SC1015 Mini-Project:

Skincare Recommendation

SC4

Amabel Lim Hui Xin
Kristine Yeo Kaiyun
Nyan Maw Htun

Practical
Motivation

Motivation

Make **better decisions** when it comes to purchasing **skincare products**

Factors include ingredients and price

Consider similar products to make informed comparisons

Problem Definition

Detection of patterns in the data to produce a personalized recommendation system for skincare products

Is there any relationship between price and rank, within products of the same categories? (i.e. within Moisturizers, Cleansers, etc.)

Are we able to recommend similar products by analysing the ingredients used?

Dataset from Kaggle

cosmeticsdata = pd.read_csv('**cosmetics.csv**')

	Label	Brand	Name	Price	Rank	Ingredients	Combination	Dry	Normal	Oily	Sensitive
0	Moisturizer	LA MER	Crème de la Mer	175	4.1	Algae (Seaweed) Extract, Mineral Oil, Petrolat...	1	1	1	1	1
1	Moisturizer	SK-II	Facial Treatment Essence	179	4.1	Galactomyces Ferment Filtrate (Pitera), Butyle...	1	1	1	1	1
2	Moisturizer	DRUNK ELEPHANT	Protini™ Polypeptide Cream	68	4.4	Water, Dicaprylyl Carbonate, Glycerin, Ceteary...	1	1	1	1	0
3	Moisturizer	LA MER	The Moisturizing Soft Cream	175	3.8	Algae (Seaweed) Extract, Cyclopentasiloxane, P...	1	1	1	1	1
4	Moisturizer	IT COSMETICS	Your Skin But Better™ CC+™ Cream with SPF 50+	38	4.1	Water, Snail Secretion Filtrate, Phenyl Trimet...	1	1	1	1	1
...	...	...	...	...	...	...	...	...	...	...	...
467	Sun protect	KORRES	Yoghurt Nourishing Fluid Veil Face Sunscreen B...	35	3.9	Water, Alcohol Denat., Potassium Cetyl Phospha...	1	1	1	1	1
468	Sun protect	KATE SOMERVILLE	Daily Deflector™ Waterlight Broad Spectrum SPF...	48	3.6	Water, Isododecane, Dimethicone, Butyloctyl Sa...	0	0	0	0	0
469	Sun protect	VITA LIBERATA	Self Tan Dry Oil SPF 50	54	3.5	Water, Dihydroxyacetone, Glycerin, Sclerocarya...	0	0	0	0	0
470	Sun protect	ST. TROPEZ TANNING ESSENTIALS	Pro Light Self Tan Bronzing Mist	20	1.0	Water, Dihydroxyacetone, Propylene Glycol, PPG...	0	0	0	0	0
471	Sun protect	DERMAFLASH	DERMAPROTECT Daily Defense Broad Spectrum SPF 50+	45	0.0	Visit the DERMAFLASH boutique	1	1	1	1	1

https://www.kaggle.com/datasets/kingabzpro/cosmetics-datasets

Data Preparation

Rows of Missing Information

Data viewed using Microsoft excel:

ENRI	Sheer Trar	38	4.2	Visit the OLEHENRIKSEN boutique
MAF	100 percei	48	4.5	Organic Argania Spinosa (Argan) Kernel Oil*. *Organic. **Natural.
SME	Your Skin	38	3.9	Water, Dimethicone, Butylene Glycol Dicaprylate/Dicaprate, Butylene Glycol, Titanium Dioxide [Nano], Tita
ÁLI	Unicorn Es	54	3.9	Water, Propanediol, Glycerin, Polysorbate 20, Glyceryl Polyacrylate, Euterpe Oleracea (Açaí) Fruit Extract, V
IGE	Water Slee	25	4.4	Water, Butylene Glycol, Cyclopentasiloxane, Glycerin, Cyclohexasiloxane, Trehalose, Sodium Hyaluronate,
IGE	Water Bar	35	4.4	Water, Glycerin, Butylene Glycol, Squalane, Dimethicone, Pentaerythrityl Tetraethylhexanoate, BIS-PEG-18
	Facial Trea	99	4.1	Galactomyces Ferment Filtrate (Pitera), Butylene Glycol, Pentylene Glycol, Water, Sodium Benzoate, Methyl
ART+	Premium I	39	4.2	#NAME?

