ML Practicum Project Update

Team Members: Zhengqi Dong, Cole Hunter, Zihua Xin, Mark Yang

Resources Used

- We have used the previous teams scrapping setup to download around 1000 images for testing purposes
- Using CRAFT for the text segmentation
- Tr-OCR model for text recognition
- Using string-grouper for species matching

- Using PP-OCR(Ultra-lightweight OCR System) for the whole pipeline: latest PP-OCRv3 model is used for inference
- 6 cutting-edge algorithms for text detection: DB(used), EAST, SAST, PSENet …
- 15 cutting-edge algorithms for text recognition: SVTR(used),CRNN, SRN, NRTR …
- And more available algorithms, click here for more details:

Text Detection
- EAST
- DB
- SAST
- PSENet
- FCENet
- ......

Text Recognition
- CRNN
- SRN
- NRTR
- SVTR
- ABINet
- ......

End-to-end
- PGNet

Progress

- Using CRAFT for image segmentation
  - Running on GPU for faster evaluation
- Implemented Tr-OCR for text recognition
  - Processing ~19 image segmentations per second
- Implemented a new string matching setup
- Currently have implemented a full pipeline using the above tools

- Implemented the PP-OCR pipeline based on [PP-OCRv3](#) model, with DB for text detection and SVTR or text recognition.

Detection and Recognition result: [bbox: 2D list, (predicted_txt: str, confidence: float)

```
[[[696.0, 26.0], [844.0, 26.0], [844.0, 84.0], [696.0, 84.0]], ('Field', 0.9904948472976685)]
[[[680.0, 56.0], [710.0, 56.0], [710.0, 77.0], [680.0, 77.0]], ('The', 0.925682544708252)]
[[[635.0, 68.0], [647.0, 68.0], [647.0, 80.0], [635.0, 80.0]], ('(', 0.6750483512878418)]
[[[692.0, 75.0], [816.0, 80.0], [815.0, 106.0], [691.0, 100.0]], ('Museum', 0.997275829315185
5)]
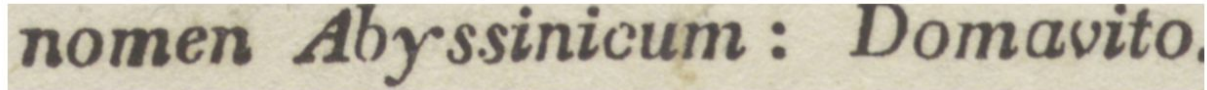```

Original image vs predicted result:



Ground truth label (top), original image (bottom)
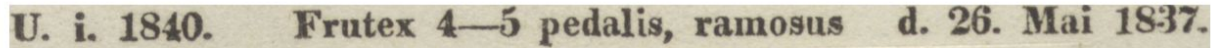
Predicted result
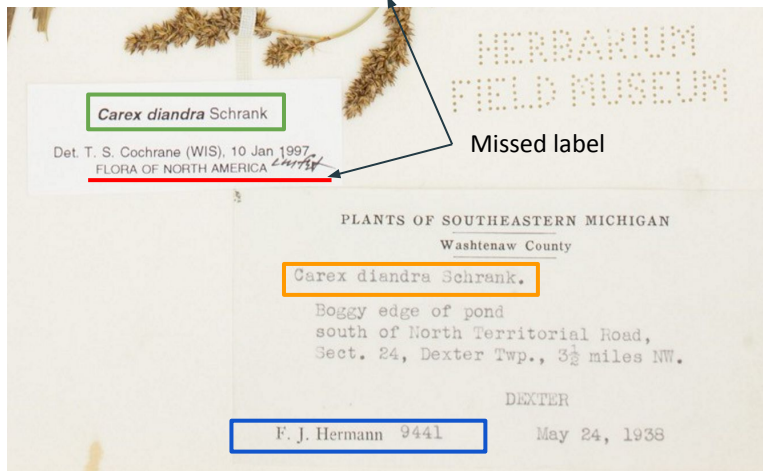
# CRAFT for Image Segmentation

After doing our literature review, we found that CRAFT remains the best option for extracting the regions in images which contain text.
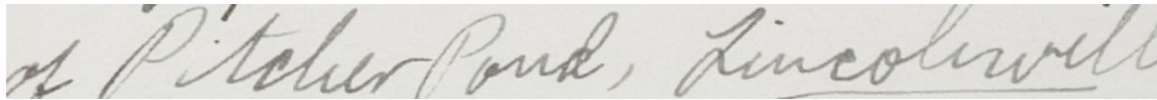
# Tr-OCR for Text Recognition

nomen Abyssinicum : Domavito.

Text: nomen Abysinicum : Domavito

U. i. 1840.    Frutex 4—5 pedalis, ramosus    d. 26. Mai 1837.

Text: U. i. 1840. Frutex 4-5 pedalis, ramosus d

a Pitcher-Pond, Lincolnwell

Text: a Pitcher-Pond, Lincolnwell

Taraxacum duplidens Lb. f.

Text: Tataxacanthus duplichens that

# String Matching

Using string-grouper to process Tr-OCR outputs

- The advantage of using this method for string matching is that it is extremely quick
  - Matching ~22000 transcriptions against the full corpus takes around 25 seconds
- Speed-to-match does not scale to lower input numbers particularly well

# String Matching

Can set your matching confidence arbitrarily high

- As expected, increasing the desired confidence tends to significantly reduce the number of images the model has strong guesses on
- Generally speaking, the higher the quality of the base image, the more accurate all aspects of the model are, including the matching
- At higher confidences, matching issues are related more to synonym differences

# String Matching

