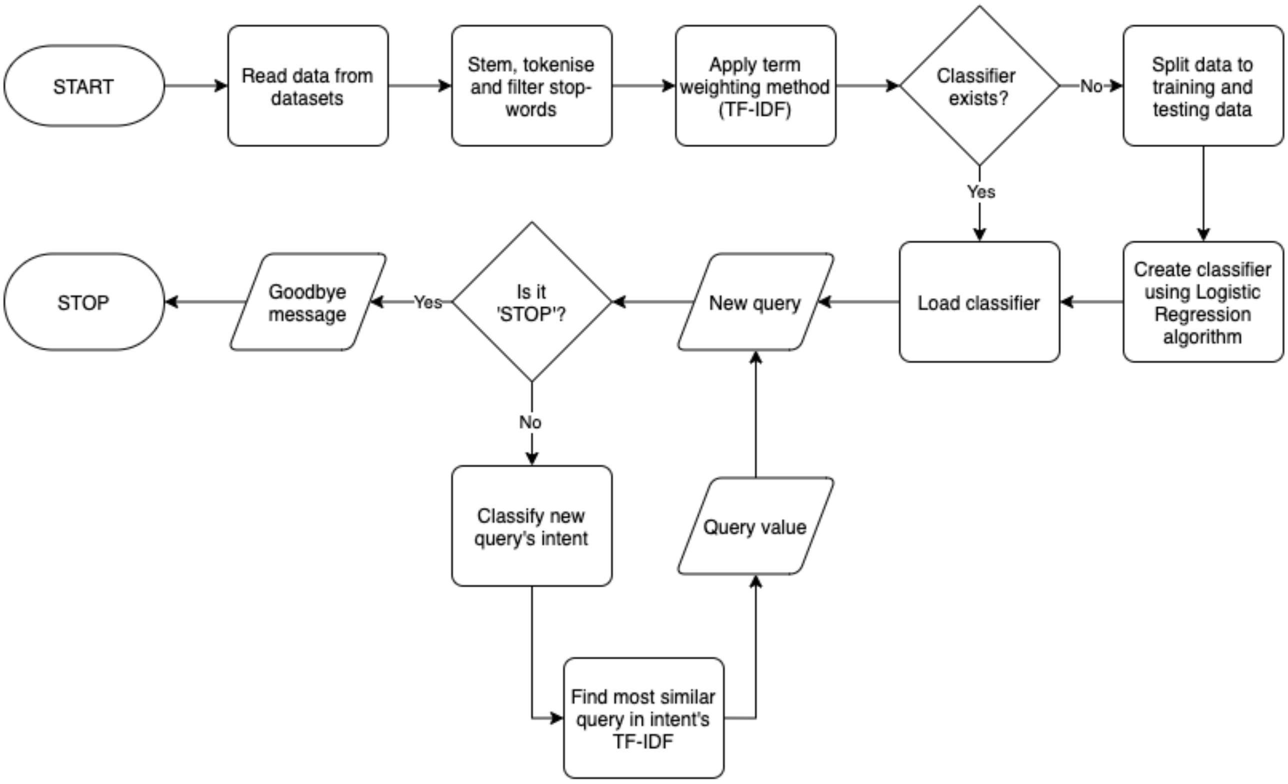
Conversational Chatbot Report

Human-AI Interaction (COMP3074 UNUK)

Luqman Hakim Bin Ariffin (14326114)

Word Count: 1,422

Design of the system


START
Read data from datasets
Stem, tokenise and filter stop-words
Apply term weighting method (TF-IDF)
Classifier exists?
No
Split data to training and testing data
Yes
Create classifier using Logistic Regression algorithm
Load classifier
New query
Is it 'STOP'?
Yes
Goodbye message
STOP
No
Classify new query's intent
Find most similar query in intent's TF-IDF
Query value

Order of processes once chatbot script begins

Before the interaction with the user begins, the dataset for each intent are read into separate dictionaries, where the 'key' is the potential query and the 'value' holds a list of possible replies for the chatbot. These replies will be randomly chosen for the user. There is no dataset for Intent 1 (Name), only for Intent 2 (Small talk) and Intent 3 (Question answering).

Each dictionary's keys are stemmed, tokenised and have English stop-words removed using a CountVectoriser. Stemming uses Snowball/Porter2 stemming algorithm. It is recognised as a better solution than the Porter stemmer, and is more reliable than the Lancaster stemmer. Lemmatisation was initially used, however, it resulted in less accurate results. It is likely due to the context of being a chatbot. As the queries are shorter and simpler by nature than longer texts, simplifying the tokens further made it more ambiguous when finding the similarity between queries.

