Case #1 - Image data analysis

Business case

Approximately 1.8 billion images are uploaded daily to social media and more than 70% of them are black boxes because they do not contain any proper tagging or text information that can be analyzed. Dashmote with the help of computer vision techniques is turning images into data and offer an extra source of customer intelligence which otherwise would have been lost.

Currently Dashmote was tasked with discovering hairstyle trends at a global scale on Instagram. By using image recognition, Instagram posts have been clustered together into hairstyles and merged with engagement data at the level of the post.

Instructions

The case was created to show off your creativity and analytical thinking in a real world situation. You have 24 hours at your disposal to complete the analysis and submit your results.

This case challenges you to:

1. Discover insights at a global level
2. Discover insights at hairstyle level
3. Explore the relationships in between the hashtags used in the posts
4. Mention other analysis that you would perform

Detail the steps you have taken in your analysis and be as creative as you want because there are no right or wrong answers.

Description of the data

The file `Dashmote_Database.json` contains a sample 10000 images mined from Instagram and clustered based on the hairstyle they showcase.

The variable `cluster` represents the hairstyle cluster that the image has been assigned to by the visual recognition algorithm.

Each row contains the variable `url` which is the link to the image and the number of `likes` together with the `comments` per image. The `user_id` is the unique id of the Instagram account from which the post comes and the variable `id` is the unique identifier associated with the post itself.

Each post contains the date(`date_unix`) in unix format when the image was posted on Instagram and additionally the date has been converted to different formats (**date_week** ->non-iso number of the week, **date_month** -> the month,**date_formated**->full date dd/mm/YY) partly for use in prior analyses. Feel free to convert that variable in a way that suits your analysis.

Additionally a classifier **influencer_flag** was added to each of the images which have more than 500 likes, flagging them as influencer posts.

Schema of the Data

cluster	STRING	NULLABLE
id	STRING	NULLABLE
influencer_flag	STRING	NULLABLE
user_id	STRING	NULLABLE
influencer_num	INTEGER	NULLABLE
url	STRING	NULLABLE
date_month	STRING	NULLABLE
hashtags	STRING	REPEATED
date_unix	INTEGER	NULLABLE
comments	INTEGER	NULLABLE
date_week	STRING	NULLABLE
likes	INTEGER	NULLABLE
date_formated	STRING	NULLABLE
inf_true	STRING	NULLABLE