How to Scrape Online Information: With Application to Online Vape Shops

Tobacco Online Policy Seminar (TOPS)

Jun. 17, 2022

Shaoying Ma, PhD; The Ohio State Center for Tobacco Research
Collaborators:

- Ce Shang, PhD; The Ohio State Center for Tobacco Research (CTR)
- Shuning Jiang; OSU Computer Science and Engineering (CSE)
- Zefeng Qiu; OSU CSE
- Jian Chen, PhD; OSU CSE
- Theodore Wagener, PhD; OSU CTR
- Bo Lu, PhD; OSU College of Public Health
- Darren Mays, PhD; OSU CTR
Disclosure:

- This project is funded by the OSU Comprehensive Cancer Center (CCC) - Center for Tobacco Research pilot study mechanism
- We declare no conflict of interests
Overview – e-cigarette data collection from online stores:

- Motivation
- Price, volume, in stock status, and price promotion
- Nicotine strength, and nicotine form (freebase vs salt)
- VG/PG ratio, flavors, and brands
- Package images
- Customer numeric ratings and review contents
Motivation – Why do we need to collect e-cigarette product information from online stores?

• Sales data are necessary to describe the marketplace and inform policies at the federal, state, and local levels.

• However, existing sales data such as Nielsen Retail Scanner data only capture a portion of the market.
  
  • Convenience stores, gas stations, grocery stores, drugstores/pharmacies, and mass merchandiser outlets

• Need surveillance of products sold in vape shops and online stores to capture the full spectrum of e-cigarette products.

Photo source: Ganz et al. 2015 Tobacco Control
E-cigarette purchase locations by consumers in the US:

<table>
<thead>
<tr>
<th>Locations</th>
<th>Adults</th>
<th>AYAs &lt; legal age</th>
<th>AYAs ≥ legal age</th>
</tr>
</thead>
<tbody>
<tr>
<td>Online</td>
<td>23%</td>
<td>16%</td>
<td>14%</td>
</tr>
<tr>
<td>Vape shops</td>
<td>40%</td>
<td>54%</td>
<td>56%</td>
</tr>
<tr>
<td>Other retail stores</td>
<td>37%</td>
<td>30%</td>
<td>30%</td>
</tr>
</tbody>
</table>

Sources: International Tobacco Control (ITC) 4 Country Smoking and Vaping Survey, and ITC Youth Tobacco and Vaping Survey

* AYAs = Adolescents and young adults
Methods to conduct surveillance of e-cigarette products sold online and in vape shops:

- Tobacco **surveys** containing product-related questions
  - Self-reported, subject to measurement errors
  - a variety of pack sizes and volumes - challenging to memorize
  - different models, e.g., disposables, rechargeables, pods, cartridges
The Standardized Tobacco Assessment for Retail Settings: Vape Shops (vSTARS) surveillance tool (Kong et al. 2017; Henriksen et al. 2016)

- Costly – train staff to visit stores and document information
- Capture popular or commonly sold brands – hard to document all products

Photo source: https://countertobacco.org/media-gallery/store-image-maps/
- **Brand websites** surveillance (Hsu et al. 2018)
  - Sampling – brand mentions

- **Social media** surveillance (Lu et al. 2020; Zhou et al., 2018)
  - Focus on marketing, instead of marketplace

- **Web scraping**
  - *Complement* alternative methodologies
  - Not subject to self-reported errors
  - Capture *a wide range of products* – everything *listed* by online stores
  - Capture *a wide range of brands* – do not need a predetermined brand list
  - Complement marketing data with *marketplace data*
  - Consumer ratings – *consumer reception of marketing*
• **Web scraping** complement **Nielsen Retail Scanner** data from **brick-and-mortar stores**
  
  • Nielsen data are quite important, though
    
    • Could take a lot of efforts to **identify product info based on UPC**
    • **Costly process** to obtain nicotine strength, flavors, etc.

