NLP of Customer Reviews : Women's Clothing Brand

ITC6010A1

Natural Language Processing

Instructor:

Lazaros Polymenakos

Alkiviadis Kariotis
Celsa Almeida
Georgios Aliferis
Konstantinos Margaritis

Deree
School of Graduate and Professional Education

THE AMERICAN COLLEGE OF GREECE
Pierce Deree Alba
1875

Presentation Overview

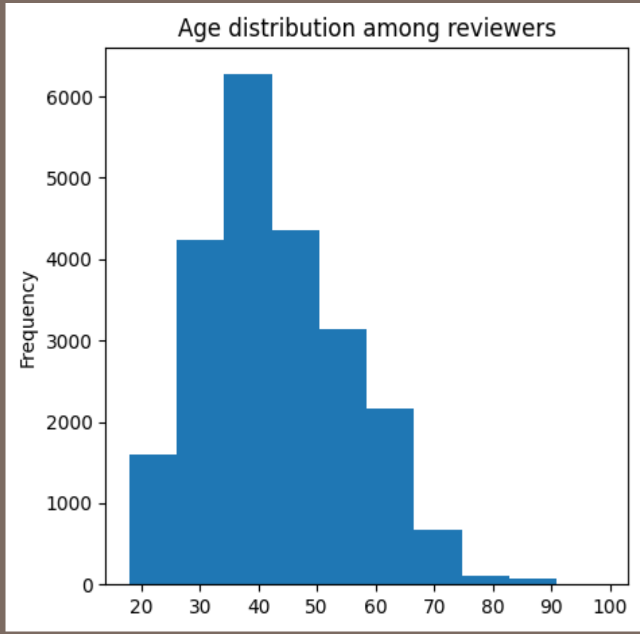
1. EDA
2. Sentiment Analysis
3. Topic Extraction
4. Text Summarization
5. LLM & DataFrame chatbot
6. Future Improvements

