Huggingface TextGen Inference

Text Generation Inference is a Rust, Python and gRPC server for text generation inference. Used in production at HuggingFace to power LLMs api-inference widgets.

This notebooks goes over how to use a self hosted LLM using `Text Generation Inference`.

To use, you should have the `text_generation` python package installed.

```
# !pip3 install text_generation
```

```
llm = HuggingFaceTextGenInference(
    inference_server_url="http://localhost:8010/",
    max_new_tokens=512,
    top_k=10,
    top_p=0.95,
    typical_p=0.95,
    temperature=0.01,
    repetition_penalty=1.03,
)
llm("What did foo say about bar?")
```