• Web scraping algorithms –
  
  • **Automate** and **streamline** the process
  • **Straightforwardly** obtain nicotine strength, flavors, etc.
  • Relatively **costless** – could easily expand without much additional costs
What this project does:

• To address the data limitation in e-cigarette marketplace
  • We scraped data from online stores
    • Price, volume, in stock status, and price promotion
    • Nicotine strength, and nicotine form (freebase vs salt)
    • The ratio of vegetable glycerin (VG) to propylene glycol (PG)
    • Flavors, and brands
    • Package images
    • Customer numeric ratings and review contents
  • Focus on e-liquid products (disposables, devices, and starter kits to be cleaned)
Methods:

• Store selection process
  • We searched Google and Reddit using the key terms “best online vaping stores in 2020” (1/26/2021)
  • Selected 3 stores from Google search and 2 stores from Reddit search
  • Confirmed that these stores sell products nationwide in the US
  • Scrapped product data from online stores between February and May 2021
• 14K unique products from five stores
Five popular online vape shops:

- To mask store identities: referred to as stores 1-5
Website example: store 1

- **Age verification** when visiting website –
  - Confirm the visitor is > 21 and of legal age of smoking in state of residence
- Once enter, *e-cigarette health warning* on homepage
- Multiple tabs – *product types*
Website example: store 1 homepage layout
Web scraping tools:

```python
In [ ]:

def dataFrame(
    'Name':[], 'TotalVolume':[], 'PackNum':[], 'VolumePerPack':[],
    'Price':[], 'OrgPrice':[], 'Nicotine':[], 'VG/PG':[], 'inStock':[],
    'Rating':[], 'ReviewCnt':[], 'Brand':[], 'ProductLink':[], 'ImgLink':[]
):
    df.index.name = 'Index'

product_lists = []
for page_num in tqdm(range(1, 36)):
    # print(f'{page_num} / {777}')
    params = params
    params['page'] = page_num
    params['startIndex'] = (page_num-1) * 20
    response = requests.get('https://www.searchkaise.com/getresults', headers=headers, params=params)
    res = eval(re.search(r'(?<=\("\.(\.|(\?\=)\))', response.text).group()).replace('\\', ' '))
    res['items'][0]['link']

    product_lists.extend(res['items'])

for product in tqdm(product_lists):
    # while True:
    #     try:
    #         product_url = product['link']
    #         product_page = BeautifulSoup(get_content(product_url), 'html5lib')

    # product_img = product_page.find('meta', {'property': 'og:image'})['content']
    # product_name = product_page.find('meta', {'property': 'og:title'})['content']
    # try:
    #     org_price = product_page.find('div', {'class': 'price--compare-at visible'}).find('span', {'class': 'money'}).text.replace('£', '£')
    # except:
    #     org_price = 'N/A'

    try:
        vpg = re.search(r'(?<=vg-pg-ratio)\d+\d+\d+(?<!a>)', str(product_page.find('div', {'class': 'product-description rte'}))).group()}
except:
    try:
        vpg = re.search(r'(?<=ratio)\d+\d+\d+(?<!a>)', str(product_page.find('div', {'class': 'product-description rte'})))
```

Chrome extension “Web Scraper”, and Python
Source code from *store 1* website:

```html
<dl class="product-name hidden-xs">
  <dt>Original Price</dt>
  <dd>$35.95</dd>
  <dt>Current Price</dt>
  <dd>$16.99</dd>
</dl>
```
Embrace a cool and sugary serving of Butter Pecan by No Hype E-Liquid. This delicious blend of crispy buttered pecans meets a rush of cool ice cream to make for a not soon put down.