Some data do not display ingredients

```
# removing rows without ingredients

cosmeticsdata = cosmeticsdata[cosmeticsdata["Ingredients"].str.contains("Visit") == False]
cosmeticsdata = cosmeticsdata[cosmeticsdata["Ingredients"].str.contains("No Info") == False]
cosmeticsdata = cosmeticsdata[cosmeticsdata["Ingredients"].str.contains("NAME") == False]
cosmeticsdata = cosmeticsdata[cosmeticsdata["Ingredients"].str.contains("product package") == False]
cosmeticsdata
```

Changing the Indexing

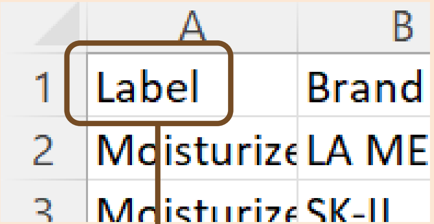
	Label	Brand	Name
0	Moisturizer	LA MER	Crème de la Mer
1	Moisturizer	SK-II	Facial Treatment Essence
2	Moisturizer	DRUNK ELEPHANT	Protini™ Polypeptide Cream
3	Moisturizer	LA MER	The Moisturizing Soft Cream
4	Moisturizer	IT COSMETICS	Your Skin But Better™ CC+™ Cream with SPF 50+
...	...	...	...
1467	Sun protect	KORRES	Yoghurt Nourishing Fluid Veil Face Sunscreen B...
1468	Sun protect	KATE SOMERVILLE	Daily Deflector™ Waterlight Broad Spectrum SPF...
1469	Sun protect	VITA LIBERATA	Self Tan Dry Oil SPF 50
1470	Sun protect	ST. TROPEZ TANNING ESSENTIALS	Pro Light Self Tan Bronzing Mist
1471	Sun protect	DERMAFLASH	DERMAPROTECT Daily Defense Broad Spectrum SPF 50+

Changed the numerical indexes of the dataset to the name of the products

Easier to locate the product

Six Categories of Labels

Data viewed using Microsoft excel:



	A	B
1	Label	Brand
2	Moisturize	LA ME
3	Moisturize	SK-II

Essential to split them up and analyze separately to fit our scope and make a fair comparison

- Moisturizer
- Face Mask
- Eye Cream
- Cleanser
- Treatment
- Sun protect

Exploratory Analysis

Price VS Rank

Moisturizer

	Price	Rank
Price	1.000000	-0.189539
Rank	-0.189539	1.000000

Sun protect

	Price	Rank
Price	1.000000	-0.015988
Rank	-0.015988	1.000000

Treatment

	Price	Rank
Price	1.000000	0.065344
Rank	0.065344	1.000000

Cleanser

	Price	Rank
Price	1.000000	-0.002363
Rank	-0.002363	1.000000

Eye cream

	Price	Rank
Price	1.000000	0.133562
Rank	0.133562	1.000000

Face mask

	Price	Rank
Price	1.00000	-0.03379
Rank	-0.03379	1.00000

Machine
Learning

- NLP
- Dimensionality Reduction

Filtering of Data

```python
# filtering out data that are cleansers for oily skin

dataset1 = cosmeticsdata[cosmeticsdata['Label'] == 'Cleanser'][cosmeticsdata['Oily'] == 1]
dataset1
```

