Welcome

Introduction

A critical task in building AI models is collecting raw, unstructured or structured data from various sources that will be used to train the model. The goal is to accumulate a large, diverse, and representative sample of data that reflects the problem the AI model is intended to solve. This data can come from many different sources, such as the humans (like you!), the web, sensors, cameras, text, or databases. Once the data has been collected, it then needs to be annotated (also called labeling or tagging) to add meaningful information to the raw data to help the AI model understand the context and features of the data.

In the Pentas Voluntary Study project, you will collect and annotate booking reservation data across a range of different use cases including travel, entertainment, healthcare, and events. This data will be used to train virtual personal assistants to better identify and interpret the information in text messages, emails, and attachments relating to reservations so this can automatically be added to a calendar, reminders, etc.

Key Terms

Before we begin, review the basic terms and definitions below.

Asset

An asset may be a text message, email, or PDF attachment that contains the type of data requested by the client for their AI model.

For this project, assets will consist of booking confirmations, cancellations, changes/modifications, and event invitations.

Metadata

Metadata describes, explains, or gives context to the main data, helping to make it easier to understand, find, organize, or use. Metadata is often referred to as "data about data" because it offers additional details beyond the actual content, such as how, when, or by whom the data was collected.

For this project, you will provide metadata about your asset, which includes data such as the number of guests in a hotel booking or whether a flight confirmation is one-way or round-trip.

Personally Identifiable Information (PII)

Refers to any information that can be used to identify, locate, or contact an individual, either directly or indirectly. PII includes a broad range of data that, on its own or when combined with other information, can uniquely identify a person.

Various laws and regulations, such as the **General Data Protection Regulation (GDPR)** in the EU and the **California Consumer Privacy Act (CCPA)** in the U.S., govern the collection, storage, and handling of PII. Organizations must ensure that they follow these laws to avoid penalties and legal consequences.

Redaction

To protect your identity, you will be tagging all PII in your asset to be removed in a process called redaction, which is the process of editing a document to remove or obscure sensitive, confidential, or private information before publication or distribution.

For instance, if your asset contains your full name, it would be replaced with something like [first_name_person] [last_name_person].

Data Collection, Redaction and Annotation

Now that you know the basic terminology of data collection, we'll look at the types of assets you'll be asked to collect, how to annotate the assets with metadata, and how to identify PII to redact.

This resource contains all the information required to get started. Please read them thoroughly before completing the quiz to test your understanding. Once you've passed the quiz, you'll be assigned to the task in LDP to start labeling!

Click to Continue

There are two ways to navigate this resource:

1. Use the left-hand menu to quickly move between topics
2. Use the buttons at the bottom of the pages