Comp Sci 839 Stage 2 Report

Fangfei Lin, Jiahan Li, Chen Qian

We have selected two shopping websites: Sephora and Ulta to extract the structured cosmetics information.

We applied manual approach to extract the cosmetics information from the structured html webpage. Since both websites are shopping websites and they are following very formulated structure, then manual approach is rather simple. By finding the regular expressions, we are able to find the target entities of each product. For example, we want to extract the ratings of a specific product from Ulta, we used the regular expression r'<label class="sr-only">(.*?) out of 5 stars</label>' to find the target information.

For each cosmetics, we extract the following five entities of cosmetics: product name, brand name, price range, ratings and number of colors/scents. We extract 3212 structured tuples from Sephora and 3089 structured tuples from Ulta.

We used Chrome developer tool, a suite of tools to help people build and debug the webpages, to inspect source codes and elements, thus we are able to locate each products shown on the website to its corresponding chunk of source codes. For each product, we can observe its characteristics of code and find the regular expression of each entity.