WORK EXPERIENCE for David Black

Speech Science and Machine Learning Team Member *(May 2018-Present)* *CaptionCall (Salt Lake City, UT)*

Wrote and maintained two-part software to normalize text representation in transcripts to spoken forms and then to score results from different transcribers. Researched the process and then was called on to implement the system into production. Oversaw dataset management, including verifying existing data, accessing data from collection efforts already in process, creating new and tailored datasets, outsourcing parts of data collection and ground-truth generation, finding and fixing problems with ground truth data, maintaining the data, and using my knowledge of the datasets to feed other users.

» **Automated Normalization and Scoring** || Used Dynamic-Programming-based alignment and scoring in concert with regular expressions and customized algorithms for these purposes || Designed the software – a Python package – to flexibly manage different scoring philosophies || Developed and deployed the software for automated scoring of a large number of daily transcripts || Helped with the integration of software features into other projects

» **Dataset Management** || Used customized hardware, existing call-recording software, and my experimental-design skills to create new and extensive datasets || Curated existing, created, and purchased datasets, with a focus on verifying ground-truth data and allowing quick finding of various data for unforeseen research needs || Two of the more-important datasets: a pair of conversational-audio datasets - on the order of hundreds of data files - personal oversight of collection strategy; a dataset of telephonic data - on the order of hundreds of thousands of data files and on the order of terabytes of data - helped to ensure legal conformity

» **General Research** || Co-inventor on a 550-page patent application || Helping team members with: model creation for machine learning (kaldi, Scikit-learn, TensorFlow), vendor benchmarking, predictive networks, voting models, ground-truth standardization, and especially file-encoding and file-decoding issues

Python, bash, Scikit-learn, TensorFlow, kaldi, Java (Weka), sclite, Windows batch, WebSockets, JavaScript, C#, Selenium/gecko webdriver, Azure, AWS, CUDA

Data Analyst, Contingent *(July 2016-July 2017)* *FamilySearch (Salt Lake City, UT)*

Found, prepared, and analyzed records and their annotations for the Advanced Research Team working on handwriting recognition using kaldi as an underlying analysis framework. I also used a Dynamic-Programming-based scoring system to analyze results. Discussed and developed algorithms, analyzed data patterns with the research team.

» **Data Curation** || Reviewing and selecting data for collections, particularly a corpus of ~20k documents in 12 languages that cover patron demand || Using my knowledge of: other languages; linguistics; paleography; codicology (bookmaking); and graphemics, spearheading and continuing finding of handwriting examples in these languages that would prove pathological for the ML models. This detection of specific problem cases, as well as further quantification thereof through my personal database searches, helped solve and prevent system problems for the research team and allowed them to make improvements to their algorithms and models, leading to a more robust architecture || Quickly finding various data for unforeseen research needs

» **Data Annotation** || Creating data sets for machine learning algorithms || Writing computer programs (Java) that allow volunteers to transcribe documents, classify documents, and split up images

» **Software Design** || for image classification (Java) || for segmentation of document images and segment classification (Java) || for various encoding, validation, and completion testing (bash, python, C++, Java)

» **Software Evaluation** || Personally wrote code to incorporate a vendor's Asian-Language OCR API using Java Native Library functionality so as to test the vendor's C++/DLL engine in our Java-based system || Scored performance for comparison to other products.

Java, JNI, bash, C++, Windows batch script, kaldi, Python, sclite

Document Specialist *(Fall 2017-April 2018)* *LDS Church History Department (Salt Lake City, UT)*

» Digitization of thousands of family-history-related images from microfilm documents, including image analysis and processing, image manipulation to ensure readability, and quality assurance.

» Served as a paleography and language-recognition resource for other workers due to expertise in these areas.

Proprietary Church Image Processing Software

Graduate Research Assistant *(September 2010-October 2013)* *UC Riverside Physics Department*

» Researched at Brookhaven National Laboratory, developing computer simulations, analyzing data from particle collisions. Statistical analysis, pattern-matching, and signal recognition (C++, shell scripting)

» Oversaw data flow and real-time data quality assurance for terabytes of collision-geometry data daily

» Helped with circuit design, analysis, and implementation at the detector.

C++, bash, C-shell, Condor batch processing, perl, PHP, Python, pSQL, ssh, ftp, Large C++ Physics Libraries

FamilySearch Volunteer - International Floor and AGES Project *(Fall 2016-Present)* *FamilySearch (Salt Lake City, UT)*

» Worked as a language and paleography expert - combined with Google Translate skills - to help Patrons with research in many areas and languages.

» Work with the AGES team to find relevant features and to design and implement different ML architectures for use in Asian-language Handwriting Recognition using Machine Learning.

Python, TensorFlow, Keras