CodeBot
Ethan Owusu
34564104

1.0 Overview

The goal of CodeBot was to create a simple chat agent, capable of functioning as an assistant or study aid to first-year computer science students. Information, recognized examples of speech, and key words are stored in text files, while java files perform computation necessary to simulate conversation.

GitHub repository: https://github.com/eowusu/COSC310Final.git

1.1Classes

Classes included and their functions:

- Codebot.java is the class that handles all interactions with the user. All other classes are used by this class. This class gathers information from the other classes and uses that information to attempt to answer user questions. This class also handles regular parts of conversation such as greetings. If the system is unable to answer a user question then this class will perform an internet search for the topic and show the user the resulting information if it is relevant.

- Populate.java is the class that reads text files and stores them as either an arraylist of words (greetings.txt, prompts.txt, verbs.txt, etc), or a hashmap with multiple words as keys, and an explanation of the key as the value (topics.txt, details.txt, etc). This allows us to store a lot of information in text files without needing to search through it all every time that a user enters information to the system, since it will be in memory when the program starts up.

- Punctuation.java formats each response to correspond with the way we store words in our libraries. That way we can properly search through them. Specifically, this class puts extra spaces around words that are near punctuation as we do no store punctuation in out libraries.

- Comparison.java searches through the libraries to determine if a term is contained in the given library. The libraries are defined in the text files and correspond to different parts of conversation, such as greetings, and different topics such as arrays.

- Matcher.java is the class that checks for possible spelling mistakes and replaces words that aren't understood by the program with words that are more likely to be interpreted correctly. I.e. it would change 'varaibles' into 'variables' (a simple spelling error one might make).

- Tag.java implements Stanford’s parts-of-speech tagging library, to take input in the form of text strings representing sentences, and delivers output in the form of arrays containing symbols representing the part of speech that each word of the original sentence was an instance of.

- Winui.java is the class responsible for building the graphical user interface, that is used by the Codebot class.

- CBnoUI.java is a version of Codebot.java, that does not use the winui.java GUI. This version is used for the networked sockets.

- Server.java and Client.java build Sockets, and ServerSockets making the program network capable, and allowing the chat agent to speak with itself, with another automated agent, or with a user in a remote location.

- ServerRun.java and ClientRun.java are the classes responsible for building instances of the Server.java and Client.java classes and simulating a conversation between two instances of the chat agent.

- Wolf.java is a class that has implemented Wolfram Alpha's developer API and is responsible for properly constructing, executing, and returning information from search queries when information can not be found in one of the local banks of knowledge.

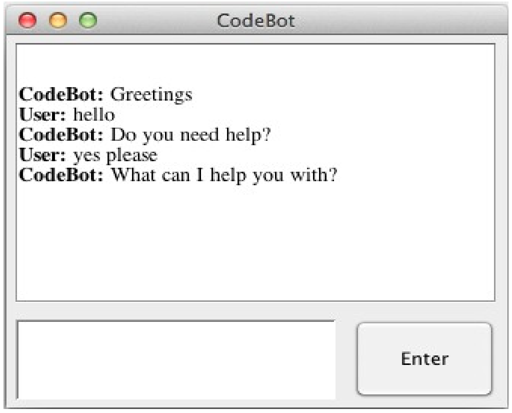
- Tweeter.java is a class that given a Twitter App's authoriation information, returns strings containing random tweets from a given Twitter handle

- WikiFinder.java is a class that returns parsed out string data from WikiPedia, when information isnt available locally

2.0Additions

Recently implemented features and improvements include:

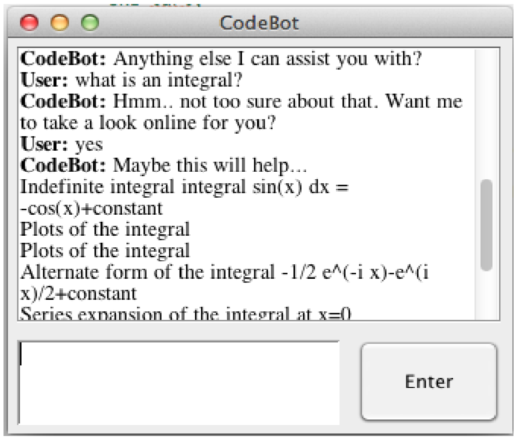
- The addition of a GUI (pictured below), making the agent easier to use, and providing easy viewing of the conversation history.



