**Challenge Scenario**

You are a Marketing Campaign Coordinator at a media company, working closely with the Marketing Manager to plan, execute, and evaluate campaigns to meet sales targets. Recently, you secured an exciting new contract with Google. As a Marketing Campaign Coordinator, you’re eager to dive into the materials that will help you familiarize yourself with the Google brand and Google brand identity as quickly as possible. Therefore, you plan to review Google’s brand guidelines, previous campaigns, product ads, customer testimonials, and financial reports by leveraging Gemini’s innovative capabilities to gain deeper insights into Google more efficiently.

In this challenge, you begin with multimodal prompts to extract information from text and visual data, generating a video description, and retrieving extra information beyond the video by using multimodality with Gemini. You also build metadata of documents containing text and images, getting all relevant text chunks, and printing citations by using Multimodal Retrieval Augmented Generation (RAG) with Gemini.

**Task 1. Generate Multimodal Insights with Gemini**

In this task, you familiarize yourself with the Google brand and Google brand identity using Gemini, which is a multimodal model that supports multimodal prompts. You include text, image(s), and video in your prompt requests and get text or code responses.

To complete this task, follow the instructions in the notebook.

**Note:**Save the notebook script before clicking on the  button for every task.

Task 1.1 Image understanding across multiple images

Image understanding across multiple images.

Task 1.2. Similarities/Differences between images

Similarities/Differences between images.

Task 1.3. Generate a video description

Complete this subtask using <https://storage.googleapis.com/spls/gsp520/google-pixel-8-pro.mp4>, and proceed to the next task.

Task 1.4. Extract tags of objects throughout the video

Complete this subtask using <https://storage.googleapis.com/spls/gsp520/google-pixel-8-pro.mp4>, and proceed to the next task.

Task 1.5. Ask more questions about a video

Complete this subtask using <https://storage.googleapis.com/spls/gsp520/google-pixel-8-pro.mp4>, and proceed to the next task.

Task 1.6. Retrieve extra information beyond the video

Generate a video description and retrieve extra information beyond the video.

**Task 2. Retrieve and integrate knowledge with multimodal retrieval augmented generation (RAG)**

To complete this task, follow the instructions in the notebook.

Available data and helper functions for Task 2

For this task, you use two different source data documents:

1. The [Terms of Service document](https://storage.googleapis.com/spls/gsp520/Google_Branding/Google_terms_of_service_en_us.pdf) for Google's services, defining the relationship between Google and its users. It covers what users can expect from Google, the rules for using the services, intellectual property rights related to content, and the procedures for resolving disputes or disagreements. This sample document contains only text.
2. A modified version of [Google-10K](https://abc.xyz/assets/investor/static/pdf/20220202_alphabet_10K.pdf) which provides a comprehensive overview of the company's financial performance, business operations, management, and risk factors. As the original document is rather large, you will be using a modified version with only 14 pages, split into two parts - [Part 1](https://storage.googleapis.com/github-repo/rag/intro_multimodal_rag/intro_multimodal_rag_old_version/data/google-10k-sample-part1.pdf) and [Part 2](https://storage.googleapis.com/github-repo/rag/intro_multimodal_rag/intro_multimodal_rag_old_version/data/google-10k-sample-part2.pdf) instead. Although it's truncated, this sample document still contains text along with images such as tables, charts, and graphs.

You also select from the following helper functions to complete the tasks (more information on these functions available on [GitHub](https://raw.githubusercontent.com/GoogleCloudPlatform/generative-ai/main/gemini/use-cases/retrieval-augmented-generation/utils/intro_multimodal_rag_utils.py)):

* For **Inspect the processed text metadata**:
  + **text**: the original text from the page
  + **text\_embedding\_page**: the embedding of the original text from the page
  + **chunk\_text**: the original text divided into smaller chunks
  + **chunk\_number**: the index of each text chunk
  + **text\_embedding\_chunk**: the embedding of each text chunk
* For **Inspect the processed image metadata**:
  + **img\_desc**: Gemini-generated textual description of the image.
  + **mm\_embedding\_from\_text\_desc\_and\_img**: Combined embedding of image and its description, capturing both visual and textual information.
  + **mm\_embedding\_from\_img\_only**: Image embedding without description, for comparison with description-based analysis.
  + **text\_embedding\_from\_image\_description**: Separate text embedding of the generated description, enabling textual analysis and comparison.
* For **Import the helper functions to implement RAG**:
  + **get\_similar\_text\_from\_query()**: Given a text query, finds text from the document which are relevant, using cosine similarity algorithm. It uses text embeddings from the metadata to compute and the results can be filtered by top score, page/chunk number, or embedding size.
  + **print\_text\_to\_text\_citation()**: Prints the source (citation) and details of the retrieved text from the get\_similar\_text\_from\_query() function.
  + **get\_similar\_image\_from\_query()**: Given an image path or an image, finds images from the document which are relevant. It uses image embeddings from the metadata.
  + **print\_text\_to\_image\_citation()**: Prints the source (citation) and the details of retrieved images from the `get\_similar\_image\_from\_query()`` function.
  + **get\_gemini\_response()**: Interacts with a Gemini model to answer questions based on a combination of text and image inputs.
  + **display\_images()**: Displays a series of images provided as paths or PIL Image objects.

Task 2.1. Build metadata of documents containing text and images

Complete this subtask, and proceed to the next task.

Task 2.2. Create a user query

Complete this subtask, and proceed to the next task.

Task 2.3. Get all relevant text chunks

Complete this subtask, and proceed to the next task.

Task 2.4. Create context\_text

Complete this subtask, and proceed to the next task.

Task 2.5. Pass context to Gemini

Retrieve and integrate knowledge with multimodal retrieval augmented generation (RAG).