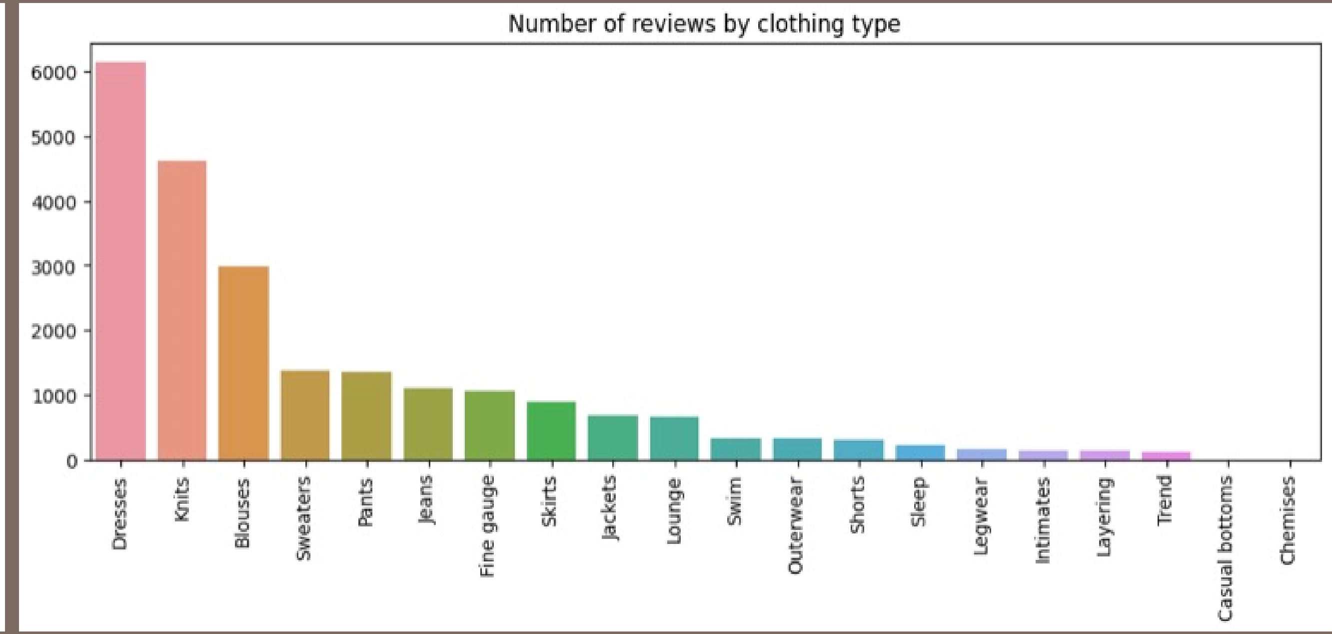
Dataset, Processing, Tokenization

Dataset Overview

Clothing ID	Age (Customer)	Title	Review Text	Rating	Recommended IND	Positive Feedback Count	Division Name	Department	Class Name

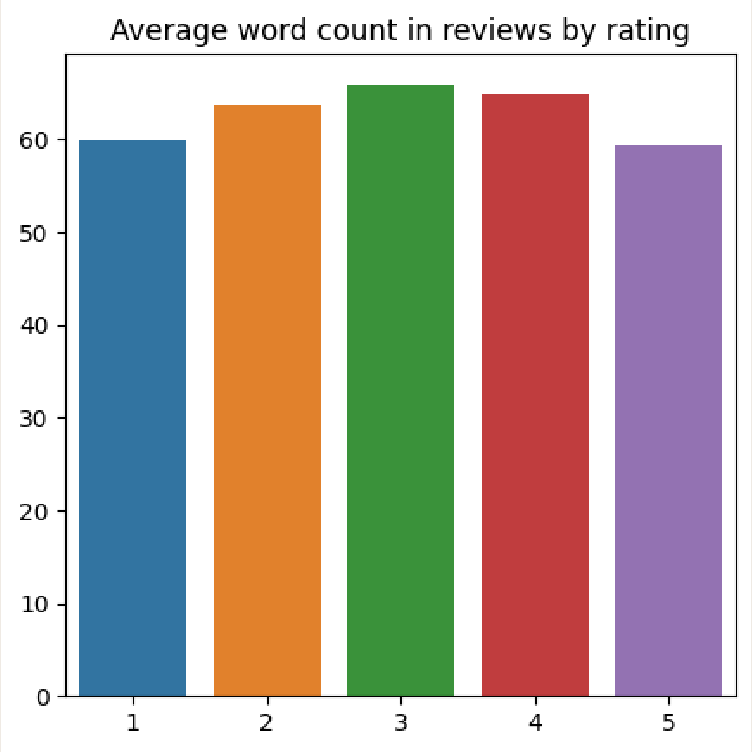
We prep the data by first checking and removing entries with missing values and then start preliminary examination as below

