



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



Delivering insight through data for a better 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
<b>4. Claim compliance assessment:</b> Supervised (compliant/non-compliant labels available) and feasibility of unsupervised (time permits)	February 2020
<b>5. Integration and deployment:</b> Prototype user interface (web application) Set up earlier?	March 2020





# Variability in Products



#### Variability in NPN







3

# Variability in Claims DIN 00782696 Non-Drowsy Claritin Loratadine Tablets USP, 10 mg Helps lower cholesterol in as little as 4 weeks\*. Supports heart health\*. Fast Relief from indoor and outdoor! allergy symptoms:







For

HEART

HEALTH

90 Tablets







#### Variability in Ingredients







Milk Thistle Seed (Silyburn marianum) - 50% Silymarin*	200mg
DL-Alpha-Lipoic Acid	200mg
Garlic Bulb (Allium sativum) 4:1 Extract (QCE 500mg)	125mg
N-Acetyl-L-Cysteine	100mg
Vitamin C (Ascorbic acid)	50mg
Artichoke Leaf (Cynara scolymus) 4:1 Extract (QCE 100mg)	25mg
Dandelion Root (Taraxacum officinale) 4:1 Extract (QCE 100mg)	25m
Turmeric Root, Rhizome (Curcuma longa) 4:1 Extract (QCE 100r	ng) 25mg
Selenium (selenomethionine)	100mc
*50% Silymarin via HPLC = 80% Silymarin via UV	
Animal Ingredients: N-Acetyl-L-Cysteine is from animal	
Renew Life does not use genetically modified ingre	dients. A
ingredients are NON-GMO / GMO-FREE.	
Directions: Take 1 capsule three times per day. Take v	vith meals
providing protein. Use for a minimum of 3 weeks to see	beneficia
effects. Consult a health care practitioner for use beyond	6 months.
Warning: Discontinue use if you develop symptoms of liv	or trouble
dark urine laundice) Consult a he	alth care
of excess stomach acid have a history of non-melar	noma ekin
you have liver or gall bladder disorder and	or howal
obstruction. Do not use if you are allergic to plant	

Recommended use: Source of Probiotics. Helps support intestinal/gastrointestinal health. Could promote a favourable gut flora. Directions: Adults take 1 capsule per day. If you are on antibiotics, take at least 2 to 3 hours before or after.	Usage recommandé: Source de probiotique Aide au soutien de la santé intestinale/gast intestinale. Pourrait promouvoir une flore intestinale saine. Mode d'emploi: Adultes: Prendre 1 capsule par jour. Si vous prenez des antibiotiques.
Each capsule contains Medicinal ingredients: Britiobacterium animals 13 billion CFU* subsp. lactis (BI-04*) Lactobacterium acidophilus 12 billion CFU* (La-14)	prenez au moins 2 ou 3 heures avant ou aprèc Chaque capsule contient Ingrédients médicinaux : Bilfidobacterium animalis 13 milliards UFC' subsp. liactis (BI-04*) Lactobacillus accidophilus 12 milliards UFC'
Lactobacillus acidophilus 1 billion CFU* (NCFM®) Lactobacillus rhamnosus 1 billion CFU* (Ur-32)	(La-14) Lactobacillus acidophilus1 milliard UFC* (NCFM*) Lactobacillus rhamnosus1 milliard UFC*
Lactobacillus casei	(Lr-32) Lactobacillus caser 1 milliard UFC* (Lc-11) Lactobacillus plantarum 1 milliard UFC* (Lp-115) Biflobacterium biflidum 1 milliard UFC*
*Colony Forming Units	(Bb-06)
Non-medicinal ingredients: Microcrystalline cellulose, magnesium stearate, silica, hydroxypropyl methyl- cellulose.	*Unités formatrices de colonies Ingrédients non médicinaux : Celulose microsistelline, stéarate de magnésium, silice, hydroxypropyliméthylosifulose.
Warning: If you have fever, womiting, bloody diamnes or severe abdominal pain, consult a health care practitioner prior to use. If symptoms of digestive upon 16,6,6 diamness (occur, worsen or preset beyond 3 days, discontinue use and consult a health care practitioner. Centra-indications: If you have an instrume-compromised condition (e.g.,	Avertissement: Si vous souffrez de fièvre, de vonissements, de danrière sangiante ou de deuteurs adominales graves, consulter un praticien de soins de santé avant d'en faire l'usage. Si des symptômes de bruobles dipestifs (p. ex., claimfele) apparaissent, saggravent ou persistent au-deid de 3 jours, cesser l'usage et consulter un praticien de soins de santé.
Aucs, lymphoma, patients undergoing long-term certicosteroid treatment, do not use this product. Keep out of reach of children. Safety sessied were.	Contre-indications : Si yous souffrez d'une maladie auto-immune (p. ex., SIDA, hymphomo, patients suivant une conticothérapie à long terme), ne pas utiliser ce produit.
use if seal is broken or missing. Store in a cool, dry place.  B-04" is a trademark of Deport. In the USA, NCFA® is a trademark of the Neth Carolina Agricultural Foundation, for.	Tonir hors de portée des enfants. Dispositif invisitable avec scesa intérieur imprimé. Ne pos utiliser al le dispositif invisitable est tried ou manquant. Hanger dans un ensired hais et acc.
Sections Appendixes Foundation, toc.  NCPM and B-Out are used under scenese.  O2018 Imported and Distributed by / Importe of distribute par: Natura's Way of Careda Ltd., Durinouth, NS Careala  BIS 0AS multrossyntamistic Careala  BIS 0AS multrossyntamistic Careala  1-800-600-3618	IN-OF* est une marque de commerce de Dupord. Acc E-U., NCTA®* est une marque de commerce de hoste Carolina Apricoltural Foundation, inc. NCFM® et BI-OF® acrd utilisés accus automation.
The state of the s	WE VEGETARIAN - YEGSTARIEM

