

Web Scraping to Build and Classify a Movie-Personality Dataset

Ryunosuke Saito (466), Jason Lin (466)

EN.601.466 Information Retrieval and Web Agents

Dr. David Yarowsky

Summary: Our project unleashes web scrapers in two domains -- www.dailyscript.com and www.personality-database.com -- to scrape dialogues of movie characters and subsequently mine the Myers-Briggs Type Indicator (MBTI) of each character. This constructed dataset is then used to 1.) ask ChatGPT to evaluate the dialogue and 2.) build and train a natural language classifier that predicts MBTI based on dialogue.

Dialogue/Personality Scraper Code:

<https://colab.research.google.com/drive/1HLgkkE-sBji9YIw3BTOZHfJtJfwLaDEJ?authuser=1>

OpenAI Q&A Code:

[Available upon request]

Binary Language Classifiers + Confidences Code:

https://colab.research.google.com/drive/1BJ0LW8udn0aB3GSSpivq0qPX_aNcFLV#scrollTo=rV0M76qzNrb9

Scraped Dialogue + Personalities CSV:  scripts.csv

OpenAI Q&A CSV:  scripts_openai

I. How to Run:

The program requires a computer that has the following installed, in addition to the packages used in the code:

- Google Chrome
- Jupyter Notebooks or a similar IDE (not Colabs)
- Reliable internet connection

The program itself can be broken up into three features: the web-scraping component, the OpenAI portion, and then the natural language classification component.

The web-scraping component uses the Beautiful Soup library to collect movie dialogue and relevant character information, and Selenium to collect MBTI and general description information for each character. This code need only be executed once to produce a 1557 row x 8 col pandas dataframe. Reliable internet connection is required, as the process is time-intensive. The output is a general NLP dataset that includes a movie, a character name, their dialogue, their MBTI types, and any general information about the character.

The OpenAI portion showcases the remarkable level of detail and richness in the scraped dialogue data. By leveraging the power of GPT, we are able to generate five comprehensive questions for each character and successfully utilize the dialogue to answer them. The existing NLP datasets in the movie industry, such as `cornell_movie_dialog` or `wiki_movies`, are deficient in both character dialogue and character descriptions. However, with the help of the OpenAI portion, we can create a comprehensive dataset that can be leveraged to train question-answering systems and dialogue summarizers.

The natural language classification component first performs post-processing on the dataframe by removing the rows in the dataframe for which an MBTI type was not found. The dialogue data is then tokenized and converted into a HuggingFace dataset. We then create a 80/20 train/test split on the dataset and train four different binary language classifiers, one for each MBTI type. Each model trains over 15 epochs, with a weight decay of 0.01 and learning rate of $3e-6$. The training loss, validation loss, and accuracy for each epoch is tabulated during training. The evaluative accuracy of both the training and test sets are returned. Additionally, the accuracy of the twenty inputs that the model classifies with the highest confidence is returned. The output of this section is four language models that are able to classify Introversion/Extroversion, Sensing/Intuition, Thinking/Feeling, and Judging/Perceiving, given a piece of dialogue.

II. Qualitative Evaluation

1. Relatively high precision on top 20 inputs that the models had highest confidence on.

2. Compiled a list of “non-character stopwords” that helped to separate non-characters (i.e. stage directions, narrators, etc.) from actual characters. Led to cleaner dataset and more effective personality scraping.
3. Created an NLP dataset that is organized by movie, character, and personality type.
4. Though movie character personality classification has been performed before, ours is novel in that we attempted to collect “description” fields of each character. The resultant dataset shows that most characters do not contain a description field, but demonstrates that other metrics besides MBT type can be easily collected using the personality scraper.

III. Limitations

1. Some movie dialogues were in the form of PDFs, which could not be scraped. This led to a smaller overall dataset.
2. Discrepancies in how movie titles are written in the script and in the personality database means that some characters are missed during scraping, leading to a smaller overall dataset.
3. Because MBTI types scraped on the personality database are based on user voting, unpopular characters have less confident evaluations of their MBTI types, which can prove to be a problem when training the language models.
4. Web Scraping with selenium is a time-intensive process (around 4 hours per run)
5. Ground truth needs to be established for OpenAI-generated answers.
6. Overall language model accuracy is lower, in part due to small dataset size.