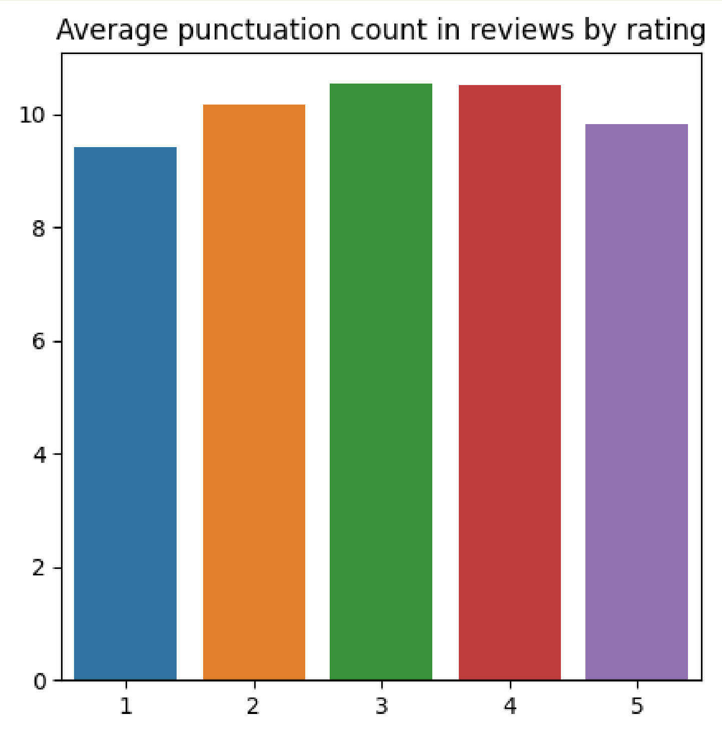




Dataset, Processing, Tokenization

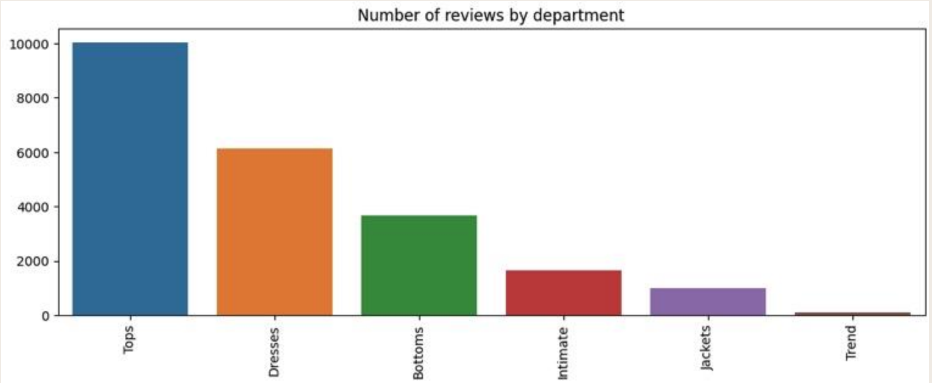
This analysis gives an understanding of whether reviewers leaving high or low ratings tend to write longer or more complex reviews. In our case this does not happen.





Next, we change all text to lowercase, removing punctuation and stopwords, lemmatize the text and Tokenize with **NLTK's RegexpTokenizer**

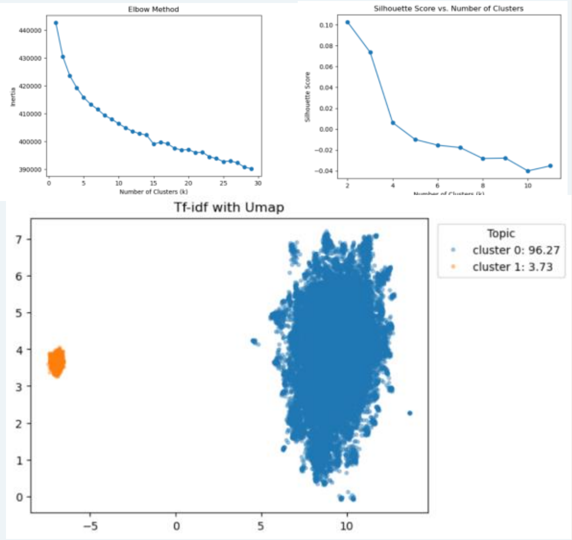
Exploring the Dataset Structure & Topic Modelling/Extraction



However, even though the clustering does not have clear/obvious significance, we know there are only 5 departments . *The 'Trend' category consists of reviews of items from other departments and may be set aside.* So , Likelihood of 5 matching topics.

In both extraction techniques, the top-10 most frequent words are determined for each topic in order of decreasing importance

Exploring Clusters



Term Freq-
Inverse Doc Freq
+
Uniform Manifold Approx. & Projection

Extracting Topics with LDA

CountVectorizer +

Latent Dirichlet Allocation

```
Extracting Topics with LDA...

Topic #0: usual blouse sheer thing bra washed fabric trying suit wash

Topic #1: arrived money product hole cheap retailer disappointed hopes package taking

Topic #2: sweater sleeves warm pockets long soft weight wear coat like

Topic #3: love great color wear fit jeans comfortable perfect soft bought

Topic #4: dress size like fit small just fabric ordered love wear
```

Tops
Intimates
(possibly)
Jackets
Bottoms
Dress

Extracting Topics with NMF

TextTfidfVectorizer +

NMF (Non-Negative Matrix Factorization)

```
Extracting Topics with NMF...

Topic #0: size small large medium ordered runs usually fit wear true

Topic #1: dress beautiful flattering love perfect slip wear comfortable dresses summer

Topic #2: love great jeans comfortable perfect soft wear color pants sweater

Topic #3: like just really fabric look nice good didn color fit

Topic #4: shirt cute super white little material boxy really underneath soft
```

