Weekly update - 2023-04-25 to 2023-05-02

Paula Jeniffer dos Santos Viriato[1][0000−0003−0900−1686]

Institute of Computing, University of Campinas, Campinas - SP, Brazil
p234831@dac.unicamp.br

1 Clustering and Analytics - Reddit Dataset

Data extracted from Reddit until the last week were analyzed, and a first clustering was performed. The first fact we can verify based on Figure 1 is that the dataset is unbalanced. There are some classes with vast amounts of data (like *photography* and *science*), while others have virtually no data (like *construction*). This fact was presented in the previous report through a pie chart on the percentage of themes in the dataset.

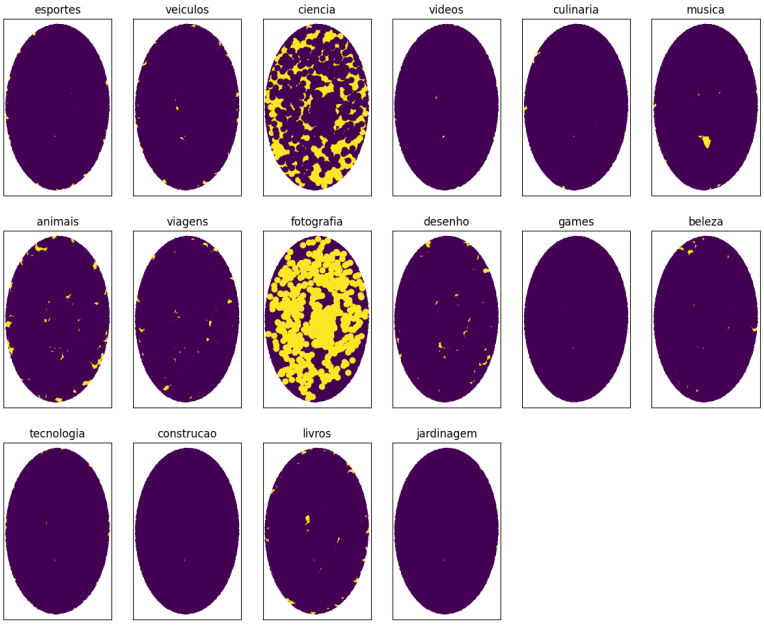

Fig. 1: Location of Data Classes in Entire Dataset

We also noticed in Figure 2 that the data are visually less grouped and more spread out. It may be due to the presence of denser data classes or because the visualization used (TSNE) is not the most suitable for the case. It will

be necessary to adjust the codes so that visualization is possible via UMAP and mainly pyLDAvis, given that both libraries seek a better visualization and pyLDAvis allows visualization interactive.

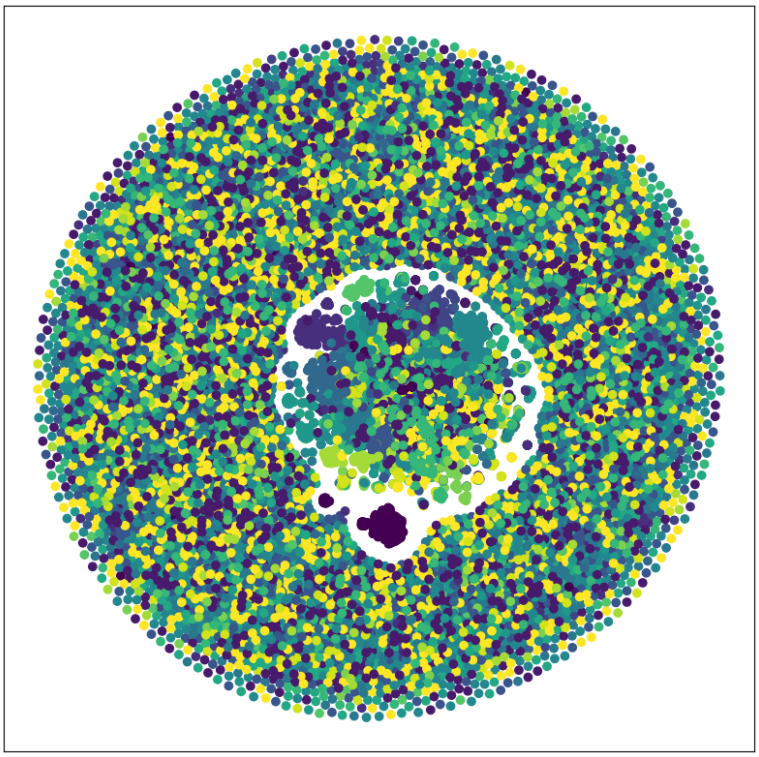

Fig. 2: Dataset Cluster Distribution

Despite the little clustering noted in Figure 2, Figures 3 and 4 show a clear trend for some clusters found. In viewing order (from left to right, top to bottom), we identified that the themes presented respectively are: music, movies, cars, football, computers, and animals.

Two more dataset facts were found. The first, shown in Figure 5, is that some posts already suggest that the Reddit user wants a recommendation. The second fact, shown in Figure 6, is that despite the selection of subreddits in Portuguese, many users post in other languages.