IV. Figures

Example Scraped Movie Script

```

TEN THINGS I HATE ABOUT YOU

written by Karen McCullah Lutz & Kirsten Smith

based on 'Taming of the Shrew' by William Shakespeare

Revision November 12, 1997

PADUA HIGH SCHOOL - DAY

Welcome to Padua High School,, your typical urban-suburban
high school in Portland, Oregon. Smarties, Skids, Preppies,
Granolas. Loners, Lovers, the In and the Out Crowd rub sleep
out of their eyes and head for the main building.

PADUA HIGH PARKING LOT - DAY

KAT STRATFORD, eighteen, pretty -- but trying hard not to be
-- in a baggy granny dress and glasses, balances a cup of
coffee and a backpack as she climbs out of her battered,
baby blue '75 Dodge Dart.

A stray SKATEBOARD clips her, causing her to stumble and
spill her coffee, as well as the contents of her backpack.

The young RIDER dashes over to help, trembling when he sees
who his board has hit.

RIDER

Hey -- sorry.

Cowering in fear, he attempts to scoop up her scattered
belongings.

KAT

Leave it

He persists.

KAT (continuing)

I said, leave it!

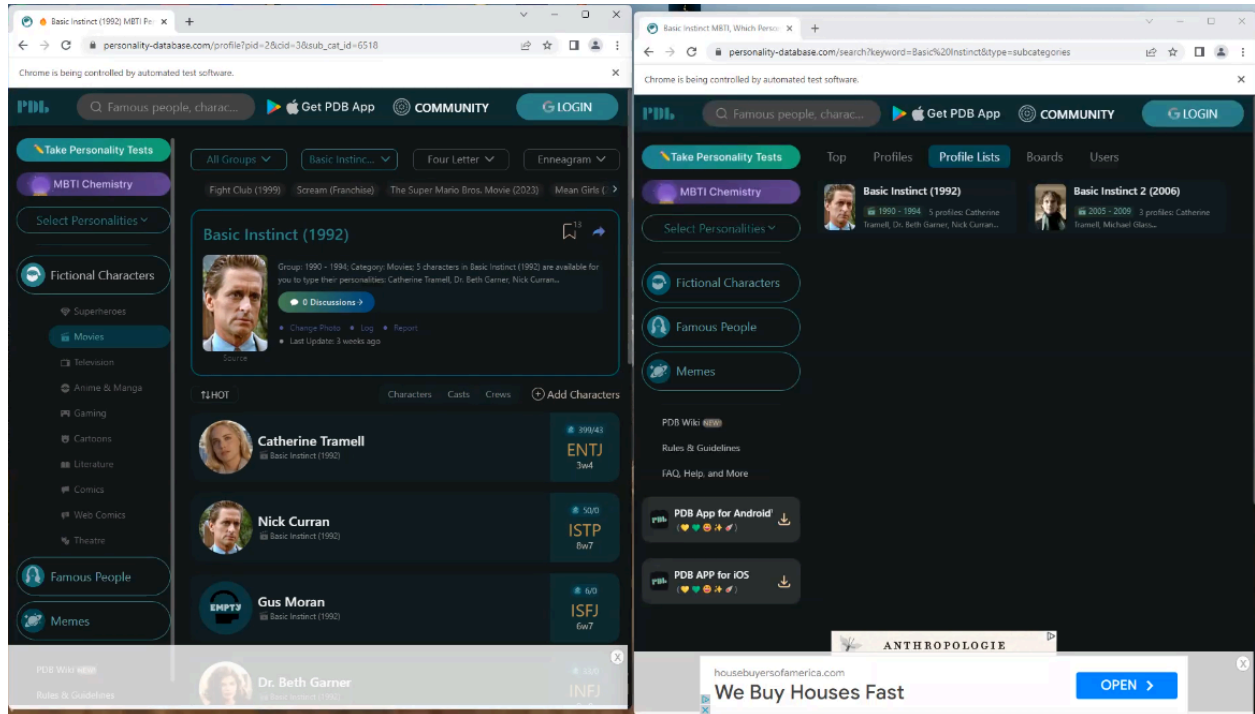
She grabs his skateboard and uses it to SHOVE him against a
car, skateboard tip to his throat. He whimpers pitifully
and she lets him go. A path clears for her as she marches
through a pack of fearful students and SLAMS open the door,
entering school.

INT. GIRLS' ROOM - DAY

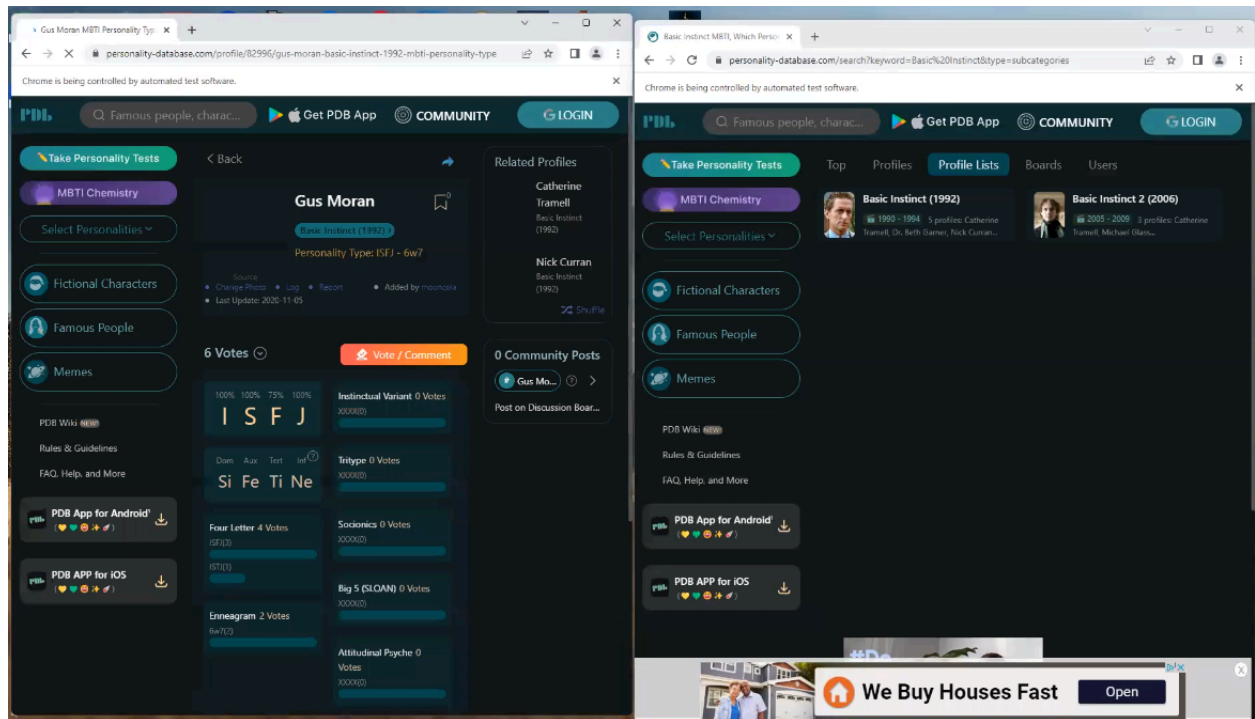
```

