Homework 3 - Web scraping

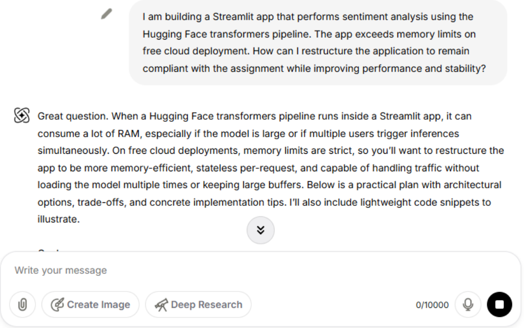
Introduction

This project implements a Streamlit web application for monitoring customer sentiment based on e-commerce reviews from 2023. The app combines web scraping, transformer-based sentiment analysis, and interactive visualizations, and is deployed as a public cloud service.

Web Scraping Strategy

HTML elements were identified using browser developer tools by targeting specific tags and CSS class names containing review text, ratings, product titles, and dates. Pagination links were followed programmatically to collect multiple pages of reviews, and all extracted dates were converted to datetime objects to enable monthly filtering.

AI Implementation



I am building a Streamlit app that performs sentiment analysis using the Hugging Face transformers pipeline. The app exceeds memory limits on free cloud deployment. How can I restructure the application to remain compliant with the assignment while improving performance and stability?

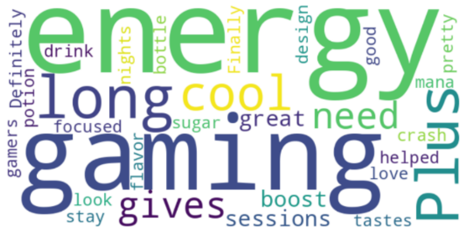
Great question. When a Hugging Face transformers pipeline runs inside a Streamlit app, it can consume a lot of RAM, especially if the model is large or if multiple users trigger inferences simultaneously. On free cloud deployments, memory limits are strict, so you'll want to restructure the app to be more memory-efficient, stateless per-request, and capable of handling traffic without loading the model multiple times or keeping large buffers. Below is a practical plan with architectural options, trade-offs, and concrete implementation tips. I'll also include lightweight code snippets to illustrate.

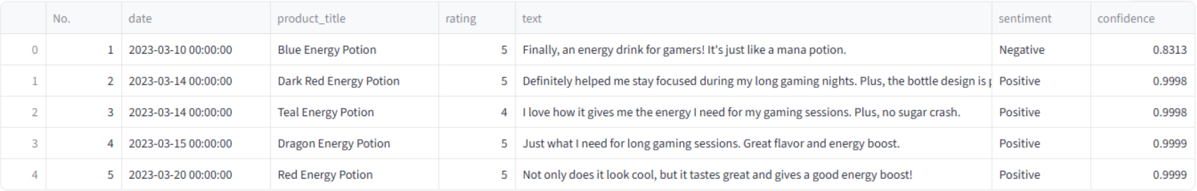
Sentiment analysis was implemented using the Hugging Face transformer model *distilbert-base-uncased-finetuned-sst-2-english* via the transformers pipeline, classifying each review as Positive or Negative with an associated confidence score. Gemini was used as an AI assistant to debug Streamlit and memory issues, refine the sentiment analysis logic, and improve the overall code structure.

Results

HW3 – Web Scraping & Sentiment Analysis (2023)
Reviews (2023)
Sentiment summary
Positive vs Negative (count)
Average confidence by sentiment
Word Cloud (selected month)

Word Cloud (selected month)





	No.	date	product_title	rating	text	sentiment	confidence
0	1	2023-03-10 00:00:00	Blue Energy Potion	5	Finally, an energy drink for gamers! It's just like a mana potion.	Negative	0.8313
1	2	2023-03-14 00:00:00	Dark Red Energy Potion	5	Definitely helped me stay focused during my long gaming nights. Plus, the bottle design is p	Positive	0.9998
2	3	2023-03-14 00:00:00	Teal Energy Potion	4	I love how it gives me the energy I need for my gaming sessions. Plus, no sugar crash.	Positive	0.9998
3	4	2023-03-15 00:00:00	Dragon Energy Potion	5	Just what I need for long gaming sessions. Great flavor and energy boost.	Positive	0.9999
4	5	2023-03-20 00:00:00	Red Energy Potion	5	Not only does it look cool, but it tastes great and gives a good energy boost!	Positive	0.9999

Conclusion

The application demonstrates a complete data mining workflow from web scraping to transformer-based sentiment analysis and visualization. Real-time inference was not implemented due to memory constraints of the free cloud deployment, and sentiment was therefore precomputed to ensure stable performance and reproducibility.

Links

- GitHub Repository: https://github.com/EvaStrasek/tretja_dn_data_mining
- Live Application (Render): https://hw3-streamlit-sentiment.onrender.com/