Dk-Fa-Cosmatics Data-set

Table of Contents

DK-FA-Cosmetics Data-set

The dk-fa-cosmatics data-set, is a collection of about 420k user comments on cosmetic products in Persian (Farsi) language. Comments are crawled from an online-shop's website. The current version is accumulated raw data, and no post processing is applied to the data yet. Some of the texts may contain emojies or characters not belonging to Farsi language.

Comments Data

Comments are stored in a structured format. And beside the comment body (The text that user posted as their comment), they also contain User's star-based rating (0-5), Comment's title, user-name, other user's reaction to this comment. Some users also have provided a list of advantages and disadvantages of the product alongside their overall opinion of the product.

The C# model classes for Comment objects, are provided in c-sharp-models directory. This code can be used to read and write objects from the data-set using any c# parser and de-serializer of your choice.

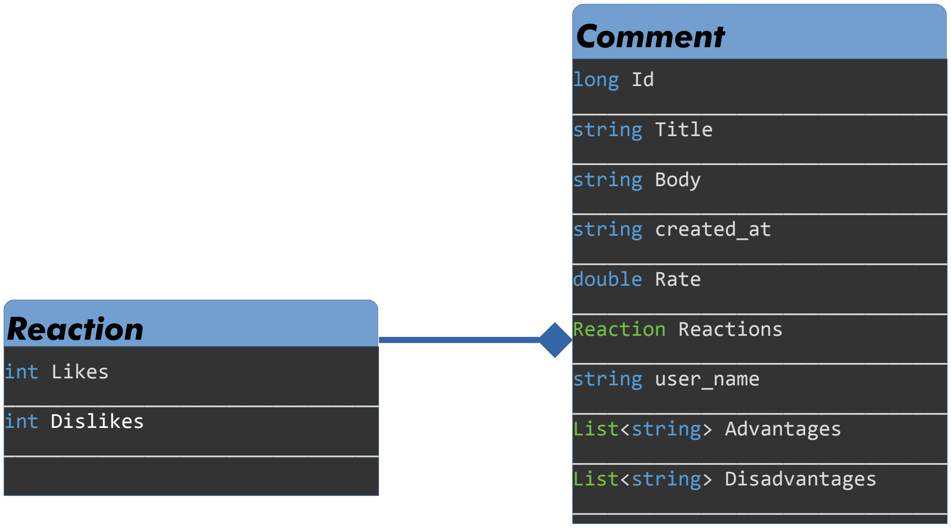

Figure 1: Diagram for comment object and information it represents.

File Formats

Data-set (and its subsets) are provided in three formats: Json, Json Lines and CSV but the data is the same. Except for CSV files which do not include collection objects like advantages/disadvantages.

The Json file, is a large array of comment objects, which may not be desirable to load into memory in one read for many cases. To address this issue, one can use the JsonLines format, which is a new-line delimited file. Each line containing exactly one comment object and can be read and deserialized line by line. Both json and json-lines formats contain all the comment belonging information.

The CSV files, have headers for the first row. Each column name in the header is associated with one property of the comment object. For nested properties, a dot notation has been used. (Ex: Reactions. Likes)

while parsing the CSV file of a set, you might want to consider:

- Strings with the comma character and the new-line characters are double-quoted but NOT escaped.
- Texts, might contain ampersand, semicolon, tabs and other characters which some CSV parser might consider as a delimiter, so you need to choose and configure your CSV parser to only use coma as delimiter for columns and only use new-line character as row delimiter.

Sets

The collection, dk-fa-cosmatics, is the main set which contains all 421,078 comments. The other sets are smaller subsets of the dk-fa-cosmatics so that it might be easier to use for cases which might need a smaller set of data. Each directory contains one set/subset in json, json-lines and CSV formats.

The following table shows subsets and number of comments on each one.

Set	Number of Comments	Number Of Products	Average Comments Per Product
dkfacs-eyeliner	30824	284	109
dkfacs-stand	83197	1738	48
dkfacs-mascara	47961	338	142
dkfacs-sun-screen	118699	772	154
dkfacs-eye-shadow	14532	634	23
dkfacs-nails	75209	3260	23
dkfacs-lipsticks	50656	1299	39
dk-fa-cosmetics	421078	8325	51

Table 1: set/subsets and number of comments in each

Crawling

The target online shop, suggests products of similar category (related products), while visiting each product. So crawler, starts scraping comments about a product. Then it lists all related products and makes them to be visited next. It keeps doing so for each listed product unless it's already visited. Until there is no more un-visited suggestions.

The current data set is the result of such strategy for 7 starting products, in different sub-categories of cosmetics. (Eyeliners, Stand box, Mascara, Sun-screens, Eyeshadows, Nail polishes, Lipsticks)

For each of the sets, the number of scraped product pages and the number of product pages queued to be scraped, are saved during the crawling process. The file <set-name>.history.csv in each set's directory contains these information. If you look at these numbers, you can see how the number of un-scraped pages grow at the beginning, and as the process goes forward, it starts to drop until the crawler has passed through all related products graph. Figure-2 shows the crawling queue history for dkfacs-eyeliner set. (dkfacs-eyeliner.history.csv file)

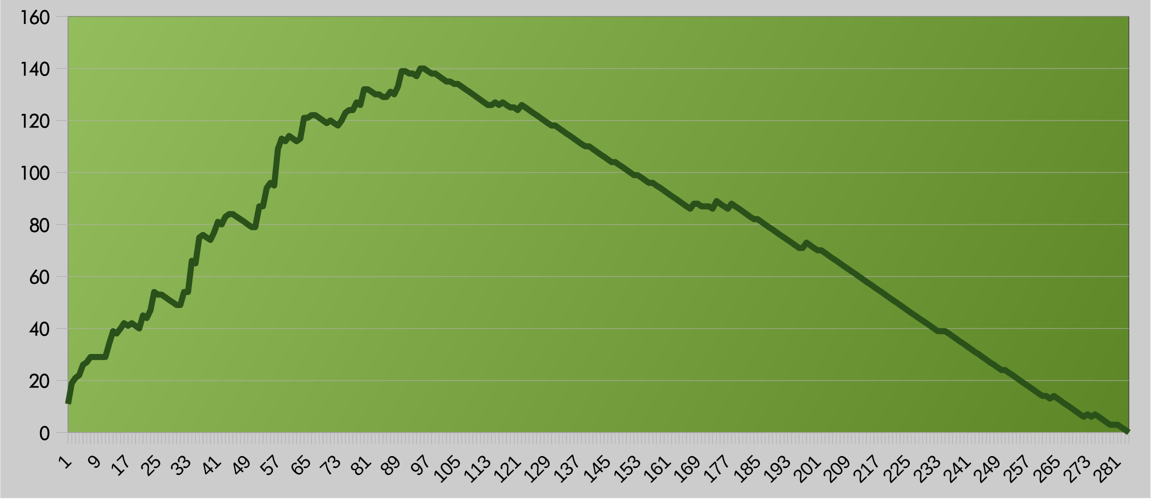

Figure 2: Crawling Queue history for dkfacs-eyeliner