AJ Prosser

Web Crawler

The topic I covered for my project was the Japanese magazine publisher Nikoli, who are better known as the creators of Sudoku, along with many other pencil and paper puzzles. I created my knowledge base with no modifications, purely through the execution of the code. I began with a list of three websites: a page on Nikoli's website listing their puzzles and instructions on how to solve them, the Wikipedia page for Nikoli, and the page for Nikoli on the U.S. based parent publisher Hachette Book Group. After lots of testing, I found that these three sites tended to give the best output and most suitable important terms.

I initialize three variables: num_pages_to_scrape, linksperdomain, and linksperpage. The first controls the total number of pages scraped/visited by the webcrawler, the second controls the number of pages allowed per each base-level web domain, and links per page controls the number of links added to the list of links by each webpage.

The program keeps track of a list of URLs aptly titled "urls". The list of 3 base URLs are added to this list, and during the crawling process, links found on each page are appended to this list according to a list of criteria which will be explained later. For each URL, it loads it with BeautifulSoup, removes unwanted tags like headers and footers, unwraps formatting like italics and bold text, and extracts the text. After doing this, it removes foreign characters (this dataset in particular has many pages with Japanese text, so they must be regex'ed away), saves the modified text to a file, and adds the URL to a list of already traversed URLs.

Next, it gathers all the links on the page by looking up all the href tags, filtering out any that do not begin with https://, and ignoring any additional tags if more than the number in linksperpage have been looked up. The program then finds a rough approximation of the domain, grabbing the fifth to the fifteenth characters, and adds that to a dictionary, incrementing the value by 1 if it already is present. This list is used to fulfill the domain check which also uses linksperdomain. The program also filters out keywords in URLs corresponding to domains such as Wikimedia and Google, which

were determined by which sites were giving me the most trouble in testing. If the link passes all of these checks, it is added to the end of the URL list, incrementing the number used to check against linksperpage (defined as the variable “linkssofar”).

Once it loops through this process enough times, outputting as many files as determined in num_pages_to_scrape, the files are then cleaned for NLP parsing. All the Nikoli text is lowercased, with all lines shorter than 15 characters being removed from the dataset (hopefully to get as many untagged titles and footers as possible). I forewent removing stopwords, since the TF-IDF process I use shows that they are common to all documents anyways. All the files are concatenated into one long nikoli_corpus string.

This nikoli_corpus string is compared to the text from three other unrelated documents: two are chapters of textbooks used in the homework, and the other is the first 15000 characters of Moby Dick, which is one of the default texts in NLTK. The TF-IDF is calculated for the Nikoli text against the three other documents, with the terms most specific to the Nikoli text being sorted to the top. This returns a large list of words, of which I picked the following for my important terms:

[“nikoli”, “puzzle”, “sudoku”, “rules”, “book”, “magazine”, “japan”, “solve”, “cells”, block”]

These terms are then looked up in the unmodified corpus for sentences containing them–sentences are segmented using the NLTK sentence tokenizer. This tokenizer isn’t perfect, however, and often it will mistag long unbroken strings as single sentences, so any sentence longer than 400 characters is thrown out, line returns are removed, and the sentence is added to a list, which is then the value for each word’s key in the knowledge base dictionary. This knowledge base is pickled and saved to a file to be used by the chatbot.

