Pipeline de Dados para Extração de Vídeos do YouTube (Engenheiro de Dados)

Objetivo: Criar um pipeline de dados simples para extrair dados textuais de vídeos do YouTube, realizar limpeza e pré-processamento, armazenar os dados em formato adequado e fornecer um exemplo de uso para aprendizado de máquina.

Cenário: Construir um pipeline de dados para alimentar um sistema de aprendizado de máquina com dados textuais extraídos de vídeos do YouTube, como transcrições e metadados.

Tarefa:

- Criar um script Python para coletar dados textuais de vídeos do YouTube (metadados e transcrições).
- Implementar limpeza e pré-processamento dos dados textuais.
- Armazenar os dados processados em formato JSON.
- Hospedar o código em um repositório GitHub com documentação.

Requisitos:

- Usar Python 3.9+.
- Coletar metadados (título, descrição) e transcrições de um vídeo do YouTube via API ou biblioteca (e.g., pytube, youtube_transcript_api).
- Realizar limpeza (remover caracteres especiais, normalizar texto) e pré-processamento (tokenização, remoção de stopwords).
- Armazenar dados em JSON em memória (evitar I/O local, mas simular armazenamento).
- Código modular, seguro e com logging básico.
- Incluir README com instruções de configuração, dependências e uso.
- Sugerir ferramentas para cada etapa do pipeline.

Ferramentas Sugeridas:

- **Coleta de Dados**:
    - pytube: Extrair metadados do YouTube (título, descrição).
    - youtube_transcript_api: Obter transcrições automáticas ou manuais.
    - Alternativa: YouTube Data API v3 (requer chave API).
- **Limpeza e Pré-processamento**:
    - re: Remover caracteres especiais.
    - nltk ou spacy: Tokenização, remoção de stopwords, normalização (minúsculas).
- **Armazenamento**:
    - json: Estruturar e salvar dados em formato JSON.
    - Alternativa: pandas para manipulação intermediária e exportação.
- **Outras**:
    - logging: Registrar etapas do pipeline.
    - python-dotenv: Gerenciar chaves API de forma segura.

Formato de Saída Esperado (JSON):

```
{
  "metadata": {
        "title": "sample video title",
        "description": "sample youtube video description"
  },
  "transcript": [
        {
        "start": 0.0,
        "text": "welcome video"
        },
        {
        "start": 2.1,
        "text": "example content"
        }
  ]
}
```

Sugestão de vídeo para extração e tratamento:
https://www.youtube.com/watch?v=N6kdD_x3v1g

Entrega:

- Disponibilizar código no github pessoal.
- Incluir README breve explicando configuração, dependências e uso.
- **Entrega em 6 dias, a contar da data do recebimento da tarefa.**