```
* Persicaria sagittata || Polygonum sagittatum || 1424774269
* Persicaria sagittata || Polygonum sagittatum || 1424774292
* Diphasiastrum sitchense || Lycopodium sitchense || 1503173489
* Silene flos-cuculi || Lychnis flos-cuculi || 1701285949
* Arabidopsis thaliana || Arabis thaliana || 1701477809
Ground Truth Not Found for: 1701985163
* Lipandra polysperma || Chenopodium polyspermum || 1702254438
Ground Truth Not Found for: 1702403767
* Micranthes nivalis || Saxifraga rivalis || 1702425560
* Solanum donianum || Solanum verbascifolium || 1928658806
* Spiraea canescens || Spiraea brachybotrys || 1930241969
* Bistorta vivipara || Polygonum viviparum || 1978791988
* Chamaenerion angustifolium || Epilobium angustifolium || 1989131227
* Phelipanche ramosa || Orobanche ramosa || 1990250633
* Platanthera flava || Habenaria habenaria || 1998968996
* Pseudolycopodiella caroliniana || Lycopodium carolinianum || 2274177724
* Gaertnera rufinervis || Psychotria psychotrioides || 2514524961
Ground Truth Not Found for: 2514610251
* Campanula persicifolia || Campanula linifolia || 2516495194
* Struthiopteris spicant || Blechnum spicant || 2572960903
* Persicaria maculosa || Polygonum persicaria || 3392122454
* Clethra alnifolia || Solidago canadensis || 3709961181
* Senna multiglandulosa || Cassia tomentosa || 437056001
Accuracy on Predicted: 0.7444444444444445
```

PP-OCRv3 Pipeline
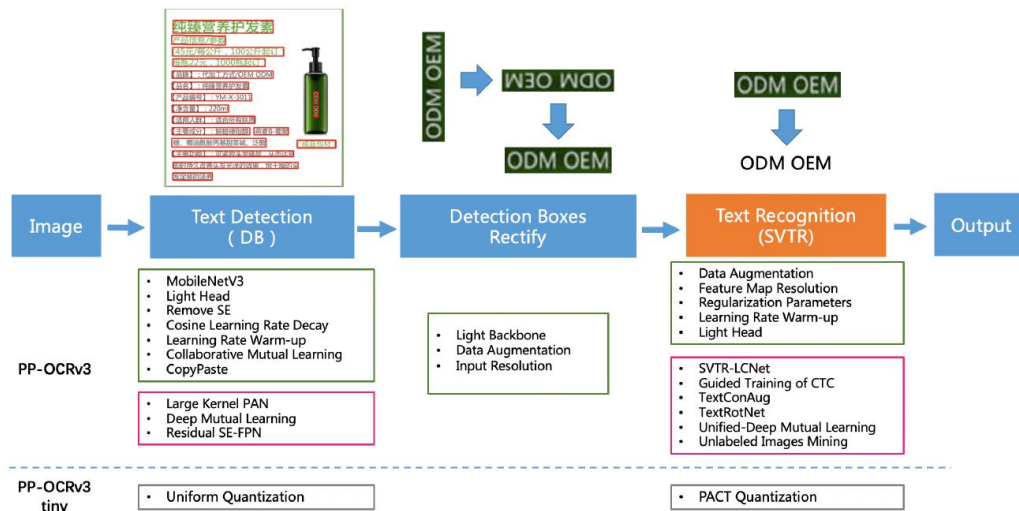
Pros:

- Lightweight model: English detection(1.1M), English recognition(3.2M), Angle classification(1.38M)
- Support for multilingual (80+ languages) detection and recognition
- Has [many cutting-edge algorithms](#) to compare with (e.g., DB/DB++, SAST, SVTR, SPIN…)

Cons

- Doesn't have any good pretrained model on IAM handwriting dataset

Current result from PP-OCRv3

