[**WM\_Paper**](https://docs.google.com/document/d/1pScl9BYPOuO5x8eytv5jMAL8JpuG9ULLsPpgjflPpz8/edit?usp=sharing) **This doc contains all steps clearly**

**Recommendation system for an e-commerce website** that uses **web usage mining**

techniques to analyze user behavior and suggest relevant products to users. The system includes several components, including a data generator, data cleaner, data analyzer, Apriori association algorithm, collaborative filtering algorithm, content-based filtering algorithm, hybrid filtering algorithm, and recommender system.

GitHub link:

<https://github.com/yashaswiniravuri/Recommendation-system-using-web-usage-mining>

Meeting-1

Tasks to do

1. Data collection
2. run the data generator to collect and preprocess the web usage data generated by the e-commerce website. [link](https://github.com/yashaswiniravuri/Recommendation-system-using-web-usage-mining/blob/main/data_gen.py) manvi
3. use the data cleaner to clean and preprocess the data to remove irrelevant information, handle missing values, and convert the data into a consistent format. [link](https://github.com/yashaswiniravuri/Recommendation-system-using-web-usage-mining/blob/main/data_cleaning.py) ishita
4. The data analyzer component is responsible for analyzing the data and extracting relevant patterns and associations using the Apriori association algorithm [link](https://github.com/yashaswiniravuri/Recommendation-system-using-web-usage-mining/blob/main/data%20analysis.py) and [link](https://github.com/yashaswiniravuri/Recommendation-system-using-web-usage-mining/blob/main/association_analysis.py) sarim and sunaina
5. collaborative filtering algorithm uses the user-item rating matrix to identify similar users and make personalized recommendations [link](https://github.com/yashaswiniravuri/Recommendation-system-using-web-usage-mining/blob/main/collaborative.py) priyesh
6. the content-based filtering algorithm recommends items that are similar to items a user has liked before based on cosine similarity matrix. [link](https://github.com/yashaswiniravuri/Recommendation-system-using-web-usage-mining/blob/main/content.py) vikas
7. The hybrid filtering algorithm combines both collaborative and content-based filtering to produce more accurate recommendations, which are then presented to the user through the recommender system. [Link](https://github.com/yashaswiniravuri/Recommendation-system-using-web-usage-mining/blob/main/hybrid.py) gaurav

supriya - documents

7 python codes + data collection → 8 people

MEETING 2

<https://www.kaggle.com/datasets/muhammadroshaanriaz/time-wasters-on-social-media>  
  
Analysis

Link prediction

Unreal data

MEETING 3

Final Project Decided after discussing with sir: to try and implement this paper for Indian cities

<https://www.sciencedirect.com/science/article/pii/S026427512300402X>

Step-1: Data Collection

Manvi, Sarim, Gaurav, Ishita

Try to scrape/mine data from official tourist websites of a few cities

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

# Methodology in Paper

Step-1 Sustainable tourism dictionary development

1. The first step (Objective clarification) allowed us to clearly define the aim of the dictionary which is to assess online sustainability communication of tourism destinations providing a tool for fast and systematic processing of large amounts of text.

"To develop a dictionary that helps assess the presence and emphasis of themes like nature, sustainability, and safety on tourism websites to make personalized travel recommendations."

1. identification of categories and categorizing entries. Sub-categories have been identified for each dimension to investigate and capture different aspects within the same category.

[WM\_project\_categories.ipynb](https://colab.research.google.com/drive/1bBuqrnZc0rJjb5YW2CA_HyU1z7-7H1Kc#scrollTo=r3cxOmpLXq50)

**Nature**Focuses on the natural attractions and outdoor experiences a state offers.

* *Sub-categories*: Flora and Fauna, Landscapes, Adventure, Water Activities
* *Examples*: Wildlife, national parks, mountains, beaches, rivers, hiking, camping

**Sustainability** Emphasizes eco-friendly practices and sustainable tourism efforts.

* *Sub-categories*: Environmental Initiatives, Community Support, Eco-friendly Travel
* *Examples*: Carbon-neutral, recycling, eco-tourism, local culture, renewable energy

**Safety** Covers information about how safe or accessible a destination is, often important for families or solo travelers.

