

Comparing e-shops' online ratings with social media footprint

Project Proposal by Adrien Boyer and Matej Maivald

Data Processing in Python

In our project, we will try to compare Czech e-shop's ratings from Heureka.cz with social media footprint - the number of posts and followers.

- The first part of the application will collect data about e-shops from Heureka.cz (<https://obchody.heureka.cz/>). The application will accept category as an argument and within the category, a (specified) number of e-shops will be scraped. Following attributes will be stored: Eshop name, total reviews, negative reviews, positive reviews, the value of rating, URL of e-shop and a list of text reviews (of specified length).
- In the second step, we'll visit the collected URLs from the first part and collect possible URLs of Instagram accounts linked to the e-shop.
- In the third step, we will collect information from Instagram accounts, including number of followers, number of posts and if possible, average number of likes per post.
- In the final part, we will try to find if there exists any meaningful relationship between quantitative data collected from Heureka and data collected from e-shop's social media account. This section will include descriptive graphics in Jupyter notebook.
- Optionally, we would like to try to classify words and phrases from text reviews based on the ratings of individual text reviews – find out, which words are typical for negative and positive reviews.

The first three parts will be designed as Python scripts, which will be triggered by the Jupyter notebook.