SWAHILI CHATBOT IMPLEMENTATION

PROJECT REPORT

Amittai Siavava
Dartmouth College
amittai.j.wekesa.24@dartmouth.edu

January 12, 2023

ABSTRACT

This project simulates an eliza bot using regular expressions. It interprets user input in Swahili language and responds accordingly.

1 Supported Statements

All statements were are adapted from the original Spanish statements in the assignment.

1.1 Initial Greeting

Prompt		**Response**	
Jina langu ni Amittai.	My name is Amittai.	Shikamoo, Amittai.	Hello, Amittai. How have you been?
Naitwa Amittai.	I am called Amittai.	Umeshindaje?	

Table 1: Greetings and Responses.

1.2 State of Mind

Prompt		**Response**	
Nimefurahi.	I am happy.	*Mbona umefurahi?*	Why are you happy?
Mimi nimefurahi.	*I* am happy.		
Nimefurahi sana.	I am very happy.		
Sijafurahi.	I am not happy.	*Mbona hujafurahi?*	Why aren't you happy?
Sijafurahi sana.	I am not very happy.		
Mimi sijafurahi.	*I* am not happy.		
Nimehuzunika.	I am sad.	*Mbona umehuzunika?*	Why are you sad?
Mi nimehuzunika kidogo.	I'm a little bit sad.		
Nimehuzunika kidogo.	I am a little bit sad.		
Nimehuzunika kiasi.	I am quite sad.		
Nimehuzunika kidogo.	I am sort of sad.		

Table 2: State-of-Mind Statements.

1.3 Characteristics of a Person

Prompt		**Response**	
Mimi ni mchafu.	I am untidy.	*Mbona wewe ni mchafu?*	Why are you untidy?
Ni kwamba mimi ni mchafu sana.	It's just that I am fairly untidy.	*Mbona wewe ni mchafu sana?*	Why are you fairly untidy?
Mimi kawaida ni mtu mchafu.	I am usually dirty	*Mbona wewe kawaida ni mtu mchafu?*	Why are you usually dirty?
Mimi kawaida si mtu safi.	I am not happy.	*Mbona wewe kawaida si mtu safi?*	Why are you usually not a clean person?
Mimi ni mwanafunzi wa Dartmouth.	I am a student at Dartmouth.	*Mbona wewe ni mwanafunzi wa Dartmouth?*	Why are you a student at Dartmouth?

Table 3: Person Descriptions and Responses.

1.4 About Family

Note: I expanded the regex implementation here to also include *kakako* ("your brother") and *dadako* ("your sister").

Prompt		**Response**	
Mamangu anapenda waridi.	My mom likes roses.	*Niambie mengine kumhusu mamako.*	Tell me more about your mother.
Babangu hapendi kakati.	My dad does not like cacti.	*Niambie mengine kumhusu babako.*	Tell me more about your father.
Kwa sababu mamangu yu hodari na mimea.	Because my mother is good with plants.	*Niambie mengine kumhusu mamako.*	Tell me more about your mother.
Kwa sababu babangu anapendelea miti.	Because my father prefers trees.	*Niambie mengine kumhusu babako.*	Tell me more about your father.

Table 4: Family Descriptions.

