Names: Ghaida Alshiddi & Bushra Rahman

Class: CS 4395.001

Assignment: Chatbot

April 15, 2023

Chatbot Report and Evaluation

System Description

This project is a chatbot coded from scratch in Python, without the use of any external frameworks. The chatbot is designed to hold conversations on the topic of sports, mainly serving as an encyclopedia or fact dispenser for various sports terms. First the chatbot reads in a knowledge base, which is given in the form of a text file containing individual terms and a lengthy corresponding entry of sentences that are facts about the term. This knowledge base is read into a dictionary, where each key is a term and each corresponding value is a list of the sentences elaborating on that term. The chatbot uses this dictionary as its primary source for generating responses, along with a number of built-in lists for handling various canned interactions, such as greetings and goodbyes. The user's input is stored in a user model file for reference.

NLP Techniques Description

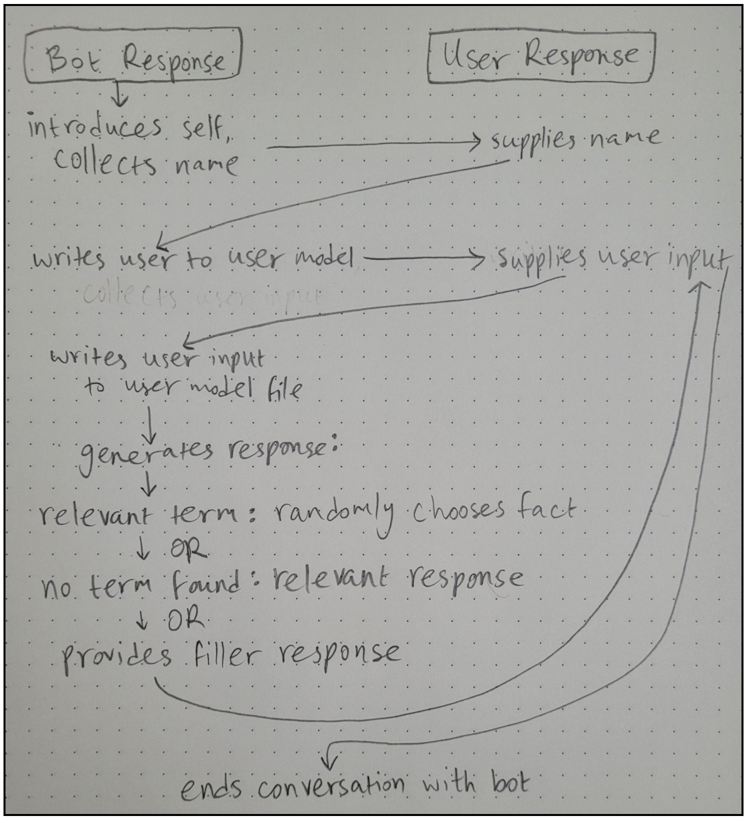
This project uses multiple NLP techniques including: user input parsing, knowledge base text parsing, term frequency measurements using vector space modeling, and cosine similarity measurements.
After parsing the knowledge base, the chatbot would store user information in the user_model.txt. The program is able to parse the knowledge base using some NLP and NLTK functionalities such as word_tokenizer and sent_tokenizer. The program used the word and sentence tokenizers to create a knowledge base dictionary that that program would be able to use later to respond to users. Then the chatbot would parse the user input to retrieve and extract important information or terms, and the program

would also use some NLTK functionalities to tokenize and extract information. When extracting or processing user input, the program needs to lower-case the text, then make sure that the text only contains words and remove all stopwords. Then the chatbot can use the extracted and preprocessed information and the knowledge base to get the term frequency, which will be used to model vector representations of the sentences in the knowledge base and ultimately get an index from cosine similarity to the user's input. To get the term frequency, the program needs to create vector representations for the user input and each sentence in the knowledge base. Then the term frequency values are going to be used to calculate the cosine similarity, then the program would get the index of the highest cosine similarity value. To calculate the cosine similarity, the program is going to use some functionalities from the numpy library to accurate cosine similarity values. Then the program would index the parsed knowledge base dictionary using the index retrieved from the cosine similarity function to get the best response for the user. Therefore, the chatbot is able to provide the user with the best response using the cosine similarity measurement.