For each character (in this screenshot, RIDER and KAT), the dialogue is scraped and assigned. For each movie, the top five characters with the most dialogue are saved into a csv.

Selenium Webscraping Screenshot



The above is a snippet of the webscraping process using Selenium. The right window first scrapes the results after inputting a movie query. If there is a regex match with the results, then the left window opens. The left window then scrapes the list of movie characters. If there is a regex match between the character we are currently searching and a character on the page, then an additional window appears:



The MBTI types and the description (if present) of the character is then scraped.

ChatGPT-Generated Q&A

The following is an example of using the ChatGPT API to generate five questions and answers for each of the first five characters with scraped dialogue:

```
[ '1. What is the character\'s opinion of Hemingway and why do
they feel that way? - "Frankly, I\'m baffled as to why we
still revere Hemingway. He was an abusive, alcoholic
misogynist who had a lot of cats."\n2. What does the character
think about expressing their opinion? - "Expressing my opinion
is not a terrorist action."\n3. How does the character feel
about their mission in life? - "My mission in life."\n4. What
was the character\'s reaction to being accused of lying about
being in jail? - "No, you weren\'t. Why\'d you lie?"\n5. How
does the character feel about being set up by someone else? -
"You set me up. What? To completely damage me? To send me to
therapy forever? What?"',
```

```
"1. What is the protagonist's job? \nAnswer: The protagonist
is unemployed and looking to earn money by taking out a
girl.\n\n2. Who does the protagonist have a problem with?
\nAnswer: The protagonist has a problem with Kat, the girl
whom they are trying to befriend.\n\n3. What does the
protagonist think about the music at a concert? \nAnswer: The
protagonist thinks the music is not great but it's not bad
either and compares it to Bikini Kill and The Raincoats.\n\n4.
Why does the protagonist ask Cameron if he likes the girl he
```

is interested in? \nAnswer: The protagonist asks Cameron if he likes the girl because he recognizes Cameron's interest in Bianca.\n\n5. What does the protagonist want to do with a girl at Dollar Night? \nAnswer: The protagonist wants to take a girl out to the track on Dollar Night, buy her some noodles, and put his hand on her ass.",

"1. Did the character change her hair? \nAnswer: No, she didn't. \n\n2. Is the character willing to stay in the girls' room? \nAnswer: No, she suggests leaving the girls' room and entering the hallway. \n\n3. Does the character have a boyfriend? \nAnswer: No, she doesn't want to date. \n\n4. Is the character reluctant to attend Bogey Lowenstein's party? \nAnswer: Yes, she is. \n\n5. Who is the character waiting for outside? \nAnswer: The character is waiting for her date to arrive in five minutes.",

"1. Where is the character from and what is their age? \n\nAnswer: The character is from North and is 32 years old.\n\n2. Who is the character interested in and why are they interested in them? \n\nAnswer: The character is interested in Bianca's sister, but it is not clear why they are interested in her.\n\n3. What are some of the things that the character knows about Bianca's sister? \n\nAnswer: The character knows that Bianca's sister hates smokers, likes Thai food and feminist prose, and enjoys indie-rock music.\n\n4. What is the character's opinion of Joey and what is Bianca's opinion of him? \n\nAnswer: The character seems to dislike Joey, but it is not clear what Bianca's opinion of him is.\n\n5. What does the character do to try to win over the person they are interested in, and does it work? \n\nAnswer: The character's attempts to win over Bianca's sister are not clear, but it is suggested that they may have humiliated her in some way. It is also suggested that they have not been successful in winning her over.",




'1. What is the character's job or role in the school? \nAnswer: The character is supposed to give the new guy a tour of the school. \n\n2. What is the character's opinion of the student body at the school? \nAnswer: The character thinks "most of them" are "evil." \n\n3. Who does the character suggest going out with in order to date Bianca Stratford? \nAnswer: The character suggests finding someone to date Kat, Bianca's sister. \n\n4. What does the character think of Joey, Bianca's ex-boyfriend? \nAnswer: The character thinks Joey is a "kiss ass." \n\n5. What play does the character mention and is anyone else a fan of it? \nAnswer: The character mentions Macbeth and confirms that both they and Kat are fans of the play.']