What's Included
- 1 x 120mL Bottle of No Hype Butter Pecan E-Liquid

Specs & Features
- 30% PG/70% VG
- Primary Flavors: Vanilla Ice Cream and Pecans
**Product Details**

**Name:** Butter Pecan by No Hype E-Liquid (120mL)

**Brand:** NO HYPE

**Volume:** 120mL

**Original Price:** $35.95

**Current Price:** $16.99

**Choose option wrapper (different Nicotine level value here)**

**Description:**

Embrace a cool and sugary serving of Butter Pecan by No Hype E-Liquid. This delicious blend of crispy buttered pecans meets a rush of cool ice cream to make for a scrumptious vape treat you will not soon put down.

**Inclusion:**

- 1 x 120mL Bottle of No Hype Butter Pecan E-Liquid

**Features:**

- **Specs & Features:**
  - 30% PG/70% VG
  - Primary Flavors: Vanilla Ice Cream and Pecans
Online store traffic:

• With data on **organic store traffic** -
  • Keep *monitoring* the popularity of stores in our sample
  • *Expand* our efforts and scrape data from
    • *More stores*
    • *Stores with the heaviest traffic*
  • *Representativeness* - reflect the rapidly changing online market
Data format and standardization:

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Image</td>
<td>Name</td>
<td>Price</td>
<td>OrgPrice</td>
<td>Volume</td>
</tr>
<tr>
<td>305</td>
<td><img src="image1.png" alt="Image" /></td>
<td>360 Triple Melon E-Liquid by Twist (180mL)</td>
<td>$21.95</td>
<td>$29.95</td>
<td>180</td>
</tr>
<tr>
<td>306</td>
<td><img src="image2.png" alt="Image" /></td>
<td>360 Triple Red E-Liquid by Twist (180mL)</td>
<td>$21.95</td>
<td>$29.95</td>
<td>180</td>
</tr>
<tr>
<td>324</td>
<td><img src="image3.png" alt="Image" /></td>
<td>Apple Twist Crisp Apple Smash by Twist E-Liquid (120mL)</td>
<td>$21.95</td>
<td>$24.95</td>
<td>120</td>
</tr>
</tbody>
</table>

- **Standardized price** = \[rac{\text{actual price of the product}}{\text{total volume of the product}}\]

- **Price promotion** = \[rac{\text{original price} - \text{actual price}}{\text{original price}} \times 100\]
E-liquid Price Distribution

- *E*-liquids are very affordable

<table>
<thead>
<tr>
<th>Percentile</th>
<th>5%</th>
<th>10%</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>90%</th>
<th>95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standardized price ($/ml)</td>
<td>0.10</td>
<td>0.12</td>
<td>0.14</td>
<td>0.21</td>
<td>0.37</td>
<td>0.43</td>
<td>0.50</td>
</tr>
</tbody>
</table>
Distribution of e-liquid volume per bottle (Kernel Density Estimates)

- 60 ml, 38%
- 100 ml, 27%
- 120 ml, 7%
- 30 ml, 25%
Price promotion strategies:

- Price promotion (% off) calculated for each e-liquid product sold in our scraped data
  
  \[ \text{Price promotion} = \frac{\text{original price} - \text{actual price}}{\text{original price}} \times 100 \]

- Data of bundled products

- We checked sitewide discounts as well as shipping policies on store websites at two time points, July 7 and September 12, 2021

- Quantity discount in the form of buy x get y free
Promotion strategies in each online vape shop:

<table>
<thead>
<tr>
<th>Promotion Strategies</th>
<th>Stores</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sales (Percentage Off)</td>
<td>All five stores</td>
</tr>
<tr>
<td>Sitewide discount</td>
<td>Stores 1, 2, 3 and 5</td>
</tr>
<tr>
<td>Free shipping</td>
<td>Stores 1, 2, 3 and 5</td>
</tr>
<tr>
<td>Product bundling</td>
<td>Stores 1, 4 and 5</td>
</tr>
<tr>
<td>Buy X get Y free</td>
<td>Store 1</td>
</tr>
</tbody>
</table>

- Average discount (% off) was 39%
- We also documented frequency of pack sizes, and percentage of bundled products in each store
E-liquid flavors: how to code it?

