Categorizing E-Commerce Products by Natural Language Processing Method

Chang Zhang

2019/08/26

43000252352	lot of 2 baker s 4 oz. sweet chocolate bar german s all natural 48 cacao	Confectionery/Sugar Sweetening Products	Confectionery Products
43000000564	mio liquid water enhancer strawberry watermelon 1.62 fluid ounce	Beverages	Non Alcoholic Beverages
43000029220	maxwell house original medium roast ground coffee 11.5 oz	Beverages	Coffee/Tea/Substitutes
43000000670	mio liquid water enhancer energy black cherry 1.08oz squirt bottle	Beverages	Non Alcoholic Beverages
29000021327	planters raw mixed nuts 5.5 oz	Fruits/Vegetables/Nuts/ Seeds Prepared/Processed	Fruit/Nuts/Seeds Combination
29000021334	planters raw cashews unsalted 5.5 ounce bag	Beverages	Coffee/Tea/Substitutes

Executive Summary

Objective:

Develop a code framework to scrape UPC codes from webpages and automate product categorization process. The results can help human l

Steps taken to achieve the objective:

1. Scraped 1,200 UPC codes from *www.barcodelookup*.com (website has a max query limit per day) .
2. Cleaned data and wrangled data into a structured .csv format.
3. Grouped UPC codes according to their brands, dimensions, size, color, etc.
4. Categorized UPC codes into family category and class category of GS1 codes using *Bag-of-Words (BOW) doc2vec* approach.
5. Evaluated model results by manually checking the categorization outcome.

Results:

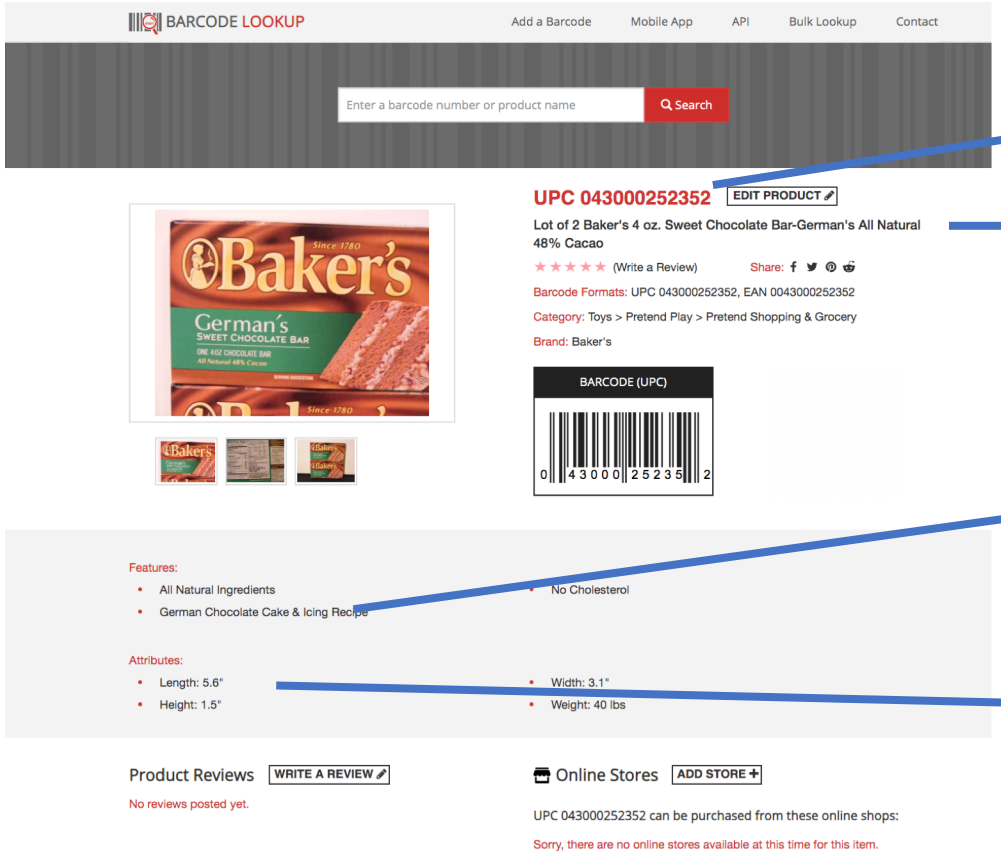
1. Successfully developed a code framework to achieve the process described above.
2. Categorization accuracy (based on a sample of 101 items, checked by human):
    - **Family category (16 categories): 92.1%**
    - **Class category (109 sub categories): 86.1%**

Outline

- 1 Webpage Scraping
- 2 Data Cleaning and Data Structuralization
- 3 Product Categorization based on doc2vec Approach
- 4 Model Evaluation
- 5 Conclusion and Recommendation

1 Webpage Scraping

I firstly used *beautifulsoup* (A python library) to scrape important information associated with each UPC code on *barcodelookup* website.



https://www.barcodelookup.com/

UPC code: 043000252352

Title: Lot of 2 Baker's 4 oz. Sweet Chocolate Bar-German's All Natural 48% Cacao

Features: All natural ingredients, no cholesterol, German chocolate cake and icing recipe.