1.5 Modal Verbs

Prompt		Response	
Nataka kujifunza kuandika msimbo.	I want to learn how to program.	*Mbona unataka kujifunza kuandika msimbo?*	Why do you want to learn how to program?
Sitaki kujifunza kuandika msimbo.	I don't want to learn how to program.	*Mbuna hutaki kujifunza kuandika msimbo?*	Why don't you want to learn how to program?
Nataka kuenda safari.	I want to go on a trip.	*Mbona unataka kuenda safari?*	Why do you want to go on a trip?
Sitaki kuenda safari.	I don't want to go on a trip.	*Mbona hutaki kuenda safari?*	Why don't you want to go on a trip?
Lazima niende kwenye sinema.	I must go to the movies.	*Mbona lazima uende kwenye sinema?*	Why must you go to the movies.
Lazima nihamie kwenye nyumba tofauti.	I must move to a different house.	*Mbona lazima uhamie kwenye nyumba tofauti?*	Why must move to a different house?
Naweza kutembea kwenye theluji.	I can walk on the snow.	*Mbona unaweza kutembea kwenye theluji?*	Why can you walk on the snow?
Naweza kula vidakuzi kumi.	I can eat ten cookies.	*Mbona unaweza kula vidakuzi kumi?*	Why can you eat ten cookies?
Rafiki yangu anazema naweza kula vidakuzi kumi.	My friend says I can eat ten cookies.	*Mbona unaweza kula vidakuzi kumi?*	Why can you eat ten cookies?

Table 5: Modal Verbs and Responses.

1.6 Thoughts and Hopes

Prompt		Response	
Nadhani rafiki zangu walifanya hivo pia.	I think my friends did that too.	*Mbona unadhani hivo?*	Why do you think so?
Natumai kuhitimu mwaka ujao.	I hope to graduate next year.	*Mbona unatumai hivo?*	Why do you hope so?

Table 6: Thoughts, Hopes, and Responses.

1.7 Asking for Examples

Prompt		Response	
Yeye husema hivo tu.	He always says that.	*Unaweza nipa mfano?*	Can you give me an example?
Kakangu hula vidakuzi.	My brother always eats cookies.	*Unaweza nipa mfano?*	Can you give me an example?

Table 7: Prompting for Examples.

1.8 Handling Insults

Prompt		Response	
Hapana, mshenzi!	No, stupid!	*Ha! Matusi hapana! Tulia na uniambie mengine.*	Hey! No insults! Relax and tell me more.
Wewe ni mjinga.	You are an idiot.		

Table 8: Handling Insults.

1.9 All Other Statements

Prompt		Response	
Ninapenda kusoma.	I like reading.	*Niambie mengine...*	Tell me more...
Mimi natazama runinga kwa zaidi.	I watch too much television.		

Table 9: Other Statements.

2 Discussion

This program is obviously very limited. After making the program, please answer the following question in the form of a comment at the end of the code: What changes would you make to the program so that it can carry out a more human-like conversation? Please make three suggestions for how to improve the program and find a section of the Jurafsky and Martin textbook that can help you implement those changes. Mention those sections explicitly. The answer should be at least 200 words long.

1. **Use a Better Model:** A main issue is that regular expressions do not have the capacity to fully understand nuances in language, such as emotionality, sarcasm, and other forms of non-literal speech. Regular expressions only account for the phrases they have been explicitly programmed to handle. A better model such as a neural network or especially a transformer would be able to understand the nuances in language better and generate more human-like responses. Transformer models and LSTM neural networks are particularly useful for language understanding and generation since they can understand the context of a sentence and the context of each word in a sentence. **Jurafsky and Martin** discuss such neural network models in sections 9 and 10.

   Aside: A transformer model would also be more computationally intensive. If there are significant computational constraints, that could be a reason to stick with regular expressions.

2. **Increase the range of responses:** The program, as is, is limited to a small set of responses to an equally small set of prompts. Increasing the repertoire of responses would make for a more human-like conversation. I think techniques such as *normalization* using *stemming* and *lemmatization* could be useful *before* matching regular expressions – so that related words (for example, "go" and "went") are matched by the same regular expression. **Jurafsky and Martin** discuss how to do such normalization in Section 2.4.

3. **Adding Variance:** Another shortfall, currently, is that the program *always* generates the same response for a given prompt. This is "OK" – but not human-like. Adding some noise in the decision-making process would be useful, such as by using a probabilistic model to choose a response *after matching a regular expression.* One way to achieve this could be to use a *Markov Chain*. **Jurafsky and Martin** do not discuss Markov Chains explicitly, but they discuss *Hidden Markov Models* in Section 8.4.