Tagging of second-hand e-commerce fashion articles

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

Second-hand fashion articles are plenty but scattered across websites. Providing a central place to search all newly posted second hand items, and at the same time extending the possible filtering possibilities could improve the shopping experience for the users.

Nicolas Freddy Pierre Claude Houriet

Motivation

Second-hand websites offer many search possibilities, but often with different search options. This makes the user search experience less consistent and practical when looking for second-hand articles.

The goal of this work is to develop an automated solution for the collection and tagging of listed items, and to utilize this information to implement an advanced filter-based search solution for second-hand items.

Architecture

The solution consists of two major parts:

* An automated scraping module
* A web application hosting advanced search capabilities

Monitoring capabilities, extensive testing and an extensible design make this scraping system a reliable way to extract a variety of data. The scraped fashion articles are stored in a relational database.

Tagging process

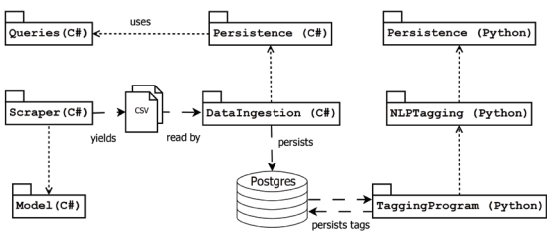
Two pre-processing steps have to be applied to the collected text data:

* multi-lingual content is analysed using fasttext and langdetect to determine the ad language.
* sentences are cleaned using regex (smileys..)

Each item within a category can then be accessed from the database and tagged using NLP matchers that make use of the Spacy library. The detected tags are persisted to the database, which can then be displayed using the web application developed using the blazor web assembly framework.

Result

The mean accuracy of the 317 detected tags samples (taken from over 35'000 detected tags) is 94%. For tag categories such as brands or colors, the performance is excellent, which could be expected due to the regex-like nature of the matchers for those tags.



Package diagram representing the overall architecture

Tag category	Language	Items	Accuracy
brand	DE	3	100%
	EN	34	100%
	FR	17	100%
color	DE	18	100%
	EN	2	86%
		12	
	FR	2	92%
		23	
motif	DE	20	100%
	EN	16	100%
	FR	1	
		22	96%
origin	DE	13	100%
	EN	12	100%
	FR	15	
		12	44%
size	DE	18	100%
	EN	14	100%
	FR	63	100%
Grand Total		**317**	**94%**

Accuracy of evaluated items per tag category and language