Jackets/
Intimates
Dress
Bottoms
Jackets/
Intimates
Tops

Sentiment Analysis

[Confide

Sentiment Intensity Analyzer (NLTK)

Quantifies sentiment through lexicon-based scoring algorithm: Compound score of overall sentiment polarity, ranges from -1 (most negative) to 1 (most positive)

```
--------------------------------------------------------------------------------
Negative comment ranked 1
--------------------------------------------------------------------------------
I agree with the other two reviews that this shirt runs large, or something? ! i ordered an xs, my usual size, xs-s. the body o
f the shirt is very loose, the upper arms are a bit snug, and i don't have large arms. the shoulder seams are in a very strange
place, but when you look at the model, it doesn't look so bad. with the shirt i received, and another reviewer, the seam was pu
ffy, making for a strange look. in theory, the shirt should have worked, but didn't. i loved the pattern, not the fit.
--------------------------------------------------------------------------------
Negative comment ranked 2
--------------------------------------------------------------------------------
I'll start by saying, over the years, i get more and more frustrated by the lowered quality of retailer's products. this dress
represents all of my frustration. cheap, cheap cheap material. low thread count. elastic!!! around the end of the sleeves that
rubs the cheap fabric against your arm. i mean, this dress is a total disaster - and looks nothing like the photograph shown.
--------------------------------------------------------------------------------
Negative comment ranked 3
--------------------------------------------------------------------------------
As another reviewer wrote.. this is not the same jean featured on the website. same cut and style but it's a light denim wash w
ith wiskering and worn look at the knees. i called customer service to see if this was a mistake and they said there was not re
ally a way to check it unless i exchanged them. i didnt want the same wrong jeans again so i opted out ... too bad they are cur
e.
```

```
Positive comment ranked 1
--------------------------------------------------------------------------------
This top is very gorgeous and chic. it is very tight at the bottom, but i feel like it will stretch a little over time, it is a
lso very long which i love, because it can be rouched up to a great length if you are curvy. both colors are great, i bought bo
th and love both. the sizing was correct, if you want it to fit like it does on the model shot. if you want it a little more fi
tted, size down, but it may be very tight on the bottom. all in all i love them and they are definitely a great find fo
--------------------------------------------------------------------------------
Positive comment ranked 2
--------------------------------------------------------------------------------
I saw this top online and fell in love with the overall casual look.
i ordered all 3 colors, and they just arrived........love love love !!!
they are soft, comfortable and a have a very flattering casual fit.
the understated ruffled details add the special "something" that is so retailer.
they are sophisticated yet pretty.
the orange is more vibrant in person and is a beautiful shade.
the plum/purple is a rich, unique and pretty color.
the navy is a perfect "denim" color.
a fantastic basic
--------------------------------------------------------------------------------
Positive comment ranked 3
--------------------------------------------------------------------------------
I rarely write reviews, but this one deserves one. this top is completely, utterly, adorable! it flows so beautifully, which ma
kes it perfect to wear with leggings/skinny jeans. i wore mine with a blazer for a work holiday party, and i received so many c
ompliments. this top makes you feel so feminine and pretty. i love the maeve brand, and this is certainly no exception! i'd lov
e to get it in multiple colors!
```

Considering we have 5 labels of sentiment in our dataset, we initially try to map the compound score to these labels or levels: 'compound' <-0.56586 :1, -0.56586 <compound<-0.17552: 2, -0.17552 <compound<0.21482: 3 0.21482 <'compound<0.60516 :4, and 5 otherwise. An initial exploratory Multinomial Naive Bayes classifier showed poor metrics for low ratings .

