Class of 2024 – **5th year Engineering cycle**

Major DIA

Leveraging AI for Sustainable Travel

Green Travel Recommendation Web Application Development

SUBMITTED BY: Limin TIAN and Wenbo SUI

WEBSCRAPING & APPLIED ML

26/01/2025

Content

Project Background

Sustainable development is a shared global goal, and green tourism is becoming increasingly important in modern society. While many travel-related websites and blogs provide abundant suggestions and guides to help travelers enjoy their trips, they often overlook the concept of sustainable tourism. Green tourism focuses on minimizing the environmental impact of travel while preserving local culture and natural resources.

However, existing blogs and websites promoting green tourism are fragmented, making it difficult for travelers to quickly access and utilize this information. Thus, creating an integrated platform that consolidates green tourism recommendations can significantly enhance user convenience and promote the widespread adoption of sustainable tourism practices.

Project Objectives

1. **Integrate Green Tourism Information**: Use web scraping techniques to gather scattered green tourism content from multiple websites and blogs.
2. **Develop an Interactive Tool**: Create a web application that enables users to interactively obtain green tourism suggestions.
3. **Promote Sustainable Tourism**: Help users understand how to achieve lightweight, eco-friendly travel while enjoying a pleasant journey.

Technology Stack

Backend Technologies

- **Python**: The core programming language for the project.
- **Flask**: Used to build the server and handle HTTP requests.
- **Data Processing and Storage**:
  Pandas: For reading and processing CSV data.
  PyTorch: To store and load embedding data.

Frontend Technologies

- **HTML**: Used to render static or dynamic web pages through Flask's template engine.

Text Embedding and Natural Language Processing

- **SentenceTransformer**: Provides sentence embedding models (e.g., all-MiniLM-L6-v2).

- **RAG**: Combines retrieval with language generation

Artificial Intelligence and Language Processing

- **GPT API (OpenAI)**: Integrated via Azure OpenAI to generate and enhance model functionalities.
- Natural language understanding and processing

Implementation Process

We implemented a RAG (Retrieval-Augmented Generation) system for sustainable tourism Q&A using Python, Flask, and modern NLP technologies. The backend architecture integrates Flask for HTTP request handling, Pandas for CSV operations, and PyTorch for embedding storage. Web content was collected through BeautifulSoup scraping, cleaned, and structured into CSV format.

When processing queries, the system employs a multi-stage pipeline that combines SentenceTransformer (all-MiniLM-L6-v2) for text embeddings with Azure OpenAI's GPT model. The semantic search component calculates cosine similarity between query and article embeddings to identify the three most relevant articles, which are then concatenated as context for GPT response generation. This retrieval-augmented approach ensures responses are grounded in factual tourism data rather than relying solely on model training. The evaluation pipeline continuously monitors system performance through comprehensive logging of question-answer pairs, search results, and similarity metrics.

1. **Requirement Analysis and Planning**

- Define the goals and features of the green tourism recommendation system.
- Identify data sources and establish the technology stack.

2. **Data Collection and Processing**

- Use web scraping tools (BeautifulSoup) to gather relevant green tourism content.
- Clean and structure the data, saving it in CSV files.

3. **Text Embedding Generation**

- Generate text embeddings using the SentenceTransformer model.
- Cache embedding files (.pt format) to improve processing efficiency.

4. **Backend Development**

- Build a server using Flask and set up routes and APIs.
- Integrate Pandas and PyTorch to handle user queries and return recommendations.

5. **Frontend Development**

- Render HTML pages using Flask's template engine.
- Implement simple user input and result display functionality.

6. **Artificial Intelligence Integration**

- Use the Azure OpenAI API to enhance query generation and recommendation capabilities.
- **RAG** (Retrieval-Augmented Generation)

7. **Testing and Optimization**

- Test the performance of web scraping, model embedding, and APIs.
- Optimize code and system resource usage.

Outcomes

Web Scraping

We completed the web crawling for the specified sustainable travel websites. We utilized the requests library for page retrieval with configured unified headers to simulate browser behavior. For content parsing, we implemented BeautifulSoup to parse HTML, using multi-level selectors to precisely target content while performing text cleaning and formatting.

