Build a prompt library for common business use cases

```
!pip install transformers accelerate torch
```

```
Downloading nvidia_cufft_cu12-11.2.1.3-py3-none-manylinux2014_x86_64.whl (211.5 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 211.5/211.5 MB 5.8 MB/s eta 0:00:00
Downloading nvidia_curand_cu12-10.3.5.147-py3-none-manylinux2014_x86_64.whl (56.3 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 56.3/56.3 MB 17.5 MB/s eta 0:00:00
Downloading nvidia_cusolver_cu12-11.6.1.9-py3-none-manylinux2014_x86_64.whl (127.9 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 127.9/127.9 MB 7.8 MB/s eta 0:00:00
Downloading nvidia_cusparse_cu12-12.3.1.170-py3-none-manylinux2014_x86_64.whl (207.5 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 207.5/207.5 MB 6.3 MB/s eta 0:00:00
Downloading nvidia_nccl_cu12-2.21.5-py3-none-manylinux2014_x86_64.whl (188.7 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 188.7/188.7 MB 6.5 MB/s eta 0:00:00
Downloading nvidia_nvjitlink_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl (21.1 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 21.1/21.1 MB 78.3 MB/s eta 0:00:00
Installing collected packages: nvidia-nvjitlink-cu12, nvidia-nccl-cu12, nvidia-curand-cu12, nvidia-cufft-cu12, nvidia-cuda-runtime-cu1
  Attempting uninstall: nvidia-nvjitlink-cu12
    Found existing installation: nvidia-nvjitlink-cu12 12.5.82
    Uninstalling nvidia-nvjitlink-cu12-12.5.82:
      Successfully uninstalled nvidia-nvjitlink-cu12-12.5.82
  Attempting uninstall: nvidia-nccl-cu12
    Found existing installation: nvidia-nccl-cu12 2.23.4
    Uninstalling nvidia-nccl-cu12-2.23.4:
      Successfully uninstalled nvidia-nccl-cu12-2.23.4
  Attempting uninstall: nvidia-curand-cu12
    Found existing installation: nvidia-curand-cu12 10.3.6.82
    Uninstalling nvidia-curand-cu12-10.3.6.82:
      Successfully uninstalled nvidia-curand-cu12-10.3.6.82
  Attempting uninstall: nvidia-cufft-cu12
    Found existing installation: nvidia-cufft-cu12 11.2.3.61
    Uninstalling nvidia-cufft-cu12-11.2.3.61:
      Successfully uninstalled nvidia-cufft-cu12-11.2.3.61
  Attempting uninstall: nvidia-cuda-runtime-cu12
    Found existing installation: nvidia-cuda-runtime-cu12 12.5.82
    Uninstalling nvidia-cuda-runtime-cu12-12.5.82:
      Successfully uninstalled nvidia-cuda-runtime-cu12-12.5.82
  Attempting uninstall: nvidia-cuda-nvrtc-cu12
    Found existing installation: nvidia-cuda-nvrtc-cu12 12.5.82
    Uninstalling nvidia-cuda-nvrtc-cu12-12.5.82:
      Successfully uninstalled nvidia-cuda-nvrtc-cu12-12.5.82
  Attempting uninstall: nvidia-cuda-cupti-cu12
    Found existing installation: nvidia-cuda-cupti-cu12 12.5.82
    Uninstalling nvidia-cuda-cupti-cu12-12.5.82:
      Successfully uninstalled nvidia-cuda-cupti-cu12-12.5.82
  Attempting uninstall: nvidia-cublas-cu12
    Found existing installation: nvidia-cublas-cu12 12.5.3.2
    Uninstalling nvidia-cublas-cu12-12.5.3.2:
      Successfully uninstalled nvidia-cublas-cu12-12.5.3.2
  Attempting uninstall: nvidia-cusparse-cu12
    Found existing installation: nvidia-cusparse-cu12 12.5.1.3
    Uninstalling nvidia-cusparse-cu12-12.5.1.3:
      Successfully uninstalled nvidia-cusparse-cu12-12.5.1.3
  Attempting uninstall: nvidia-cudnn-cu12
    Found existing installation: nvidia-cudnn-cu12 9.3.0.75
    Uninstalling nvidia-cudnn-cu12-9.3.0.75:
      Successfully uninstalled nvidia-cudnn-cu12-9.3.0.75
  Attempting uninstall: nvidia-cusolver-cu12
    Found existing installation: nvidia-cusolver-cu12 11.6.3.83
    Uninstalling nvidia-cusolver-cu12-11.6.3.83:
      Successfully uninstalled nvidia-cusolver-cu12-11.6.3.83
Successfully installed nvidia-cublas-cu12-12.4.5.8 nvidia-cuda-cupti-cu12-12.4.127 nvidia-cuda-nvrtc-cu12-12.4.127 nvidia-cuda-runtime
```

```
!pip -q install -U transformers accelerate sentencepiece
```

```
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 42.0/42.0 kB 1.1 MB/s eta 0:00:00
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 11.3/11.3 MB 29.6 MB/s eta 0:00:00
```

```
import torch, platform
print("PyTorch:", torch.__version__)
print("CUDA available:", torch.cuda.is_available())
if torch.cuda.is_available():
    print("GPU:", torch.cuda.get_device_name(0))
print("Python:", platform.python_version())
```

```
PyTorch: 2.6.0+cu124
CUDA available: True
GPU: Tesla T4
```

```
Python: 3.11.13
```

```
!pip -q install -U "transformers>=4.44" "accelerate>=0.34"
```

```
from transformers import pipeline, AutoTokenizer
import random
model_id = "TinyLlama/TinyLlama-1.1B-Chat-v1.0"
tok = AutoTokenizer.from_pretrained(model_id)

generator = pipeline(
    "text-generation",
    model=model_id,
    tokenizer=tok,
    device=-1,
    torch_dtype="auto",
)


prompt_library = {
    "communication": [
        "Write a professional email to a client explaining a project delay. Keep the tone polite, empathetic, and solution-focused.",
        "Summarize the following meeting notes into key action items and deadlines: {notes}"
    ],
    "marketing": [
        "Write 3 variations of a LinkedIn ad for a SaaS product that helps small businesses manage expenses.",
        "Generate 3 customer personas for an online fitness coaching app. Include demographics, goals, and motivations."
