

Cyclops Project

**Data Science Division, Statistics Canada
Presentation 1 - October 24, 2019**



Delivering insight through data for a better Canada



Statistics
Canada

Statistique
Canada

Canada

Project Scope

Objectives	Timeline
1. Image processing: Preprocessing: Contrast improvement, thresholding (binary, adaptive mean and Gaussian), Rotation, etc.	October 2019 - January 2020
2. OCR extraction: text detection, text recognition, end-to-end	October 2019 - January 2020
3. NLP content recognition and compliance verification: NHP number, List of ingredients (medical and non-medical), including cross-referencing alternative ingredient names, Each ingredient's amount per unit, Recommended dosage and cross-reference with HC database, Claim.	January 2020 – February 2020
4. Claim compliance assessment: Supervised (compliant/non-compliant labels available) and feasibility of unsupervised (time permits)	February 2020
5. Integration and deployment: Prototype user interface (web application) Set up earlier?	March 2020



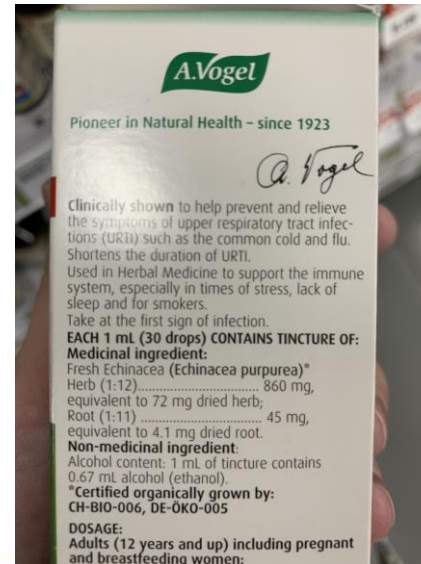
Variability in Products

Variability in NPN

100



Variability in Claims



Statistics
Canada

Statistique
Canada

Delivering insight through data for a better Canada

Canada

[illegible]

Contra-indications: Do not use if you are pregnant or breastfeeding. Keep out of reach of children.

