CS486C – Senior Capstone Design in Computer Science

Project Description

Project Title: Image Similarity Search Tool for Cline Library's SCA Digital Archives	
Sponsor Information:	**Mike Taylor** NAU Cline Library Head, Technology Strategies and Services mike.taylor@nau.edu

Project Overview:

Cline Library's Special Collections and Archives (SCA) is home to unique collections that help tell the stories of the Colorado Plateau. These images, manuscripts, diaries, maps, and other items provide a valuable resource to researchers, students, and others with an interest in the many topics contained in our physical and digital collections. Whether your interests are in river guides of the Colorado River (before the dam) or stories about trading posts across the Navajo Reservation we are the location for unique primary sources to aid in your research.

SCA archives contain a vast array of digital images that are used extensively for research and educational purposes. Over 100,000 items have been digitized and made available in a searchable platform called CONTENTdm. The image to the right can be found here. Navigating this extensive collection can be challenging. Sometimes you're not sure what you are looking for or even the best way to proceed. Maybe the researcher wants to start a search with her own photo and find other similar items. Or a researcher may want to see results grouped by how similar they are to each other. Navigate to https://nau.edu/special-collections for a look at our current search experience.

This project proposes the development of an image similarity search application that enables users to upload an image and find similar images within the SCA digital archives. Snapping a quick photo of Old Main and uploading to this application would result in images from our collection that we have of that building. This would add another dimension to the search experience that may assist users in ways they had not previously thought possible.

Image similarity searching is not uncommon. Google has had a "Search by image" feature for some time. This is a great tool, but results come from across the internet. What is required in our case is an application fine tuned to our collection. Recent strategies to accomplish this often involve the creation of "image embeddings" or a "vectorization" of the images which are then stored in a database. These vectors can be created from image metadata or the images themselves or both. This article describes such a system. There are many ways to approach this problem and the linked article is just one.

We envision a responsive, web-based application that would allow a user to upload an image as the basis for a search of SCA archives. Search results would be those images most similar to the uploaded image and could be presented in a grid. Similarity would be determined by comparing the embedding calculated on the fly of the uploaded image to all the others from the archive using a measure such as Cosine Similarity. Clicking on an image would automatically reorder the grid, so that the "seed" image would be in the top left hand side position, with the next most similar beside it, and so on, in left to right reading order. As an extension, the inverse of this type of interface could be used to remove groups of related "unwanted" images. We would be able to provide a dataset of images from which embeddings could be created.

Knowledge, skills, and expertise required for this project:

Knowledge of modern Web2.0 programming techniques, languages, and frameworks

Basic understanding of user experience (UX/UI), user design (UD), and accessible web/app interfaces

Skills in GUI design and evaluation

An understanding of image embeddings (vectors) and how to create and store them

An understanding of similarity measures like Cosine Similarity

An understanding of vector databases

Database management and interaction

Equipment Requirements:

There should be no equipment or software required other than a development platform and software/tools freely available online

Cline Library will provide a dataset of images

Software and other Deliverables:

A complete software product as outlined above, fully tested, and installed on a platform of client's choice

A strong as-built report detailing the design and implementation of the product in a complete, clear and professional manner. This document should provide a strong basis for future development of the product

Complete professionally documented codebase, delivered both as a repository in GitHub, or some other version control repository, and as a physical archive on a USB drive