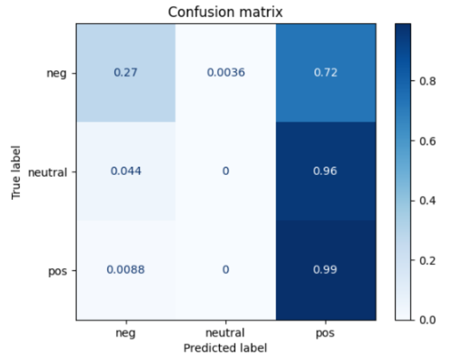
	precision	recall	f1-score	support
1	0.00	0.00	0.00	148
2	0.00	0.00	0.00	309
3	0.50	0.00	0.00	588
4	0.14	0.00	0.00	939
5	0.56	1.00	0.72	2545
accuracy			0.56	4529
macro avg	0.24	0.20	0.15	4529
weighted avg	0.41	0.56	0.41	4529

Logistic Regression

From analyzer: compound'> 0.05 'pos'. 'compound'< -0.05, it is assigned as 'neg'. Otherwise 'neutral'.
Pipeline combining text feature extractor using TF-IDF vectorization and logistic regression classifier:

Grid Search for best hyperparameters, cross-validation and evaluating accuracy.

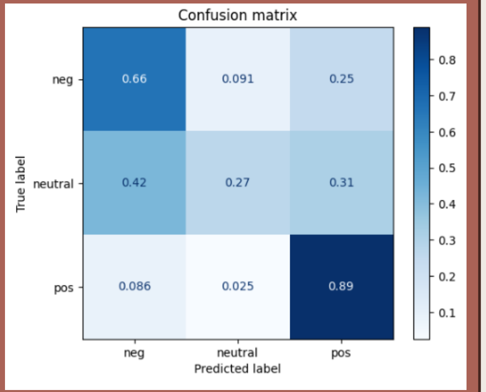
Training and Metrics



Up-sampling Minority Class

RandomOverSampler used to address dataset bias to 5-star reviews.

Logistic regression model fitted on the up-sampled data




```
Enter text: As another reviewer wrote.. this is not the same jean featured on the website. same cut and style but it's a light
denim wash with wiskering and worn look at the knees. i called customer service to see if this was a mistake and they said ther
e was not really a way to check it unless i exchanged them. i didnt want the same wrong jeans again so i opted out ... too bad
they are cure.
Predicted sentiment: neg
```

```
Enter text: Horrible fir, will return
Predicted sentiment: neg

Enter text: mediocre but like the color
Predicted sentiment: pos

Enter text: Not sure, bought 2 items, not very impressive
Predicted sentiment: neg

Enter text: Love it. So perfect.
Predicted sentiment: pos
```

Accuracy: 0.8690660189887393

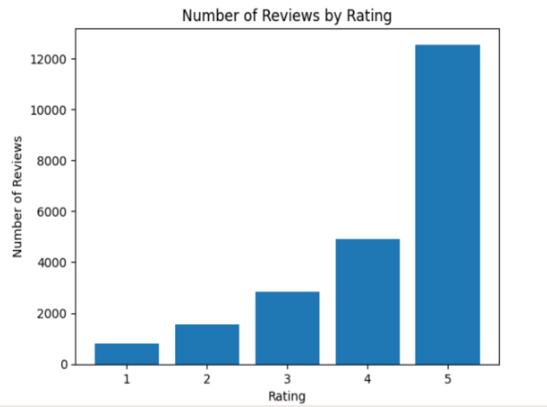
Classification Report:

	precision	recall	f1-score	support
neg	0.33	0.66	0.44	276
neutral	0.08	0.27	0.13	45
pos	0.98	0.89	0.93	4208
accuracy			0.87	4529
macro avg	0.46	0.61	0.50	4529
weighted avg	0.93	0.87	0.89	4529

Sentiment Analysis

Bert-base model+Neural Network Layers (Tensorflow)

BERT base model with layers trained to 90.2 %. Model evaluated . Text classified with rating



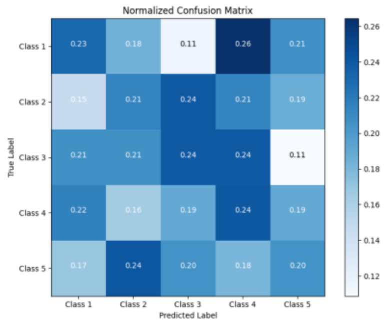
Classification Report:

	precision	recall	f1-score	support
0	0.04	0.06	0.04	84
1	0.05	0.04	0.05	157
2	0.13	0.15	0.14	295
3	0.25	0.20	0.22	493
4	0.55	0.56	0.56	1243
accuracy			0.38	2272
macro avg	0.20	0.20	0.20	2272
weighted avg	0.38	0.38	0.38	2272

Undersampling for equal no. of samples for each rating jeopardizes prediction of 5 rating. Possible overfitting.

