Enhancing Crowdsourcing 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
Reasons for 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
User Stories

- 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
    - Humanity’s community using now!
- Let transcriber choose label format
- **Typed?....go to word cloud workflow**
aOCR iDigBio – BRIT Hackathon Lichen Dataset

10,000 Lichen Label Images

200 hand-typed

Gold txt

200 hand-parsed

Gold csv

200 silver OCR

Silver txt

200 silver-parsed

Silver csv

Comparison Analysis Scoring Parsing

Comparison Analysis Scoring

Word Cloud

Crowd!

Stop/not useful words removed
Herbarium Sheet Images

100 hand-typed

100 hand-parsed

Comparison Analysis Scoring

5,000

Gold txt

Gold 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

CalBug Entomology Images

500

100 hand-typed

100 hand-parsed

Comparison Analysis Scoring

Segmentation Sorting

100 silver OCR

100 silver-parsed

100 silver OCR

Silver txt

Silver csv

Gold txt

Gold csv

Comparis on Analysis Scoring
Overall Word Cloud Workflow

Images → OCR Engine → OCR Output → DwC Parsed Output → OCR confidence (n-gram) → Web Service (Jason Davies)

OCR Output → Index (Solr) → Histogram (Google Charts, Facet Explorer)

Web Service (Jason Davies) → Word Cloud

Cluster (carrot^2) → OCR confidence (n-gram) → Index (Solr) → OCR Output

Google Charts: http://developers.google.com/chart/interactive/docs/gallery
Facet explorer: http://github.com/idigbio-citsci-hackathon/facet-explorer
Jason Davies WC: http://www.jasondavies.com/wordcloud/
Apache Solr: http://lucene.apache.org/solr/
carrot^2: http://project.carrot2.org/
Web Service–Based Word Cloud

http://aocr1.acis.ufl.edu/datasets/lichens/silver/ocr/WebrootDatasetsLichensSilverOcr.txt

Try:

Datasets Indexed:
- 100 herb silver (OCR)
- 5,000 herb silver (OCR)
- 200 lichen silver (OCR)
- 10,498 lichen ABBY (OCR)
- 10,495 lichen ocropus (OCR)
- 10,498 lichen tesseract (OCR)
- 809 Smithsonian/BVP
  - catalogNumber, collector, country, eventDate,
  - fieldNotes, fieldNumber, scientificName,
  - stateProvince, transcriberNotes, validatorNotes,
  - verbatimElevation, verbatinLatitude,
  - verbatimLongitude, verbatimLocality

Top 1,000 lichen terms
Token Histogram – Google Charts

Token occurrence in Lichen dataset.

- < 16:
- 16+:
- 24+:
- 32+:
- 40+:
- 48+:
- 56+:

lichens count: 49
iDigBio Faceted Collection Browser

- discovery
- how many documents have this issue
Based on the probability of n-grams
Estimating OCR confidence

- 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)
Visualizing OCR Confidence

https://github.com/idigbio-citsci-hackathon/OCR-Text-Confidence-Visualization
Word Cloud using Solr + Carrot²

http://ammatsun.acis.ufl.edu:5901/carrot2-webapp-3.8.1/
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
Folder View of Search

Top 100 results of about 100 for *

   NEW YORK BOTANICAL GAR©® 00040919 The lei York Sotaical Garden INSTITUTE OF ECONO
   Couoivealtk of Doailla SAPOTACEAE Manilkara bidentata (A.L. Candolle) Chevalier - t fJAn* T.
   Carib Territory. Land of Wilbur Rabbaes, Atkinson. In field 0.9km downhill from top of ridge. 15
   Cleared forest, cultivated banana field. Tree to 20n with dbb 1.5b; bark rough, thick; mature lea
   i.: Billet food (Bilisk). Balata iCreofe-Patios]. OS: lateral for coisticricio ivood. Oseed ii housebu
   August 24, 1992 with Prosper Paris Fieldwork supported by the latioaI Cancer Institute aid Mero
   http://amatsun.acis.ufl.edu:5901/carrot2-webapp-3.8.1/herbsilvertrimgram/NY_00040919.txt

   REPORT ANY REIïÌÍ TO THE INBTÌÝÌÚ 5A/.f / b-fs The Key York Botanical Garden INSTITUTE OF
   Cononiealtk of Domica uu./. (Jlj ARACEAE Anthurium -jacant) d<th // T. 93 Hooker| Kïnt
   INDIES, Dominica. St.George. Above new steel bridge on road to Freshwater Lake. 15°19′N 61
   secondary montane rainforest. Herb, terrestrial, forming a rosette; leaves 1.3m long x 0.35m w
   upward in debris between leaves. NEW YORK BOTANICAL GARDEN 00040904 WmEE ya BOTAN
   Stiffohorn 776 May 11, 1992 with G. A. Eidesen and H. T. Beck Fieldwork supported by the latic
   r.....— 6 7 8 9 10 the New York copyright reserved botanical garden ■■ 00040904
   http://amatsun.acis.ufl.edu:5901/carrot2-webapp-3.8.1/herbsilvertrimgram/NY_00040904.txt

   The New York Botanical Garden Institute of Economic Botany Plants of Dominica RHIZOPHORA
   (Sw.) Poir. det: M. Nee, 1998 NEWYOFK BOTANICAL QARDFN 0019 20S BOTANICAL GARDE**
Circles Visualization of Search

Top 100 results of about 100 for *


   REPORT ANY REIDÉ TO THE INSTITUTE S.A. F. & E-FS The Key York Botanical Garden INSTITUTE OF ECONOMIC BOTANY Plants of Cononieaff of Domica uu <./. (in) ARACEAE Anthuriu -plants- dch ckh T. 93 Hocker, k'netu - M. 4 r. T. Croq <./. WEST INDIES, Dominica. St. George. Above newly bridge on road to Freshwater Lake. 15°19'N/ 60°19'W. 765m. Advanced secondary montane rainforest. Herb, terrestrial, forming a rossette; leaves 1.5m long x 0.35m wide; roots also occur upward in debris between leaves. NEW.
Source of our Ll Ll team name!
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 Clouds using N-gram Scoring, Faceting, Solr + Carrot²
iDigBio Faceted Collection Browser

- Discovery
- How many documents have this issue
Imagine Integration with NfN/BVP
Let’s make it happen!

Use for initial sort
or validation
Using OCR output to enhance the transcription process

LI LI LI Team!