

SUMMARY OF PRODUCT REVIEWS

Are you annoyed by having to read through hundreds of reviews when you want to buy something specific? Don't worry, a solution will be available soon. Currently, I'm working on this in my bachelor thesis. I'm a perfectionist, and when something isn't perfect, I'm extremely annoyed. The same thing arises when purchasing a product that should meet my requirements. Sure, you can just read the product specifications, but the cons are always missing – that's why I often rely on reviews, where you can find everything – pros and cons.


WEB SCRAPING

If the e-shop doesn't provide you with a dataset, you need to scrape the web. What does it mean? The definition of web scraping is collecting data from a website using a program. It's important to be familiar with HTML programming language because you could get lost while browsing the code.

Another thing to consider is creating a program to accomplish this task. I used a Python package Beautiful Soup, which is also the most common library for scraping information from web pages. This library is an HTML or XML parser. It breaks down text into recognized strings of characters for further analysis, allowing you to get rid of HTML tags and other non-relevant information. Then you can save the necessary data into a CSV file as a table.

PREPROCESSING

The structure of data may become non-systematic after opening the CSV file, but another library called Pandas can solve this issue. It provides functions for analysing, cleaning, exploring, and manipulating data.

 Fun fact: The name "Pandas" has a reference to both "Panel Data", and "Python Data Analysis". However, I always imagine the panda bear from China.

Once you reached the good structure of the text, you can proceed with further data preprocessing. It often happens that some data contain errors or are even missing, and you must remove them, as well as punctuation in the text. Computers can then easily identify patterns and relationships between words.

The next step involves converting the text to lowercase and choosing stop words to eliminate – words that don't have any semantic meaning, such as “was”, “she”, “for” ... These are commonly used words in a language that don't carry useful information. There are many lists available online, especially for the English language; however, for Czech it's more challenging.

LANGUAGE MODELLING

When the dataset is cleaned, we use a language model, which is a type of machine learning model trained to predict a probability distribution over words. To achieve satisfactory results, the dataset should be extensive. I chose the fastText library created by Facebook's AI Research (FAIR) lab. It's also considered a language model for learning word embeddings and text classification.

FastText is efficient for morphologically complex languages. It learns to understand words by breaking them down into smaller parts called n-grams. From this, it creates word embeddings to determine semantic similarity in words as vectors. The closer vectors are, the more similarity they have. Therefore, it also handles spelling errors better than other language model, such as Word2Vec, which use word embeddings as entire words, not subwords.

In conclusion, fastText is much more complex than we have discussed here, but for the purposes of summarising my bachelor thesis, I believe it's sufficient. 