The veracity of these questions is yet to be evaluated, but demonstrates a proof-of-concept in using a GPT and movie dialogue to quickly build a profile of a character.

V. Quantitative Evaluation

The results are on the following pages (also accessible [here](#)):

Results

MBTI	Training/Test Performance and Validation Accuracy	Accuracy on Top 20 Most Confident Inputs																																																																																																																														
I/E	Test {'eval_loss': 0.6802179217338562, 'eval_accuracy': 0.5233644859813084, 'eval_runtime': 1.9505, 'eval_samples_per_second': 54.858, 'eval_steps_per_second': 13.843, 'epoch': 15.0}	<div>Accuracy = 0.75</div> <table><tr><th></th><th>input</th><th>confidence</th><th>actual</th><th>predicted</th><th></th></tr><tr><td>0</td><td>JEROME emerges from the incinerator room into ...</td><td>0.725790</td><td>0</td><td>0</td><td></td></tr><tr><td>1</td><td>We're going to freeze to death.\nMaybe somebod...</td><td>0.716319</td><td>0</td><td>0</td><td></td></tr><tr><td>2</td><td>ALICE comes into the MS. She is carrying a sho...</td><td>0.712792</td><td>0</td><td>0</td><td></td></tr><tr><td>3</td><td>Where's the coffee.\nDALLAS looks at his grogg...</td><td>0.709347</td><td>0</td><td>0</td><td></td></tr><tr><td>4</td><td>Elena. My name's Elena.\nELENA. He nods his he...</td><td>0.700423</td><td>0</td><td>0</td><td></td></tr><tr><td>5</td><td>...Hello.\n...What? Who is this?\nYeah, I'm st...</td><td>0.689290</td><td>0</td><td>0</td><td></td></tr><tr><td>6</td><td>RUDY BUTLER -- hard eyes and an easy smile; ob...</td><td>0.679778</td><td>1</td><td>0</td><td></td></tr><tr><td>7</td><td>You're Julia, right?\nI'm brother Frank.\nThat...</td><td>0.665829</td><td>1</td><td>0</td><td></td></tr><tr><td>8</td><td>TINA rushes away, hands over her ears.\nTINA o...</td><td>0.659402</td><td>1</td><td>0</td><td></td></tr><tr><td>9</td><td>She must have done something with the child. W...</td><td>0.652228</td><td>0</td><td>0</td><td></td></tr><tr><td>10</td><td>Where'd you come from?\nWell, don't just sit t...</td><td>0.679473</td><td>1</td><td>1</td><td></td></tr><tr><td>11</td><td>Who are they?\nYes.\nHow do you do.\nCedar, Ce...</td><td>0.655754</td><td>1</td><td>1</td><td></td></tr><tr><td>12</td><td>This was a valued rug. He elaborately clears h...</td><td>0.655414</td><td>1</td><td>1</td><td></td></tr><tr><td>13</td><td>What so you want to know?\nYeah, it's legal, b...</td><td>0.640242</td><td>0</td><td>1</td><td></td></tr><tr><td>14</td><td>Wait a minute! Who elected you leader a this o...</td><td>0.638728</td><td>1</td><td>1</td><td></td></tr><tr><td>15</td><td>Come in, Mr. Juarez. I'd stand, but, well, you...</td><td>0.637062</td><td>1</td><td>1</td><td></td></tr><tr><td>16</td><td>FRANK What's your last name?\nWhat are all them...</td><td>0.634773</td><td>0</td><td>1</td><td></td></tr><tr><td>17</td><td>JACK It's six a.m.... Ooooooo and that bed nev...</td><td>0.621986</td><td>1</td><td>1</td><td></td></tr><tr><td>18</td><td>I'll be Goddamned! I'm not dead! Sarge hollers...</td><td>0.605450</td><td>1</td><td>1</td><td></td></tr><tr><td>19</td><td>I'm busy, Floyd.\nWhich one?\nYou sure he didn...</td><td>0.594596</td><td>1</td><td>1</td><td></td></tr></table>		input	confidence	actual	predicted		0	JEROME emerges from the incinerator room into ...	0.725790	0	0		1	We're going to freeze to death.\nMaybe somebod...	0.716319	0	0		2	ALICE comes into the MS. She is carrying a sho...	0.712792	0	0		3	Where's the coffee.\nDALLAS looks at his grogg...	0.709347	0	0		4	Elena. My name's Elena.\nELENA. He nods his he...	0.700423	0	0		5	...Hello.\n...What? Who is this?\nYeah, I'm st...	0.689290	0	0		6	RUDY BUTLER -- hard eyes and an easy smile; ob...	0.679778	1	0		7	You're Julia, right?\nI'm brother Frank.\nThat...	0.665829	1	0		8	TINA rushes away, hands over her ears.\nTINA o...	0.659402	1	0		9	She must have done something with the child. W...	0.652228	0	0		10	Where'd you come from?\nWell, don't just sit t...	0.679473	1	1		11	Who are they?\nYes.\nHow do you do.\nCedar, Ce...	0.655754	1	1		12	This was a valued rug. He elaborately clears h...	0.655414	1	1		13	What so you want to know?\nYeah, it's legal, b...	0.640242	0	1		14	Wait a minute! Who elected you leader a this o...	0.638728	1	1		15	Come in, Mr. Juarez. I'd stand, but, well, you...	0.637062	1	1		16	FRANK What's your last name?\nWhat are all them...	0.634773	0	1		17	JACK It's six a.m.... Ooooooo and that bed nev...	0.621986	1	1		18	I'll be Goddamned! I'm not dead! Sarge hollers...	0.605450	1	1		19	I'm busy, Floyd.\nWhich one?\nYou sure he didn...	0.594596	1	1	
			input	confidence	actual	predicted																																																																																																																										
	0		JEROME emerges from the incinerator room into ...	0.725790	0	0																																																																																																																										
	1		We're going to freeze to death.\nMaybe somebod...	0.716319	0	0																																																																																																																										
	2		ALICE comes into the MS. She is carrying a sho...	0.712792	0	0																																																																																																																										
	3		Where's the coffee.\nDALLAS looks at his grogg...	0.709347	0	0																																																																																																																										
	4		Elena. My name's Elena.\nELENA. He nods his he...	0.700423	0	0																																																																																																																										
	5		...Hello.\n...What? Who is this?\nYeah, I'm st...	0.689290	0	0																																																																																																																										
	6		RUDY BUTLER -- hard eyes and an easy smile; ob...	0.679778	1	0																																																																																																																										
	7		You're Julia, right?\nI'm brother Frank.\nThat...	0.665829	1	0																																																																																																																										
	8		TINA rushes away, hands over her ears.\nTINA o...	0.659402	1	0																																																																																																																										