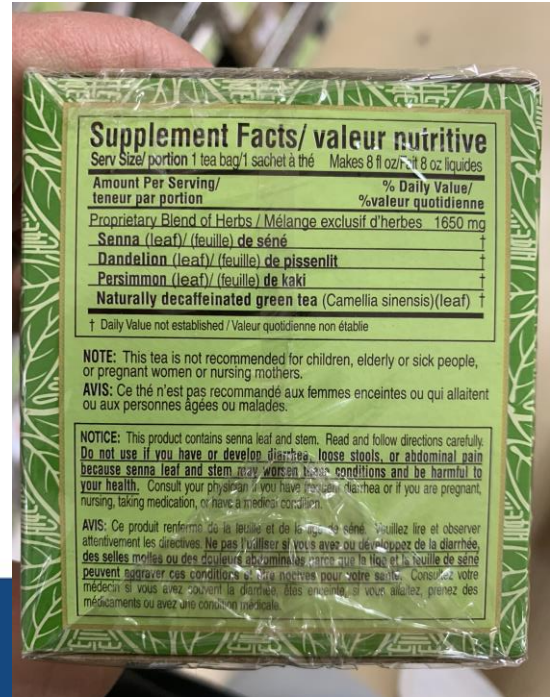
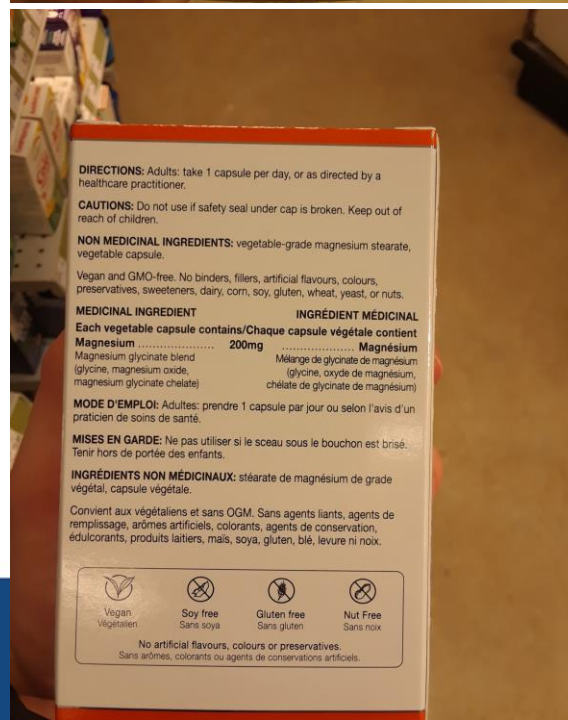
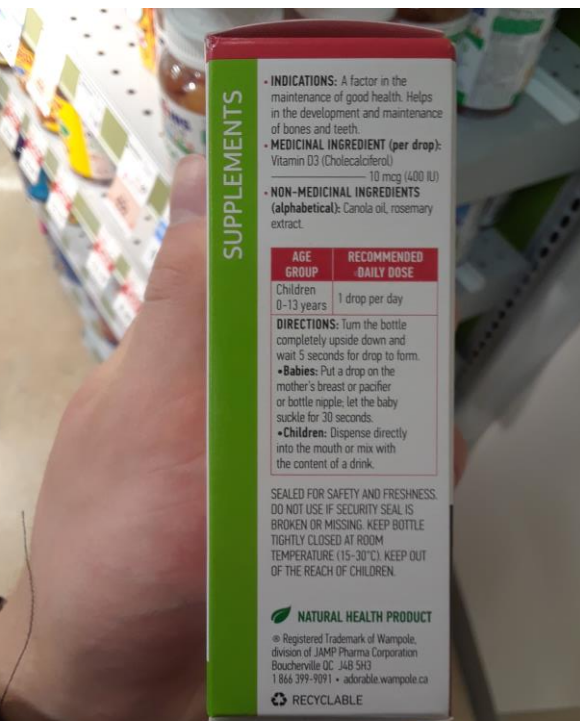
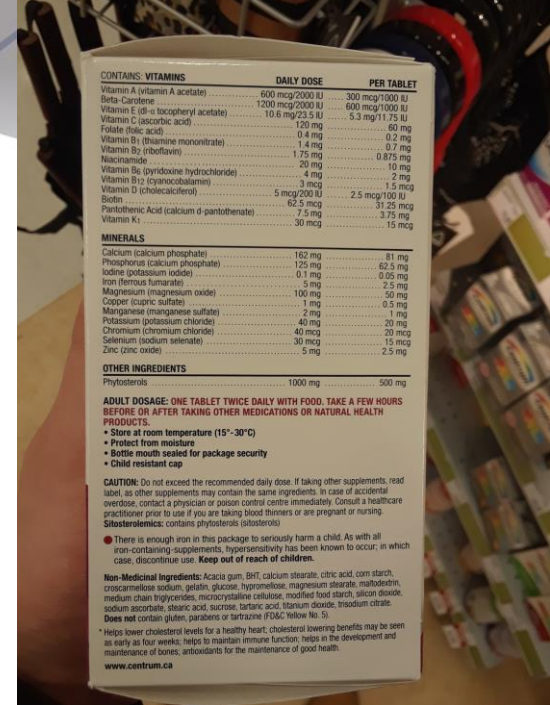
Warning: Discontinue use if you develop symptoms of liver trouble (abdominal pain, dark urine, jaundice). Consult a health care practitioner prior to use if you have diabetes, gallstones, stomach ulcers or excess stomach acid, have a history of non-melanoma skin cancer, are taking blood thinners, antiplatelet medication or protease inhibitors, if you have cystinuria, or are taking nitroglycerin, or if taking antibiotics. Do not use if you have a bile duct obstruction and/or if you have liver or gall bladder disorder, and/or bowel obstruction. Do not use if you are allergic to plants of the

Manufactured by / Fabriqué par : The Hydration Pharmaceuticals Trust, Victoria 3076 Australia.
Imported by / Importé par : ANB Canada Inc., Newmarket, ON L3X 3C7.
Phone / Téléphone : 904 550 9810, www.hydrations.com