CodeBot: Greetings
User: hello
CodeBot: Do you need help?
User: yes please
CodeBot: What can I help you with?

Figure 2.0.1 CodeBot GUI

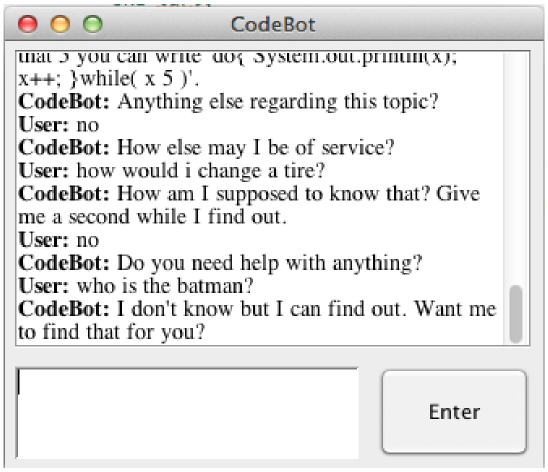
- Web based search capabilities have been added, allowing the agent to provide assistance on other topics related to science and mathematics. A user can now ask for information regarding our many topics, stored in files locally (many of which have been updated or added too) and receive answers immediately, or make requests regarding more peripheral topics, not directly related to computer science, and only wait a few seconds while Codebot retrieves relevant information from the net. A case of this is pictured below.



CodeBot: Anything else I can assist you with?
User: what is an integral?
CodeBot: Hmm.. not too sure about that. Want me to take a look online for you?
User: yes
CodeBot: Maybe this will help…
Indefinite integral integral sin(x) dx = -cos(x)+constant
Plots of the integral
Plots of the integral
Alternate form of the integral -1/2 e^(-i x)-e^(i x)/2+constant
Series expansion of the integral at x=0

Figure 2.0.2 CodeBot retrieves information from the internet

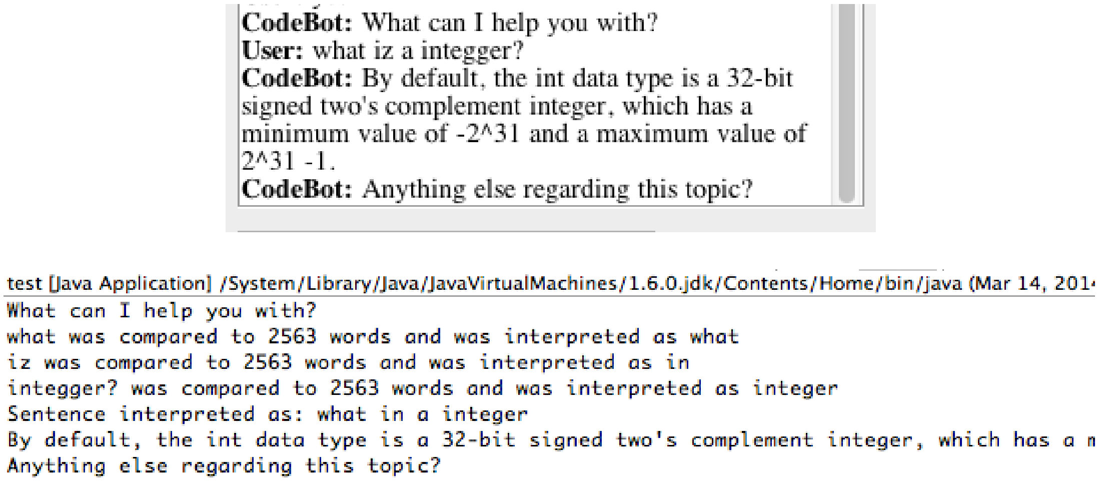
- When a topic is brought up that Codebot has no locally stored information regarding, one of 5 responses is selected at random, in an attempt to resemble the confusion humans are prone to, rather than simply repeating the same phrase ad nauseam.



that 5 you can write 'do{ System.out.println(x);
x++; }while(x 5)'.
CodeBot: Anything else regarding this topic?
User: no
CodeBot: How else may I be of service?
User: how would i change a tire?
CodeBot: How am I supposed to know that? Give me a second while I find out.
User: no
CodeBot: Do you need help with anything?
User: who is the batman?
CodeBot: I don't know but I can find out. Want me to find that for you?