	9		She must have done something with the child. W...	0.652228	0	0																																																																																																																										
	10		Where'd you come from?\nWell, don't just sit t...	0.679473	1	1																																																																																																																										
	11		Who are they?\nYes.\nHow do you do.\nCedar, Ce...	0.655754	1	1																																																																																																																										
	12		This was a valued rug. He elaborately clears h...	0.655414	1	1																																																																																																																										
	13		What so you want to know?\nYeah, it's legal, b...	0.640242	0	1																																																																																																																										
14	Wait a minute! Who elected you leader a this o...	0.638728	1	1																																																																																																																												
15	Come in, Mr. Juarez. I'd stand, but, well, you...	0.637062	1	1																																																																																																																												
16	FRANK What's your last name?\nWhat are all them...	0.634773	0	1																																																																																																																												
17	JACK It's six a.m.... Ooooooo and that bed nev...	0.621986	1	1																																																																																																																												
18	I'll be Goddamned! I'm not dead! Sarge hollers...	0.605450	1	1																																																																																																																												
19	I'm busy, Floyd.\nWhich one?\nYou sure he didn...	0.594596	1	1																																																																																																																												
	Training {'eval_loss': 0.5838929414749146, 'eval_accuracy': 0.7505882352941177, 'eval_runtime': 7.6693, 'eval_samples_per_second': 55.415, 'eval_steps_per_second': 13.952, 'epoch': 15.0}																																																																																																																															
	<table><tr><th>Epoch</th><th>Training Loss</th><th>Validation Loss</th><th>Accuracy</th></tr><tr><td>1</td><td>0.695000</td><td>0.690723</td><td>0.523364</td></tr><tr><td>2</td><td>0.685100</td><td>0.690351</td><td>0.514019</td></tr><tr><td>3</td><td>0.680700</td><td>0.682370</td><td>0.504673</td></tr><tr><td>4</td><td>0.654400</td><td>0.683732</td><td>0.532710</td></tr><tr><td>5</td><td>0.630500</td><td>0.680218</td><td>0.523364</td></tr><tr><td>6</td><td>0.584900</td><td>0.691516</td><td>0.542056</td></tr><tr><td>7</td><td>0.558600</td><td>0.699094</td><td>0.551402</td></tr><tr><td>8</td><td>0.512200</td><td>0.716574</td><td>0.532710</td></tr><tr><td>9</td><td>0.475300</td><td>0.729324</td><td>0.551402</td></tr><tr><td>10</td><td>0.433200</td><td>0.742108</td><td>0.542056</td></tr><tr><td>11</td><td>0.407700</td><td>0.757867</td><td>0.542056</td></tr><tr><td>12</td><td>0.390200</td><td>0.789310</td><td>0.542056</td></tr><tr><td>13</td><td>0.347300</td><td>0.791795</td><td>0.514019</td></tr><tr><td>14</td><td>0.326500</td><td>0.802578</td><td>0.504673</td></tr><tr><td>15</td><td>0.314600</td><td>0.806875</td><td>0.514019</td></tr></table>	Epoch	Training Loss	Validation Loss	Accuracy	1	0.695000	0.690723	0.523364	2	0.685100	0.690351	0.514019	3	0.680700	0.682370	0.504673	4	0.654400	0.683732	0.532710	5	0.630500	0.680218	0.523364	6	0.584900	0.691516	0.542056	7	0.558600	0.699094	0.551402	8	0.512200	0.716574	0.532710	9	0.475300	0.729324	0.551402	10	0.433200	0.742108	0.542056	11	0.407700	0.757867	0.542056	12	0.390200	0.789310	0.542056	13	0.347300	0.791795	0.514019	14	0.326500	0.802578	0.504673	15	0.314600	0.806875	0.514019																																																															
Epoch	Training Loss	Validation Loss	Accuracy																																																																																																																													
1	0.695000	0.690723	0.523364																																																																																																																													
2	0.685100	0.690351	0.514019																																																																																																																													
3	0.680700	0.682370	0.504673																																																																																																																													
4	0.654400	0.683732	0.532710																																																																																																																													
5	0.630500	0.680218	0.523364																																																																																																																													
6	0.584900	0.691516	0.542056																																																																																																																													
7	0.558600	0.699094	0.551402																																																																																																																													
8	0.512200	0.716574	0.532710																																																																																																																													
9	0.475300	0.729324	0.551402																																																																																																																													
10	0.433200	0.742108	0.542056																																																																																																																													
11	0.407700	0.757867	0.542056																																																																																																																													
12	0.390200	0.789310	0.542056																																																																																																																													
13	0.347300	0.791795	0.514019																																																																																																																													
14	0.326500	0.802578	0.504673																																																																																																																													
15	0.314600	0.806875	0.514019																																																																																																																													

