1. **Introduction**

In this project, we analyze the Amazon Appliances reviews dataset, focusing on extracting and summarizing the longest reviews. We then apply a pretrained Hugging Face model to demonstrate a question-answering task. This documentation provides a step-by-step guide, complete with code explanations and best practices.

1. **Project Setup**

**Environment**

* Python 3.8+
* Necessary libraries: pandas, numpy, transformers, matplotlib

**Data Source**  
We use Appliances.jsonl.gz, which contains reviews in JSON lines format. Make sure it is in the same directory as the script or adjust the code to point to its location.

1. **Data Loading and Preprocessing**

The script includes a helper function parse() to read each line of the gzipped JSON file, followed by the getDF() function, which converts the parsed JSON objects into a pandas DataFrame. After loading, rows with missing text entries are dropped to avoid errors in subsequent processing.

1. **Summarization**

We define a simple function count\_words() to measure the length of each review by word count. We then filter out the reviews that have more than 100 words, picking the first 10 of the longest. A summarizer pipeline from Hugging Face is initialized using the "sshleifer/distilbart-cnn-12-6" model. Each long review is summarized with constraints (min\_length=50, max\_length=80) to ensure a concise summary.

1. **Question-Answering Model**

Next, we initialize a question-answering pipeline with distilbert/distilbert-base-cased-distilled-squad. Given a question and a chunk of text, the model identifies the most likely answer span. This is demonstrated with:

A black screen with text

AI-generated content may be incorrect.

1. **Visualization and Reporting**

We use matplotlib to plot word-count distributions and show how many reviews exceed a certain threshold.

A graph with a blue bar

AI-generated content may be incorrect.

The graph above explains that the histogram has a large peak at or near the left side, indicating that most of the reviews are relatively short. And beyond the peak, there are fewer reviews with higher word counts (such as 500 or more words), which is typical for datasets where most reviews are short.

A graph of blue bars

AI-generated content may be incorrect.

The graph above shows the top ten reviews by word count, mostly they are ranging from 2800 words to more than 3500 words which is good for generation good summarization for question and answer task of our model.

A screenshot of a computer

AI-generated content may be incorrect.

The summaries above or key output examples is reported in a table, providing a quick overview of how effectively the model captured the main points from each review. Below are the contents and how they are summarized.

**Original Review:**

im excited to see how the coffee comes out i got 2 mason jars too so i ordered a 2nd one of these so i'll always have 2 in the fridge for the summer<br /><br />UPDATE: my first cup was a bit strong i added way too much coffee i guess but i watered it down with plenty of ice and half and half - a little cinnamon and vanilla - i ordered the 2nd filter and it will be here soon and i plan to have iced coffee all summer ... definitely cutting down on the amount of coffee though !!!!!

**Summary Review:**

i got 2 mason jars too so i ordered a 2nd one of these so i'll always have 2 in the fridge . i added way too much coffee i guess but i watered it down with plenty of ice and half and half . i ordered the 2nd filter and it will be here soon and i plan to have iced coffee all summer .

**Original Review:**

The filter did not fit in my Keurig K50. When closing the keurig, there was still a small gap that left the keurig slightly open. I had to hold down the keurig in order to get it to register that it was shut and allow me to select my cup size, and still had to hold it down while it was brewing. Then, the water barely flowed through the filter. Tried with a finer grind as described on the instructions. After a minute of holding it there, I had only less than half a cup of coffee. I tried again with a medium grind and still had the same issue. Decided it was not worth my hassle and to return it. Return was easy and was provided with a free return label.

**Summary Review:**

The filter did not fit in my Keurig K50 . After a minute of holding it there, I had only less than half a cup of coffee . Decided it was not worth my hassle and to return it . Return was easy and was provided with a free return label .

1. **Testing**

The question-answering pipeline works by breaking down the context (the review text) and analyzing it to locate the most relevant span that answers the question. The model performs this by looking at patterns in the text and using its understanding (learned from the SQuAD dataset) to find the answer. When it finds a possible answer, it uses a tokenization method to map the start and end positions of the answer span.

A screen shot of a computer code

AI-generated content may be incorrect.

In our testing, we are using the distilbert/distilbert-base-cased-distilled-squad model from Hugging Face's pipeline to perform a question-answering task on a review from the report\_summaries DataFrame. The text variable contains the review from which the model will extract an answer. The question is defined as "How many mason jars are there?", and the model will search for the most relevant span of text within the review that answers this question. The reader pipeline, initialized with the distilbert model fine-tuned on the SQuAD dataset, performs the extraction by analyzing the context (the review) and identifying the best matching text span. The model's output is a dictionary containing the answer (e.g., "3 mason jars"), the confidence score (indicating how certain the model is about the answer), and the start and end indices of the answer span within the context. This output is then converted into a pandas DataFrame for easy display and interpretation. This approach allows you to leverage the model's ability to extract specific details from text, making it particularly useful for tasks like answering questions based on product reviews.