Enhancing NfN using Text Analytics and Visualization

Deb Paul, Andrea Matsunaga, Miao Chen, Jason Best, Reed Beaman, Sylvia Orli, William Ulate

iDigBio – Notes From Nature Hackathon December 2013
Increasing Citizen Science Participation in Museum Specimen Digitization
Text Clusters

- Preprocess specimen label images with OCR
- Remove (and use!) noise from text
- Utilize OCR text
  - create word cloud linked to record ids
  - differentiate hand-written from typed labels
- Allow transcribers to choose terms from word cloud to create individual sets
- Allow validators to choose sets to clean
Reasons for Cluster Methodology

- Enhance user experience
- User Happiness!
- Leverage user expertise
- Improve speed
- Reduce Errors
- Enables ditto function
Users like ordered datasets

Transcription

- **faster** with ordered/sorted sets
- **less error prone** with sorted sets
Handwriting vs. Typed

- **Segregate hand-written from typed labels**
- Ben Brumfield code uses *regex* to sort out garbage (higher garbage = higher likelihood hand-written)
  - Read all about it at [Ben’s blog!](#)
  - Code is at [GitHub](#)
  - Humanities community using now!
- Let transcriber choose label format
- *Typed?....go to word cloud workflow*
extract character-level n-gram from a corpus (the OCR corpus + an external good corpus, ideally)

obtain a list of character-level n-gram
  ◦ e.g. bi-gram looks like th, sh, ph ...

given a word, compute its probability based on the n-gram probability

this is used for computing the final OCR confidence score

can use standard dictionary in computing the score (if time allows)
Word Cloud Workflow
aOCR iDigBio – BRIT Hackathon Lichen Dataset

Lichen Label Images → 200 hand-typed → 200 hand-parsed → Comparison Analysis Scoring Parsing

Gold txt → 200 silver OCR → Gold csv → Comparison Analysis Scoring Parsing

Silver txt → 200 silver-parsed → Silver csv

Stop/not useful words removed

Word Cloud → Crowd!

SOLR Index
aOCR iDigBio – BRIT Hackathon NYBG Herbarium Dataset

Herbarium Sheet Images

5,000

100 hand-typed

100 hand-parsed

Comparison Analysis Scoring

Golden txt

Golden csv

100 silver OCR

100 silver-parsed

Comparison Analysis Scoring Parsing

Silver txt

Silver csv

Word Cloud

Crowd!
aOCR iDigBio – BRIT Hackathon CalBug Entomology Dataset
Word Cloud using Solr + Carrot²

http://ammatsun.acis.ufl.edu:5901/carrot2-webapp-3.8.1/
Circles Visualization of Search

1. The New York Botanical Garden LICHENS OF NEW YORK STATE, U.S.A. Polystichum minutulum Kocourkové & F. Berger On 15gFilg_Silicofloides Coppins & P. James Rockland County: Harriman State Park, along Woodtown Road West near dam at 3 end of Lake Sebago along Seven Lakes Drive, 41°11′N, 74°08′W, ca. 200 m; mixed hardwood—hemlock forest with granitic erratics. 19 April 1998 Tu||W|W|W|H|H|H|H|H|W|T|W|F 0107579

2. Richard C. Harris 52756 The NCVY York Bohmica R S.11\'dQ11 LICHENS <<r: Q)\10. L-5-A - P07 corcun minutulum Kocourkové & F. Berger ___X 47 lg on Trapehosis 10a0d1′o1′des Coppins & James Scoolo County: Shawnee State Forest, along Pond Lick Run N of Forest Road 1, 0.35 mi SW of SR 125, 38°42′20″N, 30°8′25″W, ca. 190 m; sandstone boulders along stream and adjacent oak woods. 21 NEW Yorx BOTANICAL GARDEN 01075760 Collected on the 15th Tuckerman Workshop

Foam Tree Visualization of Search

1. The New York Botanical Garden LICENS OF NEW YORK STATE, U.S.A. Polycoccom minutum Kocourkové & F. Berger On Bjp58lg_21aod0101des Coppins & P. James Rockland County: Harriman State Park, along Woodtown Road West near dam at 3 end of Lake Sebago along Seven Lakes Drive, 41°11′N, 74°08′W, ca. 240 m; mixed hardwood—hemlock forest with granitic erratics. 19 April 1998 T\W/\V/W/\H/7\IT/\W1075759