S/N

Test

```
{'eval_loss': 0.6250171661376953,
'eval_accuracy': 0.6728971962616822,
'eval_runtime': 1.9623,
'eval_samples_per_second': 54.529,
'eval_steps_per_second': 13.76, 'epoch':
15.0}
```

Training

```
{'eval_loss': 0.4032510817050934,
'eval_accuracy': 0.8752941176470588,
'eval_runtime': 7.6671,
'eval_samples_per_second': 55.432,
'eval_steps_per_second': 13.956, 'epoch':
15.0}
```

Epoch	Training Loss	Validation Loss	Accuracy
1	0.693400	0.681802	0.570093
2	0.690900	0.680946	0.570093
3	0.681300	0.675720	0.588785
4	0.667000	0.664286	0.588785
5	0.641900	0.655592	0.616822
6	0.617500	0.648251	0.616822
7	0.580600	0.637726	0.635514
8	0.545700	0.628231	0.663551
9	0.493200	0.625751	0.672897
10	0.457900	0.625017	0.672897
11	0.440600	0.630354	0.654206
12	0.398100	0.637088	0.654206
13	0.385100	0.641410	0.654206
14	0.375500	0.638933	0.691589
15	0.364600	0.642250	0.682243

Accuracy = 0.80

	input	confidence	actual	predicted
0	He's been tested for drugs? RAILLY'S POV THRO...	0.828312	0	0
1	EMMA smiles. Alexander stands, just watching h...	0.792895	0	0
2	FINNEGAN, bathed in the last light of day, all...	0.762280	0	0
3	Back! Don't go near it. We don't know where it...	0.762082	1	0
4	He sounds like a real gentleman. I'm so sorry...	0.761953	1	0
5	ELLIE Dr. Grant! Dr. Grant! Grant looks up. S...	0.761208	0	0
6	As you wish. Westley is perhaps half a dozen y...	0.742651	0	0
7	OSBORN, fiftyish. Harry has already inherited ...	0.738053	1	0
8	LESTAT, a hooded figure in the corner, smiles ...	0.724831	1	0
9	VOX KIOSK VOX KIOSK The joint United Nations/...	0.718201	0	0
10	slips the postcard under the rubber bands on t...	0.894229	1	1
11	LEONARD So where are you? Leonard lifts his he...	0.888745	1	1
12	So, what's the deal? If the hospital buys ten ...	0.882648	1	1
13	Hey, hey, you start touching me there, I'm gon...	0.879142	1	1
14	AL AND ROSASHARN Hi, Tom! Howya doin'? You bu...	0.871571	1	1
15	...Hello. ...What? Who is this? Yeah, I'm st...	0.865749	1	1
16	GLORIA PEREZ walks at his side. She's a good s...	0.861561	1	1
17	You should work for yourself. Two Junior High ...	0.859280	1	1
18	TINA rushes away, hands over her ears. TINA o...	0.856094	1	1
19	LYDIA Mitch? They turn toward the door. How d...	0.853637	1	1

T/F

Test

{'eval_loss': 0.5931413173675537, 'eval_accuracy': 0.6635514018691588, 'eval_runtime': 1.931, 'eval_samples_per_second': 55.411, 'eval_steps_per_second': 13.982, 'epoch': 15.0}