[illegible]

10

Variability in Ingredients



Step 1: Optical Character Recognition

Use Pre-trained models to detect text

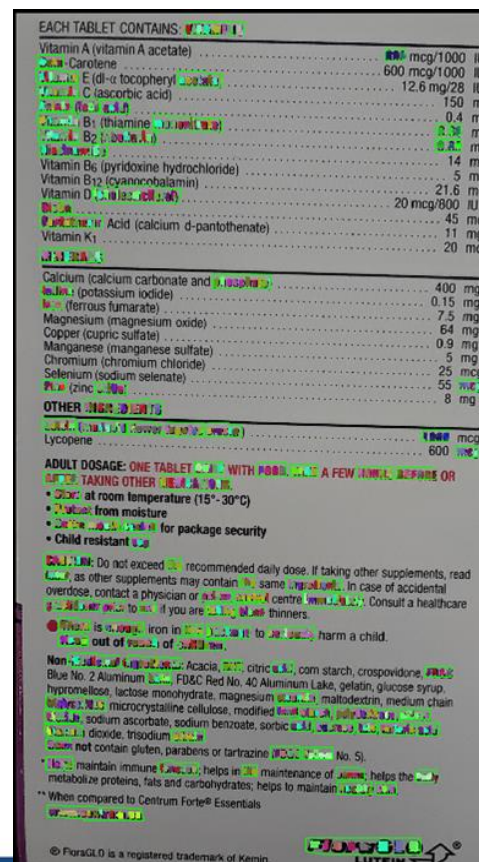
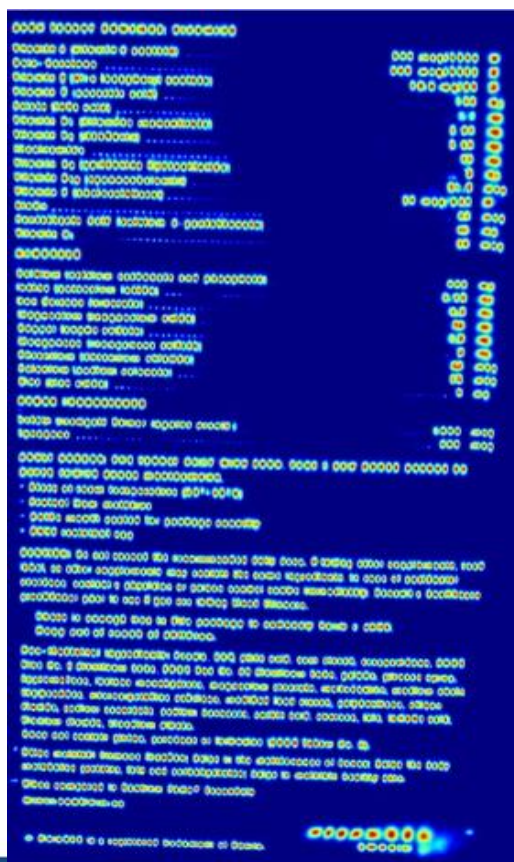
	DESCRIPTION	PROS	CONS
CRAFT-PYTORCH (DETECTOR)	Finds where the text is in the image. State of Art.	Finds alphabets and numbers well.	Breaks up the words with no formatting; just coordinates
MASK TEXT SPOTTER (DETECTOR & RECOGNIZER)	Finds where the text is in the image. & Prints out the text in the image digitally	None.	Cannot detect text in the image when the text is small or when too many texts in an image

Step 1: Optical Character Recognition

CRAFT-PYTORCH

VS.

