Recommandation System

MARCH 2024

Theo Clere & Maxime Brossard
CPE LYON | 4IRC 2024-2025

Data Mining: Recommendation System

Introduction

In a world where images abound online, it becomes crucial for platforms to recommend relevant visual content to users. Our project aims to address this need by building an image recommendation system, entirely based on Python, that adapts to each individual's preferences.

The main objective is to provide a personalized user experience by recommending images tailored to each user's interests. To achieve this, we will automate all steps, from initial data collection to the implementation of the recommendation system, including data annotation, analysis, and visualization.

Throughout the dedicated sessions of this project, we will focus on automating the process, thereby ensuring maximum efficiency and future scalability of the system. This report will detail each step of the project, highlighting the challenges, solutions, and results obtained, with the aim of providing an accurate and adaptable image recommendation system.

Data Sources and Image Licenses

The data used in this project primarily comes from Wikidata, a collaborative and free knowledge base. Specifically, we conducted a SPARQL query on Wikidata to retrieve information about objects related to the "dog" category. This query provided us with a list of 100 items inheriting from the "dog" object, including various information such as names, descriptions, and links to associated images.

To download these images, we utilize a function to automate the process of retrieving and storing all images in a local directory named "./content/images". This script has been designed to filter only the images from the data retrieved via the SPARQL query and download them directly to our system.

Regarding the licenses of the images, we have ensured to adhere to Wikidata's guidelines regarding data usage. Most images available on Wikidata are under open licenses such as Creative Commons, permitting their use for non-commercial purposes and with attribution. We are committed to respecting all licenses associated with the downloaded images and providing appropriate credits when using them in our project.

Size of your data

The total size of the data used in this project, including images downloaded from Wikidata, amounts to approximately 65 MB. This size encompasses all images extracted from the SPARQL query executed on Wikidata and stored locally in the directory "./content/images".

Managing data size is crucial in a data mining project, as it can impact the performance of the recommendation system we are using. In our case, we deliberately chose to limit our selection to only 100 images. Thus, storing and downloading these images represents the best approach to focus exclusively on processing this data, rather than handling larger volumes that could pose performance issues.

