SageMaker Endpoint Embeddings

Let's load the SageMaker Endpoints Embeddings class. The class can be used if you host, e.g. your own Hugging Face model on SageMaker.

For instrucstions on how to do this, please see here

```
!pip3 install langchain boto3
```

```
from typing import Dict
from langchain.embeddings import SagemakerEndpointEmbeddings
from langchain.llms.sagemaker_endpoint import ContentHandlerBase
import json


class ContentHandler(ContentHandlerBase):
    content_type = "application/json"
    accepts = "application/json"

    def transform_input(self, prompt: str, model_kwargs: Dict) -> bytes:
        input_str = json.dumps({"inputs": prompt, **model_kwargs})
        return input_str.encode('utf-8')

    def transform_output(self, output: bytes) -> str:
        response_json = json.loads(output.read().decode("utf-8"))
        return response_json["embeddings"]

content_handler = ContentHandler()


embeddings = SagemakerEndpointEmbeddings(
    # endpoint_name="endpoint-name",
    # credentials_profile_name="credentials-profile-name",
    endpoint_name="huggingface-pytorch-inference-2023-03-21-16-14-03-834",
    region_name="us-east-1",
    content_handler=content_handler
)
```

```
query_result = embeddings.embed_query("foo")
```

Skip to main content

```
doc_results = embeddings.embed_documents(["foo"])
```

```
doc_results
```