Classification Report:

	precision	recall	f1-score	support
0	0.25	0.23	0.24	87
1	0.19	0.21	0.20	75
2	0.27	0.24	0.25	92
3	0.20	0.24	0.22	79
4	0.23	0.20	0.22	83
accuracy			0.23	416
macro avg	0.23	0.23	0.23	416
weighted avg	0.23	0.23	0.23	416




```
Enter text: no
1/1 [==============================] - 5s 5s/step
Predicted class: 1

Enter text: Horrible fir, will return
1/1 [==============================] - 0s 440ms/step
Predicted class: 1

Enter text: Love it. So perfect
1/1 [==============================] - 0s 426ms/step
Predicted class: 5

Enter text: Not sure, bought 2 items, not very impressive
1/1 [==============================] - 0s 410ms/step
Predicted class: 2

Enter text: mediocre but like the color
1/1 [==============================] - 0s 428ms/step
Predicted class: 3

Enter text: As another reviewer wrote.. this is not the same jean featured on the website. same cut and style but it's a light
denim wash with wiskering and worn look at the knees. i called customer service to see if this was a mistake and they said ther
e was not really a way to check it unless i exchanged them. i didnt want the same wrong jeans again so i opted out ... too bad
they are cure.
1/1 [==============================] - 0s 445ms/step
Predicted class: 2

Enter text:
```

Solution: To Use pretrained sentiment model to fine-tune on our dataset: Pytorch+NLPtown (nlptown/bert-base-multilingual-uncased-sentiment)- need for discrete GPU

Text Summarization (with BART)

1. Reducing lengthy reviews to be more cost and time efficient
2. Text encoded with BART tokenizer. Generate method of the BART model uses encoded to summarize
3. Denoising encoder --> reconstruct original text after "noise" integration.

Why BART?

1. Autoencoder --> NN for unsupervised learning
2. Autoregressive --> Sequential data predictions on the previously generated outputs

```
Summarizing text with BART...

Original: I had such high hopes for this dress and really wanted it to work for me. i initially ordered the petite small (my us
ual size) but i found this to be outrageously small. so small in fact that i could not zip it up! i reordered it in petite medi
um, which was just ok. overall, the top half was comfortable and fit nicely, but the bottom half had a very tight under layer a
nd several somewhat cheap (net) over layers. imo, a major design flaw was the net over layer sewn directly into the zipper - it
c

Summary: </s><s>The top half was comfortable and fit nicely, but the bottom half had a very tight under layer and several somew
hat cheap (net) over layers. imo, a major design flaw was the net over layer sewn directly into the zipper.</s>
```

‘Conversing’ with the Data

FLAN-T5-base model

First Approach

Input : code snippet and a question

Generates: Tokenizes combo of code snippet and question and models the answer

```
code_snippet = """
def calculate_area(length, width):
    return length * width
"""
question = "What does the calculate_area function do?"
answer_question(code_snippet, question)
```

```
Based on the text, it appears that return length * width
```

```
# Use the function to answer a question
code_snippet = """
 I had such high hopes for this dress and really wanted it to work for me. i initially ordered the petite small (my usual size) but i found this to be outrageously smal
question = "What is the topic?"
answer_question(code_snippet, question)
```

```
Based on the text, it appears that dress for valentine's day
```

```
# Use the function to answer a question
code_snippet = """
As another reviewer wrote.. this is not the same jean featured on the website. same cut and style but it's a light denim wash wit
"""
question = "Did the customer call customer service?"
answer_question(code_snippet, question)
```

```
Based on the text, it appears that i called customer service to see if this was a mistake and they said there was not really a
way to check it unless i exchanged them. i didn't want the same wrong jeans again so i opted out... too bad they are cure.
```

