<CS12200> Project Proposal

Makeup product evaluation using data from Sephora.com

By On_Fire(Weijia Li, Zhuo Leng, Xinzhu Sun, Xingyun Wu)

GOAL:

The purpose of this project is to evaluate cosmetic products for a given type. We are interested in identifying the most popular items for a given category, top 5 reviewed items, the most expensive and the most economic choices. Depending on the progress, we may also compare prices on one website with another cosmetic website.

We will choose sephora.com as our initial approach and build up a system that takes a given type of makeup as input (e.g. foundation, eyeshadow, etc.) and returns lists of products with characters aforementioned. The system is designed to extract all price tags, reviews and sales volume (if accessible) for products within the given category and rank them to return the desired outputs. To make our results more intuitionistic, we plan to generate graphs to present our outcomes.

Data：

Our data will be crawled from http://www.sephora.com/ (our starting url and limiting domain). We plan to scrape price and review information from it using BeautifulSoup and urllib3.

PLAN：

We will split into two groups, one scrape and clean data and the other write the algorithm. The interface will be built after the outputs are produced. The workload is hard to forecast before seeing the data.

Algorithm:
The algorithm will be built in the light of assignment 2. After getting the data, we will then clean data using regular expression thus export a csv file. As to visualisation, we plan to generate a word-cloud to present most popular products and plot a price distribution of makeup brands with brand names on x axis and price on y axis.

TIMELINE

5th-6th Week (before first checkin)	Scrape data from website to ensure usability; preliminarily clean data if usable or change target website if not. Sketch the algorithm.
6th Week	First check-in
6th-8th Week (before second check-in)	Adjust the code as required after the first Check-in. Finish the main part of the algorithm.
8th Week	Second check-in
8th-10th Week (before presentation)	Complete the software with interface; understand the results; and prepare for the presentation.
10th Week	Update any modification if desired. Presentation.
before March 14th	Final adjustment. Write README.txt file and instruction of the software.

* This is an initial timeline and more detail to be known to provide a better time estimation. These steps may not be carried out in the same order as they appear above and a more iterative approach may be used.