Weekly update - 2023-04-11 to 2023-04-18

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

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

1 Dataset from Reddit

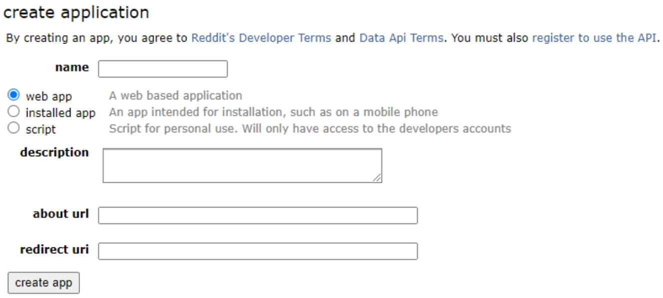
An alternative thought about a dataset in which it is possible to verify more precisely the interests of the users is the extraction of data from Reddit forums. Reddit is the self-proclaimed "front page of the internet" [2], being the world's largest online discussion forum [1]. It is organized into topics such as music, movies, gardening, computing, and more. The extraction of data from these forums focused on specific themes can help in a stronger recommendation for users.

2 Data Extraction

Initially, before you start extracting data from Reddit, you need to acquire special developer credentials. This is a simple and free process. The steps will be indicated below:

1. Create a Reddit account using the following link. It's fast and free;
2. Go to the *Authorized Applications* page, through the link;
3. On the *Authorized Applications* page, access the *create app...* button;
4. The Figure 1 page will appear at the end of the page, for creating the application. Fill in all the data, pay attention to the type of application and indicate both the URL and the URI, which can be your GitHub page, personal page, or project page;
5. In a few moments, the credentials will be created, and can be accessed through the screen in Figure 2, which will appear next.

The extraction of data from Reddit occurs as shown in Figure 3. We use a python interface for authentication called Praw, and it basically allows or disallows the user's connection to the dataset according to the credentials, which are: client_id, client_secret, user_agent and username. Credentials were previously created. After the release, we already have access to the Reddit database, and to acquire the data, we just need to indicate the subreddit we want to consult and the limit of posts returned. This data can be iterated by the application, which can also access the comments made under that post. In the application itself, the data is organized, placed in JSON format and recorded in a document for the user.



create application

By creating an app, you agree to Reddit's Developer Terms and Data Api Terms. You must also register to use the API.

name

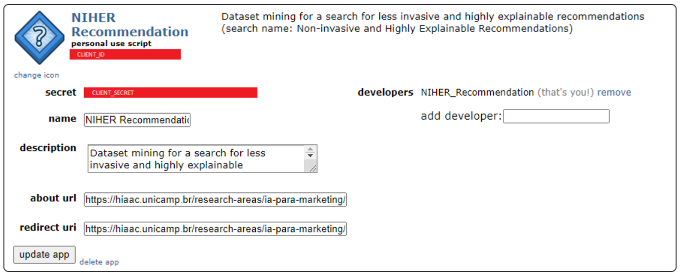
● web app — A web based application
○ installed app — An app intended for installation, such as on a mobile phone
○ script — Script for personal use. Will only have access to the developers accounts

description

about url

redirect uri

Fig. 1: Page for Generating Credentials



NIHER Recommendation
personal use script
CLIENT_ID

Dataset mining for a search for less invasive and highly explainable recommendations
(search name: Non-invasive and Highly Explainable Recommendations)

secret CLIENT_SECRET

developers NIHER_Recommendation (that's you!) remove

add developer:

name NIHER Recommendatio

description Dataset mining for a search for less invasive and highly explainable

about url https://hiaac.unicamp.br/research-areas/ia-para-marketing/

redirect uri https://hiaac.unicamp.br/research-areas/ia-para-marketing/

Fig. 2: Display of Created Credentials

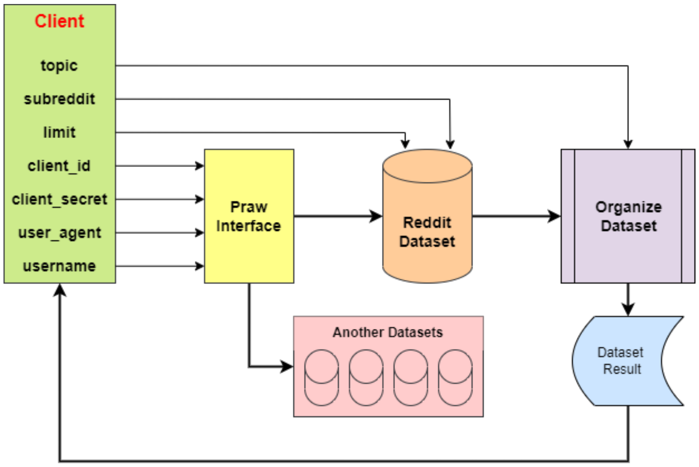


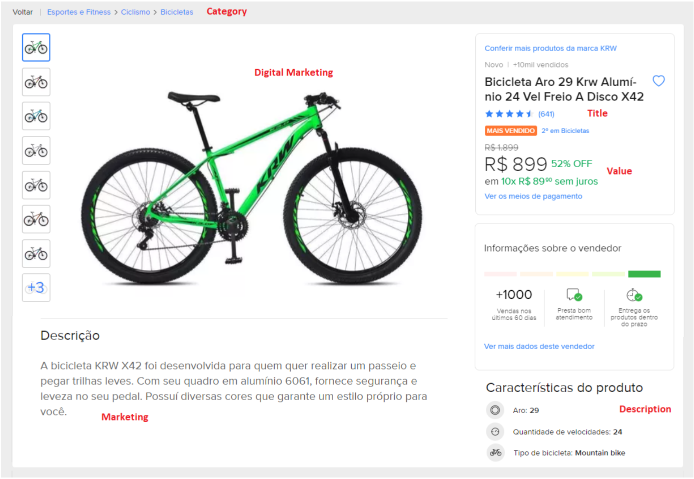
Fig. 3: Methodology for Extracting Data from Reddit

3 Digital Marketing Data Source Idea

One of the most important stages of the project, which is the entrance to the Market Segmentation phase, is the search for datasets related to Digital Marketing. There are several ways to acquire Digital Marketing data:

- Search for previously published and validated public datasets: these are rare data, hardly found on the internet, even more so in the context of digital marketing in Portuguese, but they can exist and be found with some effort;
- Search for digital marketing posts on social networks, such as Facebook, TikTok and Instagram: in this case, care must be taken with biases, as such posts already undergo a prior recommendation system;
- Generating synthetic digital marketing data: requires a lot of effort and understanding of digital marketing success criteria;
- Extraction of public and real data of products and their advertisements from various retail sites. It is a very realistic alternative, and with a lot of data in Portuguese, but it requires a large coding effort and extraction waiting time.

In all cases, for this project, priority is given to data that are **public**, **real**, **textual**, **explainable**, and that generate a recommendation system at the end of the project. invasive. As shown in the product in Figure 4, taken from a page of a well-known retail site in Brazil, and whose product link is available, it is possible to identify several textual data:



Esportes e Fitness > Ciclismo > Bicicletas

Conferir mais produtos da marca KRW

Novo | +10mil vendidos

Bicicleta Aro 29 Krw Alumínio 24 Vel Freio A Disco X42

★★★★☆ (641)

MAIS VENDIDO 2º em Bicicletas

~~R$ 1.899~~

R$ 899 52% OFF

em 10x R$ 89⁹⁰ sem juros

Ver os meios de pagamento

Informações sobre o vendedor

+1000 Vendas nos últimos 60 dias

Presta bom atendimento

Entrega os produtos dentro do prazo

Ver mais dados deste vendedor

Descrição

A bicicleta KRW X42 foi desenvolvida para quem quer realizar um passeio e pegar trilhas leves. Com seu quadro em alumínio 6061, fornece segurança e leveza no seu pedal. Possuí diversas cores que garante um estilo próprio para você.

Características do produto

Aro: 29

Quantidade de velocidades: 24

Tipo de bicicleta: Mountain bike

Fig. 4: Example of Product on the Retail Site Mercado Livre

- Category: Esportes e Fitness > Ciclismo > Bicicletas;
- Title: Bicicleta Aro 29 Krw Alumínio 24 Vel Freio A Disco X42;
- Value: R$899;

- Description: Aro 29, 24 velocidades, Mountain bike;
- Marketing: A bicicleta KRW X42 foi desenvolvida para quem quer realizar um passeio e pegar trilhas leves. Com seu quadro em alumínio 6061, fornece segurança e leveza no seu pegal. Possuí diversas cores que garante um estilo próprio para você.
- Link.

4 Next steps

The main next step of this work is to use the tool for extracting posts from Reddit, with the objective of building an own and public dataset, which can be used in the **Customer Segmentation** phase. 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.

- Extract posts and comments from Reddit;
- Search for a textually descriptive marketing dataset;
- Project schedule generation;
- MC750 Course Planning for Digital Marketing Assessment.
- Research on *Invisible Machines*;
- Read the paper *Predicting the Need for Xai from High-Granularity Interaction Data*;

References

1. Medvedev, A.N., Lambiotte, R., Delvenne, J.C.: The anatomy of reddit: An overview of academic research. Dynamics On and Of Complex Networks III: Machine Learning and Statistical Physics Approaches 10 pp. 183–204 (2019)
2. Singer, P., Flöck, F., Meinhart, C., Zeitfogel, E., Strohmaier, M.: Evolution of reddit: from the front page of the internet to a self-referential community? In: Proceedings of the 23rd international conference on world wide web. pp. 517–522 (2014)