MASK TEXT SPOTTER



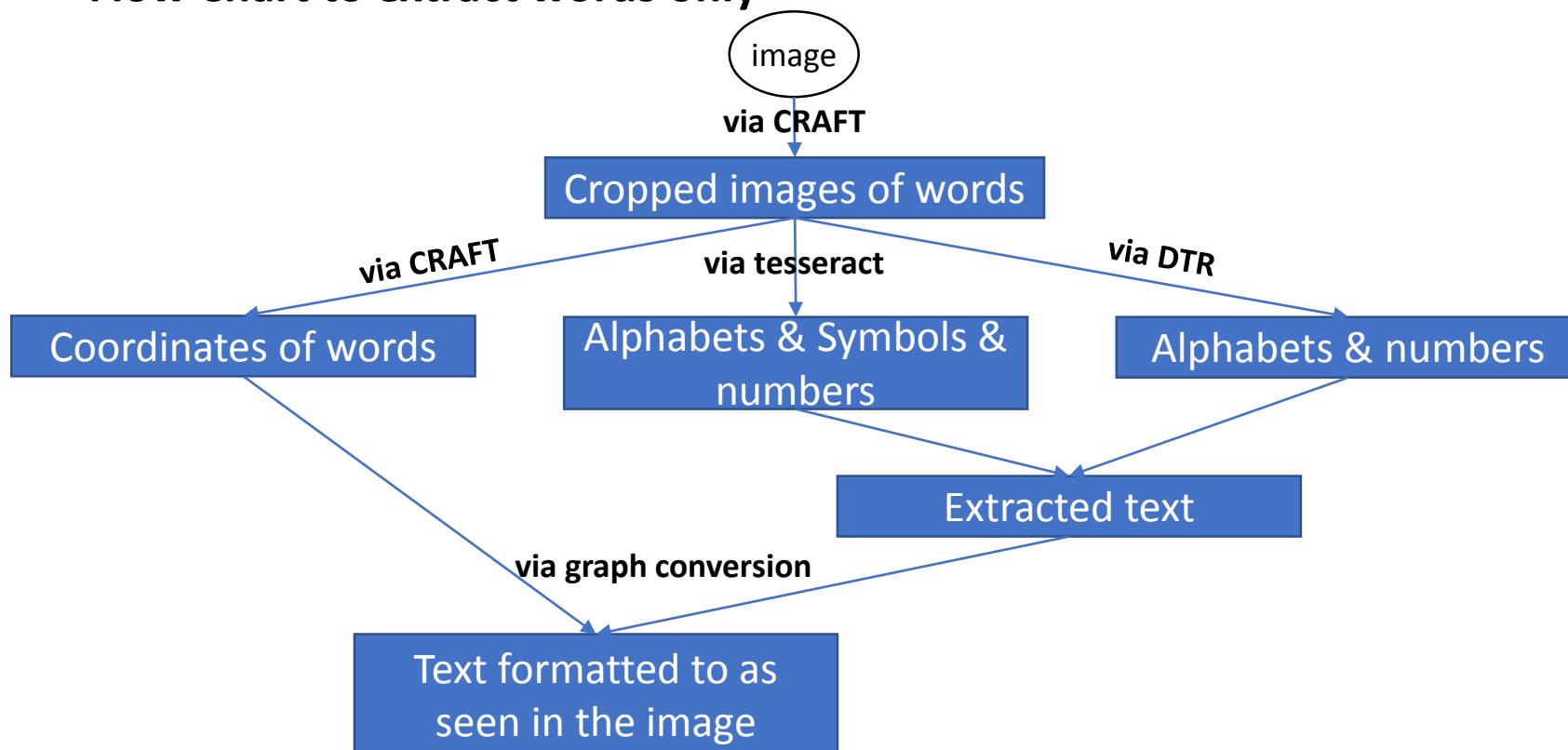
Step 1: Optical Character Recognition

Pre-trained models that we have tried: image vs.

	DESCRIPTION	PROS	CONS
DEEP TEXT RECOGNIZER (RECOGNIZER)	Prints out the text in the image digitally. State of Art.	Recognizes alphabets and number very well	Not trained on symbols. Decimal points are ignored
TESSERACT (RECOGNIZER)	Prints out the text in the image digitally.	Easily customizable. Recognizes symbols and spaces well.	Confuses characters more frequently than DTR
MASK TEXT SPOTTER (DETECTOR & RECOGNIZER)	Finds where the text is in the image. & Prints out the text in the image digitally	None.	Cannot detect text in the image when the text is small or when too many texts in an image