In short, the program is using four major NLP functionalities: parsing and tokenizing user input and knowledge base file, getting the term frequency of the parsed tokens, and lastly calculating the cosine similarity using the term frequency results.

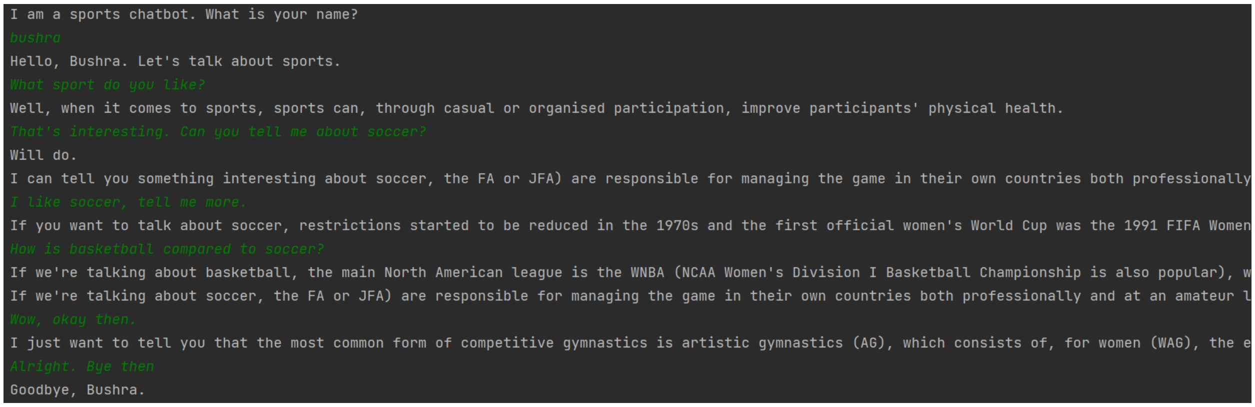
Diagram of Dialog Logic

The following diagram explains the program's logic and flow of control:

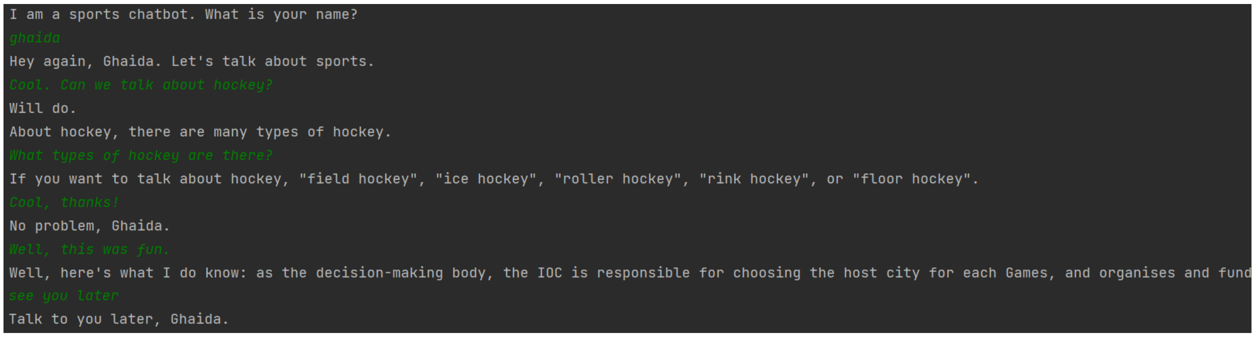

Bot Response
User Response
introduces self, collects name
supplies name
writes user to user model
supplies user input
collects user input
writes user input to user model file
generates response:
relevant term: randomly chooses fact
OR
no term found: relevant response
OR
provides filler response
ends conversation with bot