    ],
    "customer_support": [
        "Draft a polite and concise response to a customer who is upset about a billing error.",
        "Generate a FAQ section for a subscription-based e-learning platform."
    ],
    "analysis": [
        "Summarize the strengths and weaknesses of Competitor X in a SWOT table.",
        "List top 5 AI trends in retail with examples."
    ],
    "project_management": [
        "Break down the goal 'launch a new company website' into phases, tasks, and estimated timelines.",
        "List 10 potential risks for an IT system migration project, along with likelihood and mitigation strategies."
    ],
}

def generate_from_prompt(category, idx=None, custom_input=None):
    """Pick a prompt from the library and generate output"""
    prompts = prompt_library.get(category, [])
    if not prompts:
        raise ValueError(f"No prompts found for category: {category}")


    prompt = prompts[idx] if idx is not None else random.choice(prompts)

    if custom_input:
        prompt = prompt.format(**custom_input)


    messages = [{"role": "system", "content": prompt}]
    input_text = tok.apply_chat_template(messages, tokenize=False, add_generation_prompt=True)


    out = generator(
        input_text,
        max_new_tokens=200,
        do_sample=True,
        temperature=0.8,
        top_p=0.9,
        repetition_penalty=1.2,
        pad_token_id=tok.eos_token_id,
        eos_token_id=tok.eos_token_id,
        return_full_text=False,
    )
    return {"prompt": prompt, "response": out[0]["generated_text"]}

if __name__ == "__main__":

    result = generate_from_prompt("communication", idx=0)
    print("Prompt:", result["prompt"])
    print("Response:", result["response"])
```

```


result2 = generate_from_prompt("communication", idx=1, custom_input={"notes": "Discussed budget cuts, timeline change to Q4, assign
print("\nPrompt:", result2["prompt"])
print("Response:", result2["response"])

```

```
/usr/local/lib/python3.11/dist-packages/huggingface_hub/utils/_auth.py:94: UserWarning:
The secret `HF_TOKEN` does not exist in your Colab secrets.
To authenticate with the Hugging Face Hub, create a token in your settings tab (https://huggingface.co/settings/tokens), set it as secre
You will be able to reuse this secret in all of your notebooks.
Please note that authentication is recommended but still optional to access public models or datasets.
  warnings.warn(
tokenizer_config.json:   1.29k/?  [00:00<00:00,  30.2kB/s]
tokenizer.model: 100%   500k/500k  [00:01<00:00,  347kB/s]
tokenizer.json:   1.84M/?  [00:00<00:00,  26.3MB/s]
special_tokens_map.json: 100%   551/551  [00:00<00:00,  20.0kB/s]
config.json: 100%   608/608  [00:00<00:00,  15.9kB/s]
model.safetensors: 100%   2.20G/2.20G  [00:32<00:00,  111MB/s]
generation_config.json: 100%   124/124  [00:00<00:00,  8.42kB/s]
Device set to use cpu
Prompt: Write a professional email to a client explaining a project delay. Keep the tone polite, empathetic, and solution-focused.
Response: Subject: Project Delay Reason and Solutions Proposed

Dear [Client Name],

I am writing this email to you with great concern regarding the delay in completing our project as per our agreed timeline. I understand

As you know, we have been working on this project since January last year, and we had initially planned to complete it by June. However,

We realize how challenging these circumstances must have been for all involved parties, including ourselves. We believe

Prompt: Summarize the following meeting notes into key action items and deadlines: Discussed budget cuts, timeline change to Q4, assign
Response: During a meeting, it was summarized that the following action items and deadlines were discussed:

- Budget cuts have been identified for Q1
- The proposed timeline has changed from Q3 to Q4 to accommodate these changes
- A revised task assignment has been assigned to John regarding a specific project

The meeting agenda included discussions on budget cuts, assigning tasks, and ensuring projects are completed within their designated tim
```