Figure 2.0.3 CodeBot demonstrates various when faced questions outside of it's realm of expertise

- Spelling mistakes and plurality of words is handled through an implementation of the Levenshtein distance string metric. Due to hardware and runtime constraints, I have limited the system to only correcting spelling on words less than 8 characters in length, but if enough time is available, it will work to recognize strings of any size. The matcher class uses this to determine what an input sentence should be best interpreted as, changing phrases like “i can haz cheezbuggers.” to “I can has cheeseburger”, which although still grammatically incorrect, provides a better chance to help the user than the initial input did. These interpretations of input data are printed to the console.

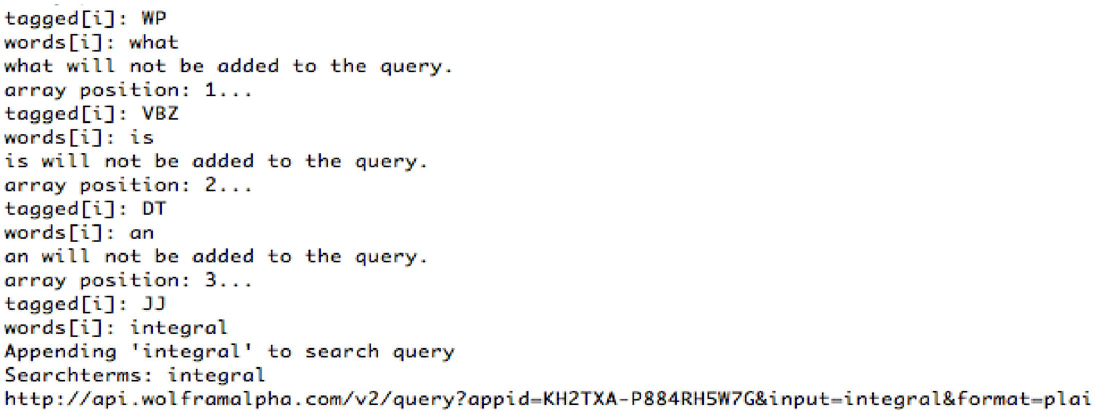


CodeBot: What can I help you with?
User: what iz a integger?
CodeBot: By default, the int data type is a 32-bit signed two's complement integer, which has a minimum value of -2^31 and a maximum value of 2^31 -1.
CodeBot: Anything else regarding this topic?

```
test [Java Application] /System/Library/Java/JavaVirtualMachines/1.6.0.jdk/Contents/Home/bin/java (Mar 14, 201
What can I help you with?
what was compared to 2563 words and was interpreted as what
iz was compared to 2563 words and was interpreted as in
integger? was compared to 2563 words and was interpreted as integer
Sentence interpreted as: what in a integer
By default, the int data type is a 32-bit signed two's complement integer, which has a m
Anything else regarding this topic?
```

Figure 2.0.4 CodeBot receives misspelled input and interprets it as what the Matcher class determines, is a more likely sequence of words

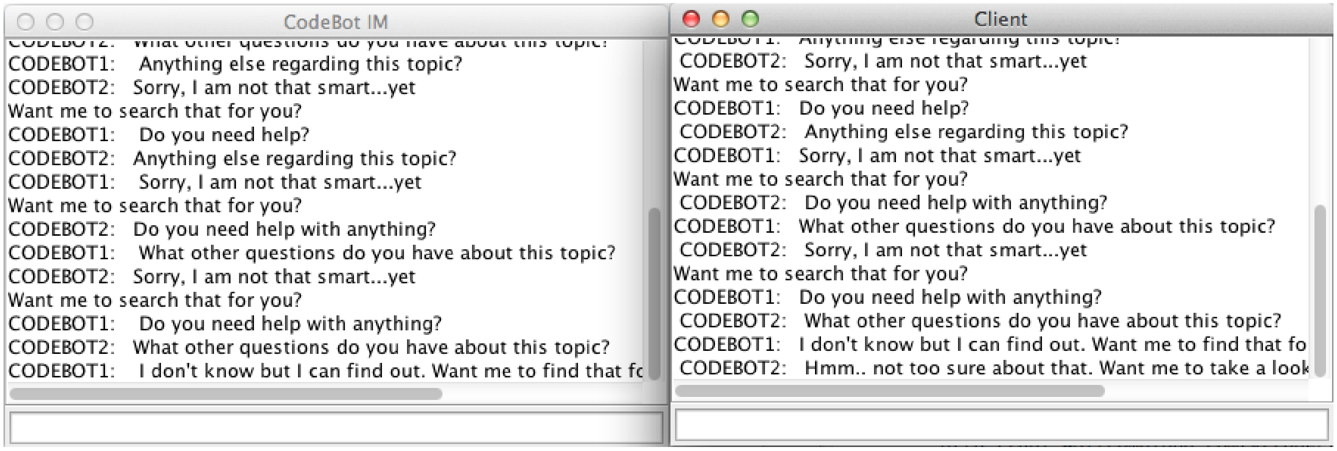
- When information is not available locally and web based resources are consulted, parts-of-speech tagging is used to identify the nouns, verbs and adjectives, that would provide the best search terms for the user's query. This is achieved via Stanford's parts-of-speech tagging library and API.



```
tagged[i]: WP
words[i]: what
what will not be added to the query.
array position: 1...
tagged[i]: VBZ
words[i]: is
is will not be added to the query.
array position: 2...
tagged[i]: DT
words[i]: an
an will not be added to the query.
array position: 3...
tagged[i]: JJ
words[i]: integral
Appending 'integral' to search query
Searchterms: integral
http://api.wolframalpha.com/v2/query?appid=KH2TXA-P884RH5W7G&input=integral&format=plai
```