Sample Dialog Interactions

Below are some sample conversations that a user may hypothetically have with the chatbot. These conversations showcase both what the bot is and isn’t capable of. The bot is capable of recognizing action statements starting with ‘can’, grateful sentiments containing ‘thanks’, and different terms in the knowledge base. In the base case scenario, the bot uses cosine similarity to output a sentence that has any degree of similarity to what the user input.



```
I am a sports chatbot. What is your name?
bushra
Hello, Bushra. Let's talk about sports.
What sport do you like?
Well, when it comes to sports, sports can, through casual or organised participation, improve participants' physical health.
That's interesting. Can you tell me about soccer?
Will do.
I can tell you something interesting about soccer, the FA or JFA) are responsible for managing the game in their own countries both professionally
I like soccer, tell me more.
If you want to talk about soccer, restrictions started to be reduced in the 1970s and the first official women's World Cup was the 1991 FIFA Women
How is basketball compared to soccer?
If we're talking about basketball, the main North American league is the WNBA (NCAA Women's Division I Basketball Championship is also popular), w
If we're talking about soccer, the FA or JFA) are responsible for managing the game in their own countries both professionally and at an amateur l
Wow, okay then.
I just want to tell you that the most common form of competitive gymnastics is artistic gymnastics (AG), which consists of, for women (WAG), the e
Alright. Bye then
Goodbye, Bushra.
```

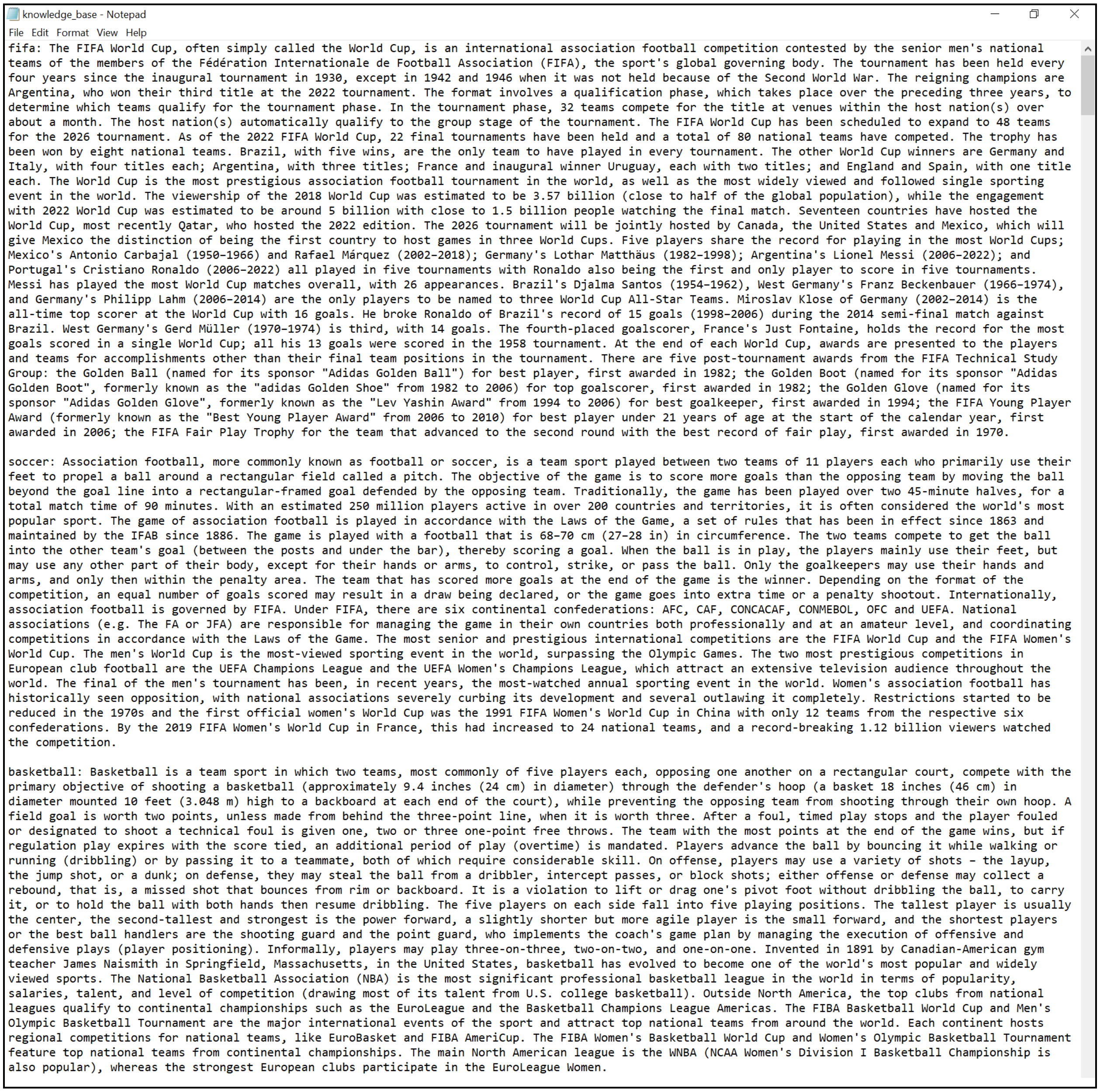


```
I am a sports chatbot. What is your name?
ghaida
Hey again, Ghaida. Let's talk about sports.
Cool. Can we talk about hockey?
Will do.
About hockey, there are many types of hockey.
What types of hockey are there?
If you want to talk about hockey, "field hockey", "ice hockey", "roller hockey", "rink hockey", or "floor hockey".
Cool, thanks!
No problem, Ghaida.
Well, this was fun.
Well, here's what I do know: as the decision-making body, the IOC is responsible for choosing the host city for each Games, and organises and fund
see you later
Talk to you later, Ghaida.
```