Knowledge base screenshots

```
{
'nikoli': [
    'nikoli book shop page “Puzzle Communication Nikoli Vol.186” is now on sale.',
    'All our magazines and books are available at the “nikoli book shop” page.',
    'You can even access Nikoli puzzles online at nikoli.com.',
    'Nikoli Video / SNS We will deliver the latest information on nikoli.',
    'Let’s Begin nikoli Puzzles!',
    'Selected 200 good Sudoku puzzles from Nikoli, the original home of Sudoku.nikoli Place light bulbs (circles) according to the following rules.',
    'All our magazines and books are available at the “nikoli book shop” page.',
    'You can order Nikoli books from Nikoli formal web site, nikoli.co.jp .',
    'We have another exciting puzzle site, nikoli.com, where anyone, anywhere can access and play our unique challenging puzzles.',
    'Let’s Begin nikoli Puzzles!',
    'Learn more about this authornikoli Place balloons (white circles) and iron balls (black circles) according to the following rules.',
    'How YouTube works Test new features NFL Sunday Ticket © 2024 Google LLC  - YouTubenikoli Place numbers in squares (parts of snakes) so that the figures fr
    'With Nikoli logic puzzles you need just paper, pencil, and an active mind, no expensive programs or training…… nikoli Puzzles page nikoli book shop “Puzzl
    'Learn more about this authornikoli Divide the grid into blocks of five cells.',
    'nikoli book shop page “Puzzle Communication Nikoli Vol.186” is now on sale.',
    'We are confident it will be Nikoli puzzles.nikoli A rectangle, bordered by bold lines, is called a “room”.',
    'Nikoli Video / SNS We will deliver the latest information on nikoli.',
    'All Rights Reservednikoli Fill in the cells under the following rules.',
    'Effective Date: March 31, 2008 Last Updated: November 1, 2023nikoli Fill in all cells, each with a — or a |.',
    'With Nikoli logic puzzles you need just paper, pencil, and an active mind, no expensive programs or training…… nikoli Puzzles page nikoli book shop “Puzzl
'puzzle': [
    'The joy of solving puzzles is more than a passing fad, it enriches lives.',
    'Creating a puzzle does not necessarily require any special skill either.',
    'You also have a chance to see Nikoli puzzles through video games and books in your country.',
    'What makes solving puzzles fun?',
    'Nikoli - The leading company of the puzzle-world.',
    'An example of a -dependent genre of puzzle would be the , which relies on a specific language and .',
    'Don’t settle when it comes to your Sudoku: Choose puzzles hand-crafted by Nikoli, the Japanese company that invented the game.',
    'These puzzles with numbers and simple symbols do not depend on a particular language for solving.',
    '’“Nikoli Puzzle Book Reprint Edition Slitherlink 2” The puzzles in this volume are all the puzzles of “Pencil Puzzle Book Slitherlink Vol.11”.',
    'A unique aspect of Nikoli puzzles is that a puzzle that is challenging to a beginner and yet solvable, can also give the expert solver pleasure in solving
    'Our carefully reviewed puzzles will not cause solvers unnecessary stress.',
    '’“Nikoli Puzzle Book Reprint Edition Kakuro 2” The puzzles in this volume are all the puzzles of “Pencil Puzzle Book Kakuro Vol.11”.',
    'The magazine has invented several new genres of puzzles, and introduced several new games to Some of the popular Nikoli puzzles, along with their names; t
    'They hold the trademark on sudoku in Japan, and supply 100 percent of the puzzles published in Japanese newspapers and magazines.',
    'It is said that more than a hundred million people are solving sudoku puzzles every day.',
    'You can solve these puzzles without knowing Japanese or mathematical techniques.',
    'The bond between puzzle creator and solver.',
    'We just wish to share the joy and excitement of puzzles with our puzzle creators and solvers.',
    'The Nikoli pages on Wikipedia are very useful to understand the rules of Nikoli puzzles for English speakers.',
```

```
'rules': [
    'Fill in all empty cells with numbers under the following rules.',
    'The Nikoli pages on Wikipedia are very useful to understand the rules of Nikoli puzzles for English speakers.',
    'Fill in cells under the following rules.',
    'Akari has very simple rules and guide you to a logical complex world YosenabeNikoli has been creating puzzles for 40 years and is surely
    'Learn more about this authornikoli Place balloons (white circles) and iron balls (black circles) according to the following rules.',
    'Here on our homepage you get the basic rules at each puzzle.',
    'Paint some of the cells black under the following rules.',
    'All Rights Reservednikoli Fill in the cells under the following rules.',