‘Conversing’ with the Data

FLAN-T5-base model

Second Approach

Only takes a question as input and directly tokenizes the question and generates the output answer using the model.

```
# Use the function to answer a question
question = "What is the name of the sea around Greece?"
answer_question2(question)
```

```
Answer: aegean sea
```

```
# Use the function to answer a question
question = "How are you doing today?"
answer_question2(question)
```

```
Answer: I'm fine.
```

```
# Use the function to answer a question
question = "Who was president of U.S that got assassinated?"
answer_question2(question)
```

```
Answer: john f kennedy
```

```
# Use the function to answer a question
question = "Write a tagline for the AI"
answer_question2(question)
```

```
Answer: The AI's mission is to make the world a better place to live, and to make the world a better place to live, and to make
the world a better place to live, and to make the world a better place to live, and to make the world a better place to live, a
nd to make the world a better place to live, and to make the world a better place to live, and to make the world a better place
to live, and to make the world a better place to live, and to make the world a better place to live, and to make the world a be
tter place to live, and to make the world a better place to live, and to make the world a better place to live, and to make the
world a better place to live, and to make the world a better place to live, and to make the world a better place to live, and t
o make the world a better place to live, and to make the world a better place to live, and to make the world a better place to
live, and to make the world a better place to live, and to make the world a better place to live, and to make the world a bette
r place to live, and to make the world a better place to live, and to make the world a better place to live, and to make the wo
rld a better place to live, and to make the world a better place to live, and to make the world a better place to live, and to
make the world a better place to live, and to make the world a better place to live, and to make the world a better place to li
ve, and to make the world a better place to live, and to make the world a better place to live, and to make the world a better
place to live, and to make the world a better place to live, and to make the world a better place to live, and to make the worl
d a better place to live, and to make the world a better place to live, and to make the world a better place to live, and to ma
ke the world a better place to live, and to make the world a better place to live, and to make the world a better place to liv
e, and to make the world a better place to live, and to
```

'Conversing' with the Data

DataFrameChatBot

How we approached it?

We used spacy to load their nlp model.

We created a function that would search for the answer for the premade question.

We created another function that will structure the answer for the premade question.

```python
import spacy
import pandas as pd

class DataFrameChatBot:
    def __init__(self, df):
        self.df = df
        self.nlp = spacy.load("en_core_web_sm")

    def get_columns(self):
        return self.df.columns.tolist()
```

```python
def process_question(self, question):
    doc = self.nlp(question.lower())

    # Identify if question is about columns
    if 'column' in [token.lemma_ for token in doc]:
        # Extract column name directly from the text
        column_names = [column for column in self.df.columns if column.lower() in question.lower()]

        if not column_names:
            return "Please mention the column name in your question."

        # Check for column type related questions
        if 'type' in [token.lemma_ for token in doc]:
            return self.get_column_type(column_names[0])

        # Check for null values in column related questions
        elif 'null' in question.lower() or 'missing' in question.lower():
            return self.get_nulls_in_column(column_names[0])
```

```python
questions = [
    "What is the type of the Title column?",
    "How many null values are in the Age column?",
    "What is the average of the Age column?",
    "What is the total sum of the Age column?",
    "What is the maximum value in the Age column?",
    "What is the minimum value in the Age column?",
    "How many unique values are in the Age column?",
    "Does the Age column exist?",
    "What is the correlation between Age and Salary columns?",
    "Does the Age column has null values?",
    "What is the most frequent category in the Age column?",
    "Can I have a summary of the dataframe?",
    "What is the type of the Rating column?",
    "How many unique values are in the Division Name column?",
    "Does the Recommended IND column have any null values?",
    "What is the most frequent category in the Class Name column?",
    "What is the correlation between Rating and Recommended IND columns?"
]

for question in questions:
    print(f"Q: {question}")
    print(f"A: {chatbot.process_question(question)}")
    print("---")
```

```
Q: What is the type of the Title column?
A: object
---
Q: How many null values are in the Age column?
A: 0
---
Q: What is the average of the Age column?
A: 43.198543813335604
---
Q: What is the total sum of the Age column?
A: 1014561
```

Summary & Future Improvements

- Create a full data pipeline through connecting the results of the methods.
- Take the answers from the LLM model and categorize them, so reviews can automatically assigned. Categorize fresh reviews w.r.t. topics/departments and update dataset/DB.
- Predict behaviour of neutral reviewers (subset of reviews + recommend/not recommend sentiment binary classification)
- Limitations of Dataset: Unusually strong bias to 5 star ratings

Thank you