heuritech

Senior Data Scientist - technical test

NB: This technical case is not only meant to evaluate your ability to answer tech-related questions, but also to illustrate some topics owned and tackled by Data Science, so you can have a good idea of what are the tools and problems we face in the DS team. Do not hesitate to share your feedback!

Contextualization - Heuritech Methodology

Heuritech builds the largest fashion trends dataset, based on social media. To do so, we apply a machine learning solution including 3 main steps:

- First, we leverage several computer vision AI models split in 2 stages. We first apply an object detection model to spot the elements useful to build fashion trends (humans, shoes, pants, tops, coats, etc.) and then feed downstream models specialized on a given scope. For instance, when we detect a shoe, a model specialized in detecting the type of shoe (Is it a pump? a sneaker?) and their attributes (the shape of the toebox, the heel height, etc.). We also feed more generic models to spot the color(s) or the textures
- By crossing the different elements detected on a single fashion item, we are able to define the data point of what we call a trend. For instance, if on an Instagram post we detect a woman wearing a red t-shirt, we can combine these 3 attributes to define the trend “Red T-Shirt for Women”. Same goes for any kind of fashion item (Long sleeve shirts for Men, etc.)
- At last, we can associate this trend data point to the date the associated social media post has been posted. By aggregating every single data point per week, we can define a time series describing the visibility of this trend throughout time on social media

Propose a new segmentation

We build Heuritech signals based on a methodology that we designed internally to gather social media accounts that look similar and that are relevant to study fashion. Indeed, it is crucial for us to distinguish the type of users to understand their behavior. For instance, we have a specific signal for edgy accounts that can describe a trend in a different way that *mainstream* accounts do. This segmentation is currently based on the amount of followers an account has and is done using our authors database and we follow these rules:

- A mainstream account has at most 12,000 followers
- A trendy account has between 12,000 and 40,000 followers
- An edgy account has more than 40,000 followers

To build insights, we store our data in a Snowflake data warehouse. Notably, we use these main tables:

- MART_AUTHORS where we store all the metadata about authors (one line per author)
- MART_AUTHORS_SEGMENTATIONS with the data that we currently use for segmenting authors (1 author per line)
- MART_IMAGES_LABELS storing some predictions made on social media post images by our deep learning models (1 line per model detection)
- MART_IMAGES_OF_POSTS that tells us for every single image which social media post it refers to (1 image per line, an image can belong to several posts, and a post can have several images)

A few notes regarding the data in these tables

The detection of our deep learning algorithms are represented in the table MART_IMAGES_LABELS. The field TYPE is the detected fashion item by our object detection algorithm. POSITION_IN_IMAGE represents the associated coordinates of this item, and LABEL_NAME one detection made a downstream model.

In the table MART_AUTHORS_SEGMENTATIONS, the field FASHION_INTEREST_SEGMENT represents whether the account has an important value fashion-wise. When it is set to TRUE, it means that our fashion teams detected it as having a heavy weight in the world of fashion and that it belongs to our “fashion forward” panels of accounts, supposed to show behaviours ahead of the mainstream behaviour.

A sample of it is provided in database TECHTEST, schema TECHTEST.

1. *What kind of analysis would you do to evaluate the quality of a segmentation methodology? What biases can you identify in the panel creation or in the accounts selection that might distort insights and how would you correct them?*
   Feel free to use the data at your disposal to support your answer.

2. *What do you think about the current segmentation? What advantages / drawbacks do you see in the methodology / in the way the information is stored? You can play around with the provided data to support your statements.*

3. *We would like to push our segmentation one step further. The current methodology refrains us from spotting interesting (sub)categories of accounts like fashion enthusiasts, luxury, sportswear, etc. Several families of methods can be considered to improve this methodology:*
   - *Statistical approach*
   - *Machine Learning approaches (supervised and/or unsupervised)*
   - *Hybrid approach combining deterministic rules and machine learning*

   Which approach would you prefer and why? What would be the advantages/limits of each of these methods? You are free to exploit the author database the way you want. You can also generate new tables to develop your approach.

   [Optional] In case you have ideas of approaches that would require more time that you are given to perform this test, it will be appreciated if you mention them and explain their added value / limitations.

Expected format

To showcase your work, you can provide a git project with the documents you produce (python code, queries, notebooks, etc.). **The quality of the whole project (formatting, git history, easiness to install and to reproduce, documentation, etc.) is greatly appreciated, as well as the clarity of the approach.**

NB: think in terms of scalability of your approach in the way you handle the data.

Good luck !