    'Selected 200 good Sudoku puzzles from Nikoli, the original home of Sudoku.nikoli Place light bulbs (circles) according to the following r
'book': [
    'Then there is the new generation, young people becoming Nikoli fans by discovering the puzzles on the bookshelves of their parents.',
    'With Nikoli logic puzzles you need just paper, pencil, and an active mind, no expensive programs or training…… nikoli Puzzles page nikoli
    'Each puzzle is handcrafted (not computer generated, like most generic sudoku books) by the experts at Nikoli Publishing,\xa0the Japanese
    'nikoli book shop page “Puzzle Communication Nikoli Vol.186” is now on sale.',
    'Novel Suspects Logo Product image pagination Open the full-size image Nikoli Publishing Formats and Prices Also available from: The ultir
    'All our magazines and books are available at the “nikoli book shop” page.',
    'With Nikoli logic puzzles you need just paper, pencil, and an active mind, no expensive programs or training…… nikoli Puzzles page nikoli
    'You also have a chance to see Nikoli puzzles through video games and books in your country.',
    'All our magazines and books are available at the “nikoli book shop” page.',
    'nikoli book shop page “Puzzle Communication Nikoli Vol.186” is now on sale.',
    'You can order Nikoli books from Nikoli formal web site, nikoli.co.jp .',
    'We publish a quarterly magazine called Puzzle Communication Nikoli along with many other books.',
    'And unlike Workman’s previous small-sized sudoku books, The Biggest Bestest Book of Sudoku\u200b will come in a larger 8"x10" format and
'magazine': [
    '’“Puzzle Communication Nikoli Vol.186” Nikoli is the most advanced pencil puzzle magazine in Japan.',
    'We started the puzzle magazine Puzzle Communication Nikoli in 1980, and Nikoli has always specialized in logic puzzles.',
    'All our magazines and books are available at the “nikoli book shop” page.',
    'They hold the trademark on sudoku in Japan, and supply 100 percent of the puzzles published in Japanese newspapers and magazines.',
    '’“Puzzle Communication Nikoli Vol.186” Nikoli is the most advanced pencil puzzle magazine in Japan.',
    'We publish a quarterly magazine called Puzzle Communication Nikoli along with many other books.',
    'The magazine has invented several new genres of puzzles, and introduced several new games to Some of the popular Nikoli puzzles, along wi
    'We also provide puzzles to major Japanese newspapers and magazines.',
    'All our magazines and books are available at the “nikoli book shop” page.',
    'They hold the trademark on sudoku in Japan, and supply 100 percent of the puzzles published in Japanese newspapers and magazines.'],
'Japan': [
    'They hold the trademark on sudoku in Japan, and supply 100 percent of the puzzles published in Japanese newspapers and magazines.',
    'Don’t settle when it comes to your Sudoku: Choose puzzles hand-crafted by Nikoli, the Japanese company that invented the game.',
```

```
    'A number indicates the number of cells of the area containing this cell.'],
'block': [
    'A chain cannot contain more than one block of one size and shape.',
    'Divide all of the board into blocks.',
    'Lines must surround a cell and cannot enter a block.',
    'The number written on a cell indicates the number of black cells that make up the "block" containing that cell.',
    'Blocks are considered to have the same shape even if a block matches the rotated or mirror image of another block in a chain.',
    'There can be no blocks with fewer (0 - 4) or more (six and over) cells.',
    'A number in a cell shows how many lines of a block there are around it, the number includes the lines of the outer frame.',
    'Same sized blocks cannot touch each other, horizontally or vertically.',
    'Fill each block with the same number horizontally or vertically.',
    'Consecutive black cells are called a "block".',
    'Every block must contain one number or one "?".',
    'All blocks must be part of a chain.',
    'A block must contain a pair of areas of white cells and gray cells having the same form (size and shape).',
    'When two or more blocks are in contact at a corner it is called a "chain".',
    'A block can contain any number of cells with the number.',
    'Divide the grid into blocks by drawing solid lines over the dotted lines.',
    'Learn more about this authornikoli Divide the grid into blocks of five cells.',
    'Each block contains as many cells as the number in the block.',
    'A number indicates the number of cells of that color in the block; the number corresponds to half the number of cells of the block.',
    'There may be as many as 5 cells with numbers in a block.',
    'If there is a cell with a "?", you do not know the number of black cells that make up the block containing that cell.']
}
```