Appendix: Knowledge Base

The knowledge base is a text file called `knowledge_base.txt`. It was created by manually scraping text off of Wikipedia articles on each included term. The knowledge base therefore functions like an encyclopedia, making this chatbot best suited for regurgitating relevant facts as per the user's request.

The knowledge base contains the following sports terms: fifa, soccer, basketball, cricket, football, sports, olympics, tennis, esports, super bowl, baseball, hockey, rugby, swimming, gymnastics. Below are some sample entries from the knowledge base. Hence, the knowledge base is mainly about sports and athletics.



fifa: The FIFA World Cup, often simply called the World Cup, is an international association football competition contested by the senior men's national teams of the members of the Fédération Internationale de Football Association (FIFA), the sport's global governing body. The tournament has been held every four years since the inaugural tournament in 1930, except in 1942 and 1946 when it was not held because of the Second World War. The reigning champions are Argentina, who won their third title at the 2022 tournament. The format involves a qualification phase, which takes place over the preceding three years, to determine which teams qualify for the tournament phase. In the tournament phase, 32 teams compete for the title at venues within the host nation(s) over about a month. The host nation(s) automatically qualify to the group stage of the tournament. The FIFA World Cup has been scheduled to expand to 48 teams for the 2026 tournament. As of the 2022 FIFA World Cup, 22 final tournaments have been held and a total of 80 national teams have competed. The trophy has been won by eight national teams. Brazil, with five wins, are the only team to have played in every tournament. The other World Cup winners are Germany and Italy, with four titles each; Argentina, with three titles; France and inaugural winner Uruguay, each with two titles; and England and Spain, with one title each. The World Cup is the most prestigious association football tournament in the world, as well as the most widely viewed and followed single sporting event in the world. The viewership of the 2018 World Cup was estimated to be 3.57 billion (close to half of the global population), while the engagement with 2022 World Cup was estimated to be around 5 billion with close to 1.5 billion people watching the final match. Seventeen countries have hosted the World Cup, most recently Qatar, who hosted the 2022 edition. The 2026 tournament will be jointly hosted by Canada, the United States and Mexico, which will give Mexico the distinction of being the first country to host games in three World Cups. Five players share the record for playing in the most World Cups; Mexico's Antonio Carbajal (1950–1966) and Rafael Márquez (2002–2018); Germany's Lothar Matthäus (1982–1998); Argentina's Lionel Messi (2006–2022); and Portugal's Cristiano Ronaldo (2006–2022) all played in five tournaments with Ronaldo also being the first and only player to score in five tournaments. Messi has played the most World Cup matches overall, with 26 appearances. Brazil's Djalma Santos (1954–1962), West Germany's Franz Beckenbauer (1966–1974), and Germany's Philipp Lahm (2006–2014) are the only players to be named to three World Cup All-Star Teams. Miroslav Klose of Germany (2002–2014) is the all-time top scorer at the World Cup with 16 goals. He broke Ronaldo of Brazil's record of 15 goals (1998–2006) during the 2014 semi-final match against Brazil. West Germany's Gerd Müller (1970–1974) is third, with 14 goals. The fourth-placed goalscorer, France's Just Fontaine, holds the record for the most goals scored in a single World Cup; all his 13 goals were scored in the 1958 tournament. At the end of each World Cup, awards are presented to the players and teams for accomplishments other than their final team positions in the tournament. There are five post-tournament awards from the FIFA Technical Study Group: the Golden Ball (named for its sponsor "Adidas Golden Ball") for best player, first awarded in 1982; the Golden Boot (named for its sponsor "Adidas Golden Boot", formerly known as the "adidas Golden Shoe" from 1982 to 2006) for top goalscorer, first awarded in 1982; the Golden Glove (named for its sponsor "Adidas Golden Glove", formerly known as the "Lev Yashin Award" from 1994 to 2006) for best goalkeeper, first awarded in 1994; the FIFA Young Player Award (formerly known as the "Best Young Player Award" from 2006 to 2010) for best player under 21 years of age at the start of the calendar year, first awarded in 2006; the FIFA Fair Play Trophy for the team that advanced to the second round with the best record of fair play, first awarded in 1970.