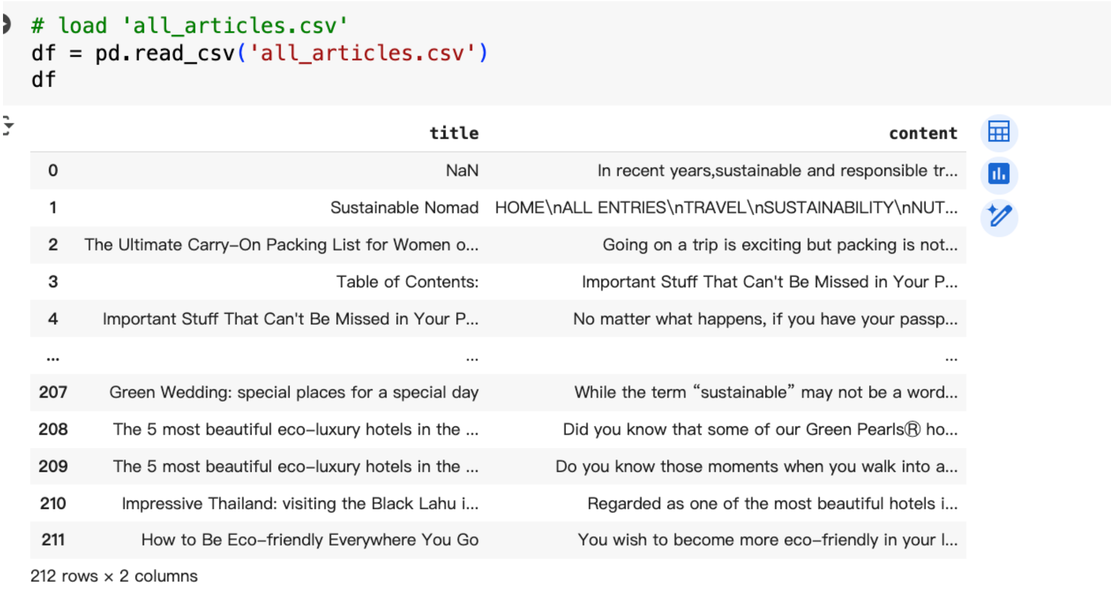
Target Websites

1. greenglobaltravel.com
2. sustainablenomad.blog
3. sustainabletravel.org
4. nationalgeographic.com/travel
5. greensuitcasetravel.com
6. greenerasmus.org
7. nettzero.com.au
8. green-travel-blog.com
9. thegreenpick.com
10. traveldifferently.org

For technical implementation, we developed a retry mechanism to handle intermittent failures and implemented delay controls to comply with server access limitations. We designed flexible parsing strategies to accommodate varying HTML structures across websites. All collected data was stored in CSV format with a two-column structure (title and content) using UTF-8 encoding to ensure proper special character display.

Data processing

We processed the data using Python's natural language processing libraries, including NLTK and spaCy, to analyze the content of our sustainable travel articles.



```
# load 'all_articles.csv'
df = pd.read_csv('all_articles.csv')
df
```

	title	content
0	NaN	In recent years,sustainable and responsible tr...
1	Sustainable Nomad	HOME\nALL ENTRIES\nTRAVEL\nSUSTAINABILITY\nNUT...
2	The Ultimate Carry-On Packing List for Women o...	Going on a trip is exciting but packing is not...
3	Table of Contents:	Important Stuff That Can't Be Missed in Your P...
4	Important Stuff That Can't Be Missed in Your P...	No matter what happens, if you have your passp...
...	...	...
207	Green Wedding: special places for a special day	While the term “sustainable” may not be a word...
208	The 5 most beautiful eco-luxury hotels in the ...	Did you know that some of our Green Pearls® ho...
209	The 5 most beautiful eco-luxury hotels in the ...	Do you know those moments when you walk into a...
210	Impressive Thailand: visiting the Black Lahu i...	Regarded as one of the most beautiful hotels i...
211	How to Be Eco-friendly Everywhere You Go	You wish to become more eco-friendly in your l...

