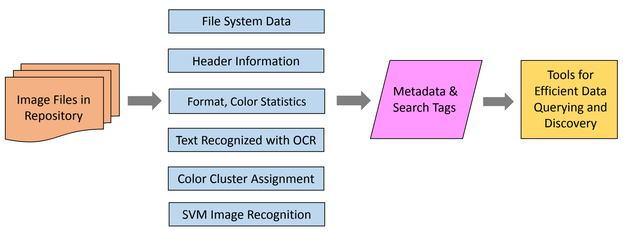
Applying Image Feature Extraction to Cluttered Scientific Repositories

Emily Herron, Kyle Chard, Tyler Skluzacek, Ian Foster
Mercer University; Computation Institute, University of Chicago and Argonne National Laboratory

Introduction

- Poor organization and clutter in large scientific repositories and file systems complicates data discovery and use
- Skluma [1] provides an automated pipeline for extracting metadata and inferring contextual relationships that can be used for organization and discovery
- We present a module for Skluma for extracting and processing feature and content-based metadata from images

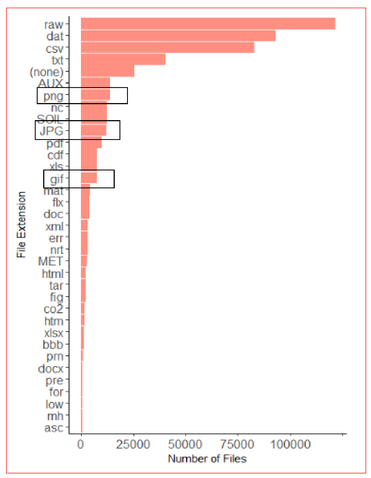
Image Metadata Extraction Pipeline



- Metadata is collected from file system, image headers, text, and content using a variety of techniques
- Result: metadata for organizing and querying images in scientific repositories

Example Data Repository

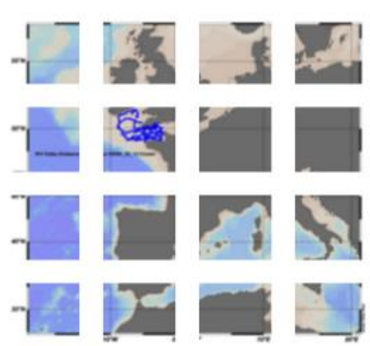
- DOE Carbon Dioxide Information and Analysis Center's climate data repository contains thousands of image files, e.g., JPEG, PNG, BMP

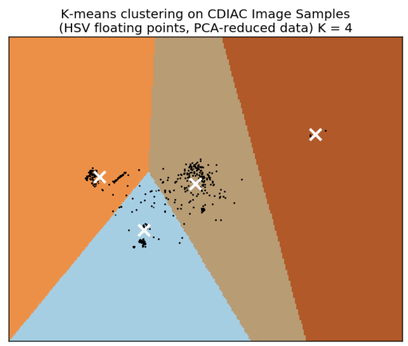


Extracting File System Header Metadata

File System Data: Crawled repository to extract file name, path, extension, and size.
Image Header Information: Used the Python Image Library (PIL) to extract image mode, resolution, encoding, creation date, etc.

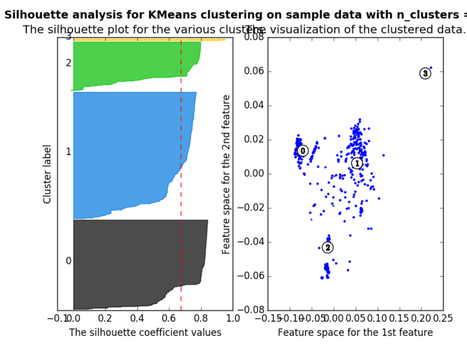
Identifying Image Clusters

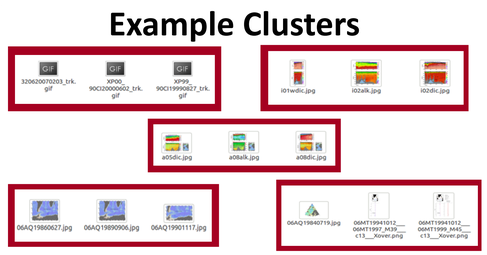




Color-Based Clusters: Clustered color feature data using K-Means
- RGB and RGBA mode images resized, divided into 4 by 4 grids
- Mean floating point RGB values calculated grid sections
- PCA-reduced feature vectors clustered K-Means