Expand on automatic flavor identification and categorization

• Expand the flavor taxonomy using existing database: WordNet (Princeton University 2010 https://wordnet.princeton.edu/)

• Develop algorithms to extract flavors from the following:
  • Source code of product webpage
  • Flavor filter provided by each online vape shop
  • Product description box on list page (aided by keyword matching)
Flavor data hierarchy

• $N = 833$ key terms
Flavor marketing observation:

- Multiple flavored are often mentioned in different placements: flavor description, flavor filter, and marketing description
  - Example: Berries, Menthol
- Primary vs. secondary flavors are not distinguishable in marketing description or flavor filter
- Concept flavors: e.g., ice, blue razz, refresher
E-liquid flavor – rich information

Summer Vibes By Ripe Vapes
60ml

$23.99  $13.99

Envision palm trees swaying in the wind, tropical sunsets and moonlit beaches while you vape by grabbing Summer Vibes by Ripe Vapes.

NICOTINE
3 MG

In stock

ADD TO CART
**Summer Vibes By Ripe Vapes 60ml Review**

Do you want to envision palm trees swaying in the wind, tropical sunsets and moonlit beaches while you vape? If this sounds too good to be true, you've got to grab Summer Vibes by Ripe Vapes. This island-inspired treat will have you feeling blissed out with its authentic tropical goodness that will make you want to do the hula in your bedroom. A thirst-quenching trio of juicy strawberries, creamy coconut, and sweet bananas is spritzed with the perfect amount of zesty lime to make your palate go wild. Talk about a truly euphoric ADV experience in the comfort of your own home.

At first, that strawberry's brightness lifts your mood, making it seem as though you are indulging in something refreshing. Then, exquisite banana slides down the tongue, teasing your sweet tooth along the way. A bath of coconuts soak the tongue before that citrus splendor liven things up once enough vapor has escaped. You won't be able to resist taking another draw, that's for sure.

Summer Vibes vape juice from Ripe Vapes comes in a medium-sized bottle with a delectable cloud and flavor chasing blend of 75/25 VG/PG that makes it even more blissful for vapers everywhere to have with their favorite mods.

So, what are you waiting for? Now is the time to finally feel those 'vibes' of 'summer' and just enjoy your own fruity vaping paradise, all when simply taking a few pulls of this paradise-filled E-Liquid sensation today. Trust us, you'll be glad you took this getaway as those cravings will also be on vacation as well. Go grab some today while it lasts!

**Package Contents Include:**

- 1 x 60ml bottle of Summer Vibes by Ripe Vapes

**VG/PG: 75/25**
**Flavor: Strawberry, Coconut, Banana, Lime**
Frequencies of E-liquids Containing a Certain Flavor (Product Total N = 14,477)
Double-coding policy-relevant packaging attributes

- Placement of warnings
- Colors
- Flavor descriptions

Warning placement (store 1)
Data visualization tools and algorithms

- Automatically parse out information
- Extract package colors and layout
- Machine coding + **human coding** (two coders who code information independently)

Extracting information from package images
Web data by e-cigarette product types

- Scrape and analyze *e-liquids* as first-step
  - Not well captured in Nielsen data; hard to code volume
  - Web scraping could make the greatest contribution
- Other product data that we have obtained –
  - Disposables, and cartridge-based e-cigarettes
    - Measured in Nielsen data
  - Starter kits, and devices that use with e-liquids
  - CBD products
Nicotine strength in mg/ml, $N = 14,427$

Average nicotine strength is 12 mg/ml.
Nicotine form, and VG/PG ratio:

- Nicotine salts are less bitter and harsh than freebase
- Higher the VG/PG ratio, bigger the vapor clouds
Estimating willingness to pay:

- **Stated preference** data - discrete choice models; hypothetical setting

- **Revealed preference** -
  - Hedonic pricing model; observational data
    - Measure *relative importance* of product attributes
  - \( \log(\text{StandardizedPrice}_{ij}) = \beta_0 + \beta_1 \cdot \text{NicotineStrength}_i + \beta_2 \cdot \text{NicotineForm}_i + \beta_3 \cdot \text{VGPG}_i + \beta_4 \cdot \text{Flavor}_i + s_j + \varepsilon_{ij} \)
    - \( i \) denotes product, and \( j \) denotes store; \( s_j \) are store fixed effects; and \( \varepsilon_{ij} \) are the error terms
  - Coefficients of interest, \( \beta_1 \) through \( \beta_4 \), indicate the % change in the standardized price in response to the change in a certain product attribute
Findings: associations between e-liquid prices and nicotine form

* Generalized estimating equation (GEE) model; both statistically significant at 1% level

Consistent with market observation: uptake of nicotine salts sold by JUUL
Findings: associations between e-liquid prices and flavors, among nicotine salts

- Tobacco/unflavored: 2.77%*
- Menthol: 2.85%*
- Fruit: 2.04%

* Generalized estimating equation (GEE) model; all statistically significant at 10% level
Flavor categories used in the analysis

<table>
<thead>
<tr>
<th>Flavors</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tobacco/unflavored</td>
<td>4%</td>
</tr>
<tr>
<td>Fruit</td>
<td>36%</td>
</tr>
<tr>
<td>Sweets</td>
<td>8%</td>
</tr>
<tr>
<td>Menthol</td>
<td>27%</td>
</tr>
<tr>
<td>Nut/spice/alcohol/other beverages</td>
<td>25%</td>
</tr>
</tbody>
</table>

- 1) containing fruity flavor, without menthol/mint; 2) sweets flavor, without menthol or fruit; 3) containing menthol/mint flavor(s); 4) containing any flavor(s) that are not menthol/mint, fruits, sweets, or tobacco; 5) tobacco flavor only, or unflavored
Findings:

• No statistically significant associations between e-liquid prices and *flavors*, among *nicotine-free* or *freebase nicotine* e-liquids
  • Most e-liquids in our data contain *fruity* flavors
Policy Environment:

• In Jan, 2021, FDA issued warnings to firms that produced and sold e-liquids online without a premarket tobacco product application (PMTA) by Sep. 9, 2021 deadline.

• In our scraped data from online stores (Feb. - May 2021), 241 different e-liquid brands remained in the market, and only 23% of over 14K unique products are nicotine-free.

• Keep monitoring brand availability in online stores and provide data on PMTA enforcement and changes in the online market.
Policy Environment:

- In Apr. 2022, FDA issued two proposals that would *ban menthol flavor in cigarettes* and *all non-tobacco flavors in cigars*
  - If these proposals lead to policy actions, the *flavor availability in e-cigarettes* may *incentivize the transition* from combustible smoking to e-cigarette use.
- Flavor is a main reason of **AYAs experimenting** with e-cigarettes
  - FDA banning flavors other than menthol/mint and tobacco in *cartridge-based* e-cigarettes
  - **Sales growth in disposables** such as puff bars
- Assess importance of flavors to product popularity, and inform about the impact of flavor bans or restrictions
Limitation:

• Unlike **sales** data from Nielsen, web scraping does not allow us to obtain information on **transactions**

Web scraping could serve as a cheaper alternative tool, to help us obtain data that complements existing data sources
Future directions:

• **Online store traffic** will allow us to **improve the sampling**
• Use web scraping tools to **obtain a great amount of data**
  • rich and in-depth product-level information
• **Expand** efforts to scrape data from **more stores**
• **Web data** serve as a **complementary** source to retail scanner data
• **Content analysis** on consumer ratings and **reviews**
  • consumer preferences - which **attributes** are important to them
• Use **algorithms** to extract information from **package images**
Questions and comments

shaoying.ma@osumc.edu
https://shaoyingma.com/

Scan to keep following our web scraping efforts

Scan to read our paper and download the price data

After data cleaning and analysis, our team will publish all data with our research papers