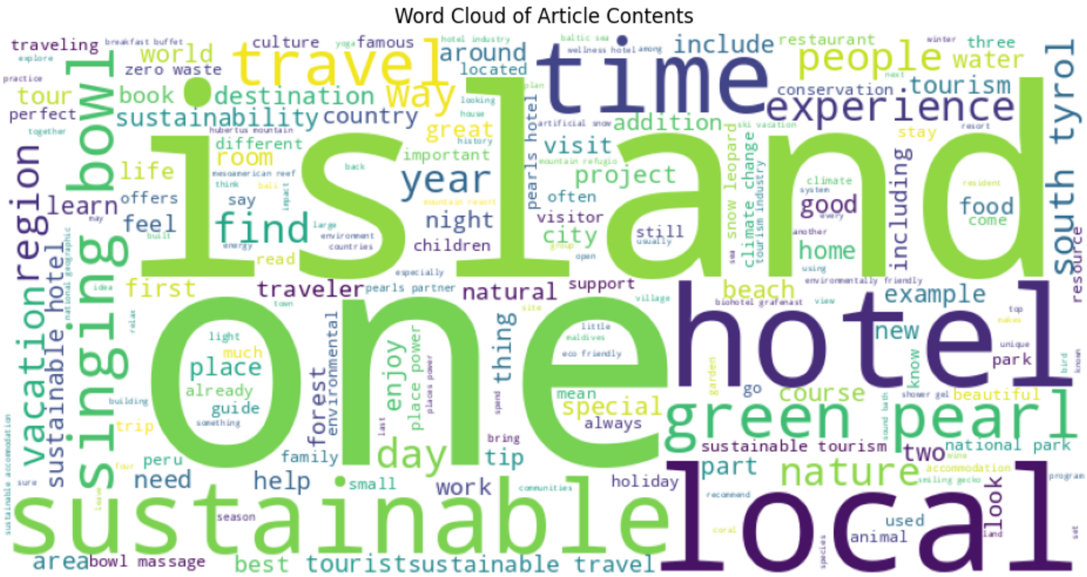
212 rows × 2 columns

Text preprocessing pipeline

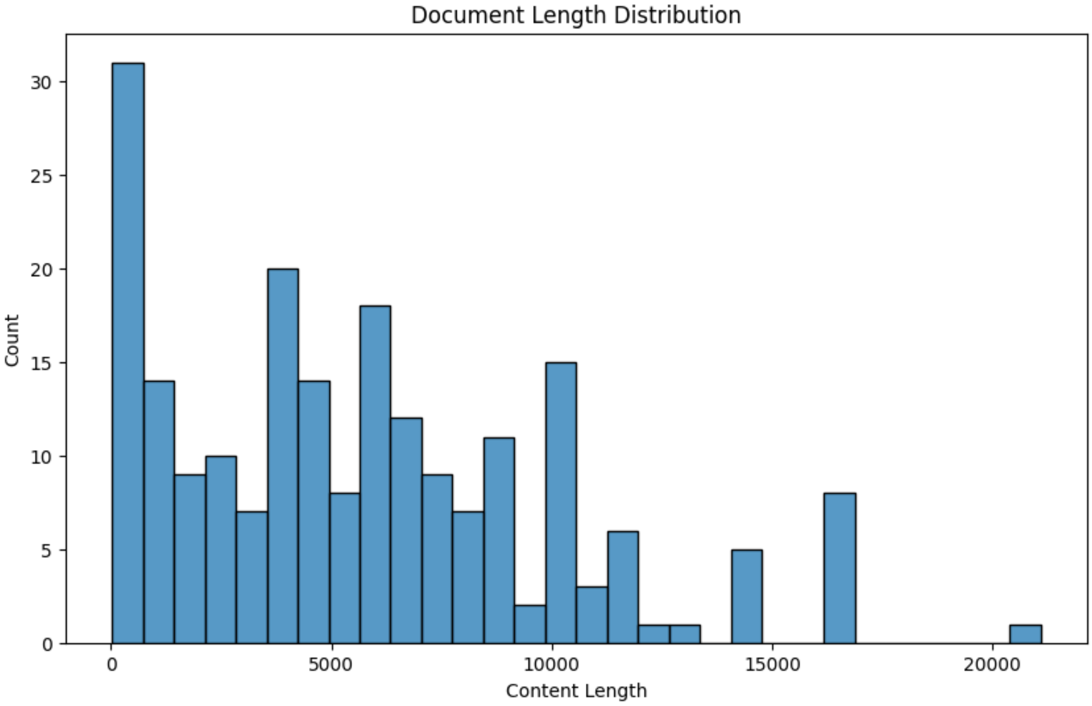
- Custom stopwords removal combining NLTK's English stopwords and domain-specific terms like 'offer', 'even', 'without'
- Text lemmatization using both NLTK's WordNetLemmatizer and spaCy's built-in lemmatizer
- Special character removal and text standardization using regex
- Case normalization to lowercase

Analyze key patterns

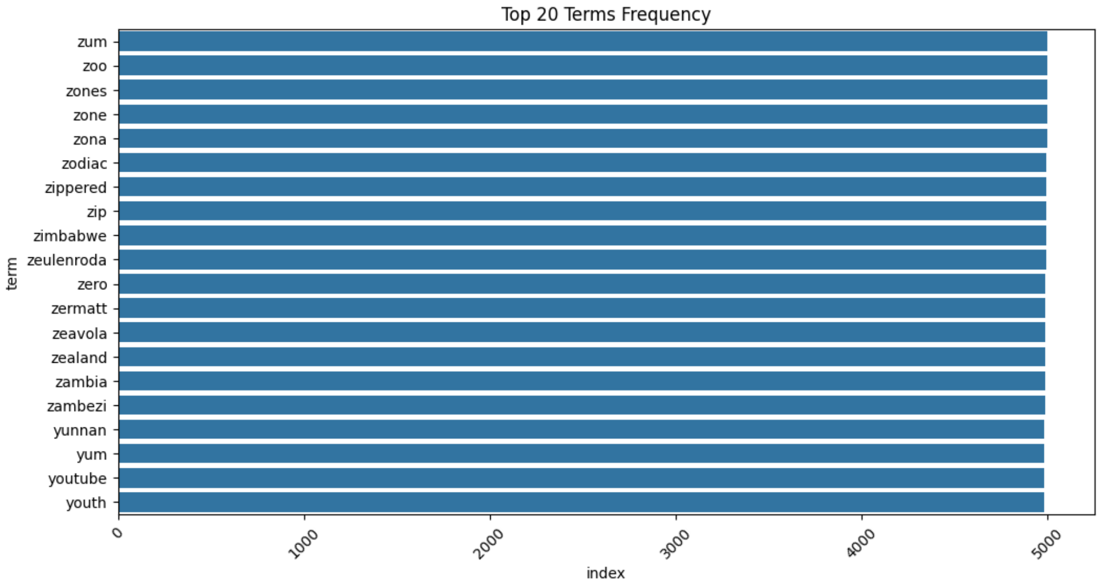
The word cloud visualization revealed dominant themes in our content, with "sustainable," "local," "hotel," and "travel" appearing as the most prominent terms. This aligned with our content strategy focusing on eco-friendly and local tourism experiences.