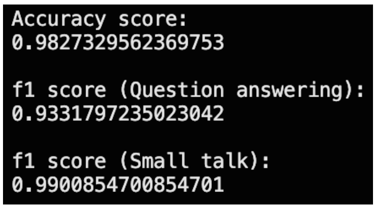
The output of the CountVectoriser is then applied to a TfidfTransformer, creating the TF-IDF for each intent. The method has the parameters 'use_idf', 'smooth_idf', and 'sublinear_idf' set as true.

Interaction begins once a classifier is loaded. The diagram shows the general pattern of communication, however, the chatbot will begin the conversation by first prompting the user to enter their name. The user will also be informed to type 'STOP' in order to properly stop the chatbot. Afterwards, the user will be in a loop, where they enter a query and will be given the most appropriate response available, until they enter 'STOP', which will trigger a goodbye message before quitting itself.

Intent matching

User queries are classified first to know which dataset to compare with. The script will first try loading in a classifier within the same folder. If not found, it will create a new classifier and save it for future cases the chatbot is run. This slightly decreases performance load on future startups.

The classifier is trained using the Logistic Regression algorithm. The test size is set to 30%, much lower or higher led to lower accuracy scores. This was tested in intervals of 5%. Logistic regression was initially used and, since the results were favourable, no other classification algorithm was tested. Since other classification algorithms were not tested, there may possibly be a better solution. However, in comparison to the achieved results, improvements will likely be marginal.



```
Accuracy score:
0.9827329562369753

f1 score (Question answering):
0.9331797235023042

f1 score (Small talk):
0.9900854700854701
```

Classifier's accuracy scores

Name management

As stated, this intent does not have a dataset and the chatbot begins with prompting the user for their name. This intent is recognised by checking if the user query contains any substring from a list of substrings. Within this intent, there is the intent to retrieve the name and to change the name, each having their own list of substrings associated.

This intent's approach is more primitive, as it simply recognises substrings and has one reply for each sub-intent. Also, it does not recognise if the user provides a name when requesting a 'changing name' intent. It will simply re-prompt the user for a name. The likelihood to enter this intent is also much stricter, as the substring to match is generally a phrase and the list is quite short. Therefore, it is unlikely to hinder the performance of classifying the other intents.

The name itself is stored in a variable.

Small talk and Question answering

The TF-IDF which the query will be compared to depends on which intent was classified, however, the process is the same. The user query will be compared to each query in the intent's TF-IDF using cosine similarity. If the similarity result for a query does not meet a certain threshold, this is set to 0.5, then it will not be considered at all. Once all queries have been checked, a list is made which contains all queries which met the threshold. If there are no queries in the list, a general response is shown to indicate that there is no related data associated. If there are queries, the list will be sorted and the one with the highest similarity index will be used. The query will be used as a key to access the value. Some keys have multiple responses, in which case, a response will be randomly chosen.

For the question answering dataset, multiple responses were collated when the query matched an existing dictionary key. In this case, the response would be appended to the value list. If no dictionary key existed, it would initialise the value to contain an empty list which would then be appended.

The small talk dataset was externally sourced. There were multiple versions of the dataset. To create multiple responses, a document was made which contained three versions of responses for each query. As it is all contained in one document, the performance load stays the same. The link to the source: https://github.com/Microsoft/BotBuilder-PersonalityChat/tree/master/CSharp/Datasets

Evaluation - test queries

The following table shows the results of the given test queries, where the chatbot would have picked a potential response from the highest similarity index returned. The threshold set for this test was 0.5, thus, no results less than 0.5 are shown. The test query “What’s my name?” is not included in the table as its similarity is binary, such that it either exists or does not exist in the substring array. If it exists, it returns the name given by the user.

Test query	Matched query in dataset	Similarity index	Potential responses
Hello, how are you?	Hello	0.761667514655298	[‘Hello there!’, ‘Hi!’, ‘Hello.’]
	How are you?	0.647968052544894	I'm great, thanks for asking!
			Awesome! Thanks for asking.
			Great, thanks.
How many people play in a Hockey team?	How many professional hockey teams in canada	0.610364955333909	It started with four teams and, through a series of expansions, contractions, and relocations, the league is now composed of 30 active franchises.
How many people live in Atlantis	How many people live in atlanta georgia	0.683773900216148	Atlanta (, stressed , locally) is the capital of and the most populous city in the U.S. state of Georgia , with an estimated 2011 population of 432,427.
			Atlanta is the cultural and economic center of the Atlanta metropolitan area , home to 5,457,831 people and the ninth largest metropolitan area in the United States.
	How many people live in memphis tennessee	0.673467278186622	Memphis had a population of 672,277 in 2011 making it the largest city in the state of Tennessee , the largest city on the Mississippi River , the third largest in the Southeastern United States , and the 20th largest in the United States.
What are the big ten?	How many schools are in the big ten	0.759320939282834	Its twelve member institutions (which are primarily flagship research universities in their respective states, well-regarded academically, and with relatively large student enrollment) are located primarily in the Midwest , stretching from Nebraska in the west to Penn State in the east.