* *Sub-categories*: Health and Hygiene, Security, Family-Friendliness
* *Examples*: Safety measures, clean, emergency services, family-friendly, safe trails

**Culture & Heritage** Highlights unique cultural aspects, historical sites, and local traditions.

* *Sub-categories*: Historical Landmarks, Festivals, Local Traditions
* *Examples*: Museums, heritage sites, festivals, traditional crafts, local foods

**Adventure & Activities** Focuses on thrill-seeking opportunities and sports.

* *Sub-categories*: Extreme Sports, Outdoor Activities, Guided Experiences
* *Examples*: Rafting, skydiving, mountain biking, scuba diving, safaris

**Local Cuisine & Dining** Covers food-related aspects, a key interest for many tourists.

* *Sub-categories*: Regional Dishes, Food Tours, Specialty Markets
* *Examples*: Street food, farm-to-table, wine tasting, regional dishes, food festivals

**Accessibility** Focuses on ease of travel for all types of tourists, including those with mobility needs.

* *Sub-categories*: Transportation, Facilities for Disabled, Infrastructure
* *Examples*: Accessible, public transport, ramps, walkways, visitor centers

**Affordability** Emphasizes budget-friendly travel options.

* *Sub-categories*: Budget Accommodations, Affordable Dining, Deals and Discounts
* *Examples*: Cheap, budget, affordable, discounts, hostels

1. we identified the corpus on which the dictionary is developed. We adopted a set of documents that include textual content related to sustainable tourism
   1. a sustainable tourism dictionary (in Italian) previously developed by Marchi et al. (2021).....web content mining
   2. guidelines and indicators systems recognized at the international level,
   3. textual contents of the top 10 European tourism cities by the number of bed nights in 2019 (..... web scraping

Urls ready

Code almost ready [WM\_crawler.ipynb](https://colab.research.google.com/drive/1YkD073VPRJFZPI6vStB-xSZBa7OqbnT_)

1. we prepared the textual contents of the 10 European urban destinations for further analysis through the preprocessing operation (such as stop words removal, reduction of all words to lowercase).

After corpus

1. validation of the dictionary entries. This study adopted the keyword-in-context (KWIC) method, which is an automatic system that allows the search of a particular keyword in the text and analyzes its local meaning in relation to a number of words immediately preceding and following it

Step-2 Sustainable tourism dictionary application

Dictionary ready

Corpus ready

Rank these

<https://scholarworks.lib.csusb.edu/jitim/vol27/iss3/7/>

Yes, the paper **"Inside the Black Box of Dictionary Building for Text Analytics: A Design Science Approach"** outlines a structured methodology for building dictionaries for text analytics. It focuses on using a **Design Science Research (DSR) approach** and provides actionable steps to systematically develop and validate domain-specific dictionaries.

Here are the key steps typically outlined in the paper:

### **1. Clarify the Objective:**

* Define the purpose of the dictionary
  + Making different classes like : Nature, safety, culture, adventure, affordability
* Specify the research questions or analytical goals the dictionary aims to address.
  + Rank states based on these measures based on their official tourism websites’ content

### **2. Develop the Dictionary:**

* **Keyword Identification**:
  + Use domain knowledge, literature reviews, or expert consultation to identify initial terms.
    - Nature: "forest," "beach," "waterfall," "wildlife," "national park," "eco-friendly," "flora."
    - "low crime," "security," "police," "emergency services," "safe zones," "family-friendly."
    - "heritage," "museum," "festival," "tradition," "local art," "historical sites."
    - "hiking," "rafting," "skydiving," "mountain climbing," "extreme sports," "safaris."
    - "cheap," "budget," "discounts," "cost-effective," "affordable hotels," "low-cost travel."
  + Mine existing datasets for relevant terms using techniques like word frequency analysis, topic modeling, or co-occurrence networks.
* **Sub-Categorization**:
  + Organize the dictionary into meaningful subcategories (e.g., group terms by themes like safety, sustainability, or adventure for tourism analysis).
* **Thesaurus Expansion**:
  + Add synonyms, variations, and related terms to ensure broader coverage.
* **Iterative Refinement**:
  + Continuously revise the dictionary as new insights emerge.

### **3. Validate the Dictionary:**

* **Expert Review**:
  + Validate terms and categories with subject-matter experts.
* **Quantitative Validation**:
  + Test the dictionary on a sample dataset and assess its performance (e.g., precision, recall, and coverage metrics).
* **Cross-Validation**:
  + Apply the dictionary to multiple datasets or contexts to ensure robustness and generalizability.

### **4. Apply the Dictionary:**

* Use the dictionary in text mining or analytics tasks such as keyword extraction, sentiment scoring, or trend analysis.
* Compare dictionary performance with other analytical approaches to evaluate its effectiveness.

### **5. Document the Process:**

* Maintain detailed documentation of all steps taken to ensure transparency and reproducibility.
* Include a record of sources, decisions, and any iterative refinements.

### **Key Insights:**

* The process emphasizes both qualitative and quantitative validation to ensure the dictionary is relevant and effective for the intended purpose.
* It leverages the **design science paradigm**, which ensures the dictionary is a rigorously developed and practically useful artifact.

Would you like guidance on applying any specific step to your project?

Cultural Tourism

Guided Toursm

Stay & Local Services

Adventure

Festivals

Religious

Trekking

Cuisines

Hill

Eco-tourism

Coastal Attractions

Beach

Waterfalls