	Active ingredients / Ingrédients ac Sodium chloride / Chlorure de sodium Potassium citrate / Citrate de potass Citric acid (anhydrous / Acide citriqu Glucose (anhydrous / anhydre) In 4.9 g (average mass provider) made up to Dens 4.9 g (pouder de masse moyenne) pos	m ilum (monohydrate) ile (anhydre) 200 mL	2.6 mg/mL 2.2 mg/mL 4.4 mg/mL 14.5 mg/mL
	Non-medicinal ingredients / Ingréd Sucralose, natural flavour, / Sucralos	dients non médicinai	ıx:
	Hydralyte is free from / Hydralyte e gluten, lactose. Hydralyte is low in car socique, gluten, lactose. Hydralyte es Edectrolytes per litre / Électrolytes pp. 20 mEq., chloride / chlorure 45 mEq. ci	st exempt de : Sodiur bohydrates (glucose), t également faible en g ar litre : Sodium 45 mb	/ Saccharine plucides (glucose). Eq. potassium
	Recommended dosage / Posolog Vomiting and diarrhea / Vomissen	de recommandée	on long to 202 mg.
	Age / Âge	Take in first 6 hours / Prendre dans les 6 premières heure	Maximum per day / Maximum s par jour
	Under 12 months / Moins de 12 mois	400-600 mL	1300 mL
	1-3 years / 1 an à 3 ans	700-1000 mL	1900 mL
	3-6 years / 3 à 6 ans	1000-1400 mL	2600 mL
8	6-12 years / 6 à 12 ans	1800-2400 mL	3900 mL
	12 years-Adult / 12 ans et plus	3000-3600 mL	6600 mL
	Dehydration due to other causes / Drink Hydratyte when thirsty or exhibiting lorsque vous ressentez la soif ou démon	a the signs of dehydratic	on, / Buyez Hydraly
	10-13 years / 10 à 13 ans	Take 1-4 sachets per Prendre 1 à 4 sachet	
	13 years-Adult / 13 ans à adulte	Take 1-7 sachets per Prendre 1 à 7 sachet	
	For up to 3 days. For prolonged deh Jusqu'à 3 jours. Pour une déshydrate	ydration see your doc ation prolongée, consu	tor. Itez votre médeci
Ш	Store below 30 °C. Rangez à moins de 30 °C		HLSALM510C/S