Chatbot

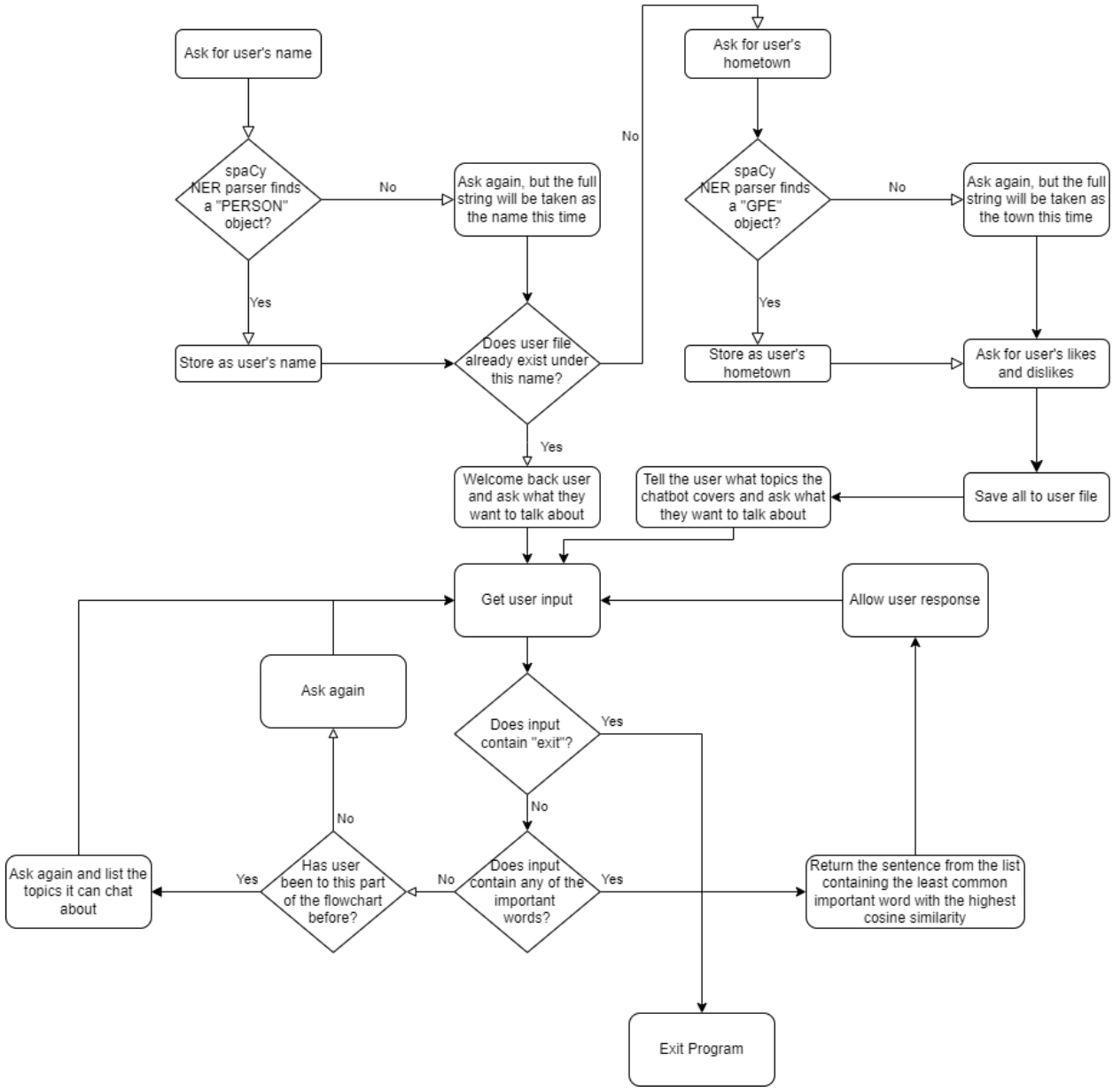
System Description

The chatbot consists of a rules-based beginning, which eventually becomes a query/response lookup with the knowledge base. The rules-based beginning uses named entity recognition (NER) to identify the user's name and their hometown, however it defaults to just using the whole string as some names and locations always get misidentified regardless of phrasing (from testing, I could rarely get it to pick up my own name, "AJ", and there are doubtless others).

