ETL Project Proposal

Team: Two Sunnys (Thidar Swe Tin, Sunmin Lee)

Extract

- Data source 1 (Sunmin): Web scraping product reviews from Sephora https://www.sephora.com/brand/patrick-ta?icid2=meganav_brands_newbrands_patrickta_d_link1
- Data source 2 (Thidar): Web scraping product reviews from Strawberrynet https://www.strawberrynet.com/en-us/main.aspx?gclid=CjwKCAjwy7vlBRACEiwAZvdx9sylZ2e9-xdzlPwXtjlWxigsawtIebRR7fzzVRIdAavucLNu7_pycxoC5vQQAvD_BwE
- Comparing top 10 brands by rating and top 3 products from each of those brands from two sources
    - Categories:
        - Skin care:
            - Eye care / Lip treatments vs Eye & lip
            - Cleansers vs Cleansers
            - Masks vs Masks
            - Moisturizers / Treatments vs Moisturizers & Treatments
            - Sun Care vs Sun Care
    - Brand name
    - Product name / picture
    - Ratings (star)
    - Reviews (count, date) sort by newest (If more than 100, use 100. If less than 100, use counted number)
    - Prices
- Store scraped raw data in MongoDB

Transform - pandas dataframe

- Data cleaning
- Aggregating
    - Average ratings by brands and categories
    - Number of review by brands and categories
    - Average price by brands and categories
- Summarization
    - Top ratings -> top 10 brands -> top 3 products
- Analysis
    - Topic modeling

Load

- Load transformations back to MySQL
    - Database: sunny_db

- Table: sephora