Name	Label	Brand	Price	Rank	Ingredients	Combination	Dry	Normal	Oily	Sensitive
T.L.C. Sukari Babyfacial™	Cleanser	DRUNK ELEPHANT	80	4.5	Water, Glycolic Acid, Hydroxyethyl Acrylate/So...	1	1	1	1	0
T.L.C. Framboos™ Glycolic Night Serum	Cleanser	DRUNK ELEPHANT	90	4.3	Water, Glycolic Acid, Butylene Glycol, Glyceri...	1	1	1	1	0
Green Clean Makeup Meltaway Cleansing Balm with Echinacea GreenEnvy™	Cleanser	FARMACY	34	4.6	Cetyl Ethylhexanoate, Caprylic/Capric Triglyce...	1	1	1	1	1
Purity Made Simple Cleanser	Cleanser	PHILOSOPHY	24	4.5	Water, Sodium Lauroamphoacetate, Sodium Tridec...	1	1	1	1	1
The Rice Polish Foaming Enzyme Powder	Cleanser	TATCHA	65	4.4	Microcrystalline Cellulose, Oryza Sativa (Rice...	1	1	1	1	1
...	...	...	...	...	...	...	...	...	...	...
Rosa Centifolia™ No.1 Purity Cleansing Balm	Cleanser	REN CLEAN SKINCARE	32	4.2	Prunus Amygdalus Dulcis (Sweet Almond) Oil, Ce...	1	1	1	1	1
Blue Herbal Acne Cleanser Treatment	Cleanser	KIEHL'S SINCE 1851	22	3.5	Water, Coco-Glucoside, Propylene Glycol, Ammon...	1	0	0	1	0
Pore Refining Detox Double Cleanse	Cleanser	ERNO LASZLO	55	5.0	Water, Propanediol, Sodium C14-16 Olefin Sulfo...	1	1	1	1	1
Herbal-Infused Micellar Cleansing Water	Cleanser	KIEHL'S SINCE 1851	28	3.7	Water, Glycerin, Propanediol, Melissa Officina...	1	1	1	1	1
Refreshing Gel Cleanser	Cleanser	CLARISONIC	19	5.0	Water, Glycerin, Coco-Betaine, Sodium Cocoyl G...	1	1	1	1	1

147 rows × 10 columns

Lexical Analysis (Tokenization)

- Splitting the ingredient list into single word items

```
# tokenisation of the ingredients list

index = 0
ingredient_dict = {}
corpus = []

for i in range(len(dataset1)):
    ingredients = dataset1['Ingredients'][i]
    ingredients_lower = ingredients.lower()        # change all to lower case
    tokens = ingredients_lower.split(', ')         # split up the ingredients from the string
    corpus.append(tokens)

    for ingredient in tokens:
        if ingredient not in ingredient_dict:      # prevents duplication
            ingredient_dict[ingredient] = index
            index += 1
```

Ingredients
Algae (Seaweed) Extract, Mineral Oil, Petrolat...
Galactomyces Ferment Filtrate (Pitera), Butyle...
Water, Dicaprylyl Carbonate, Glycerin, Ceteary...
Algae (Seaweed) Extract, Cyclopentasiloxane, P...
Water, Snail Secretion Filtrate, Phenyl Trimet...

One-hot Encoding

1 – Present , 0 – Absent

Categorical, nominal (named categories) data

Example

Color
Red
Red
Yellow
Green
Yellow

Red	Yellow	Green
1	0	0
1	0	0
0	1	0
0	0	1

https://www.kaggle.com/code/dansbecker/using-categorical-data-with-one-hot-encoding/notebook

Ingredients

```
array([[1., 1., 1., ..., 0., 0., 0.],
       [1., 1., 0., ..., 0., 0., 0.],
       [1., 0., 0., ..., 0., 0., 0.],
       ...,
       [1., 0., 0., ..., 0., 0., 0.],
       [1., 0., 0., ..., 0., 0., 0.],
       [1., 0., 0., ..., 1., 1., 1.]])
```

Dimensionality Reduction

- **Unsupervised machine learning:**
  - only input data of ingredients to train the model
  - no output variables to predict
- Methods include **UMAP**, PCA, t-SNE