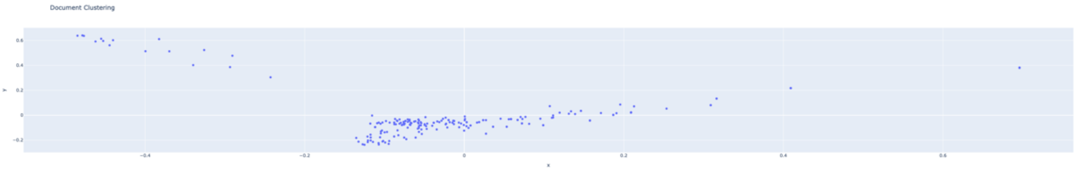
Document length analysis showed that most of our articles fell between 2,000 to 3,000 characters, with a clear right-skewed distribution. A small number of comprehensive guides and in-depth reviews exceeded 15,000 characters, representing our long-form content.



Our term frequency analysis highlighted the vocabulary distribution, with sustainability-related terms appearing consistently across articles. The top 20 terms frequency chart showed a diverse range of travel-related vocabulary, indicating well-balanced content coverage.

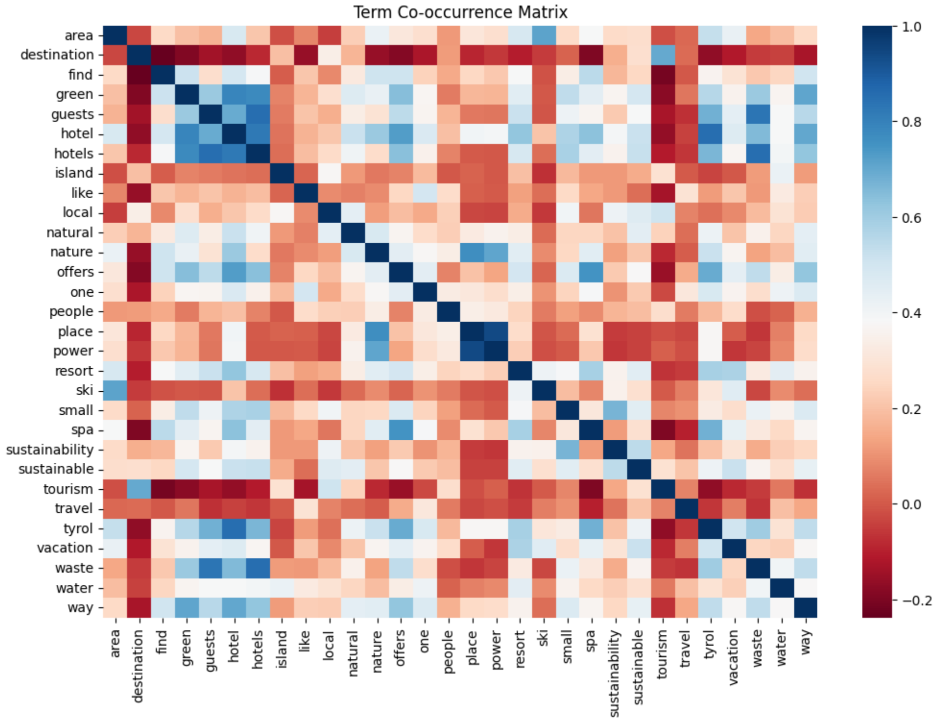


The document clustering visualization revealed distinct content groupings. The main cluster centered around (0,0) represented our core sustainable travel content, while outlier clusters indicated specialized topics like eco-luxury hotels and green wedding destinations.



The term co-occurrence matrix demonstrated strong relationships between key concepts. We found significant correlations between "sustainable" and "tourism," as well as "hotel" and "local," reflecting our focus on promoting environmentally conscious local travel experiences.

Token analysis comparing NLTK and spaCy processing showed similar patterns but with slightly different counts (NLTK mean: 530, spaCy mean: 499), validating the robustness of our text processing approach.



Data Exploration

We conducted exploratory data analysis on our preprocessed sustainable travel content dataset containing 212 articles. The analysis focused on content length, keyword distribution, readability metrics, and sentiment patterns.