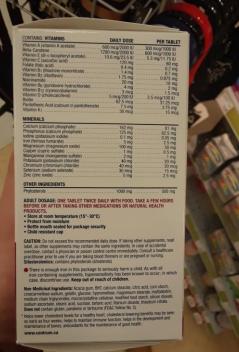


## Variability in Ingredients











#### **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

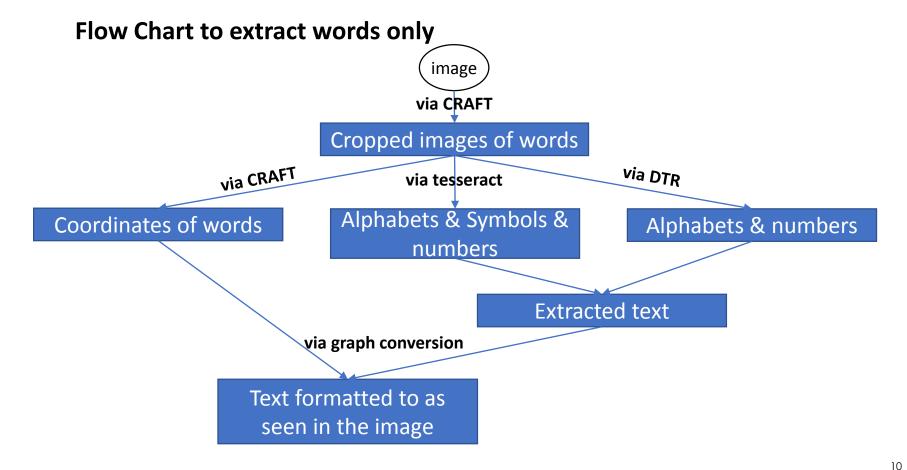
#### CRAFT-PYTORCH VS. MASK TEXT SPOTTER

0000 000000 00000000 0000000	
Occasio e estocato e occasione Occas- Occasiono	000 montooo @
GEORGE G EDRO BOCCOCCUS COCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOC	000 G000000 G 000 G0000 G
Otomos de attomados atomadostas Otomos de attomadostas	600
Microscopico Otopias de pocificaco diposcopicados Otopias de pocificación diposcopicados	90 00 00 00 00 00
GROOMS OF RECEIVED BUSINESS OF CONTROL OF CO	00 000 000 G
GOCCOCCC	8 8
CONCRETE SECURIOR CONTROLOGY AND ADDROCOURS	
Elos Eciscos Roiscocción	600 G
#00000000 #00000 000000 #0000000 #00000 000000	8
SCHOOLOG MOCKOOM CONCOUNTS	ි ල න න
00000 40000000000	. 0 00
BODOGOCO BODOGO BODOGO COCCUDA	8000 mag
00000 000000 000 000000 0000 0000 0000 0000	000 00000 000000 00
00000000 00 000 0000000 000 0000000000	1000 00000 congressors, cond 20 to cone el codicional 10000 000000 o codicion
Scott in cocces was at their pockage to company by Scott and at special at postage.	
900-80000000 10000000000000000000 000 000 00	A GOSCON (MISCOS) ESINON CECCOSTRIAL MICORONI COCAL A SCEPCOSTRICO, STRICON XCOS, KIRA, ENGICE DORD.
Code extensive entrope to controller at the entrope code to the entrope entrope to extensive entrope e	
Gree companie to decimal strate decimals and a memor	a promis core
9000000 to a expensed bodogges of Speaks.	00000