UMAP

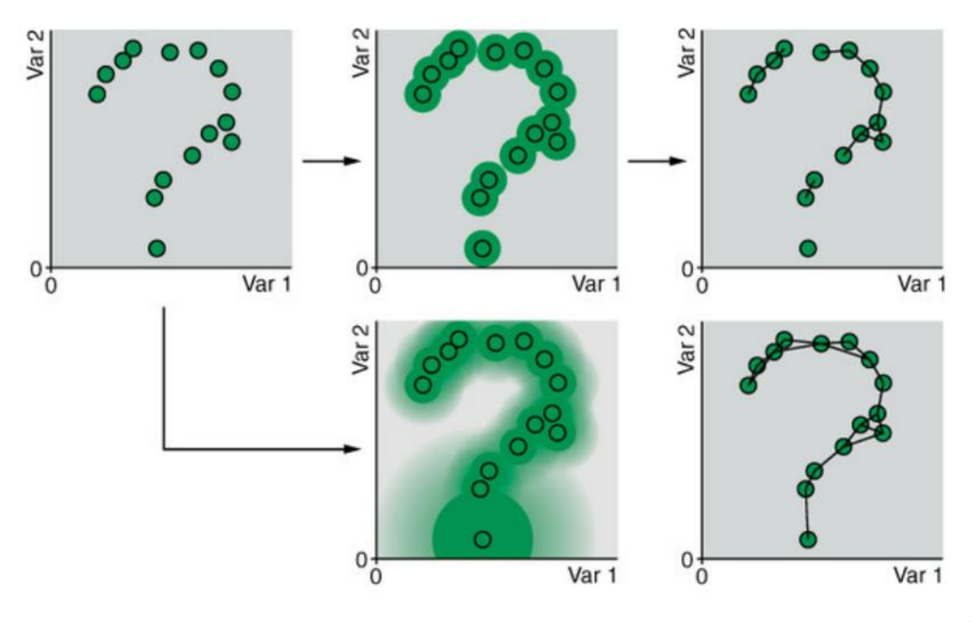
```
# installing UMAP

!pip install umap-learn

import umap

import numba
```

- Predicts a manifold
- A search region around a point to detect neighboring points
- Additional search region (Fuzzy) – larger for lower-density areas
- Iteratively shuffle this manifold until distances are like the original

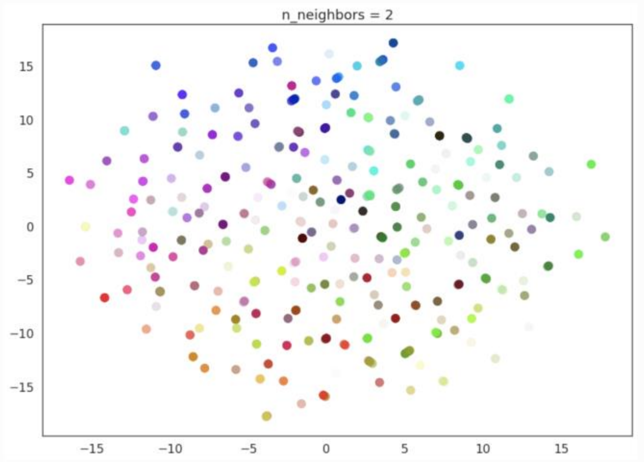


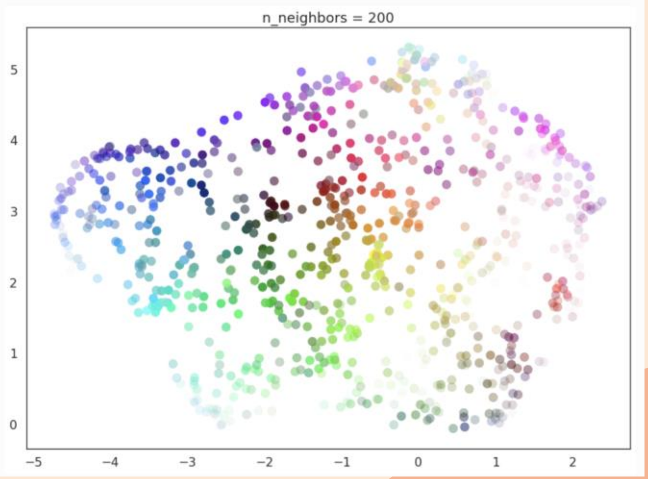
https://livebook.manning.com/book/machine-learning-for-mortals-mere-and-otherwise/chapter-14/73

