Weekly update - 2023-04-18 to 2023-04-25

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

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

1 Data Extracted from Reddit

Extraction of posts and comments from Reddit according to theme. The first topics worked on in Portuguese were: animals, beauty, science, construction, cooking, design, sports, photography, games, gardening, books, music, technology, vehicles, travel, and videos. In this first application of the Reddit data extractor, present in the following link, we obtained 17.386 posts. The merging of posts from all themes into a single dataset and some statistics can be acquired by the check_data code. The number of posts extracted for each of the themes is shown in Table 1.

Table 1: Number of Posts Acquired by Subreddit

animals: 2403	cooking: 590	games: 205	technology: 1080
beauty: 180	drawing: 866	gardening: 200	vehicles: 1329
science: 2068	sports: 2101	books: 1396	trips: 293
build: 10	photography: 987	music: 1997	videos: 1681

Figure 1 shows how much each theme affects the overall design of the dataset so far, by means of a pie chart. Here we have already managed to analyze that some themes participate more in the composition of the Dataset than others, such as animals, science, books, music, videos and sports. This could be happening for three reasons. The first one due to a real trend on Reddit for certain subjects. The second justification would be an imbalance of Reddit as a whole for Reddit in Portuguese. The third justification would be that not enough data has been mined yet. For the second case, an alternative would be to translate subreddits from other topics, using libraries already implemented for this purpose; and for the third case, the alternative would be to continue mining more and more data.

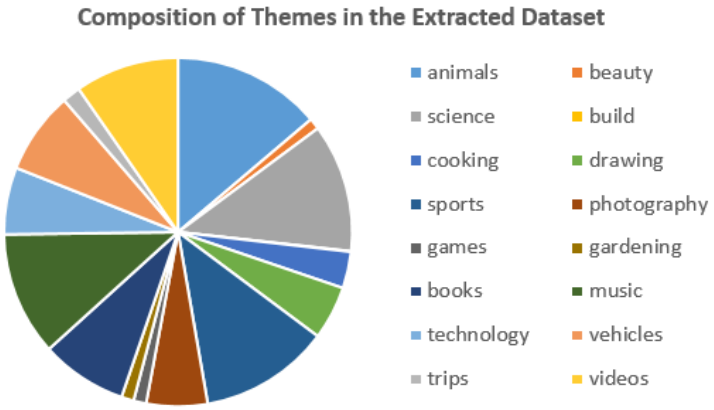

Fig. 1: Composition of Themes in the Extracted Dataset

2 Ideal Data Proposal

As presented in the attached Overview, data is required for the Customer Segmentation and Market Segmentation phases, and the results of these phases affect the project as a whole.

The company Eldorado pointed to the possibility of generating synthetic data, and in this way, two data needs arise here: Consumer Data and Marketing Data. Figure 2 presents the logical data model for the ideal dataset. The explanations for each field are presented below.

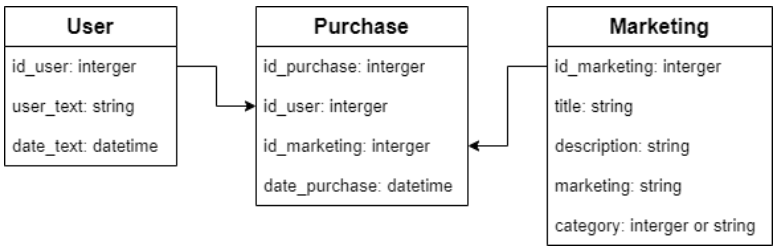

Fig. 2: Logical Data Model for the Ideal Dataset

Consumer Data

- **Fields**: [id_user, user_text, date_text], [id_purchase, id_user, id_marketing, date_purchase];
- **id_user**: anonymized identification of the users;

- **user_text**: textual data generated by users. Possible sources are: posts on social networks, comments on posts or forums, messages to other users, or sentences applied on search engines;
- **date_text**: date and time of sending the text (datetime);
- **id_purchase**: unique identifier for the purchase made;
- **id_marketing**: unique identification of the product or service;
- **date_purchase**: date and time of purchase (datetime);
- Textual data together with date and time data (datetime) are essential in both cases, for texts generated by the user and for purchases made;
- General public, users of digital media over 18 years old.

Marketing Data

- **Fields**: [id_marketing, title, description, marketing, category];
- **id_marketing**: unique identification of the product or service;
- **title**: title or name of the product or service;
- **description**: a textual description of the product or service;
- **marketing**: textual presentation or advertisement of the product or service;
- **category**: product or service category (label);
- Digital marketing data for products or services;
- Possible sources of data: social networks, various retail sites, Google Ads.

3 Initial Schedule

The figure 3 represents the initial schedule, still need to add some tasks. Featured tasks are those that are currently in focus. During the next week this schedule will be more filled, as well as Trello, becoming more complete.

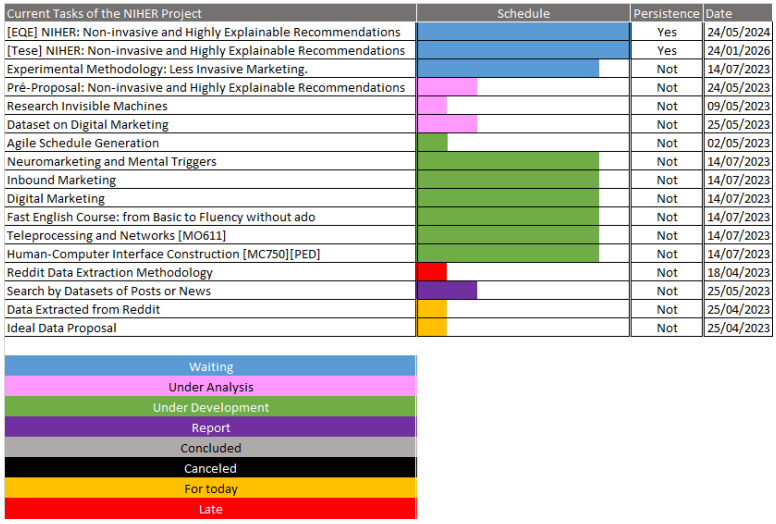

Fig. 3: Initial Schedule until August 1, 2023

4 Next steps

This week we already started the processes of analysis and clustering of the data extracted from the dataset, but these are still under development. The expectation is that this process will end by April 28, 2023 (Friday). Another task that is being done together is the verification 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.

As seen in Section 2, 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. 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*;