Training

{'eval_loss': 0.3855719268321991, 'eval_accuracy': 0.8658823529411764, 'eval_runtime': 7.6247, 'eval_samples_per_second': 55.74, 'eval_steps_per_second': 14.033, 'epoch': 15.0}

Epoch	Training Loss	Validation Loss	Accuracy
1	0.694400	0.688204	0.551402
2	0.683700	0.677762	0.616822
3	0.665000	0.659612	0.682243
4	0.634200	0.636408	0.672897
5	0.590500	0.634663	0.644860
6	0.551400	0.598493	0.700935
7	0.531000	0.618571	0.663551
8	0.497700	0.594219	0.682243
9	0.470200	0.603671	0.682243
10	0.430200	0.593141	0.663551
11	0.411300	0.618296	0.691589
12	0.402100	0.618630	0.691589
13	0.369600	0.613723	0.682243
14	0.365700	0.615908	0.682243
15	0.367900	0.623889	0.672897

Accuracy = 0.85

	input	confidence	actual	predicted
0	Where's the coffee.\nDALLAS looks at his grogg...	0.910493	0	0
1	Time for another one.\nThe small matter of a s...	0.894505	0	0
2	THIBADEAUX, a Cajun who carries himself with a...	0.875652	0	0
3	Turn the goddamn thing off, will ya.\nGlad to ...	0.870671	0	0
4	DOYLE is at the wheel. BLACK PUSHER is sitting...	0.869121	0	0
5	MAX Just overnight is all. Tomorrow I'll\nMAX ...	0.867767	0	0
6	Second Foot, wheel right, advance quick step.....	0.864412	0	0
7	I'm busy, Floyd.\nWhich one?\nYou sure he didn...	0.862973	0	0
8	How much you gonna pay me? Huh? I'd be doing y...	0.844138	0	0
9	How much you gonna pay me? Huh? I'd be doing y...	0.844138	0	0
10	Hello.\nYou're living at home now. Is that rig...	0.854150	1	1
11	- and today it is right that we should ask our...	0.849012	0	1
12	Elena. My name's Elena.\nELENA. He nods his he...	0.847736	1	1
13	AMBER ATKINS - naturally pretty blonde, sweet ...	0.845886	1	1
14	We're going to freeze to death.\nMaybe somebod...	0.841355	1	1
15	Okay.\nWhat class?\nOh...\nYou suck.\nI've bee...	0.840448	1	1
16	Tripp\nBrimming. Say hello to my new friend, ...	0.832067	0	1
17	PEGGY SUE GOT MARRIED An Original Screenplay b...	0.828447	1	1
18	Oh, he's not getting away with anything.\nWith...	0.828130	0	1
19	Look, my son is allergic to the material in th...	0.827116	1	1

J/P

Test
 {'eval_loss': 0.6672698259353638,
 'eval_accuracy': 0.6074766355140186,
 'eval_runtime': 1.9403,
 'eval_samples_per_second': 55.147,
 'eval_steps_per_second': 13.916, 'epoch':
 15.0}

Training
 {'eval_loss': 0.5323470234870911,
 'eval_accuracy': 0.8023529411764706,
 'eval_runtime': 7.6783,
 'eval_samples_per_second': 55.351,
 'eval_steps_per_second': 13.935, 'epoch':
 15.0}

Epoch	Training Loss	Validation Loss	Accuracy
1	0.691000	0.681992	0.570093
2	0.684400	0.680340	0.570093
3	0.674200	0.677962	0.570093
4	0.664800	0.677483	0.588785
5	0.652100	0.670326	0.579439
6	0.626300	0.668275	0.588785
7	0.587100	0.667270	0.607477
8	0.549700	0.668403	0.598131
9	0.519500	0.677436	0.551402
10	0.489500	0.678381	0.588785
11	0.457700	0.689076	0.598131
12	0.424700	0.691064	0.598131
13	0.410700	0.695733	0.588785
14	0.401200	0.698877	0.588785
15	0.383600	0.700164	0.579439

Accuracy = 0.70

	input	confidence	actual	predicted
0	STARGHER'S HANDS COME AT HER FROM BEHIND! One ...	0.679632	1	0
1	Hello, Mr. Deckard. My name is Rachael. Deckar...	0.672443	0	0
2	You're Julia, right?\nI'm brother Frank.\nThat...	0.656160	1	0
3	You're Julia, right?\nI'm brother Frank.\nThat...	0.650364	1	0
4	We're going to freeze to death.\nMaybe somebod...	0.647148	0	0
5	LARRY COTTON, the other his wife JULIA. Clearl...	0.646854	0	0
6	Yes, with Fedens and the children. He wants so...	0.633817	0	0
7	Geoffrey gave me your monograph\nLove? Romanti...	0.618266	0	0
8	LYDIA - a dowdy, waif-like sparrow of a thing....	0.602772	1	0
9	That is of no interest. And now he turns and I...	0.589014	0	0
10	You guys know my cousin Mikey Sullivan?\nWell ...	0.757074	1	1
11	What do you want me to say? That I'm\nSEBASTIA...	0.738762	1	1
12	DUKE We were somewhere around Barstow on the e...	0.733254	1	1
13	DUKE VOICEOVER We were somewhere around Barsto...	0.732039	1	1
14	Neighbor, I'd feel better about the damned inc...	0.730712	0	1
15	AL AND ROSASHARN Hi, Tom! Howya doin'?\nYou bu...	0.729697	1	1
16	You didn't have to come to Cleveland to get be...	0.726678	1	1
17	How much you gonna pay me? Huh? I'd be doing y...	0.724883	1	1
18	How much you gonna pay me? Huh? I'd be doing y...	0.724883	1	1
19	- Okay now, tell me about the hash bars?\nWell...	0.720606	0	1