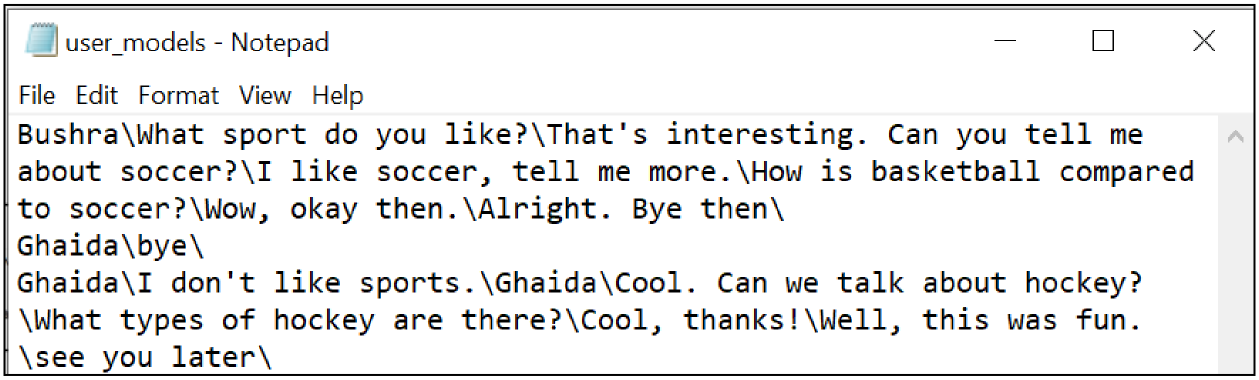
soccer: Association football, more commonly known as football or soccer, is a team sport played between two teams of 11 players each who primarily use their feet to propel a ball around a rectangular field called a pitch. The objective of the game is to score more goals than the opposing team by moving the ball beyond the goal line into a rectangular-framed goal defended by the opposing team. Traditionally, the game has been played over two 45-minute halves, for a total match time of 90 minutes. With an estimated 250 million players active in over 200 countries and territories, it is often considered the world's most popular sport. The game of association football is played in accordance with the Laws of the Game, a set of rules that has been in effect since 1863 and maintained by the IFAB since 1886. The game is played with a football that is 68–70 cm (27–28 in) in circumference. The two teams compete to get the ball into the other team's goal (between the posts and under the bar), thereby scoring a goal. When the ball is in play, the players mainly use their feet, but may use any other part of their body, except for their hands or arms, to control, strike, or pass the ball. Only the goalkeepers may use their hands and arms, and only then within the penalty area. The team that has scored more goals at the end of the game is the winner. Depending on the format of the competition, an equal number of goals scored may result in a draw being declared, or the game goes into extra time or a penalty shootout. Internationally, association football is governed by FIFA. Under FIFA, there are six continental confederations: AFC, CAF, CONCACAF, CONMEBOL, OFC and UEFA. National associations (e.g. The FA or JFA) are responsible for managing the game in their own countries both professionally and at an amateur level, and coordinating competitions in accordance with the Laws of the Game. The most senior and prestigious international competitions are the FIFA World Cup and the FIFA Women's World Cup. The men's World Cup is the most-viewed sporting event in the world, surpassing the Olympic Games. The two most prestigious competitions in European club football are the UEFA Champions League and the UEFA Women's Champions League, which attract an extensive television audience throughout the world. The final of the men's tournament has been, in recent years, the most-watched annual sporting event in the world. Women's association football has historically seen opposition, with national associations severely curbing its development and several outlawing it completely. Restrictions started to be reduced in the 1970s and the first official women's World Cup was the 1991 FIFA Women's World Cup in China with only 12 teams from the respective six confederations. By the 2019 FIFA Women's World Cup in France, this had increased to 24 national teams, and a record-breaking 1.12 billion viewers watched the competition.