Once the query section begins, however, the input from the user is modified using WordNetLematizer and NLTK, tokenizing, lowercasing and lemmatizing the words, as well as removing stopwords. The input is then checked against the 10 critical terms–if the input contains none of them, the chatbot will identify the topics it knows and ask the user to repeat. The critical words are listed in order of decreasing frequency and the least common word is used as a base. This operates off the assumption that if a sentence contains a less common word, it is probably based around that sentence. So the chatbot looks at all the sentences from the knowledge base and compares them using cosine similarity to the query sentence. The most similar sentence is then returned to the user, and the user is allowed to ask another question.

AJ Prosser

Flowchart



Sample Dialogue:

First time user:

Hello, what is your name?
>Jonathan
[Jonathan]
Nice to meet you! What is your hometown?
>Phoenix
[Phoenix]

Alright! What is something you like?
>Video Games
Okay, what is something you dislike?
>Twitter
Nice to meet you, Jonathan! I am a chatbot with information on the Japanese puzzle publisher, Nikoli, which is most well known for creating Sudoku.
What would you like to talk about? (type "exit" to quit.)
>Have they made video games?
Sorry, I don't understand. I can talk about the company Nikoli, their puzzles (including Sudoku), their rules and how to solve them, Nikoli's books and magazines, or common puzzle elements like cells and blocks.
>What is the best game of sudoku ever played?
“What’s next after the sudoku craze?” People are looking to Nikoli for answers.
>exit

Returning user:

Hello, what is your name?
>AJ
Sorry, I didn't quite get that. What is your name?
>AJ
Oh, hello, AJ! I remember you--you're the person from Phoenix who likes Food and dislikes Bugs.
What would you like to talk about? (type "exit" to quit.)
>Where can I buy Nikoli's magazines?
“Puzzle Communication Nikoli Vol.186” Nikoli is the most advanced pencil puzzle magazine in Japan.
>where can I buy nikoli's books?
“Puzzle Communication Nikoli Vol.186”,“Nikoli Puzzle Book Reprint Edition Kakuro 2” and “Nikoli Puzzle Book Reprint Edition Slitherlink 2” are now on sale.
>how might a puzzle use cells?
Balloons are light and float, so they must be placed in one of the cells at the top, or in a cell right under a black cell (filled-in cell) or under other balloons.
>exit

As you can see, the parser could not pick up “AJ” as a name.

Hello, what is your name?
>hi my name is matthew
Sorry, I didn't quite get that. What is your name?
>My name is Matthew
Nice to meet you! What is your hometown?
>Murietta California
Alright! What is something you like?
>I like to play Sudoku on my phone.
Okay, what is something you dislike?

```
>Cold food
Nice to meet you, My name is Matthew! I am a chatbot with information on the Japanese puzzle publisher, Nikoli, which is most well known for creating Sudoku.
What would you like to talk about? (type "exit" to quit.)
>I want to learn about Nikoli
With Nikoli logic puzzles you need just paper, pencil, and an active mind, no expensive programs or training…… nikoli Puzzles page nikoli book shop “Puzzle Communication Nikoli Vol.186”,“Nikoli Puzzle Book Reprint Edition Kakuro 2” and “Nikoli Puzzle Book Reprint Edition Slitherlink 2” are now on sale.
>Tell me more about the creation of sudoku
Each puzzle is handcrafted (not computer generated, like most generic sudoku books) by the experts at Nikoli Publishing, the Japanese creators of the game, so solving one is like pitting your mind against that of a sudoku master.
>Tell me about other games published by Nikoli
With Nikoli logic puzzles you need just paper, pencil, and an active mind, no expensive programs or training…… nikoli Puzzles page nikoli book shop “Puzzle Communication Nikoli Vol.186”,“Nikoli Puzzle Book Reprint Edition Kakuro 2” and “Nikoli Puzzle Book Reprint Edition Slitherlink 2” are now on sale.
>Tell me about the founder of the company
Sorry, I don't understand. I can talk about the company Nikoli, their puzzles (including Sudoku), their rules and how to solve them, Nikoli's books and magazines, or common puzzle elements like cells and blocks.
>Tell me about the founder of Nikoli
With Nikoli logic puzzles you need just paper, pencil, and an active mind, no expensive programs or training…… nikoli Puzzles page nikoli book shop “Puzzle Communication Nikoli Vol.186”,“Nikoli Puzzle Book Reprint Edition Kakuro 2” and “Nikoli Puzzle Book Reprint Edition Slitherlink 2” are now on sale.
>exit
AJs-MacBook-Air:Spring 2024 NLP AJ$ /usr/local/bin/python3 "/Users/AJ/VC/Spring 2024 NLP/Chatbot/chatbot.py"
Hello, what is your name?
>My name is Matthew
Nice to meet you! What is your hometown?
>^CKeyboardInterrupt
AJs-MacBook-Air:Spring 2024 NLP AJ$ /usr/local/bin/python3 "/Users/AJ/VC/Spring 2024 NLP/Chatbot/chatbot.py"
Hello, what is your name?
>matthew
Sorry, I didn't quite get that. What is your name?
>My name is Matthew
Oh, hello, My name is Matthew! I remember you--you're the person from California who likes I like to play Sudoku on my phone. and dislikes Cold food.
What would you like to talk about? (type "exit" to quit.)
>
```