Step 1: Optical Character Recognition

Flow Chart to extract words only



Step 1: Optical Character Recognition

Example walk through

image

Cropped images of words



res_
FishermansFriend-
Lemon_7.jpg



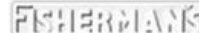
res_
FishermansFriend-
Lemon_8.jpg



res_
FishermansFriend-
Lemon_9.jpg



res_
FishermansFriend-
Lemon_11.jpg



res_
FishermansFriend-
Lemon_12.jpg



res_
FishermansFriend-
Lemon_13.jpg



res_
FishermansFriend-
Lemon_14.jpg



res_
FishermansFriend-
Lemon_15.jpg



res_
FishermansFriend-
Lemon_16.jpg



res_
FishermansFriend-
Lemon_17.jpg



res_
FishermansFriend-
Lemon_18.jpg

Propose a method

Extracted text and the Coordinates



the, 309, 340, 259, 355
 was, 333, 364, 148, 269
 and, 372, 406, 285, 439
 i, 286, 345, 908, 931
 the, 420, 453, 137, 287
 i, 344, 409, 880, 902
 i, 348, 424, 907, 929
 -?ate, 445, 476, 212, 314
 5, 412, 444, 880, 900
 i, 425, 478, 907, 928
 a, 447, 532, 208, 165
 lofthouses, 536, 569, 403, 630
 fishermen, 593, 665, 216, 598
 friend, 593, 664, 606, 819
 pastilles, 696, 721, 651, 756
 lozenges, 699, 724, 268, 395
 menthol, 699, 726, 411, 529
 (0.4%, 699, 726, 532, 598
 wlw, 700, 724, 595, 636
 02230539, 732, 750, 275, 344
 lozenges, 732, 757, 728, 823
 npn, 734, 748, 241, 275
 22, 737, 771, 681, 720
 nasal, 747, 763, 372, 412
 colds., 747, 764, 630, 672
 for relief of, 750, 766, 198, 280
 throats., 750, 764, 316, 368
 congestion, 750, 769, 414, 491
 coughs, 750, 769, 528, 580
 due, 750, 766, 582, 609
 to, 750, 764, 611, 627
 sore, 752, 764, 281, 313
 and, 752, 768, 497, 526
 pastilles, 755, 780, 728, 809
 pour, 768, 780, 200, 232
 le, 766, 780, 289, 304
 de gorge., 766, 784, 331, 390
 la, 768, 782, 395, 408
 soulager, 769, 782, 233, 289
 ma, 768, 780, 304, 326
 rhume, 768, 784, 633, 676
 congestion, 771, 784, 409, 478
 toux due au, 769, 784, 556, 633
 nasale, 772, 784, 480, 523

Step 1: Optical Character Recognition

Example walk through

image

Extracted text



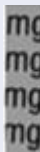


EACH TABLET CONTAINS: VITAMINS		
Vitamin A (vitamin A acetate)	300 mcg/1000 IU	IU
Beta-Carotene	600 mcg/1000 IU	IU
Vitamin E (dl- α tocopheryl acetate)	12.6 mg/28 IU	IU
Vitamin C (ascorbic acid)	150 mg	mg
Folate (folic acid)	0.4 mg	mg
Vitamin B ₁ (thiamine mononitrate)	3.85 mg	mg
Vitamin B ₂ (riboflavin)	3.85 mg	mg
Niacinamide	14 mg	mg
Vitamin B ₆ (pyridoxine hydrochloride)	5 mg	mg
Vitamin B ₁₂ (cyanocobalamin)	21.6 mcg	mcg
Vitamin D (cholecalciferol)	20 mcg/800 IU	IU
Biotin	45 mcg	mcg
Pantothenic Acid (calcium d-pantothenate)	11 mg	mg
Vitamin K ₁	20 mcg	mcg
MINERALS		
Calcium (calcium carbonate and phosphate)	400 mg	mg
Iodine (potassium iodide)		

