Nishanth Lam

CS 4395.001

Chatbot Project Report

**System Description**

I made a chatbot that can answer questions from users about tigers. It can also make some basic conversation at the start such as asking the user for their name, likes, and dislikes. The bot mainly works using a “Bag of words” and TF-IDF approach. Bag-of-words is mainly used to represent the input corpus after preprocessing. The bag-of-words allows us to represent the text in a way that describes the occurrence of words within the corpus. It mainly gives us the information of the entire possible vocabulary and how high a presence has in the corpus. This is useful, because it allows us to make the basic reasoning that two documents are similar if they have identical content. This is crucial, because for my bot’s implementation, it depends on being able to identify keywords and then spit out sentences that are similar in hope that it will answer the user’s question.

I made some basic improvements on the bag-of-words using TF-IDF Vectorization. This is way to reduce the dominance of some words by penalizing and giving less weight to words that appear too frequently. This way, common words such as “is” will not dominate our knowledge base. It also helps to give weight to longer documents since they are usually the ones that will contain more useful information. TF-IDF will generally help us to evaluate how important each word is.

Finally, we can use cosine similarity on two vectors to decide which possible answer vector is closest to the user vector, and then choose that as the bot’s response to the user.

**NLP Techniques Used**

* I searched the web to identify good possible documents about tigers
* I used web requests and html parsing to grab text content from several different websites about tigers.
* I used file io operations to save extracted text easily, and then be able to read the text into tokenizers easily as well.
* I parsed the input text using several NLTK functions as tokenizing to words and sentences. I also lemmatized the text using WordNet lemmatizer to improve model training
* I used to normalize the input data such as removing stop words, removing punctuation, and making everything lowercase
* I used the TfidfVectorizer from sklearn to build a basic machine learning model. I also made use of the cosine similarity function to measure the similarity between two documents.
* I implemented some simple rules-based approach in order to reply to possible user greetings
* I gave my bot the ability to make some basic conversation by reading in user input and spitting back some info about them.
* I performed POS tagging to better identify user information from user responses
* I implemented a default response statement in case the bot is not able to find any possible answer to the user’s question.

**Conversation Diagram Flow**

Diagram

Description automatically generated

**Sample Dialogue Interactions**

Text

Description automatically generated

Text

Description automatically generated

**Knowledge Base Appendix**

I mainly used different websites about tigers for my knowledge base. The websites are below:

* "https://en.wikipedia.org/wiki/Tiger",
* "https://www.britannica.com/animal/tiger",
* "https://kids.nationalgeographic.com/animals/mammals/facts/tiger",
* "https://www.worldwildlife.org/species/tiger",
* "https://www.livescience.com/27441-tigers.html",
* "https://a-z-animals.com/animals/tiger/",
* "https://nationalzoo.si.edu/animals/tiger",
* "https://www.theanimalfiles.com/mammals/carnivores/tiger.html",
* "https://awionline.org/content/tigers",
* "https://worldanimalfoundation.org/advocate/wild-animals/params/post/1297938/tigers",
* "https://www.animalfactsencyclopedia.com/Tiger-facts.html",
* "https://animalia.bio/tiger",
* <https://animals.fandom.com/wiki/Tiger>

These websites mainly contained various information about tigers. Some example text is below:

The tiger (Panthera tigris) is the largest living cat species and a member of the genus Panthera. It is most recognisable for its dark vertical stripes on orange fur with a white underside. An apex predator, it primarily preys on ungulates, such as deer and wild boar. It is territorial and generally a solitary but social predator, requiring large contiguous areas of habitat to support its requirements for prey and rearing of its offspring. Tiger cubs stay with their mother for about two years and then become independent, leaving their mother's home range to establish their own.

**User Models**

I stored different user information as variables that the bot could then use in conversation. In the examples above, you can see the bot makes use of the user’s name and also holds a basic conversation about their likes and dislikes

**Evaluation**

I was surprised at the effectiveness of my chatbot at talking about tigers. Considering that I used a relatively simple algorithm and only trained on about a dozen documents, I was not expecting my bot to be able to stay on topic at all. However, to my surprise it was generally able to provide relevant information about simple tiger fact queries. It was able to answer very basic questions such as “do tigers swim” and do “tigers eat birds”. However, it was often unable to provide useful information when asked more advanced scientific questions or questions that could have many answers such as “what are tigers”.

I think the approach I used could be helpful if the task is a simple information index. It works almost like a search engine, in which you can provide some basic information and get some facts that are of similar relevance. However, it works poorly for more complicated tasks such as holding an actual scientific discussion about tigers, or even to have a more advanced conversation.