Attributes: Length 5.6", Width 3.1", Height 1.5", Weight 40lbs

2 Data Cleaning and Data Structuralization

Title: Lot of 2 Baker's 4 oz. Sweet Chocolate Bar-German's All Natural 48% Cacao

Attributes: Length 5.6", Width 3.1", Height 1.5", Weight 40lbs

Cleaning text by removing special chars and punctuations

Key word matching

UPC code	Product Name	Brand	Description	Dimensions					
43000252352	lot of 2 baker s 4 oz. sweet chocolate bar german s all natural 48 cacao	Baker's	all natural ingredients no cholesterol german chocolate cake icing recipe length 5.6" width 3.1" height 1.5" weight 40 lbs	1.5"	3.1"	5.6"	40lbs	NA	NA
43000000564	mio liquid water enhancer strawberry watermelon 1.62 fluid ounce	Mio	make your mio. flip it. unlock the flavor. tip it. each squeeze into water adds more flavor. sip it. your drink your way. click it twice to lock it tight contains 0g sugar. 0 juice. 0mg caffeine serving each bottle makes 24 servings 8 ounce kosher length 4.5" width 2.2" height 1.1" weight 15 lbs	1.1"	2.2"	4.5"	15lbs	NA	NA
43000029220	maxwell house original medium roast ground coffee 11.5 oz	Kraft Foods	weight 0.88 lbs size 11.5 oz color brown	NA	NA	NA	0.88lbs	brown	11.5oz
43000000670	mio liquid water enhancer energy black cherry 1.08oz squirt bottle	Mio	mio liquid water enhancer energy black cherry 1.08oz squirt bottle	NA	NA	NA	NA	NA	NA
29000021327	planters raw mixed nuts 5.5 oz	Planters	recloseable packaging 0g trans fat 6 servings per package 170 calories per serving length 2.5" width 5.25" height 8" weight 34 lbs	8"	5.25"	2.5"	34lbs	NA	NA

Then I structuralize the scraped text into a structured .csv file with each column containing one attribute.

3 Product Categorizing based on *doc2vec* Approach

- BOW doc2vec

		red	green	apple	pepper	juice	salad
UPC	**green apple juice**	(0,	1,	1,	0,	1,	0)
GS1	**apple juice**	(0,	0,	1,	0,	1,	0)
GS1	**apple salad**	(0,	0,	0,	0,	0,	1)

More similar

Less similar

- I converted each UPC and GS1 code into vectors V_{UPC} and V_{GS1} that represents the occurrence of its words (BOW).

- Then I check the similarity between V_{UPC} and each V_{GS1}

$$\text{Similarity} = \cos(V_{UPC}, V_{GS1})$$

- The GS1 code that has the highest similarity determines the category of UPC code.

Red apple juice: -> beverage, ready to drink

3 Product Categorizing based on *doc2vec* Approach

Sample results:

UPC	title	family	class
43000252352	lot of 2 baker s 4 oz. sweet chocolate bar german s all natural 48 cacao	Confectionery/Sugar Sweetening Products	Confectionery Products
43000000564	mio liquid water enhancer strawberry watermelon 1.62 fluid ounce	Beverages	Non Alcoholic Beverages
43000029220	maxwell house original medium roast ground coffee 11.5 oz	Beverages	Coffee/Tea/Substitutes
43000000670	mio liquid water enhancer energy black cherry 1.08oz squirt bottle	Beverages	Non Alcoholic Beverages
29000021327	planters raw mixed nuts 5.5 oz	Fruits/Vegetables/Nuts/ Seeds Prepared/Processed	Fruit/Nuts/Seeds Combination
29000021334	planters raw cashews unsalted 5.5 ounce bag	Beverages	Coffee/Tea/Substitutes

4 Model Evaluation

- I manually checked categorization results of 101 data points. The mis-categorized UPC codes are marked **red** in the table on the right. (the manually checked table is in the package.)