EACH TABLET CONTAINS: VILLEP 1	
Vitamin A (vitamin A acetate)	*** w #000 W
Carotene	600 mcg/1000 IU
J. J. C (ascorbic acid)	
Tang (te a suit)	0.4 mg
Security By (thiamine and policy)	2.3 mg
In the party of	
Vitamin B <sub>6</sub> (pyridoxine hydrochloride)	14 mg
Vitamin B12 (cyanocobalamin)	21.6 mg
Vitamin D (Sin leased)	20 mcg/800 /U
Purchas If Acid (calcium d-pantothopate)	
Vitamin K <sub>1</sub>	11 mg
ara ma (	20 mcg
Calcium (calcium carbonate and p.uspim)	
to.lin.: (potassium iodide)	400 mg
(ferrous furnarate)	7.5 mg
Magnesium (magnesium oxide)	64 mg
Copper (cupric sulfate)	0.9 mg
Chromium (chromium chloride)	5 mg
Selenium (sodium selenate)	25 mcg
	8 mg
OTHER MAN DESCRIPTION	
ADULT DOSAGE: ONE TABLET OF A WITH POR	
• 315r: at room temperature (15°-30°C)	
Child resistant use	
Commence of the Commence of th	
CALLIN: Do not exceed to recommended daily time, as other supplements may contain to sam overdose, contact a physician or as an arrival of a _ all_ner or as to as if you arel blue th	entre www.l. in case of accidental entre www.l. Consult a healthcare
**Can out of result of military and	harm a child.
Non-Lodie Lof Copposition La Acadia, RAT, citric un Blue No. 2 Aluminum Blue, FDSC Red No. 40 Alum	C com starch connecident Their
hypromellose, lactose monohydrate, magnesium	Minum Lake, gelatin, glucose syrup, Minum Lake, gelatin, glucose syrup, maltodextrin, medium chain
1.3.1.1 dignore trisodium contrate, sort	HC M. A.J., DML PRODUCTOR, THE LANGE MENTS
Last not contain gluten, parabens or tartrazine (* Last maintain immune flast u.); helps in 3 mai metabolize proteins, fals and carbohydrates; help	
metabolize proteins, fals and carbohydrates; help: "When compared to Centrum Forte® Essentials	s to maintain the 19 stat.
W. L. C. S.	
S FloraGLO is a registered trademark of Kemin	MONOGEO So

## 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



## **Example walk through**





## Cropped images of words



FishermansFriend-Lemon\_7.jpq



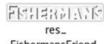
Lemon\_11.jpg

res\_

res\_ FishermansFriend-Lemon\_15.jpg



FishermansFriend-Lemon\_8.jpq



FishermansFriend-Lemon\_12.jpg



FishermansFriend-Lemon\_16.jpq



FishermansFriend-Lemon\_9.jpq



FishermansFriend-Lemon\_13.jpg



res\_ FishermansFriend-Lemon\_17.jpq



res\_ FishermansFriend-Lemon\_14.jpg



FishermansFriend-Lemon\_18.jpq







# 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 fishermans, 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





#### **Example walk through**



#### RI ET CONTAINS, VITABILIO

itamin A (vitamin A acetate)	
eta-Carotene 300 mcg/1000	IU
eta-Carotene 300 mcg/1000 itamin F (dlag tocophend accetate) 600 mcg/1000	IU
itamin Bs (pyridnyine bydrochloride) 3.85 f	mg
fitamin B <sub>12</sub> (cyanocobalamin)	mg
(Itamin D (cholecalciferol) 21.6 n	ne
(itamin D (cholecalciferol)	U
Pantothenic Acid (calcium d contatt 45 m	nco
Siotin 20 mcg/800 II Pantothenic Acid (calcium d-pantothenate) 45 m //tamin K1 11 m	ng
20 m	nca
MINERALS	4.3

#### Extracted text

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-arotene, 338, 370, 205, 398 600 mcg/1000,354,394,1034,1239 vitamin e (dl-a 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 bl (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

## Where it fails

	WHAT HAPPENS	IMAGE	
SIDE TEXT (NOT HORIZONTAL)	Can't do	2.	<u></u>
SIMILAR TEXT VERTICALLY	Detector groups them together	mg mg mg mg	
DTR vs. Tessaract	Must follow heuristics when joining two algorithms	416 vs. a% de v	<b>e</b> 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 mcg (400 UI)

INGRÉDIENT MÉDICAMENTEUX (par comprimé à croquer): Bifidobacterium longum subsp. longum 35624™ ...... 1 x 10<sup>7</sup> 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 INCREDIENTS.