2. Richard C. Harris 52766 The NC long Borneal: S,1\1\1\011 LICENS <>... Q"1)101 L-5-A - P07 cordic minutum Kocourkové & F. Berger _X On Trachyonopsis 10a010I1des Coppins & James Scboto County: Shawnee State Forest, along Pond Eagle Run II of Forest Road 1, 0.35 mi SW of SR 125, 38°42.5′N, 83°08.25′W, ca. 190 m; sandstone boulders along stream and adjacent oak woods. 21 NEW Jour Botanical Garden 111111 01075760 Collected on the 15th Tuckerman Workshop

3. The New York Botanical Garden LICENS OF NEW YORK, U.S.A. Polycoccom opulentum (Th. Fr.) Arnold on Indet. crust Clinton County: Town of Peru, Valcour Island in Lake Champlain, Buff Point, 44°37′27″N, 73°25′30″W. dolomite rock along lake shore and adjacent Wyuna forest. 18 May 2007 NEW YORK BOTANICAL GARDEN 01075761 Collected on the 4th Crum Bryophytic Workshop Richard C. Harris 52766
REPORT ANY REIDÈ TO THE INSTITUTE
SA=f / h:fs
The Key York Botanical Garden INSTITUTE OF ECONOMIC BOTANY Plats of Cononiealteti of
Domi
Wu. 1. Att. ARACEAE
Anthurium -jacant» d<th /// T. 93
HooKer\ K'intU - M.4.r. T. Croq << ?3
WEST INDIES, Dominica. St.George. Above new steel bridge on road to Freshwater Lake.
15°19'N' 61°01'W. 765m. Advanced secondary montane rainforest.
Herb, terrestrial, forming a rosette; leaves 1.3m long x 0.35m wide; roots also growing
upward in debris between leaves.
NEW YORK BOTANICAL GARDEN
00840904
WmEE
ya
BOTANICAL
GARDE++,
Fie&k Stiffohorn 776 May 11, 1992
with G. A. Eidesen and H. T. Beck
Fieldwork supported by the National Cancer Institute
in T7 r....-
6 7 8 9 10 the New York copyright reserved botanical garden
00840904
Token Histogram – Google Charts

Token occurrence in Lichen dataset.

lichens count: 49
Using OCR output to enhance the transcription process

LI LI Team!
Word Clouds using N-gram Scoring, Faceting, Solr + Carrot²

- Data Sets
  - silver (ABBYY) OCR output from 200 lichen packet images
  - all ABBYY OCR output from 10,000 lichen packet images
  - silver and gold (ABBYY, Tesseract) OCR output from 100 herbarium sheets
    - OCR output from 5000 herbarium sheets
  - SI BVP dataset
Word Cloud using Solr + Carrot²

- Index OCR text output using SOLR
  
  http://ammatsun.acis.ufl.edu:5900/solr/#/lichenssilver/schema-browser?field=content

- Using Carrot² to visualize data
  
  http://ammatsun.acis.ufl.edu:5901/carrot2-webapp-3.8.1/
Imagine Integration with NfN/BVP

Use for initial sort or validation
10,000+ records sample
iDigBio Faceted Collection Browser

- discovery
- how many documents have this issue
iDigBio Faceted Collection Browser

- discovery
- how many documents have this issue
Visualizing OCR Confidence

- early triage
  - score each document
  - score transcriptions
    - low scores to human
    - high scores to automated parsing
- humans check
- human correct
  - transcription errors
  - ocr errors

Why
PLANTS OF BAHAMA ISLANDS

Training Set

Herb gold
100 docs

N-grams probability model

Test Set

Herb silver
100 docs

PLA AMA
LAN MA
ANT A I
NTS IS
TS ISL
S O SLA
OF LAN
OF AND
F B NDS
BA ...
BAH ...
AHA 3-grams
HAM ...

...
Visualizing OCR Confidence
5000 OCR html files
See it for yourself!

- herb_allsilver_trigram_html.zip
Let’s make it happen!

Use for initial sort or validation
Happy Hollidays!

Team!