Content Length Analysis

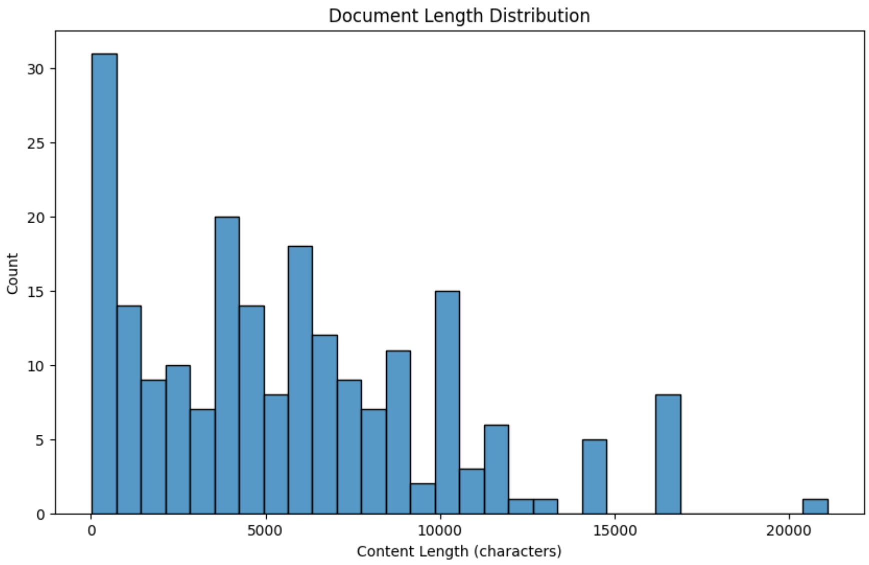
- Total articles: 212

- Average length: 5,687 characters

- Length range: 22 to 21,092 characters

- Standard deviation: 4,384 characters

- Most articles fell within 2,000-8,000 characters



Keyword Distribution

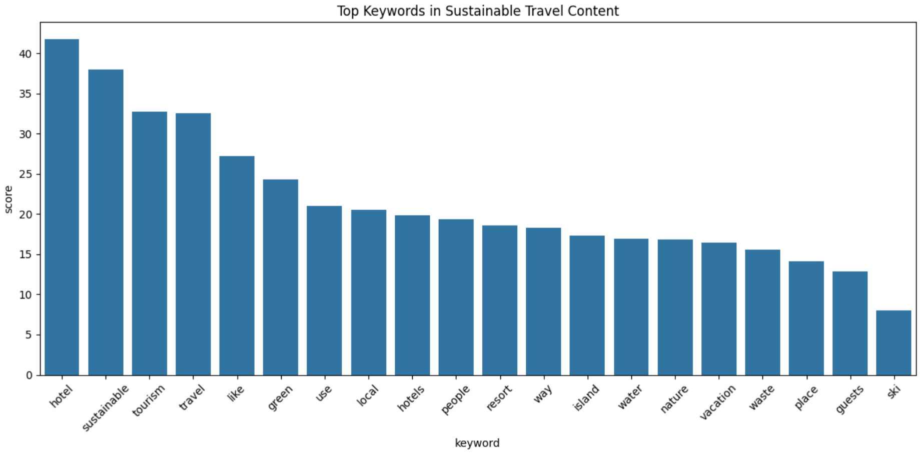
- Primary terms: hotel (41), sustainable (38), tourism (32)

