workshop.md

Medical Image De-Identification Workshop

Introduction

In this workshop we will be using machine learning (ML) services to identify and remove protected health information (PHI) from the pixels of medical images. While medical image file formats such as DICOM also store PHI as metadata within the file itself, we will not be addressing the metadata within this workshop.

In Module 1 we will deploy and run a fully-integrated Python solution that detects and redacts PHI in images stored in an S3 bucket. This solution is run as a pre-supplied Jupyter notebook running on a SageMaker instance, using the Boto3 Python SDK to call the ML services.

In Module 2 we will deploy an API-based solution exposing six endpoints, each of which runs one or more steps in the masking process. This solution is fronted by API Gateway and Lambda, with the same ML services providing the text extraction and recognition, and is deployed by a CloudFormation template.

With both solutions we'll enforce security by setting up some necessary permissions using Identity and Access Management (IAM) policies and roles.

Course Resources

Download this zip file, and extract it to your local computer. We'll use files from this directory at several points during the workshop. It contains the contents of this GitHub Repo. The extracted directories contain:

- `doc` : Markdown source code for the course guides
- `iam` : Text file with the IAM policy for module 2
- `images` : Sample images for de-identification
- `pdf` : Course guides in PDF format
- `python` : Jupyter notebook for module 1, Python source code for module 2

Open `module_1.pdf` and `module_2.pdf` from the `pdf` directory, and keep them open for reference during the workshop.