The NER has trouble recognizing lowercase names, and is not set up to handle parsing likes/dislikes.

Knowledge base:

The format of the knowledge base is a dictionary, with the important word as the key and a list of sentence options as the value. The entire knowledge base is over 200 lines long, so I will not paste it here, but the format is as follows:

```
{
'nikoli': [
       'nikoli book shop page “Puzzle Communication Nikoli Vol.186” is now on sale.',
       'All our magazines and books are available at the “nikoli book shop” page.',
       ...
],
'puzzle': [
       'The joy of solving puzzles is more than a passing fad, it enriches lives.',
       ...
],
...
{Sections for all important words}
...
'block': [...]
```

User profiles:

The format of the user profiles is a simple dictionary in the following format:

AJ.p
{“hometown”: “Phoenix”, “likes”: “Food”, “dislikes”: “Bugs”}

This format could be easily extended to include more information about the user, however this is all this iteration of the program stores.

Analysis

Since this chatbot was limited to being rules-based as well as some other challenging restraints (must utilize a relatively small knowledge base scraped from the internet, this knowledge base must largely be cleaned up by the computer rather than by a human, and created within a relatively short development cycle by one person), the

output is not very good or convincing. Some inquiries it succeeds with (it seems to be the most successful on asking where to play puzzles or where to buy books/magazines), whereas others it often outputs nonsense or incorrect information (such as asking for the rules of a specific puzzle). A more robust system of detecting likes and dislikes using sentiment analysis would also be ideal–however developing a system to differentiate between sentences that state a preference, a query, or a combination of both was a bit out of scope for the scale of this project. More development time and a more carefully curated/manually written knowledge base likely could improve the quality of this system, however this would no longer fit the mechanical requirements of this specific project.

One particular challenge was picking suitable important words–it is easy to tell when a word is important, but it's another challenge to try and pick words that will output useful information, and I tried to pick the best out of the 50 most relevant TF-IDF words. However, some words are more prone to outputting nonsense or nonsequiturs–the "block", "cells", and "Japan" words being the worst offenders, with the first two missing most of their context most of the time, and the second appearing incidentally a lot and returning information that may not be all that relevant.

For outside evaluation, I asked my friends Jonathan, Ethan, and Matthew and charted their responses in this table.

	Strongly disagree	Partially disagree	Neither agree nor disagree	Partially agree	Strongly agree
The chatbot answered my questions accurately.					
		Jonathan, Ethan	AJ, Matthew		
The chatbot remembered who I was on a second run.					
		Matthew			AJ, Jonathan, Ethan
The chatbot gave reasonable responses to my queries.					

		AJ	Matthew	Jonathan, Ethan	
The chatbot was a convincing conversation partner.					
	AJ, Jonathan, Matthew		Ethan		

Overall, the replies are only somewhat relevant to the queried topic. A larger database, a more robust data cleaning process, and word similarity checking rather than just simple word-by-word cosine similarity might help diminish these issues. The strongest part is the more rules-based beginning, but that too occasionally runs into issues if the user types in all lowercase or if they phrase their likes and dislikes as sentences.