- Secondary themes: green, local, nature

- Geographic terms: island, water

- Service terms: resort, guests showed lower frequency

- Keywords reflect strong alignment with sustainable tourism focus



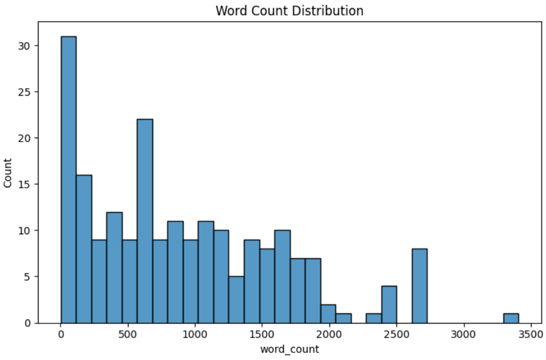
Content Metrics

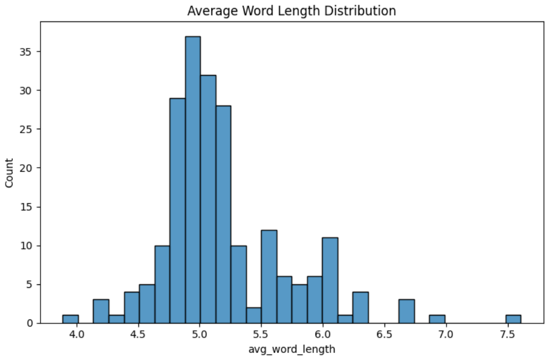
- Average word count: 921 words/article

- Mean word length: 5.19 characters

- Word count distribution showed normal curve around 900-1000 words

- Complexity level remained consistent across corpus





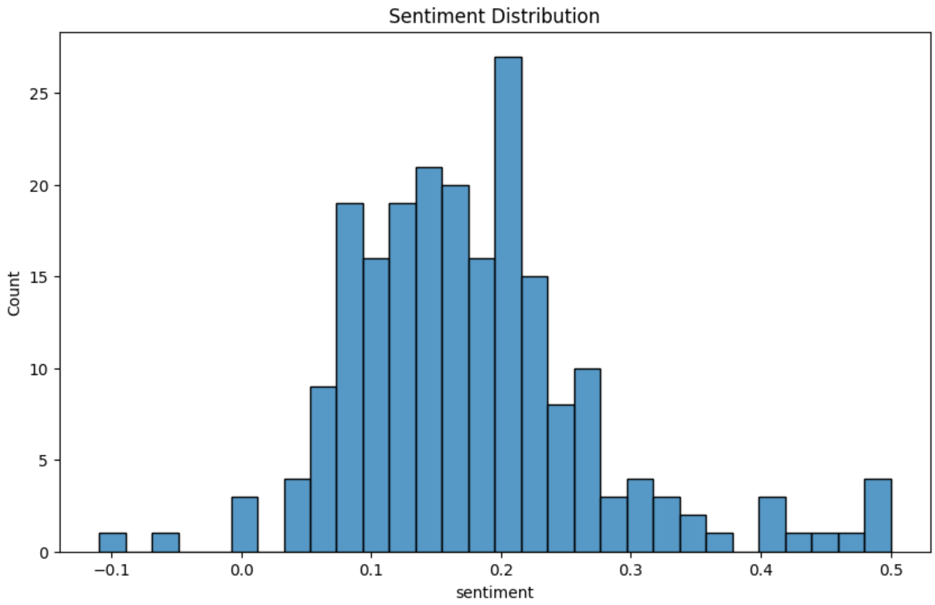
Sentiment Patterns

- Mean sentiment: 0.17 (positive bias)

- Distribution peak: 0.2

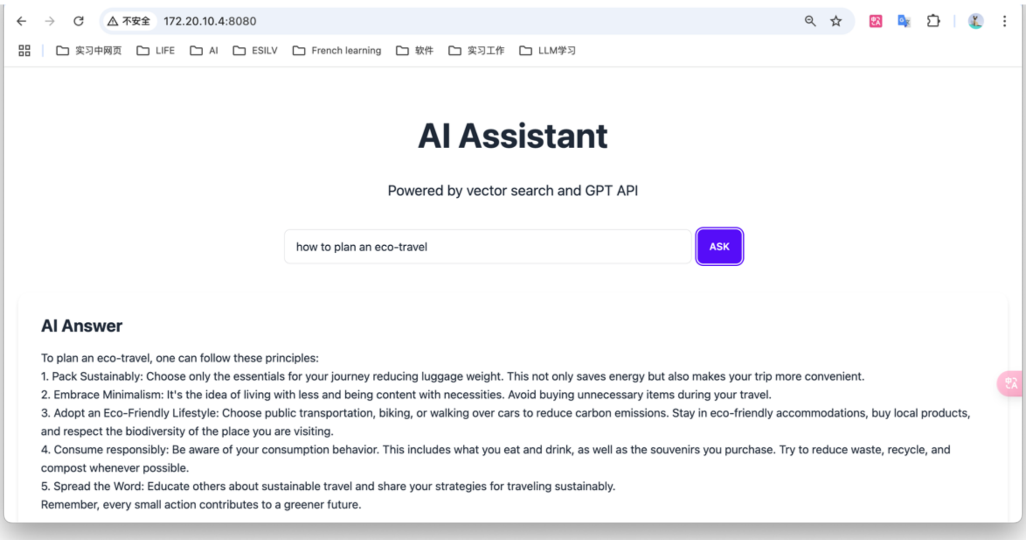
- Average subjectivity: 0.40

- Content maintained professional tone while incorporating experiential elements



Evaluation the performance

To evaluate our sustainable tourism Q&A system's performance, we designed a comprehensive test set of 20 questions and analyzed response quality using semantic similarity metrics. This evaluation aimed to assess the system's ability to provide accurate, relevant answers across key sustainable tourism topics.



AI Assistant

Powered by vector search and GPT API

how to plan an eco-travel

AI Answer

To plan an eco-travel, one can follow these principles:
1. Pack Sustainably: Choose only the essentials for your journey reducing luggage weight. This not only saves energy but also makes your trip more convenient.
2. Embrace Minimalism: It's the idea of living with less and being content with necessities. Avoid buying unnecessary items during your travel.
3. Adopt an Eco-Friendly Lifestyle: Choose public transportation, biking, or walking over cars to reduce carbon emissions. Stay in eco-friendly accommodations, buy local products, and respect the biodiversity of the place you are visiting.
4. Consume responsibly: Be aware of your consumption behavior. This includes what you eat and drink, as well as the souvenirs you purchase. Try to reduce waste, recycle, and compost whenever possible.
5. Spread the Word: Educate others about sustainable travel and share your strategies for traveling sustainably.
Remember, every small action contributes to a greener future.