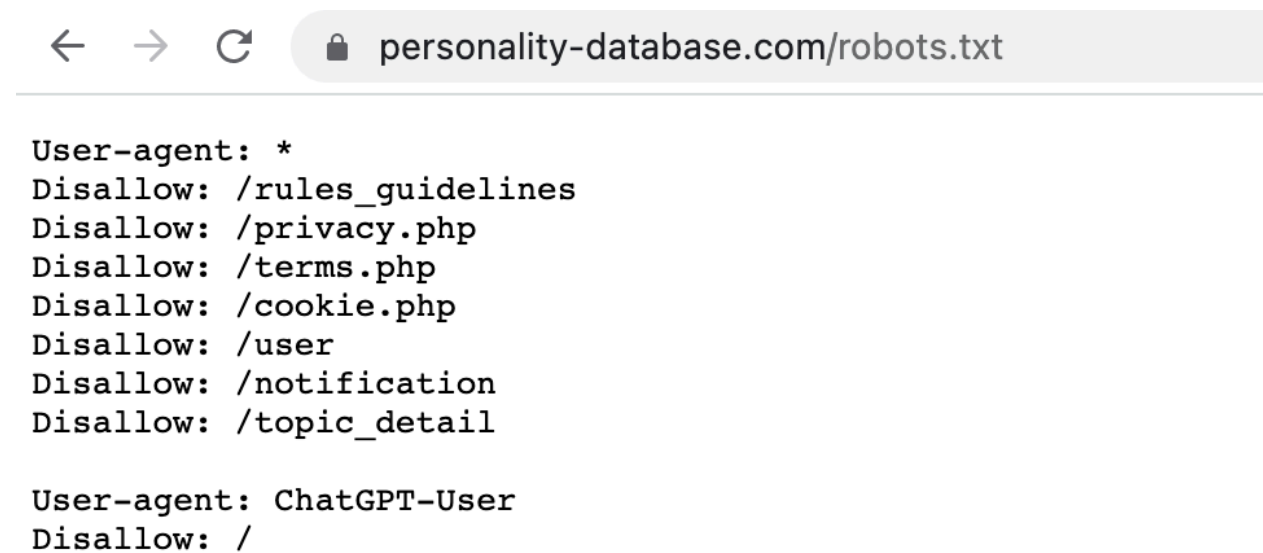
For all four dimensions of MBTI -- Introversion/Extroversion, Sensing/Intuition, Thinking/Feeling, Judging/Perceiving -- the performance on the training data was quite poor. The models classified with 52%, 67%, 66%, and 60% accuracy on the test set respectively. From this perspective, it appears that using dialogue to classify personality type is a difficult task. The model does not seem to be learning any meaningful features about the dialogue. Indeed, Sang et. al (who performed a similar experiment) notes that addition of other features -- facial expressions, body language, and other contexts -- is important in determining personality as well

[1]. Compilation of a multimodal dataset presents an opportunity for further investigation of using movie dialogue to classify personality types.

Although the initial evaluation of our model's accuracy on both classes was not satisfactory, we observed a significant improvement in accuracy when focusing on the top 10 confident predictions from each class. The accuracy increased to 75%, 80%, 85%, and 70%, respectively, indicating that our model has the potential to perform well if more data points are available and if the MBTI scores on the website are less noisy. To further enhance our model's training, we could consider filtering out data with a small number of votes from users and incorporating the percentage of vote splits in the training loop. With these improvements, we believe that pursuing this task is a viable option.

VI. Web Scraper Risk Assessment

The agents unleashed in this project were restricted only to the domains mentioned in the project. The scrapers followed the rules outlined by the robots.txt of each site (thedailyscript.com did not feature a robots.txt):



```
User-agent: *
Disallow: /rules_guidelines
Disallow: /privacy.php
Disallow: /terms.php
Disallow: /cookie.php
Disallow: /user
Disallow: /notification
Disallow: /topic_detail

User-agent: ChatGPT-User
Disallow: /
```

Upon inspection of the code, it is clear that none of the disallowed domains were visited.

References

- [1] Yisi Sang, Xiangyang Mou, Mo Yu, Dakuo Wang, Jing Li, and Jeffrey Stanton. 2022. MBTI Personality Prediction for Fictional Characters Using Movie Scripts.
DOI:<https://doi.org/10.48550/ARXIV.2210.10994>