Hyperparameters

n_neighbors

- controls the radius of the fuzzy search region
    - range from 1 – # of data
    - larger values = more focused on global structure
    - smaller values = more focused on local structure



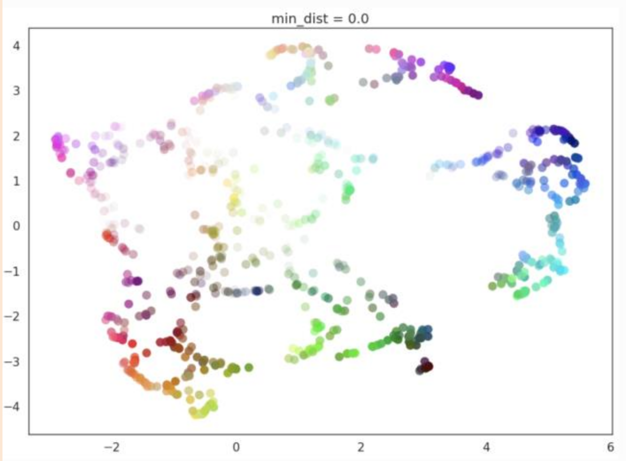


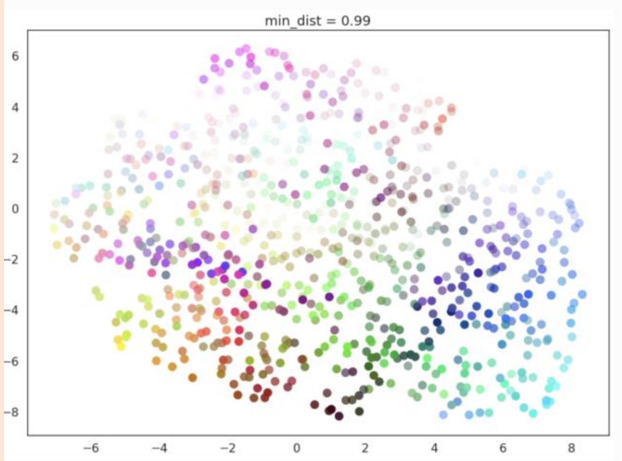
https://umap-learn.readthedocs.io/en/latest/parameters.html#:~:text=provided%20by%20n_neighbors%20.,min_dist,will%20result%20in%20clumpier%20embeddings.

Hyperparameters

min_dist

- controls the minimum distance apart to select data points to be used in the lower-dimensional representation
    - Ranges from 0 - 1
    - Low values – More clustered
    - High values – Sparser out



