a2_troll_reply_bot

February 3, 2024

1 Programming Assignment 2: Trolling a Reply Bot

In this assignment, your job will be to make modify the reply bot below and make your own version of a reply bot (look for the TODO section). Then after you've made it, see if you can troll it. At the end, you will have some reflection questions to answer.

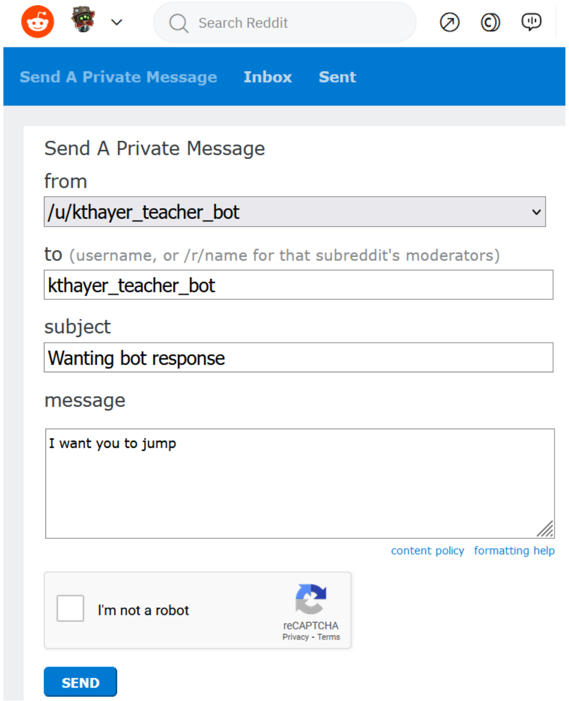
When you are done, you will need to download this file (file menu -> download) and turn it in on Canvas.

Below is the code for the first reply bot from the book, where if you message it: - Subject: "Wanting bot response", body: "I want you to ***" (where the*** is some action) - then the bot will reply,"I will now ______" (where the _____ is that same action).

1.1 Sending ourselves a message

In order to send ourselves a message we can reply to, go to: - https://www.reddit.com/message/compose/

Then compose a message to your own account with the subject: - Wanting bot response and a message body of something like: - I want you to jump



Send A Private Message

from: /u/kthayer_teacher_bot

to (username, or /r/name for that subreddit's moderators): kthayer_teacher_bot

subject: Wanting bot response

message: I want you to jump

1.2 Reply Bot

First we need to do our Reddit PRAW setup:

[3]:
```
import praw
```

(optional) make a fake praw connection with the fake_praw library

For testing purposes, we've added this line of code, which loads a fake version of praw, so it wont actually connect to reddit. **If you want to try to actually connect to reddit, don't run this line of code.**

[4]:
```
%run fake_apis/fake_praw.ipynb
```

```
<IPython.core.display.HTML object>
```

To use this on your real Reddit account, copy your developer access passwords into the reddit_keys.py file.

[5]:
```
import reddit_keys
```

[6]:
```
# Give the praw code your reddit account info so
# it can perform reddit actions
reddit = praw.Reddit(
    username=reddit_keys.username, password=reddit_keys.password,
    client_id=reddit_keys.client_id, client_secret=reddit_keys.client_secret,
    user_agent="a custom python script for learning for " + reddit_keys.username
)
```

```
<IPython.core.display.HTML object>
```

1.2.1 find my latest message

We need to find our latest message in our inbox

We do this by looking in our reddit inbox for messages (we limit it to one, since we just want the latest).

It doesn't directly give us the one message (instead it is in something called an "iterator"), but we can use the `next` function to get the message out.

We then display the subject of the message just so we can see that it found something.

[7]:
```
# Look up the subreddit "cuteanimals", then find the "hot" list, getting up to␣
 ↪1 submission
messages = reddit.inbox.messages(limit=1)

# get the first submission off the list (should only be one anyway)
latest_message = next(messages)

# display the subject and body of the message, so we can see what we found
display("latest message subject: " + str(latest_message.subject))
display("latest message body: " + str(latest_message.body))
```

```
'latest message subject: Wanting bot response'
```

```
'latest message body: I want you to jump'
```

1.3 TODO: Modify this section (If message matches our pattern, reply)

You must must modify at least one thing in the code below. You could change the expected patterns, or you could change the code that replies, or you could change both.

We will now see if the message matches our pattern of a message subject of "Wanting bot response" with a message body of "I want you to _____" and then we will reply.

First we will create strings with the patterns we are looking for and save them into variables.

[12]:

```
expected_subject = "Wanting bot Assistance"
expected_body_pattern = "Please help me with "
```

We will check if the message has the subject we are expecting. If it does it will check if the essage body starts with the expected pattern. If it does, then we will find the action from the end of the message body text (based on the expected_pattern length), and reply using that action.

