Words of that sentence are marked with either ‘+’ or ‘–’. POS-tagging is then conducted on each word. Now, each word is a unit of triplets. The words are filtered down to only verbs and nouns. The Python dictionary’s keys are the part-of-speech and the value is a list of tuples (word, +/-). So when the users sign back in using their unique filename/userID, the pickle file is unloaded and a random word is selected. Based on its part of speech and user sentiment toward it, a question is formed. These questions were statically encoded. A quick (but personal by callback) exchange is shared before transitioning into the second component.

| I like to watch movies. I am a fan of the Cowboys. I like to play with Lucy. I enjoy learning Spanish.  I do not like video games. Nor do I like rude people, nor Santa. |
| --- |
| Sample Additional Information |



|  |
| --- |
| User model given by Approach 1 |

The GPT approach collects additional information more directly, as the chatbot asks questions such as, “What are your dislikes?” Then, it uses NER and noun chunks (bigrams, etc) to extract inferred likes/dislikes. So, the Python dictionary has keys such as name, likes, and dislikes whose values are a string of comma-separated entities.

|  |
| --- |
| User model given by Approach 2 |

The second main component of the chatbot is an informative nature about recent developments in AI/ML/NLP via a knowledge base. To do this, word tokenization had to be done. Then, standard preprocessing (lemmatization, stopword removal) was used to get a simplified raw text. Then the words were put under a union operation to get a vocabulary across all documents. Then, each document was compared against the entire vocabulary to create vectors. The bag-of-words technique is used to gather the token counts and helps quantify the comparison. That way, when the system is presented with a question, it can vectorize the question and then calculate cosine similarity to retrieve the most relevant documents.

**Strengths/weaknesses.** At times there were bigrams such as ‘learning Spanish’ that lose its full meaning when only ‘Spanish’ is captured by the system. SpaCy’s NER also failed to recognize some names and date formats given to it. However, a strength in the model was leveraging OpenAI’s large language model (LLM) which covers the generative requirement of a chatbot .Another weakness of this model is in its inability to remember previous exchanges or conversations. It would probably require a lot of cross prompt engineering to create a prompt for the model that reflects messages just discussed. An advantage of this model is that it uses sentiment analysis to provide an appropriately empathetic (though generic) response to the user. The final weakness worth mentioning is that it cannot truly know the semantic of what the user is saying. If the user simply tells the chatbot about something (maybe unenjoyable) they did that day instead of their interest, it will be registered as an interest. The system design could have been more elaborate and streamlined had we figured out a better way to extract user interests.

**Appendices of examples.**

| Chatbot: Hello, I am Chatbot. Have we spoken before? (Y/N)  User: N  Chatbot: What is your full name and date of birth (yyyymmdd)?  User: Brian Cox 20000414  Chatbot: Date not detected. Please use format yyyymmdd (no delimiters, pad left 0)  User: 20000414  Chatbot: What are your likes and interests?  User: I like to watch movies. I am a fan of the Cowboys. I like to play with Lucy. I enjoy learning Spanish.  Chatbot: What are your dislikes?  User: I do not like video games. Nor do I like rude people, nor Santa.  Chatbot: Nice to meet you! Your userID is briancox2000041420000414 for future conversations. What do you want to talk about?  User: What have you been up to today?  Chatbot: I’ve been researching the latest news about Microsoft’s Edge browser and its AI image generator.  User: Tell me more about this Microsoft Edge browser and its AI image generator.  Chatbot: Microsoft Edge is a web browser developed by Microsoft. It is based on the Chromium open source project and was released in January 20202. The AI image generator can create images given a prompt.  User: That’s cool. Thank you  Chatbot: You’re welcome! Is there anything else I can help you with?  User: No thank you. Good bye  Chatbot: Goodbye! It was nice talking to you  User: q |
| --- |
| Sample sign up interaction |

| Chatbot: Hello, I am Chatbot. Have we spoken before? (Y/N)  User: Y  Chatbot: Welcome back! What is your userID?  User: briancox2000041420000414  Chatbot: Welcome back, Brian Cox 20000414. What is your favorite movie, Brian?  User: I really like Spiderman 2. What about you?  Chatbot: That’s tough… What would you like to talk about today?  User: What are your thoughts about AI?  Chatbot: I think AI has the potential to be a powerful tool for good, but is it important to ensure that it is used responsibly and ethically. We need to ensure AI systems are safe for all parties involved.  User: Are you aware of your existence?  Chatbot: I am aware that I’m a chatbot, but I’m not sure what my existence means. I don’t know.  User: Okay good bye.  Chatbot: Goodbye! It was nice talking to you.  User: q |
| --- |
| Sample sign in interaction |

|  |
| --- |
| Sample of knowledge base |

* Knowledge base consists of AI related new articles
* Knowledge base is an xlsx file which can be found on eLearning and Github