```
collector(15/17): Moore, John; Moore, Marjorie || Moore, John; Moore, Marjorie || 2265361450
taxon       Chenopodiastrum murale || Kanahia laniflora || 2513962787
geography   Portugal || Tanzania, United Republic of || 2513962787
collector   Kostermans AJGH; Kruyt W || H. Shriver || 2513962787
taxon(14/19):   Salvia rosmarinus || Salvia rosmarinus || 2514610251
geography   Spain || Sweden || 2514610251
collector   Elias H || H. Shriver || 2514610251
taxon       Rinorea guianensis || Sagina nivalis || 2515453969
geography   Brazil || Angola || 2515453969
collector   Saldanha CJ || H. Shriver || 2515453969
taxon(15/21):   Koeleria macrantha || Koeleria macrantha || 2516377361
geography   Netherlands || Russian Federation || 2516377361
collector   Bruyn L de || H. Shriver || 2516377361
taxon       Eriodictyon angustifolium || Kanahia laniflora || 2516548361
geography   United States of America || Estonia || 2516548361
collector   Epling CC; Robinson WM || H. Shriver || 2516548361
taxon       Ochna leucophloeos || Sagina nivalis || 2517498948
geography   Zimbabwe || Norway || 2517498948
collector(16/23): Lanjouw (Africa) J || Lanjouw (Africa) J || 2517498948
taxon(16/24):   Poa compressa || Poa compressa || 2900391670
geography   Sweden || United States of America || 2900391670
collector   M. Sondén || Moore, John; Moore, Marjorie || 2900391670
taxon       Taraxacum ostenfeldii || Rhododendron calendulaceum || 3016518613
Ground Truth Not Found for:3016518613
collector(17/25): Albert Üksip || Albert Üksip || 3016518613
taxon       Thuidium delicatulum || Carex diandra || 3416707636
Ground Truth Not Found for:3416707636
collector   Fulford, Margaret Hannah || H. Shriver || 3416707636
taxon(17/27):   Ardisia crenata || Ardisia crenata || 436989369
geography   Viet Nam || Angola || 436989369
collector(18/27): Poilane, E. || Poilane, E. || 436989369
taxon       Swertia marginata || Eleocharis acicularis || 437989544
geography   Russian Federation || Spain || 437989544
collector(19/28): Mission Pelliot-Vaillant || Mission Pelliot-Vaillant || 437989544
taxon(18/29):   Carex saxatilis || Carex saxatilis || 438312026
geography   Canada || Estonia || 438312026
collector(20/29): Fernald, M.L.|Wiegand, K.M. || Fernald, M.L.|Wiegand, K.M. || 438312026
taxon(19/30):   Ixora euosmia || Ixora euosmia || 574701445
geography   Cameroon || Tanzania, United Republic of || 574701445
collector(21/30): Zenker, Georg August || Zenker, Georg August || 574701445


*******************************Report Statistics
[0.6333, 0.1852, 0.7]
Total Images had been processed for evaluation: 30
Total Prediction: taxon: 30, geography: 27, collector: 30
Correct Prediction: taxon: 19, geography: 5, collector: 21
ACC: taxon: 0.6333, geography: 0.1852, collector: 0.7
Average Acc: 0.5172413793103449
```

Next Steps

- Containerize each of the pipeline processes as functions
- Finetune the Tr-OCR model on more handwritten examples
- Implement synonym matching for better recognition
- Extend the matching to geography and collector

Next Steps (Continued)

- Implement a more general evaluation function that can be used to measure the performance for PP-OCR pipeline
- Use cutting-edge pre-trained model(e.g., DB++/SAST, ABINet/VIsionLAN/SPIN) provided on PP-OCR to retrain on IAM dataset for better performance for both text detection/recognition
- Using W&B to measure and compare the performance for different models that we trained
- Experiment with Paddle Inference for deployment

Questions

An open question: if the ground truth label is not exactly the format as what is written down in the image, how should we perform the evaluation? What evaluation metric we should use? In another word, how we should define a correct prediction? a wrong prediction? missing?

- One approach: We don't evaluate on the exact string matching, but only on the similarity between two string, and the customer can configure the min_similarity that they think is acceptable for their application. Then, we can compute precision = (number of correct pred on taxon/geograph/collector) / (total number of pred on taxon/geography/collector)
- The formula:

$$Precision = \frac{Correct\ matches}{The\ number\ of\ the\ detected\ words}$$

$$Recall = \frac{Correct\ matches}{The\ number\ of\ the\ ground\ truth\ words}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}.$$