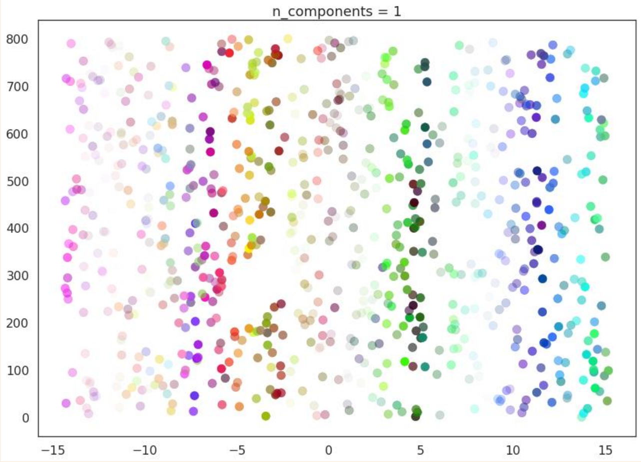


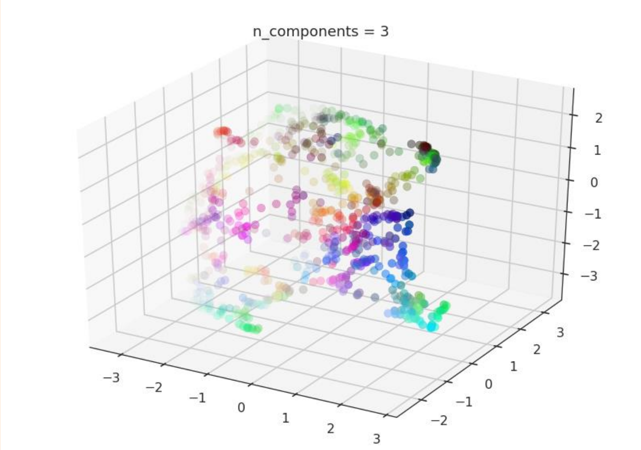
https://umap-learn.readthedocs.io/en/latest/parameters.html#:~:text=provided%20by%20n_neighbors%20.-,min_dist,will%20result%20in%20clumpier%20embeddings.

Hyperparameters

n_components

- determines the dimensionality of the reduced dimension space





https://umap-learn.readthedocs.io/en/latest/parameters.html#:~:text=provided%20by%20n_neighbors%20.,min_dist,will%20result%20in%20clumpier%20embeddings.

Our UMAP

- **2-dimension**
- Easy comparison and visualization of data

- **Smaller value**
- Filtered data -> more focus on the fine details of the data points (local structure > global structure)

```
# dimension reduction with UMAP

umap_data = umap.UMAP(n_components = 2, min_dist = 0.7, n_neighbors = 5, random_state = 1).fit_transform(matrix)

# adding 2 new columns X and Y to the dataset

dataset1['X'] = umap_data[:, 0]
dataset1['Y'] = umap_data[:, 1]

dataset1
```

- **Larger value** closer to 1
- Prevent the clustering of points for more accurate comparisons

Why UMAP?

	UMAP	t_SNE
Learn non-linear patterns	√	√
Make predictions on new data	√	
Preserves both local and global distances	√	(only local structure)
Time-efficient	√	