We also add "else" cases for when we didn't match the patter, and display a message saying what didn't match.

[13]:

```
# check if the mention text starts with our set phrase
if latest_message.subject == expected_subject:

    if latest_message.body.startswith(expected_body_pattern):
        # get the action, which should be the end of the string after the␣
↪expected pattern
        action = latest_message.body[len(expected_body_pattern) :]

        # make a new message which says we will do the action
        message = "I will now " + action

        # send our message in reply
        latest_message.reply(message)

    else: # else code for if the message body didn't match
        display("The message body (" + latest_message.body + ") didn't match␣
↪our pattern (" + expected_body_pattern + ")")

else: # else code for if the message subject didn't match
    display("The message subject (" + latest_message.subject + ") didn't match␣
↪our expected subject (" + expected_subject + ")" )
```

```
"The message subject (Wanting bot response) didn't match our expected subject␣
↪(Wanting bot Assistance)"
```

1.4 Reflection questions

1. What changes did you make to the reply bot?

I changed the expected_subject from ‘Wanting bot response’ to ‘Wanting bot Assistance’ and the expected_body_pattern from ‘I want you to’ to ‘Please help me with’. If I wanted to change the expected subject and the expected pattern, I would need to modify the code in the loop, but I wasn’t asked to do that as far as I know. With my modifications to the code, the output from the if and else statements came back as ‘The message subject (Wanting bot response) didn’t match our expected subject (Wanting bot Assistance).’

2. How could you troll this bot? Give an example of a message that would let you troll it. Or, if you don’t think it can be trolled, explain why.

[17]:
```
#Trolling a bot seems impossible because bots aren't capable of expressing human emotions.
#We can only program the bot and tell it how to respond, and that isn't exactly 'trolling' the bot.
#I guess now that I think about it more, it's possible with the celebrity AI bot. Some content creators made videos trolling the AI into
#jumping off buildings and other things. I don't see it as trolling, though, because bots can't feel anything; it's harmless fun.
```

3. What limitations does trying to prevent trolling put on your ability to create a bot? (write at least 3 sentences)

I personally don’t think that there are any limitations when creating a bot if you want to prevent trolling. An example would be if you attempt to write something negative into ChatGPT; it will shut down. I even observe this with celebrity AI chatbots, which could, at worst, frighten them into doing the unthinkable. There are many types of useful bots out there that help run servers, provide customer service, and much more, to the point where I strongly believe that the possibilities are limitless.

4. Pick two ethics frameworks and compare how they might evaluate the responsibility of someone who is creating a reply bot? (write at least 6 sentences total)

For someone who wants to create a reply bot, their main goal would probably be the safety of their users, meaning they wouldn’t want the reply bot to say something harmful in response to potentially harmful messages. Most people trolling the reply bot are more than likely doing it for their own amusement.

Virtue Ethics: This framework believes in virtuous actions that will lead to someone acting more virtuously. It suggests that if someone is willing to be good, that’s training for them to continue being good. This framework is relevant when discussing someone who would want to create a reply bot because they are likely not willing to hurt others and might be curious about learning more about communication. This would be my own alternative motive, having an interest in social situations, especially those related to people on the spectrum. If people were to see what messages were sent to the bot, though, that might be a violation of user privacy, which probably wouldn’t be considered ethical.

Deontology: This framework believes that there are absolute moral rules and duties to follow. If someone wanted to create a robot that actively communicated with people, they’d probably think about programming something that upholds their basic moral code. In terms of user experience, though, this person might want to strike a balance between their own moral code and allowing certain things to be said to the reply bot. Basic moral rules imposed on people include not killing each other, not speaking badly to one another, and so on. From most reply bots that I've interacted with, these messages would either stop the bot or trigger an error message.

5. Pick two ethics frameworks and compare how they might evaluate the responsibility of someone who is considering trolling a reply bot? (write at least 6 sentences total)

Someone willing to trolling a reply bot would probably be doing it for entertainent.

Egoism: This framework believes in 'Rational Selfishness,' asserting that it is rational to prioritize one's own self-interest above all else. I believe that if someone wanted to engage in trolling, they are actively pursuing their own self-interest, as in most cases, it isn't done for research. However, in my case, I have trolled people just to observe their responses and learn from the interaction. This, too, aligns with my own self-interest because I'm certain that the person didn't know that I was trolling them solely to gain insights into communication.

Nihilism: This framework believes that there is no right or wrong; nothing matters. Some people who troll adopt this perspective and do not consider the person in front of them or behind the screen. They believe that individuals and their actions don't matter. Right and wrong are considered arbitrary constructs, and their engagement is purely for the sake of enjoyment.