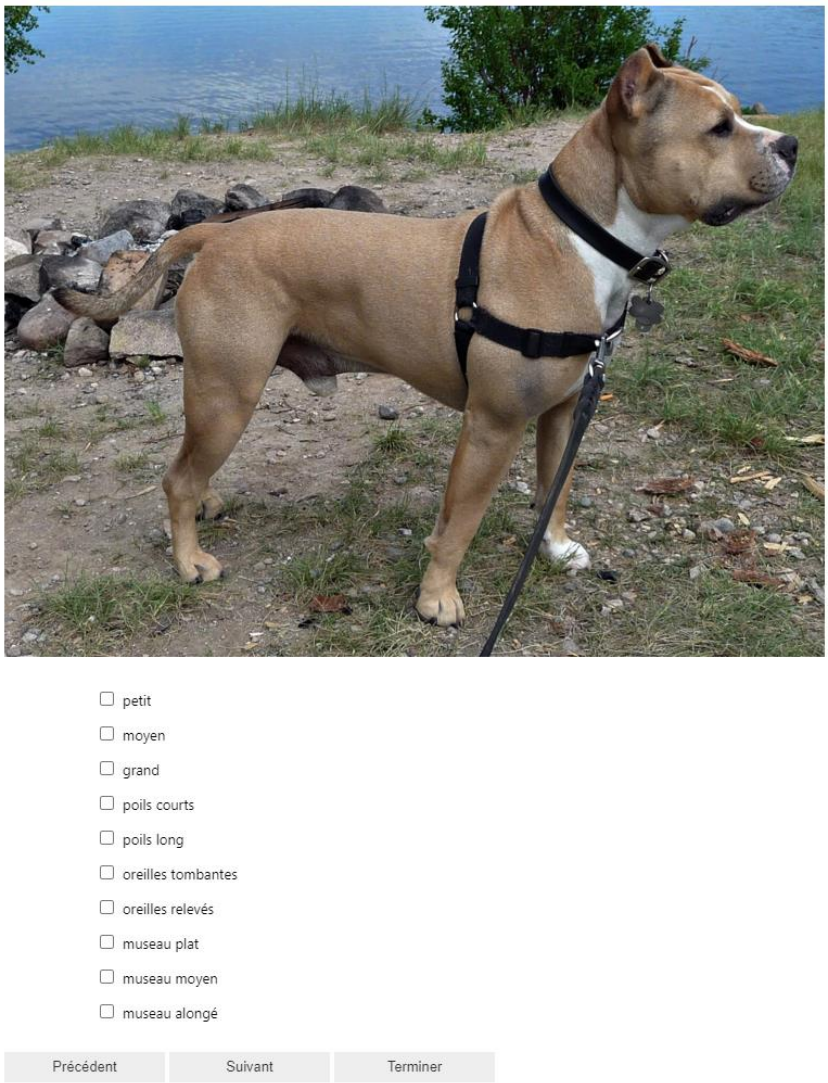
Stored Information

For each image we processed in this project, we took care to collect and store several types of information to enrich our recommendation system. Initially, we extracted the Exif data from each image. These data allowed us to obtain some relevant descriptions as well as information on the orientation of the image. However, we found that this information was quite limited and did not provide a complete overview of each image.

To complement this data, we undertook a second step of extracting the dominant colors from all images. This approach allowed us to add an important visual dimension to our dataset, enabling us to capture aesthetic and thematic aspects of the images that are not always evident from Exif data alone.

An essential phase of our project was also the task of annotating each image. We developed a dedicated user interface (using ipywidgets) to assign specific characteristics to each dog depicted in the images. This step was crucial for enriching our dataset by

adding contextual information about the animals, such as their size, fur, and other relevant features.



1 – Tagger Ui

All this information, whether it be Exif data, extracted dominant colors, or manually assigned tags, was stored in JSON files for ease of data management and manipulation. By using this approach, we were able to effectively structure our data, thereby facilitating

2 – Json files extraction

its integration into our recommendation system and its use in providing relevant and personalized image recommendations to users.

User Profile

Regarding user preferences, we've chosen to adopt an interactive approach by developing a user interface (UI) allowing the user to express their opinions on a subset of the 100 available images. In our case, this sample consists of 6 randomly selected images. This approach provides us with an effective means of gathering valuable insights into individual user preferences.

3 – User Like/Dislike Ui

The user interface enables users to provide feedback on each image by indicating whether they like it or not. This process gathers data on specific characteristics of the images that appeal to or displease the user. For example, users may express a preference for certain dog sizes, specific colors, or particular features.

By analyzing user responses, we can better understand their likes and preferences regarding dog images. These insights are then used to refine and personalize the recommendation system, providing image suggestions that better align with each user's individual preferences.

Data exploration models

For our data exploration and machine learning approach, we have adopted a logistic regression model to predict user preferences based on image features. After the user has expressed their preferences by liking or disliking the 6 selected images, we use this data to train our model.

Each image is represented by a data vector composed of 8 distinct features, including size, fur type, ear type, muzzle type, image orientation, as well as the RGB (red, green, blue) components of the dominant color. By using user responses (like/dislike) as the target variable, we thus create a dataset comprising feature vectors and associated preferences.

Once the model is trained, we evaluate its performance using various metrics such as accuracy, recall, and F-measure. These measures enable us to assess the model's ability to accurately predict user preferences for new images.

Ultimately, this model allows us to provide personalized recommendations by evaluating other images and estimating whether they are likely to appeal to our user profile, based on the training data provided by the user themselves.

0.9897312549399508

0.9896756574608203

0.9884291892360776

4 – Result Ui

Self-assessment

We consider this project to have been a rewarding experience in the field of data mining. In addition to achieving conclusive results, it has allowed us to discover several models and data analysis techniques.

Throughout the project, we were able to explore different methods for collecting, annotating, analyzing, and visualizing data. This has given us a thorough understanding of the key steps involved in the image recommendation process, as well as the challenges and opportunities associated with data mining.

Conclusion

It would have been beneficial to start the project in parallel with the practical sessions. This would have allowed students to better manage their time and have more flexibility to delve into each aspect of the project. By starting earlier, we could have devoted more effort to each phase of the project.