Data Engineering Applied to the Swiss classified Ads Market

Degree programme : BSc in Computer Science | Specialisation : Data Engineering
Thesis advisor : Prof. Dr. Erik Graf
Expert : Daniel Voisard

Searching for secondhand products online can be a painful and frustrating experience. This work introduces an automated data engineering pipeline, that is designed to improve this experience. It collects raw data from a website, extracts the useful structured elements, mines the unstructured data with the help of Natural Language Processing (NLP), and finally displays the results to the user in a web application.

Idea

The idea for this bachelor thesis was based on the personal experience of spending hours on second-hand ad platforms, like tutti.ch and anibis.ch, trying to find the right product without much success. To improve on the current situation the concept for a comparison website like toppreise.ch, but for second-hand products out of classified ads, from different platforms was conceived. The aim for this website consists of building a complete data engineering pipeline that is capable of automatically retrieving, enriching, and displaying data.

Goals

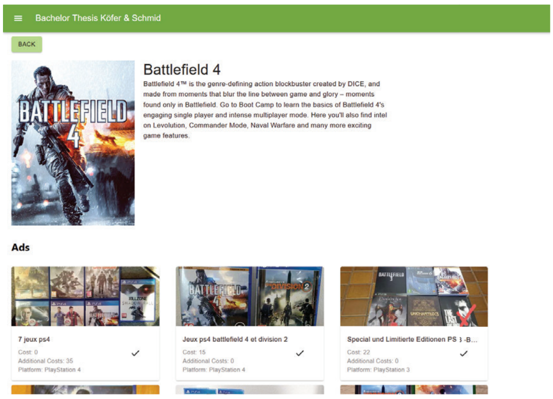
The original idea as described represents a large-scale project. To fit the workload into the context of a bachelor thesis, a smaller scope had to be defined. However, the overarching idea and end goal should be kept the same. The top priority is to offer the end-user an easier search, and a practical comparison of the offers. The solution chosen to reduce and adjust the workload was to limit the scope of work to only one initial platform (anibis.ch) and to focus on just one product category („Games & Consoles" of anibis.ch). This allows for the implementation of the whole data engineering pipeline: Collecting secondhand classified ads from a Swiss website, extracting the important, interesting, and structured data elements, mining additional information from the collected data, and making the results available to the end-users.

Solution Overview

A Scrapy spider is used to collect all ads from the website. From the scraped HTML ads, the important information is extracted, and then stored as JSON documents in an Azure Cosmos DB. On the unstructured data, like ad descriptions and titles, spaCy is used to do the NLP to exactly recognize games, platforms, and costs. This is necessary because many times there are several products with different prices in one ad. Finally, the results are shown on a web application hosted on Azure.

Conclusion

Building an automated platform requires a complex data engineering pipeline. Developing this pipeline involves solving many challenging technical tasks of many different topics. These tasks include the configuration and deployment of the scraping engine, building a performant matching of keywords, and integration of ML models. Evaluating if a ML solution or a matching based approach are better suited, formed an important part of this work. This project demonstrates that building a meta-search platform for secondhand ads is very much possible. Nonetheless, providing additional information to the user proves to be very complex and domain-specific. As such many tasks and problems are unique and must be implemented separately. It can be concluded that applying a data engineering pipeline holds great potential for improving the user experience. However, it has become evident that scalability and maintainability are challenging aspects.



Battlefield 4

Battlefield 4™ is the genre-defining action blockbuster created by DICE, and made from moments that blur the line between game and glory – moments found only in Battlefield. Go to Boot Camp to learn the basics of Battlefield 4's engaging single player and intense multiplayer mode. Here you'll also find intel on Levolution, Commander Mode, Naval Warfare and many more exciting game features.

Ads

7 jeux ps4
Cost: 0
Additional Costs: 35
Platform: PlayStation 4

Jeux ps4 battlefield 4 et division 2
Cost: 15
Additional Costs: 0
Platform: PlayStation 4

Special und Limitierte Editionen PS 3 -B...
Cost: 22
Additional Costs: 0
Platform: PlayStation 3

Web Application: Game Page with available Ads

Philipp Köfer
philipp.koefer@bluewin.ch

Raphael Josua Schmid
raphael.j.schmid@gmail.com