basketball: Basketball is a team sport in which two teams, most commonly of five players each, opposing one another on a rectangular court, compete with the primary objective of shooting a basketball (approximately 9.4 inches (24 cm) in diameter) through the defender's hoop (a basket 18 inches (46 cm) in diameter mounted 10 feet (3.048 m) high to a backboard at each end of the court), while preventing the opposing team from shooting through their own hoop. A field goal is worth two points, unless made from behind the three-point line, when it is worth three. After a foul, timed play stops and the player fouled or designated to shoot a technical foul is given one, two or three one-point free throws. The team with the most points at the end of the game wins, but if regulation play expires with the score tied, an additional period of play (overtime) is mandated. Players advance the ball by bouncing it while walking or running (dribbling) or by passing it to a teammate, both of which require considerable skill. On offense, players may use a variety of shots – the layup, the jump shot, or a dunk; on defense, they may steal the ball from a dribbler, intercept passes, or block shots; either offense or defense may collect a rebound, that is, a missed shot that bounces from rim or backboard. It is a violation to lift or drag one's pivot foot without dribbling the ball, to carry it, or to hold the ball with both hands then resume dribbling. The five players on each side fall into five playing positions. The tallest player is usually the center, the second-tallest and strongest is the power forward, a slightly shorter but more agile player is the small forward, and the shortest players or the best ball handlers are the shooting guard and the point guard, who implements the coach's game plan by managing the execution of offensive and defensive plays (player positioning). Informally, players may play three-on-three, two-on-two, and one-on-one. Invented in 1891 by Canadian-American gym teacher James Naismith in Springfield, Massachusetts, in the United States, basketball has evolved to become one of the world's most popular and widely viewed sports. The National Basketball Association (NBA) is the most significant professional basketball league in the world in terms of popularity, salaries, talent, and level of competition (drawing most of its talent from U.S. college basketball). Outside North America, the top clubs from national leagues qualify to continental championships such as the EuroLeague and the Basketball Champions League Americas. The FIBA Basketball World Cup and Men's Olympic Basketball Tournament are the major international events of the sport and attract top national teams from around the world. Each continent hosts regional competitions for national teams, like EuroBasket and FIBA AmeriCup. The FIBA Women's Basketball World Cup and Women's Olympic Basketball Tournament feature top national teams from continental championships. The main North American league is the WNBA (NCAA Women's Division I Basketball Championship is also popular), whereas the strongest European clubs participate in the EuroLeague Women.

Appendix: User Models

User models were written to a text file called `user_models.txt`. Each user's name was written to a newline, followed by every input response they provided until they quit the conversation. Input responses were separated by a backslash delimiter character. The chatbot is capable of remembering user's names and greeting them differently depending on whether they are returning users or new users.

Below are some sample entries from the user model.



```
Bushra\What sport do you like?\That's interesting. Can you tell me
about soccer?\I like soccer, tell me more.\How is basketball compared
to soccer?\Wow, okay then.\Alright. Bye then\
Ghaida\bye\
Ghaida\I don't like sports.\Ghaida\Cool. Can we talk about hockey?
\What types of hockey are there?\Cool, thanks!\Well, this was fun.
\see you later\
```

Evaluation of the Chatbot

This chatbot combines naive versions of rules-based approaches and NLP techniques to language generation, overall resulting in a naive chatbot. The chatbot is capable of holding a limited conversation with the user so long as user input remains strongly within the confines of the chatbot’s knowledge base. The NLP techniques incorporated into the chatbot allow it some form of learning responses, but not to a great degree. However, the chatbot performs surprisingly well when treated simply as a fact-retrieving tool for the topic of its domain.