Evaluating Clusters with Silhouette Analysis

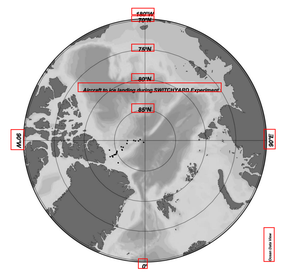




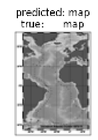
- Silhouette analysis used to select optimal number of clusters
  - Measures average distances between neighboring clusters
- Optimal k value of 4 selected based on the high mean score

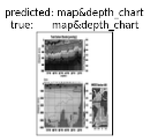
Image Text Recognition

- Text extracted from images using Python-tesseract, an optical character recognition tool and wrapper for Google's Tesseract-OCR
- Images converted to grayscale with PIL and passed to the OCR function



Classification Model





Class Label	Number Predicted in CDIAC
map	2877
map&depth_chart	22
map&plot	395
map&histogram	144
other	200
total	3638

TABLE IV

Class Label	Precision	Recall	f1 Score	Support
map	0.94	0.95	0.95	88
map&depth_chart	1.00	1.00	1.00	99
map&plot	1.00	0.50	0.67	6
map&histogram	1.00	0.43	0.60	7
other	0.00	0.00	0.00	1
avg / total	0.97	0.94	0.95	201

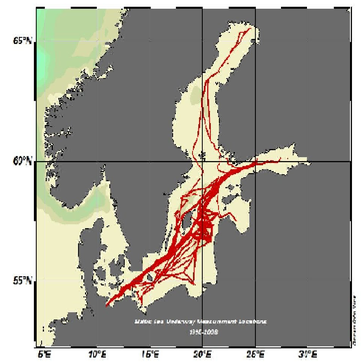
TABLE II

Class Label	Precision	Recall	f1 Score	Support
map	1.00	0.98	0.99	44
map&depth_chart	1.00	1.00	1.00	51
map&plot	1.00	1.00	1.00	2
map&histogram	0.75	1.00	0.86	3
avg / total	0.99	0.99	0.99	100

TABLE III

Support vector machine (SVM) model trained to classify images into map, figure,
- Images are resized and converted to grayscale arrays
- Dimensions reduced with PCA
- Uses Scikit-Learn c-support classification model (SVC)
- Precision, recall, F-measure, measured for 2:1 and 1:2 ratio split of test and training sets (see tables II & III).

Example Extracted Metadata



```
{
'header': {
        'header_info': {
        'jfif': 257,
        'jfif_unit': 1,
        'exif': 'Exif\x00\x00MM\x00*\...',
        'jfif_version': (1, 1),
        'jfif_density': (96, 96),
        'dpi': (96, 96)},
    'image_mode': 'RGB',
    'dimensions': '930x825'},
'color': {
        'mean_pixel_value': 144.818...,
        'extrema': ((0, 255), (0, 255), (0, 255)),
        'mode_pixel_value': 107.0,
        'median_pixel_value': 107.0,
        'std_dev_pixel_value': 69.529...},
'system': {
        'path': '/media/emily/USB DISK1/images',
        'extension': 'jpg',
        'file': 'BS_Underway_Map.jpg',
        'size': 115811},
'image_text': ['Baltic Sea Underway Measurement Locations', '1995-2008'],
'name_tags': ['BSUnderwayMap'],
'SVM_class_tags': ['map'],
'mean_colors_cluster': 1}
```

Sample JSON document containing metadata extracted from *BS_Underway_Map.jpg*. Includes header information, color statistics, file system data, extracted text, key words extracted from title, tags from SVM classification, and clusters

References

[1] P. Beckman, T. J. Skluzacek, K., Chard, I. Foster. Skluma: A Statistical Learning Pipeline for Taming Unkempt Data Repositories. Computation Institute, University of Chicago and Argonne National Laboratory, Chicago, IL. 2017.
[2] S. Hoffstaetter, et al. "pytesseract 0.1.7." Python Software Foundation. 1990-2017. https://pypi.python.org/pypi/pytesseract
[3] "Carbon Dioxide Information and Analysis Center." U.S. Department of Energy. Oak Ridge National Laboratory. 2017. ftp://cdiac.ornl.gov

Acknowledgements

This work is supported by the National Science Foundation NSF- 1461260 (BigDataX REU).