- Based on the sample, the **categorization accuracy on family category is 92.1%;** and the **categorization accuracy on class category is 86.1%.**

UPC_code	title	brand	description	height	width	length	weight	color	size	family	class
4.3E+10	lot of 2 baker s 4 oz. sweet chocolate bar german s all natural 4	Baker's	all natural ir	1.5"	3.1"	5.6"	40lbs	NA	NA	Confectioner	Confectionery Products
4.3E+10	mio liquid water enhancer strawberry watermelon 1.62 fluid ou	Mio	make your n	1.1"	2.2"	4.5"	15lbs	NA	NA	Beverages	Non Alcoholic Beverages ,Äì
4.3E+10	maxwell house original medium roast ground coffee 11.5 oz	Kraft Foods	weight 0.88	NA	NA	NA	0.88lbs	brown	11.5oz	Beverages	Coffee/Tea/Substitutes
4.3E+10	mio liquid water enhancer energy black cherry 1.08oz squirt bott	Mio	mio liquid w	NA	NA	NA	NA	NA	NA	Beverages	Non Alcoholic Beverages ,Äì
2.9E+10	planters raw mixed nuts 5.5 oz	Planters	recloseable	8"	5.25"	2.5"	34lbs	NA	NA	Fruits/Veget	Fruit/Nuts/Seeds Combination
2.9E+10	planters raw cashews unsalted 5.5 ounce bag	Kraft Heinz	unsalted but	2.3"	5.2"	7.3"	35lbs	NA	NA	Beverages	Coffee/Tea/Substitutes
2.9E+10	planters raw almonds 5.5 ounce	Kraft Heinz	recloseable	1.3"	5.1"	7.3"	35lbs	NA	NA	Beverages	Alcoholic Beverages
2.1E+10	kraft cheese medium cheddar bar 8 oz	Kraft Foods	kraft mediu	2.25"	5.75"	1"	NA	cheese	NA	Milk/Butter/	Cheese/Cheese Substitutes
2.1E+10	kraft cheese mild cheddar chunk 8 oz	Kraft Foods	kraft mild ch	2.46"	5.87"	1.2"	NA	cheese	NA	Milk/Butter/	Cheese/Cheese Substitutes
4.3E+10	kool aid drink mix tropical punch 1.62 fl oz 1 count	Kraft	weight 0.15	NA	NA	NA	0.15lbs	NA	NA	Beverages	Non Alcoholic Beverages ,Äì
4.3E+10	kool aid drink mix cherry 1.62 fl oz 1 count	Kraft Foods (	weight 0.15	NA	NA	NA	0.15lbs	NA	1.62fl	Beverages	Non Alcoholic Beverages ,Äì
4.3E+10	mio energy liquid water enhancer tropical fusion 1.62 fl oz 1 cou	Kraft Heinz F	weight 0.15	NA	NA	NA	0.15lbs	NA	1.62fl	Beverages	Non Alcoholic Beverages ,Äì
4.3E+10	crystal light liquid iced tea drink mix mandarin orange flavor 1.6	Crystal Light	0 calories pe	NA	NA	NA	NA	NA	NA	Beverages	Coffee/Tea/Substitutes
4.3E+10	crystal light liquid drink mix strawberry green tea	Kraft Foods	length 1.16"	3.73"	1.79"	1.16"	10lbs	NA	NA	Beverages	Coffee/Tea/Substitutes
4.3E+10	crystal light liquid energy drink tropical coconut 1.62 ounce	Crystal Light	package ima	3.73"	1.79"	1.16"	10lbs	NA	NA	Beverages	Non Alcoholic Beverages ,Äì
4.3E+10	crystal light drink mix blackberry lemonade 1.62 fl oz 1 count	Kraft Heinz F	weight 0.16	NA	NA	NA	0.16lbs	NA	23oz	Beverages	Non Alcoholic Beverages ,Äì
4.3E+10	maxwell house the original roast ground coffee refill	Maxwell Hou	weight 0.74	NA	NA	NA	0.74lbs	NA	11.5oz	Beverages	Coffee/Tea/Substitutes
4.3E+10	maxwell house 100 colombian medium 10.5 oz	Kraft Foods	size 10.5 oz	NA	NA	NA	NA	NA	10.5oz	Beverages	Alcoholic Beverages
4.3E+10	maxwell house breakfast blend ground coffee 11 oz. brick	Kraft Foods	size 11 oz	NA	NA	NA	NA	NA	11oz	Beverages	Coffee/Tea/Substitutes
4.3001E+10	maxwell house light roast coffee grounds master blend 11.5 oz	Kraft Heinz F	size 11.5 oz	NA	NA	NA	NA	NA	11.5oz	Beverages	Coffee/Tea/Substitutes