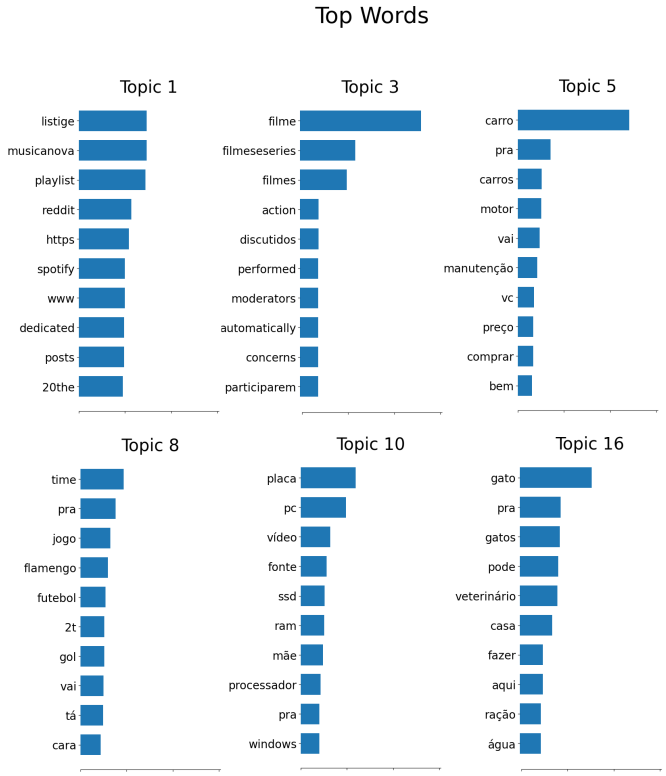

Fig. 3: Most used words in six identified cluster

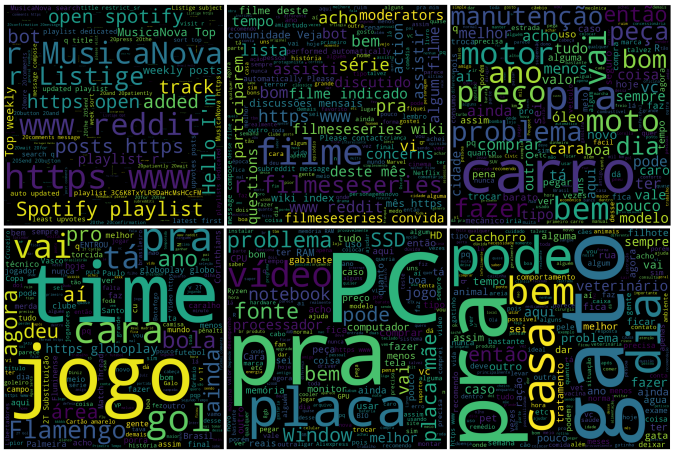

Fig. 4: Comparing the word cloud of six topics

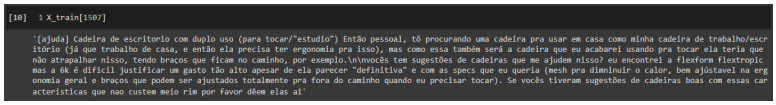

Fig. 5: Example Reddit User Needing Recommendation

Fig. 6: Example Post in Spanish on a Brazilian Subreddit

Some activities are still expected for a better conception of this dataset:

1. Balance the dataset;
2. Test supervised learning with the content;
3. Also get images and use the Image Captioning library, from Hugging-Face, to convert images into text;
4. Find better views of the dataset;
5. Discard or translate content that is not in Portuguese.

2 Textual Digital Marketing Dataset

We are checking of possible sources of marketing data, and such data can come from:

1. Public textual data previously published and validated;
2. Synthetic textual data generated by the company Eldorado;
3. Public and real textual data extracted from various retail sites.

There is an ideal data type for this research, but this does not impede or stagnate research advances. The main thing is that here we are dealing with **public data**, **textual**, **explainable**, and that they generate a **non-invasive** recommendation system.

Starting this week, we will work on extracting data via retail sites. As shown in Figure 7 this data can be category, title, value, marketing, description, and image. Digital marketing images will be captured and stored, and functions from Hugging-Face's Image Captioning library will be applied to these images.

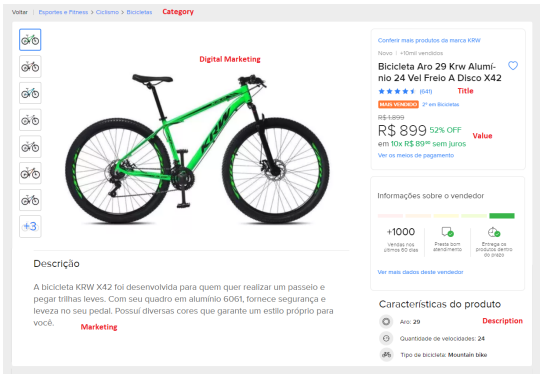

Fig. 7: Example of Product on the Retail Site Mercado Livre

3 Next steps

A Kanban is being built to present and schedule the activities of this project, and is available at the following link. Below we present the next steps of the research, in priority order.

- Analysis of Acquired Data from Reddit;
- Search for a textually descriptive marketing dataset;
- Improve the Project Schedule;
- MC750 Course Planning for Digital Marketing Assessment.
- Research on *Invisible Machines*;
- Read the paper *Predicting the Need for Xai from High-Granularity Interaction Data*;