Figure 2.0.5 CodeBot ignores words that would not be useful when searching for online descriptions or definitions

- Network capabilities have been added via Java sockets, making the system accessible remotely in future versions. This could lead in time to the capability of a student at home being able to access code bot on a school server.



```
CODEBOT2:   What other questions do you have about this topic?
CODEBOT1:    Anything else regarding this topic?
CODEBOT2:   Sorry, I am not that smart...yet
Want me to search that for you?
CODEBOT1:    Do you need help?
CODEBOT2:   Anything else regarding this topic?
CODEBOT1:    Sorry, I am not that smart...yet
Want me to search that for you?
CODEBOT2:   Do you need help with anything?
CODEBOT1:    What other questions do you have about this topic?
CODEBOT2:   Sorry, I am not that smart...yet
Want me to search that for you?
CODEBOT1:    Do you need help with anything?
CODEBOT2:   What other questions do you have about this topic?
CODEBOT1:    I don't know but I can find out. Want me to find that fc
```

```
CODEBOT1:   Anything else regarding this topic?
 CODEBOT2:   Sorry, I am not that smart...yet
Want me to search that for you?
CODEBOT1:   Do you need help?
 CODEBOT2:   Anything else regarding this topic?
CODEBOT1:   Sorry, I am not that smart...yet
Want me to search that for you?
 CODEBOT2:   Do you need help with anything?
CODEBOT1:   What other questions do you have about this topic?
 CODEBOT2:   Sorry, I am not that smart...yet
Want me to search that for you?
CODEBOT1:   Do you need help with anything?
 CODEBOT2:   What other questions do you have about this topic?
CODEBOT1:   I don't know but I can find out. Want me to find that fo
 CODEBOT2:   Hmm.. not too sure about that. Want me to take a look
```

Figure 2.0.6 Two CodeBot agents communicate with each other,
after using java sockets to establish a connection to the localhost

3.0 Sample Output

Below is the example output from the current version of CodeBot:

CodeBot: Howdy
User: hi
CodeBot: Do you need help with anything?
User: yes
CodeBot: What area do you need help with?
User: whats an integer?
CodeBot: By default, the int data type is a 32-bit signed two's complement integer, which has a minimum value of -2^31 and a maximum value of 2^31 -1.
CodeBot: Anything else regarding this topic?
User: no
CodeBot: How else may I be of service?
User: what is a double?
CodeBot: The double data type is a double-precision 64-bit IEEE 754 floating point. For decimal values, this data type is generally the default choice. As mentioned above, this data type should never be used for precise values, such as currency.
CodeBot: Anything else regarding this topic?
User: how can i declare one?
CodeBot: You can declare a double variable called x using 'double x;'.
CodeBot: What other questions do you have about this topic?
User: how can i initialize one?
CodeBot: You can initialize a double variable x using 'x = 4;' or 'x = 3.5;'
CodeBot: What other questions do you have about this topic?
User: none
CodeBot: Anything else I can assist you with?
User: yes
CodeBot: What area do you need help with?
User: what is quantum computing?
CodeBot: I don't know but I can find out. Want me to find that for you?
User: yes
CodeBot: Maybe this will help...
Results The qubit |psi> = a|0>+b|1> can be represented as a point (theta, phi) on a unit sphere called the Bloch sphere. Define the angles theta and phi by letting a = cos(theta/2) and b = e^(iphi) sin(theta/2). Here, a is taken to be real, which can always be made real by multiplying |psi> by an overall phase factor (that is