Created 20 questions across sustainable tourism domains

- Core principles and definitions

- Environmental impact reduction

- Transportation and accommodations

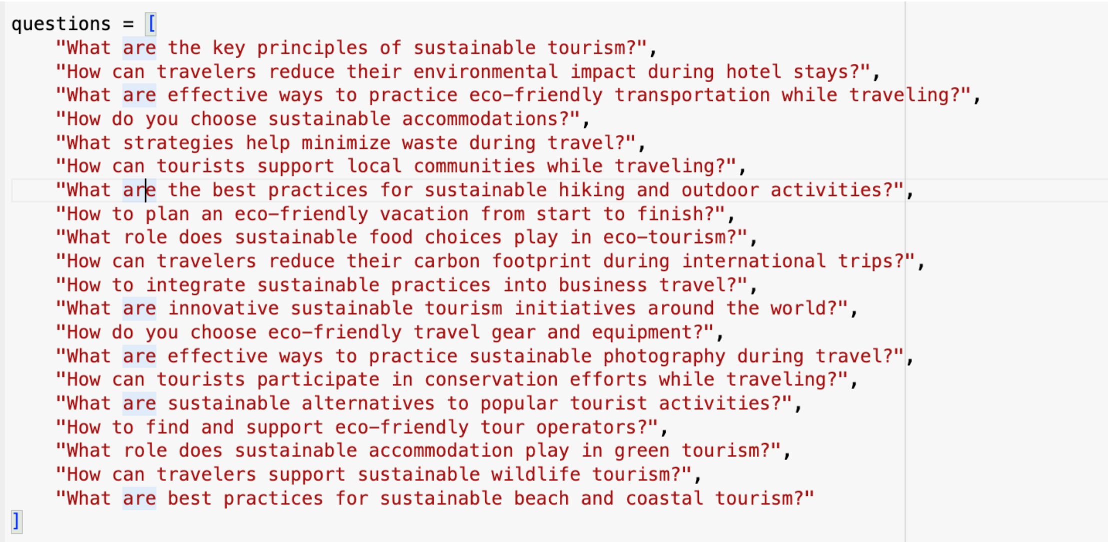
- Local community engagement

- Conservation and wildlife

- Beach and coastal tourism

- Tour operations

- Food and dining practices



```
questions = [
    "What are the key principles of sustainable tourism?",
    "How can travelers reduce their environmental impact during hotel stays?",
    "What are effective ways to practice eco-friendly transportation while traveling?",
    "How do you choose sustainable accommodations?",
    "What strategies help minimize waste during travel?",
    "How can tourists support local communities while traveling?",
    "What are the best practices for sustainable hiking and outdoor activities?",
    "How to plan an eco-friendly vacation from start to finish?",
    "What role does sustainable food choices play in eco-tourism?",
    "How can travelers reduce their carbon footprint during international trips?",
    "How to integrate sustainable practices into business travel?",
    "What are innovative sustainable tourism initiatives around the world?",
    "How do you choose eco-friendly travel gear and equipment?",
    "What are effective ways to practice sustainable photography during travel?",
    "How can tourists participate in conservation efforts while traveling?",
    "What are sustainable alternatives to popular tourist activities?",
    "How to find and support eco-friendly tour operators?",
    "What role does sustainable accommodation play in green tourism?",
    "How can travelers support sustainable wildlife tourism?",
    "What are best practices for sustainable beach and coastal tourism?"
]
```

System Performance Metrics

- Average similarity score: 0.75 (benchmark: 0.80)

- High-performing answers (>0.80): 20%

- Score distribution: 0.675-0.875

- Normal distribution with peak at 0.75-0.80

Strengths

- Environmental impact questions (0.845)

- Transportation practices (0.815)

- Basic principles (0.757)

- Wildlife tourism (0.762)

Areas for Improvement

- Accommodation selection (0.695)

- Waste management strategies (0.760)

- Local community engagement (0.723)

- Food sustainability (0.715)

Recommendations

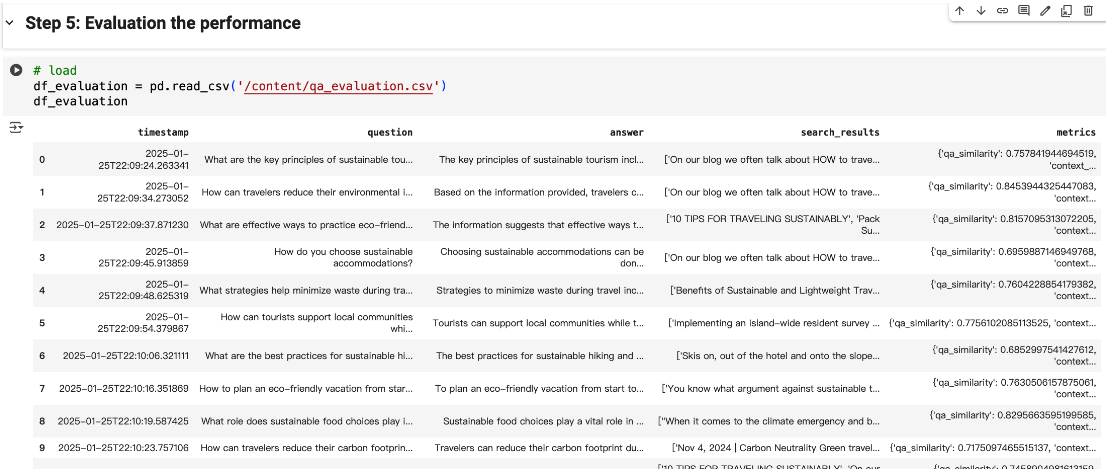
1. Enhance content coverage for accommodation topics