8.7684E+10	capri sun 100 juice apple 6 fl oz 10 ct	Capri Sun	size 6 fl oz c	NA	NA	NA	NA	other	6fl	Beverages	Non Alcoholic Beverages ,Äì
2.1E+10	kraft macaroni cheese star wars shapes 19.0 oz	Kraft Foods	size 1.9 oz	NA	NA	NA	NA	NA	1.9oz	Milk/Butter/	Cheese/Cheese Substitutes
4.3E+10	gevalia mocha latte espresso coffee k cup packs froth packets 9	Gevalia Kaffe	size 23 oz	NA	NA	NA	NA	NA	23oz	Beverages	Coffee/Tea/Substitutes
4.3E+10	maxwell house caf collection single serve coffee house blend co	Maxwell Hou	size med co	NA	NA	NA	NA	coffee	medcolor	Beverages	Coffee/Tea/Substitutes
2.1E+10	kraft natural sharp cheddar cheese chunk 8oz	Kraft	weight 0.51	NA	NA	NA	0.51lbs	NA	NA	Milk/Butter/	Cheese/Cheese Substitutes
2.1E+10	kraft cheese chunk cheese monterey jack 8 oz	Kraft Foods	natural chee	2.25"	5.75"	1"	NA	NA	NA	Milk/Butter/	Cheese/Cheese Substitutes
4.3E+10	mccaf french roast ground coffee 12 oz. bag	Kraft Heinz F	color coffee	NA	NA	NA	NA	coffee	NA	Beverages	Coffee/Tea/Substitutes
4.3E+10	mccaf french roast ground coffee 12 oz. bag	Kraft Heinz F	color coffee	NA	NA	NA	NA	coffee	NA	Beverages	Coffee/Tea/Substitutes
4.3E+10	mccaf premium roast ground coffee 12 oz. bag	Kraft Heinz F	weight 0.8 l	NA	NA	NA	0.8lbs	coffee	medcolor	Beverages	Coffee/Tea/Substitutes
4.3E+10	mccaf premium roast ground coffee 12 oz. bag	Kraft Heinz F	weight 0.8 l	NA	NA	NA	0.8lbs	coffee	medcolor	Beverages	Coffee/Tea/Substitutes
4.3E+10	mccafe breakfast blend light roast ground coffee 12 oz 340g	Kraft Heinz F	size 12 oz co	NA	NA	NA	NA	coffee	12oz	Beverages	Coffee/Tea/Substitutes
4.3E+10	maxwell house original medium roast ground coffee 11.5 oz	Kraft Foods	weight 0.88	NA	NA	NA	0.88lbs	brown	11.5oz	Beverages	Coffee/Tea/Substitutes
4.3E+10	maxwell house original medium roast ground coffee 11.5 oz	Kraft Foods	weight 0.88	NA	NA	NA	0.88lbs	brown	11.5oz	Beverages	Coffee/Tea/Substitutes
4.3E+10	maxwell house french roast ground coffee 11 oz. canister	Kraft Heinz F	weight 0.85	NA	NA	NA	0.85lbs	NA	11oz	Beverages	Coffee/Tea/Substitutes
4.3E+10	maxwell house breakfast blend ground coffee 11 oz. canister	Kraft Heinz F	size 11.0 oz	NA	NA	NA	NA	NA	11.0oz	Beverages	Coffee/Tea/Substitutes
4.3E+10	maxwell house breakfast blend ground coffee 11 oz. canister	Kraft Heinz F	size 11.0 oz	NA	NA	NA	NA	NA	11.0oz	Beverages	Coffee/Tea/Substitutes
4.3E+10	maxwell house lite ground coffee	Maxwell Hou	size 11 oz	NA	NA	NA	NA	NA	11oz	Beverages	Coffee/Tea/Substitutes

5 Conclusions and Recommendations

- We have successfully developed a code framework to scrape UPC codes from webpages and , which can significantly reduce human work.
- As we are dealing with a large dataset (>30 MB), it takes a long time to finish the code on a personal computer. Thus, I recommend take the following steps to speed up the process:

1. Label part of the codes manually and then train the model on the labeled data to improve accuracy.

2. Build an image classifier and use images to assist categorization process.

3. Purchase web API from *www.upcitemdb.com* to enable unlimited fast UPC code scraping.

4. Purchase *AWS cloud computing service* to speed up the categorization process.