Test query	Matched query in dataset	Similarity index	Potential responses
What is single malt scotch?	How is single malt scotch made	0.909844003906947	Single Malt Scotch is single malt whisky made in Scotland using a pot still distillation process at a single distillery , with malted barley as the only grain ingredient.
			As with any Scotch whisky , a single malt Scotch must be distilled in Scotland and matured in oak casks in Scotland for at least three years (most single malts are matured longer).
Who is Isaac Newton	What did isaac newton do	0.822379378404327	Sir Isaac Newton (25 December 164220 March 1727) was an English physicist and mathematician who is widely regarded as one of the most influential scientists of all time and as a key figure in the scientific revolution .
	What year did isaac newton die	0.73741264104234	
What is the best Dim Sum?	How does a dim sum restaurant work	0.541146941271735	Dim sum () refers to a style of Cantonese food prepared as small bite-sized or individual portions of food traditionally served in small steamer baskets or on small plates.
			Dim sum is also well known for the unique way it is served in some restaurants, wherein fully cooked and ready-to-serve dim sum dishes are carted around the restaurant for customers to choose their orders while seated at their tables.
			Eating dim sum at a restaurant is usually known in Cantonese as going to "drink tea" (yum cha , 飲茶), as tea is typically served with dim sum.
What is mustard made of?	What is the ingredient in mustard	0.568297295035335	Mustard (or yellow sauce) is a condiment made from the seeds of a mustard plant (white or yellow mustard, Sinapis hirta ; brown or Indian mustard, Brassica juncea ; or black mustard, B. nigra).
What is 奶茶?	N/A	N/A	I'm sorry, I don't know how to answer... Anything else?

Table contains approximately 500 words (not included in the mentioned word count)

Functionality

The chatbot functioned as expected for most of the queries. Certain queries returned an answer even though it was slightly irrelevant, such as the dim sum and hockey questions. The result was due to a shared topic, however, the answer did not match the question. In these cases, increasing the threshold would increase its functionality.

However, errors did occur. In ‘Hello, how are you?’, the chatbot returned the incorrect result. This is due to having two separate existing queries put together. This was tested to be the reason by inserting the mixed query into the dataset, in which it functioned properly. ‘How many people live in Atlantis?’, also returned a false result due to its high similarity to existing queries. This could also be solved by increasing the threshold to disregard these answers, however, this could lead to the threshold being too high for other queries and would result in the chatbot returning less answers from the dataset.

Performance

The chatbot does not produce the same answers once typos are introduced. As the similarity largely relies on token-matching, if the tokens differ then the similarity decreases. As an example, if instead the query was ‘Helo, how are you?’, the chatbot would return answers to ‘How are you?’, as it no longer recognises ‘helo’. Grammatical errors do not have as large an effect, however, increasingly deliberate mistakes decrease its accuracy in reading similarity.

As a side effect, if the threshold were raised higher to increase functionality, it will likely hold a worse performance.

Affect

The small talk is basic. The chatbot relies on the user continuing the conversation as it can only reply. Its answers do not ask the user questions in return. Thus, its answers are natural, however, they are not conversational.

Shortcomings

The classifier’s accuracy and f1 scores are likely inflated due to lack of variation between the train and test data. These scores give the impression that the chatbot is more reliable than it truly is when tested. This is seen in the following example. If the user entered ‘who are the Jonas brothers’, it would correctly classify as question answering, and reply with the correct answer. However, if it were entered slightly different as ‘who are Jonas brothers’, it would classify this as small talk and thus return an inappropriate answer. This could possibly be improved by introducing a much wider dataset within small talk and question answering. Currently, the datasets share similar sentence structure and, since the dataset size is very limited, it introduces vagueness between similar new queries.

Also, the chatbot is overly basic due to being template-based and relying on set answers for each query. If a new query were similar to an existing, however, requires a different answer then it would either return an inappropriate answer or answer that it does not know. This should only be solved once classification has improved, as it should not affect the question answering intent. The solution could be to switch to a dynamic generation when it is small talk, as these interactions are generally for entertainment and less need to be fully accurate sentences. As it is high noise, it is possibly important to inform the user of its error rates.