BELBACHIR Yassine - MARTIN Maxence

Project Data Mining

The Goal of Our Project

The objective of this project was to develop an **automated image recommender system** based on **open-access image data retrieved from Wikidata via SPARQL queries**. The system was designed to follow a structured pipeline, from **data acquisition to user preference modeling and recommendation generation**.

However, during the implementation, **several limitations emerged** that prevented the system from functioning correctly. Issues related to **metadata extraction, classification accuracy, and recommendation reliability** significantly impacted the project's success.

Data Sources of Our Images and License

1. Image Sources

Images were retrieved using **SPARQL queries on Wikidata**, specifically targeting images **hosted on Wikimedia Commons**. The dataset was centered around **animals** and was manually structured to ensure a sufficient number of images.

The images were categorized based on **taxonomic classification**, including:

- **Mammals and Birds**
- **Reptiles and Fish**
- **Invertebrates**

2. Image Licensing

All images were obtained via **Wikimedia Commons**.

Size of Our Data

We collected **approximately 400 images** using **Wikidata queries**, representing different **animal taxonomic groups**.

Information That We Decided to Store for Each Image

For each retrieved image, we stored the following information:

- **Image URL** – Direct link to the image hosted on **Wikimedia Commons**.
- **Taxonomic Classification** – Classification of the subject (e.g., species, family).
- **EXIF Metadata** – Image attributes such as format, size, and orientation (when available).
- **Dominant Colors** – Extracted using **K-Means clustering** to detect the **three most prominent colors**.

These details were stored in **JSON format** for further analysis.

Information Concerning User Preferences

The system attempted to build a **user profile** based on **selected images** by storing:

- **Favorite colors** (extracted using K-Means clustering).
- **Preferred image orientation** (landscape, portrait, or square).
- **Most frequently selected tags** (either system-generated or user-defined).
- **Chosen images** (explicit selections by the user).

Although this information was successfully stored, it **was not effectively used in the recommendation system due to classification issues**.

Data Mining and Machine Learning Models Used

1. K-Means Clustering (Color Extraction)

- **Purpose**: Extract dominant colors from images.
- **Implementation**: **K-Means clustering (k=3)** was used to group pixel colors.
- **Outcome**: Color data was successfully extracted, but its use in recommendations was **limited**.

2. Content-Based Filtering (Recommendation System)

Approach: The system attempted to **recommend images based on similarity in colors and tags**.

The filtering **did not work properly**, might be due to **Metadata inconsistencies**.

Self-Evaluation of Our Work

The workload was **evenly distributed** between the two of us.

Despite the many issues, the project provided **valuable experience** in **data mining, SPARQL queries, and image processing techniques**.

Remarks Concerning the Practical Sessions, Exercises, and Scope for Improvement

All practical sessions focused on **analyzing pre-written code rather than writing our own**. While this was useful for understanding **existing implementations**, we believe that **actively writing and debugging our own code** would have been a more effective learning experience.

Conclusion

This project attempted to implement an **automated image recommender system** using **Wikidata queries and machine learning techniques**. While the **image retrieval and annotation steps were successfully implemented**, the **classification and recommendation system did not work as intended**.

Despite the **unsuccessful outcome**, the project provided insights into **automated data processing, SPARQL queries, and the challenges of machine learning in recommendation systems**.