each tablet contains: vitamins, 245, 285, 201, 704
 vitamin a (vitamin a acetate), 301, 338, 201, 575
 300 mcg/1000, 317, 358, 1038, 1239
 ly, 317, 430, 1251, 1288
 beta-carotene, 338, 370, 205, 398
 600 mcg/1000, 354, 394, 1034, 1239
 vitamin e (dl- α tocopheryl acetate), 374, 414, 201, 652
 12.6, 390, 426, 1078, 1147
 mg/28, 395, 438, 1140, 1242
 vitamin c, 410, 442, 201, 330
 ascorbic acid), 410, 446, 342, 531
 150, 426, 466, 1179, 1235
 all, 429, 659, 1251, 1295
 folate (folic acid), 446, 483, 201, 426
 0, 474, 499, 1191, 1211
 vitamin b1 (thiamine mononitrate), 487, 523, 201, 652
 3.85, 507, 539, 1171, 1231
 vitamin b2 (riboflavin), 519, 555, 201, 491
 3.85, 543, 575, 1171, 1231
 niacinamide, 555, 586, 209, 370
 14, 579, 652, 1195, 1231
 vitamin b6 (pyridoxine hydrochloride), 591, 631, 201, 6
 vitamin b12 (cyanocobalamin), 631, 668, 205, 595



Step 1: Optical Character Recognition

Where it fails

	WHAT HAPPENS	IMAGE
SIDE TEXT (NOT HORIZONTAL)	Can't do	 
SIMILAR TEXT VERTICALLY	Detector groups them together	
DTR VS. TESSARACT	Must follow heuristics when joining two algorithms	  416 vs. a% de vs. (+)



Initial Results with Ingredients Table Extraction



• **INDICATIONS** : aide au maintien d'une bonne santé. Aide au développement et au maintien des dents et des os.

• **INGRÉDIENT MÉDICINAL (par goutte)** : Vitamine D3 (Cholécalciférol) — 10 mcq (400 UI)



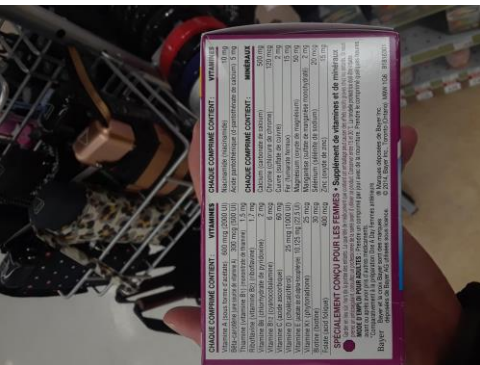
INGRÉDIENT MÉDICAMENTEUX (par comprimé à croquer) : *Bifidobacterium longum subsp. longum 35624™* 1 x 10⁹ bactérie souche/capsule

Align a une teneur en probiotiques efficace pendant toute la durée de conservation du produit.



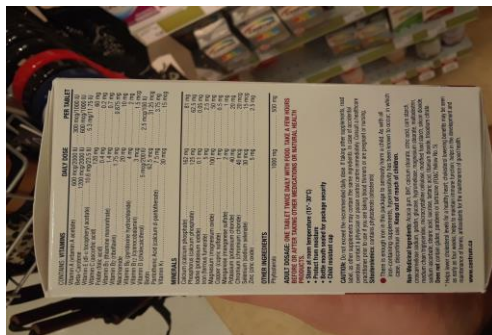
DIRECTIONS: ADULTS (13 YEARS OF AGE AND OLDER): Pour 1 stick pack (3.2 g) into at least 125 mL of any beverage or soft food. Shake or stir well until dissolved. Take 2 to 3 times daily. Take a few hours before or after taking other medicines.

MEDICINAL INGREDIENTS:



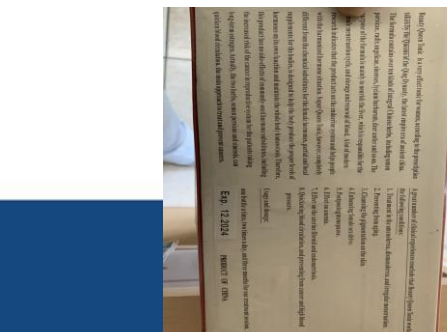
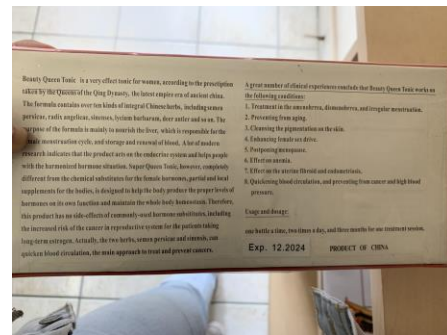
EACH TABLET CONTAINS:	VITAMINS	EACH TABLET CONTAINS:	VITAMINS
Vitamin A (as acetate)	600 mcg (2,000 IU)	Niacinamide (nicotinamide)	10 mg
Beta-Carotene (a source of Vitamin A)	300 mcg (500 IU)	Pantothenic Acid (calcium d-pantothenate)	5 mg
Thiamine (Vitamin B1) (thiamine mononitrate)	1.5 mg		
Riboflavin (Vitamin B2) (riboflavin)	1.7 mg		
Vitamin B6 (pyridoxine hydrochloride)	2 mg		
Vitamin B12 (cyanocobalamin)	6 mcg		
Vitamin C (ascorbic acid)	60 mg		
Vitamin D (cholecalciferol)	25 mcg (1,000 IU)		
Vitamin E (dl-alpha tocopheryl acetate)	10,125 mg (22.5 IU)		
Vitamin K1 (phytonadione)	25 mcg		
Biotin (biotin)	30 mcg		
Folate (folic acid)	400 mcg		

SPECIAL FORMULA FOR WOMEN • Vitamin and Mineral Supplement



CONTAINS: VITAMINS	DAILY DOSE	PER TABLET
Vitamin A (vitamin A acetate)	600 mcg/2000 IU	300 mcg/1000 IU
Beta-Carotene	1200 mcg/2000 IU	600 mcg/1000 IU
Vitamin E (dl- α tocopheryl acetate)	10.6 mg/23.5 IU	5.3 mg/11.75 IU
Vitamin C (ascorbic acid)	120 mg	60 mg
Folate (folic acid)	0.4 mg	0.2 mg
Vitamin B1 (thiamine mononitrate)	1.4 mg	0.7 mg
Vitamin B2 (riboflavin)	1.75 mg	0.875 mg
Niacinamide	20 mg	10 mg
Vitamin B6 (pyridoxine hydrochloride)	4 mg	2 mg
Vitamin B12 (cyanocobalamin)	3 mcg	1.5 mcg
Biotin	5 mcg/200 IU	2.5 mcg/100 IU
Pantothenic Acid (calcium d-pantothenate)	62.5 mcg	31.25 mcg
Vitamin K1	7.5 mg	3.75 mg
	30 mcg	15 mcg

MINERALS



[illegible]

BELL CUSTOMERS TELL THEIR STORIES ONLINE!
"The best help to kick the nicotine habit! I have been smoking for over forty years. For three years, I have been taking lozenges costing sixty dollars per month without losing my craving for cigarettes. I've been using Bell Lifestyle's Stop Smoking Help for two weeks and for the first time, my urges are almost gone completely." Freida Cruz, 75, Rogers, AR

[illegible][illegible]

Next Steps

1. Continue with more pre-trained OCR models,
2. Add preprocessing to the OCR to check the improvement
3. Get more photos – how many more for the model to improve
4. Bounding box for in the annotation tool
5. Start designing an annotation tool to evaluate OCR performance and capture labels for:
 1. Npn number,
 2. claims,
 3. ingredients,
 4. their amounts,
 5. recommended dosage,



Thank you!

References (github links)



- <https://github.com/hwalsuklee/awesome-deep-text-detection-recognition>
- <https://github.com/MhLiao/MaskTextSpotter>
- <https://github.com/ayumiymk/aster.pytorch>
- <https://github.com/clovaai/deep-text-recognition-benchmark>
- <https://github.com/clovaai/CRAFT-pytorch>
- <https://github.com/clovaai/deep-text-recognition-benchmark>
- <https://github.com/openfoodfacts/off-nutrition-table-extractor>
- <https://github.com/tesseract-ocr/tesseract>
- https://github.com/qjadud1994/Text_Detector
- https://github.com/matterport/Mask_RCNN