Limitations of UMAP

Hyperparameters

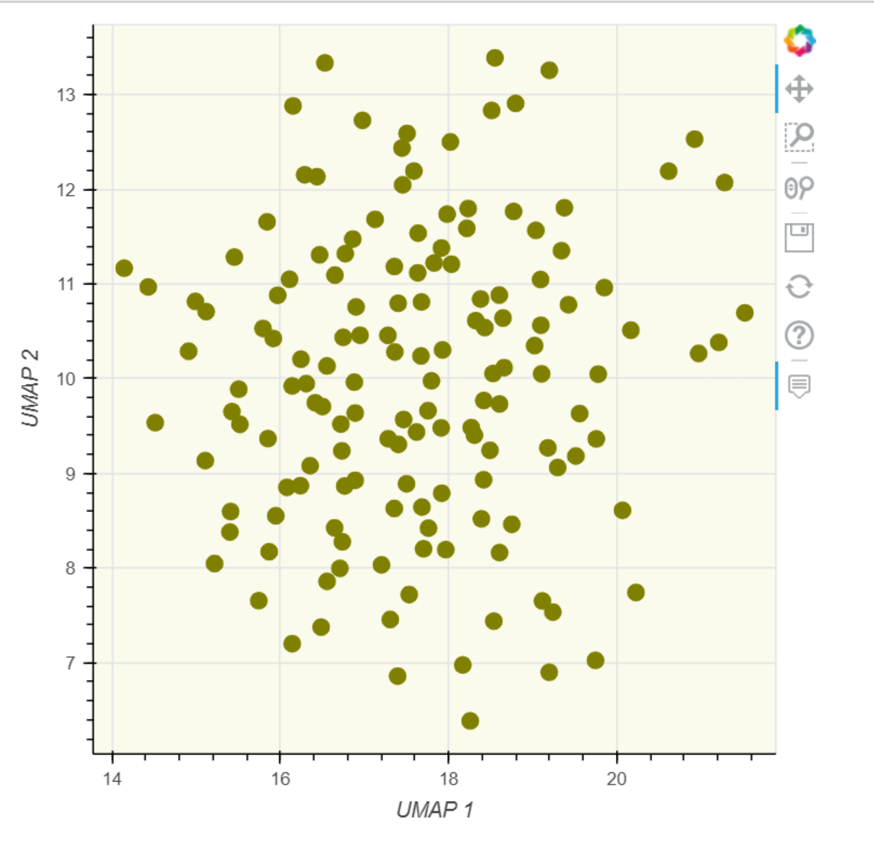
Choosing right values is not easy

Stochastic

Different runs could yield different results
BUT faster execution time

Analytic Visualisation

Bokeh Graph



Statistical Inference

Euclidean Distance

```
dataset1['Distance'] = 0.0

from math import dist

# using Greek Yoghurt Foaming Cream Cleanser as an example

myItem = dataset1.loc[['Greek Yoghurt Foaming Cream Cleanser']]

point1 = np.array([myItem['X'], myItem['Y']])
point1
```

```
array([[16.978378],
       [12.727462]], dtype=float32)
```

```
# other items

for i in range(len(dataset1)):
    point2 = np.array([dataset1['X'][i], dataset1['Y'][i]])
    dataset1.Distance[i] = dist(point1, point2)
```

```
# sorting data in ascending order

dataset1 = dataset1.sort_values('Distance')
dataset1.head(6)
```

Name	Label	Brand	Price	Rank	Ingredients	Combination	Dry	Normal	Oily	Sensitive	X	Y	Distance
Greek Yoghurt Foaming Cream Cleanser	Cleanser	KORRES	26	4.6	Water, Sodium Cocoyl Isethionate, Cocobetaine,...	1	1	1	1	1	16.978378	12.727462	0.000000
Treatment Cleansing Foam	Cleanser	AMOREPACIFIC	50	4.5	Water, Glycerin, Stearic Acid, Myristic Acid, ...	1	0	1	1	0	17.507780	12.589053	0.547196
ExfoliKate® Intensive Exfoliating Treatment	Cleanser	KATE SOMERVILLE	24	4.4	Water, Lactic Acid, Silica, Glycine Soja (Soyb...	1	1	1	1	0	17.448727	12.434882	0.553923
Soy Face Cleansing Milk	Cleanser	FRESH	38	3.9	Water, Caprylic/Capric Triglyceride, Caprylic/...	1	1	1	1	1	16.531837	13.334742	0.753782
New Day Gentle Exfoliating Grains	Cleanser	FARMACY	30	4.5	Sodium Cocoyl Isethionate, Zea Mays (Corn) Sta...	1	1	1	1	1	16.437346	12.132984	0.803816
Fresh Pressed Renewing Powder Cleanser with Pure Vitamin C	Cleanser	CLINIQUE	29	4.9	Maltodextrin , Sodium Lauryl Sulfoacetate , So...	1	1	1	1	1	17.590328	12.192686	0.812692

Conclusion

Outcome & Insights

- Low correlation between products' rank and price → more expensive products ≠ better products
- Created a personalized recommendation system that recommends products that are similar to consumer's choice of product

Learning outcomes

- Natural Language Processing (NLP)
    - Tokenisation
- Dimensionality reduction (UMAP)

Verification of UMAP

- trial and error of hyperparameter values
- Trustworthiness:

$$T(k) = 1 - \frac{2}{nk(2n - 3k - 1)} \sum_{i=1}^{n} \sum_{j \in U_i^{(k)}} \left(r(i,j) - k\right)$$

- Continuity:

$$C(k) = 1 - \frac{2}{nk(2n - 3k - 1)} \sum_{i=1}^{n} \sum_{j \in V_i^{(k)}} \left(\hat{r}(i,j) - k\right)$$

Source: “Semantically Controlled Adaptive Equalisation in Reduced Dimensionality Parameter Space”, Stasis et al 2016

Thank You