unobservable). Then |psi> is represented by the unit vector (cosphisintheta, sinphisintheta, costheta) called the Bloch vector.
Results A Bloch vector is a unit vector (cosphisintheta, sinphisintheta, costheta) used to represent points on a Bloch sphere.
Results A qubit (or quantum bit) is the analog of a bit for quantum computation. Unlike an ordinary bit, which may only assume two possible values (usually called 0 and 1), a qubit may assume a continuum of values of the form a|0>+b|1> where a and b are arbitrary complex numbers satisfying |a|^2+|b|^2 = 1.
CodeBot: Do you need help with anything?
User: yes
CodeBot: What topic is giving you trouble?
User: binary trees
CodeBot: Binary is a base-2 number system. Modern computer systems store information as strings of binary digits (bits).
CodeBot: What other questions do you have about this topic?
User: whats an example?
CodeBot: If we take the number 4 in Deciaml, what we normally count in, we can represent that number as 10 in binary. This is because each place in binary has a value of 2^i where i starts at zero next to the decimal point an increases as you move to the left. This is the same as in Decimal where each place has a value of 10^i.
CodeBot: What other questions do you have about this topic?
User: none
CodeBot: How else may I be of service?
User: you're the best
CodeBot: No, thank you!
User: cheers
CodeBot: I live to serve
User: tell me about functions
CodeBot: A named group of declarations and programming statements that can be invoked (executed) when needed. A function that is part of a class is called a method. Java has no functions because all code is part of a class.
CodeBot: Anything else regarding this topic?
User: no
CodeBot: Anything else I can assist you with?
User: what does implementation mean?
CodeBot: The process of translating a design into source code. Implementation is also refered to the source code that defines a method, class, abstract data type, or other programming entity.
CodeBot: Anything else regarding this topic?
User: what?
CodeBot: An implementation is the idea and understanding of a technical specification or algorithm as a program, software component, or other computer system through computer programming.
CodeBot: What other questions do you have about this topic?
User: how about interfaces?
CodeBot: I don't know but I can find out. Want me to find that for you?
User: no
CodeBot: Do you need help?
User: I need to know about data structures
CodeBot: Data Structure is a programming construct, either defined in the language or by a programmer, used to organize data into a format to facilitate access and processing. Arrays, linked lists, and stacks can all be considered data structures.
CodeBot: Anything else regarding this topic?
User: no
CodeBot: Anything else I can assist you with?
User: who is the batman?

CodeBot: I don't know but I can find out. Want me to find that for you?
User: yes
CodeBot: Maybe this will help...
Description Batman is the fictional Gotham City superhero whose real identity is the billionaire industrialist and playboy Bruce Wayne. (according to DC Comics)
Works featuring Batman year | title | medium 1953 | Is Your Honeymoon Really Necessary? | movie 1966 | Batman | television 1986 | Batman: The Dark Knight Returns | book | Batman: Chaos in Gotham | video game
CodeBot: Do you need help with anything?
User: good stuff!
CodeBot: Your welcome
User: haha
CodeBot: Anytime
User: goodbye
CodeBot: Farewell

4.0Final Additions (Final Submission)

Each of the packages has been made into an easy to use JAR file

- The Comparison jar allows the search for the presence for more than one String in a single Array, HashMap or String

- The Matcher jar allows fuzzy string matching, which handles things like spelling mistakes and pluralization, and compares the words entered to the set of words that can be recognized by the system

- The MyWiki jar searches WikiPedia for pages related to a given query, and uses a series of regular expression to parse the returned text data, and identify general information that might be useful to the user. The text initally returned by the MediaWiki API is almost unreadable and doesnt seem to match any conventional String structure (HTML, JSON etc.)

- The MyWolf jar searches Wolfram Alpha, and returns String data that only falls into specific categories such as "Definition" or "Description" with priorities assigned to each type of returned data

- The Punctuate jar formats the input sentence so that there are spaces between all words and punctuation, or so that there is no punctuation at all

Other APIS that have been incorporated include:

- Google translate, allowing for language speakers to use the program
- MediaWiki, allowing for retrieval of information that was not already available from Wolfram Alpha.
- Twitter4j, allowing the agent to spur the conversation when the user no longer needs help, suggesting interesting links and things to read from @CompSciFact