Skin Care Search Engine Final Report

Yujin Zhang (yujinz2@illinois.edu) Siyu Bian(siyub2@illinois.edu) ShuSong(shusong2@illinois.edu)

1) Overview of the Function (what this software does and what it can be used for).

We developed a specialized search engine for skin care products on Amazon. The user can input some keywords of the effect they hope the product can achieve, such as "dark spot serum" or "moisturizing cream", and our system will return several top rated products based on the reviews related to this effect. Basically, this is a information retrieval system based on the content and the rating of the product reviews rather than simply basing on the products' name or products' introduction.

The motivation of this system is that as there are a massive amount of skin care products on Amazon, for one thing, it is hard for users to find a suitable product when they only know their needs but do not know which specific product they are going to buy; for another, it's time consuming to go over the product reviews to figure out whether different concerns on the same product leads to different ratings, for example, a product may perform good on brightening while performing bad on moisturizing.

2) Software Implementation Details.

We first used a dataset gather in 2014 about Amazon beauty[1] to get the list of ASIN's (unique product ID of Amazon) of beauty products. Then we implemented a crawler on our own, using the list of ASIN's to crawl the latest product reviews, while filtering out products that do not belong to the skin care subcategory (perfumes and etc.) as well as removing the ones that are no longer available. In this way we guarantee that reviews are all up-to-date and relevant. Below is an example of the tuples of information we crawled after formatting.

```
{"asin": "B000052YMR", "overall": "5", "reviewText": "Great  for very dry skin.", "summary": "Great f
{"asin": "B000052YMR", "overall": "5", "reviewText": "Literally the best moisturizer that does what i
{"asin": "B000052YMR", "overall": "5", "reviewText": "Our favorite face and body cream!", "summary": "
```

As crawling takes time, if you would like to skip this step, please download the data crawled by us from here:
https://drive.google.com/drive/folders/1T5dN_SA_izfnTOgeB1og3kyntZcEvl2l?usp=sharing

Then we pre-process the data. Besides a file that contains all review texts, in our case, the documents used to build the inverted index, we also generated an auxiliary data file that contains two columns per line including the unique product IDs and overall rating in Amazon of different products, which will be used to improve our searching result and retrieve product ID based on the reviews' index.

For retrieving the appropriate products from our dataset, we implemented three methods to with different ranking methods and compared them:

The first method recommends product based on each single review. That is, it treats one single review as a product. We first used the TF-IDF weighting method, BM25, to perform document ranking basing on each review, and then weight the TF-IDF result of a review with the rating score of this review. For example, the same keyword will be weighed more in a 5-star review, while being weighed

less in another 2-star review, or maybe even being weighed negatively. The formula we used to integrate the weighting scheme is

$$\alpha(tf - idf\ score)\ + (1 - \alpha)(review\ score)$$

We returned the top three products based on the final score. Below is the result of the recommendation of "facial scrub acne"

```
Enter your query: facial scrub acne
Top 3 valid results are...
B007004PZO
B000UVWGGA
B003Z4OD24
```

1 result for **"B007004PZO"**

Salicylic Deep Gel Exfoliating Cleanser - Enhanced with Tea Tree Oil & Green Tea Extract (Professional) - Peel Prep
by Perfect Image
$24^{95} ($6.24/Fl Oz)
✓prime
Get it by **Friday, May 11**
FREE Shipping on orders over $25 shipped by Amazon
★★★★☆ ▾ 433

1 result for **"B000UVWGGA"**

St. Ives Fresh Skin Face Scrub, Apricot, 10 oz
by St. Ives
More options available:
$6.09 Other Sellers
★★★★☆ ▾ 3,638

1 result for **"B003Z4OD24"**

Acure Brilliantly Brightening Facial Scrub, 4 Ounces (Packaging May Vary)
by Acure
$7^{19} ($1.80/Ounce) $9.99
Add-on Item
Add to a qualifying order to get it by **Friday, May 11**
FREE Shipping on orders over $25 shipped by Amazon
More options available:
$6.37 Other Sellers
★★★★☆ ▾ 1,500

The second method recommends products basing on all of the reviews and the average rating score of a product. We pre-processed the data by calculating the average rating score of all reviews of

a single product and merged all of its reviews into one single tuple basing on the unique product ID on Amazon. Then we basically used the same method as method 1 to recommend the product.

Below is the result of the recommendation of "facial scrub acne"

The third method uses a different strategy to recommend beauty product: It first calculates the score of each review as like in method 1 using the formula shown below:

$$tf - idf \; score \; * \; review \; score$$

Then we calculate the score of each product basing by integrating the scores of all its reviews, using the below formula:

$$(\sum(\mathit{tf-idf\ score\ *\ review\ score}))/(\mathit{number\ of\ reviews\ of\ the\ specific\ product})$$

We also filtered to the result after applying BM25. We eliminated the products which have less than 10 reviews to avoid biased reviews. Because if a product has only one or two reviews on the effect that queried by the user, the resulted overall score might be biased and less valuable.

Below is the result of the recommendation of "facial scrub acne"

The first recommended product wins the award of **Best Facial Scrub - Allure Best of Beauty Awards 2015 & 2016. And this proves that this recommendation systems works well. As we can see that this product has raised from top3 to top1 from method 1 to method 3.**

The second and the third recommendations also look good, as we can see from the overall review rating and the products fit our requirements. This implies the gradual improvement from method 1 to method 2.

3) Usage Documentation

Step 1: crawl the infomation from amazon beauty product.

`python reader_updated.py` // this generatse the unique product ID file based on the old dataset

`python review_crawler.py asin.txt min_num_reviews max_num_reviews` // this generates the json file of the crawled information. Setting the range of the number of review enable users to parallely run the crawler on several machines, as well as filtering products with less reviews.

Step2: pre-process data

`python merge_json.py` // This is to merge the information (in json) we crawled from amazon beauty

`python transform_format.py` // This is to process the data into a better format for later use

`python product_based_data.py` // This is only needed if you wish to run method 2. This will merge all of the reviews of one single product and calculate the overall ratings.

`python reader_updated.py` // This is to generate auxiliary data file and review content file change review content file into dat format and put it in amazonbeauty folder

Step3: run search engine

`python search_rank.py config.toml (>> and then input your query)` // method 1 or 2

`python search_rank_updated.py` // if using method 3

4) Description of contribution

We came up with this project idea together and discussed the implementation choice all together. Yujin Zhang implemented the crawler. Siyu Bian and Shu Song implemented the text information retrieval part and pre-processed the data.

Reference

[1] http://jmcauley.ucsd.edu/data/amazon/