Bandly Otton Halls. It is your different and for worse, considerate the projection of the STATE STATE OF CONTROLLED OF The State STATE AND CONTROLLED OF THE STATE OF THE STAT	Agrica anthri of direct reportmen match the from from the terror the filtering rations:  The filtering rations:  The restrict the restr
--	--

EACH TABLET CONTAINS:	VITAMINS	EACH TABLET CONTAINS:	VITAMINS
Vitamin A (as acetate) 600 mi	cg (2,000 IU)	Niacinamide (niacinamide)	10 mg
Beta-Carotene (a source of Vitamin A) 300	mcg (500 IU)	Pantothenic Acid (calcium d-pantothenate)	5 mg
Thiamine (Vitamin B1) (thiamine mononitral Riboflavin (Vitamin B2) (riboflavin)	te) 1.5 mg	EACH TABLET CONTAINS:	MINERALS
Vitamin B6 (pyridoxine hydrochloride)	2 mg	Calcium (calcium carbonate)	500 mg
Vitamin B12 (cyanocobalamin)	6 mcg	Chromium (chromic chloride)	120 mcc
Vitamin C (ascorbic acid)	60 mg	Copper (cupric sulfate)	2 mg
Vitamin D (cholecalciferol) 25 m	cg (1000 IU)	Iron (ferrous fumarate)	15 mg
Vitamin E (dl-alpha tocopheryl acetate) 10.125	mg (22.5 IU)	Magnesium (magnesium oxide)	50 mg
Vitamin K1 (phytonadione)	25 mcg	Manganese (manganese sulfate monohydra	te) 2 mg
Biotin (biotin)	30 mcg	Selenium (sodium selenite)	20 mg
Folate (folic acid)	400 mcg	Zinc (zinc oxide)	15 m

CONTAINS: VITAMINS	DAILY DOSE	PER TABLET
Vitamin A (vitamin A acetate)	600 mcg/2000 III	200
Vitamin C (ascorbic acid)	10.6 mg/23.5 lU	5.3 mg/11.75 IU
Vitamin C (ascorbic acid)	120 mg	60 mg
Folate (folic acid) Vitamin B <sub>1</sub> (thiamine mononitrate) Vitamin B <sub>2</sub> (riboflavin)	0.4 mg	····· 0.2 mg
Vitamin B) (ullamine mononitrate)	1.4 mg	0.7 mg
Vitamin B <sub>2</sub> (riboflavin)	1.75 mg	0.875 mg
Niacinamide Vitamin Be (pyridovine hydrochlorido)	20 mg	10 mg
Biotin	3 mcg/200 10	2.5 mcg/100 IU
Pantothenic Acid (calcium d-pantothenate)	62.5 mcg	31.25 mcg
Vitamin V.		3.75 mg
Vitamin K <sub>1</sub>	30 mcg	15 mcg
2000000000		

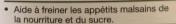












- Un soutien naturel pour les glandes surrénales.
- Une grande combinaison d'herbes et de minéraux qui aident à nourrir et à renforcer des organes importants dont le pancréas et le foie.

#### BellLifestyle.ca 🕡 💟 😉 😉

CONSEILS D'UTILISATION: Adultes: Prendre 1 capsule, deux fois par jour, INGRÉDIENTS MÉDICINAUX: Lobelia (lobelia inflata) (fleuille) 130 mg, cassia extrait de cannelle 10:1 (cinnamomum aromaticum) (ecorce) 100 mg, extraît de tussilage 10:1 (flussilago de farfara) (fleuille) 100 mg, extraît kudzu (pueraria montana, var lobata) (racine) (40% d'isoflavones) 100 mg, extraît de passifiora incarnata) (fleur) 100 mg, extraît de café (coffea arabica) haricots) (50% d'acides chiorogénique) 50 mg, extraît d'avoine 10:1 (avena sativa) (grain) 50 mg, INGRÉDIENTS NON MÉDICINAUX: Capsules vinétaziennes. Mutraroynorum athylusalibuses auti

#### 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 agne completely. "Freila Cityz. 75. Ropers, AR.





Statistics

17

## **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!

19

## 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



20