2. Expand waste management examples

3. Add more specific local community interaction cases

4. Include detailed food sustainability guidelines

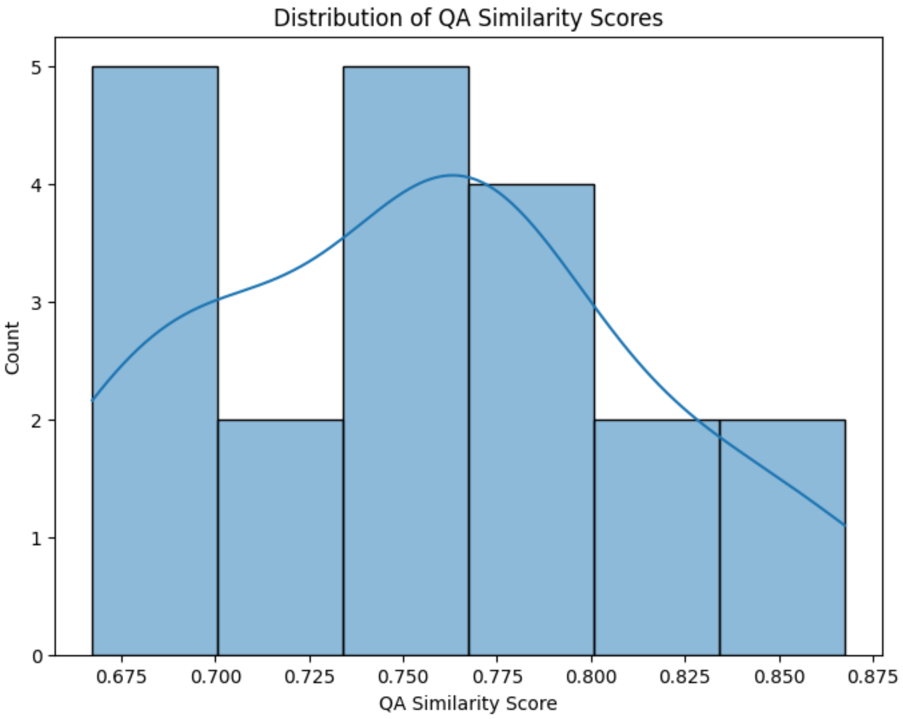
The QA system shows promise in core sustainable tourism concepts but needs refinement in practical implementation areas.



Step 5: Evaluation the performance

```
# load
df_evaluation = pd.read_csv('/content/qa_evaluation.csv')
df_evaluation
```

	timestamp	question	answer	search_results	metrics
0	2025–01–25T22:09:24.263341	What are the key principles of sustainable tou...	The key principles of sustainable tourism incl...	['On our blog we often talk about HOW to trave...	{'qa_similarity': 0.757841944694519, 'context_...
1	2025–01–25T22:09:34.273052	How can travelers reduce their environmental i...	Based on the information provided, travelers c...	['On our blog we often talk about HOW to trave...	{'qa_similarity': 0.8453944325447083, 'context...
2	2025–01–25T22:09:37.871230	What are effective ways to practice eco–friend...	The information suggests that effective ways t...	['10 TIPS FOR TRAVELING SUSTAINABLY', 'Pack Su...	{'qa_similarity': 0.8157095313072205, 'context...
3	2025–01–25T22:09:45.913859	How do you choose sustainable accommodations?	Choosing sustainable accommodations can be don...	['On our blog we often talk about HOW to trave...	{'qa_similarity': 0.6959887146949768, 'context...
4	2025–01–25T22:09:48.625319	What strategies help minimize waste during tra...	Strategies to minimize waste during travel inc...	['Benefits of Sustainable and Lightweight Trav...	{'qa_similarity': 0.7604228854179382, 'context...
5	2025–01–25T22:09:54.379867	How can tourists support local communities whi...	Tourists can support local communities while t...	['Implementing an island–wide resident survey ...	{'qa_similarity': 0.7756102085113525, 'context...
6	2025–01–25T22:10:06.321111	What are the best practices for sustainable hi...	The best practices for sustainable hiking and ...	['Skis on, out of the hotel and onto the slope...	{'qa_similarity': 0.6852997541427612, 'context...
7	2025–01–25T22:10:16.351869	How to plan an eco–friendly vacation from star...	To plan an eco–friendly vacation from start to...	['You know what argument against sustainable t...	{'qa_similarity': 0.7630506157875061, 'context...
8	2025–01–25T22:10:19.587425	What role does sustainable food choices play i...	Sustainable food choices play a vital role in ...	["When it comes to the climate emergency and b...	{'qa_similarity': 0.8295663595199585, 'context...
9	2025–01–25T22:10:23.757106	How can travelers reduce their carbon footprin...	Travelers can reduce their carbon footprint du...	['Nov 4, 2024 | Carbon Neutrality Green travel...	{'qa_similarity': 0.7175097465515137, 'context...
				['10 TIPS FOR TRAVELING SUSTAINABLY', 'On our	{'qa_similarity': 0.7458904981613159



Distribution of QA Similarity Scores
Count
QA Similarity Score