Custom String Evaluator

You can make your own custom string evaluators by inheriting from the `StringEvaluator` class and implementing the `_evaluate_strings` (and `_aevaluate_strings` for async support) methods.

In this example, you will create a perplexity evaluator using the HuggingFace evaluate library. Perplexity is a measure of how well the generated text would be predicted by the model used to compute the metric.

```
# %pip install evaluate > /dev/null
```

```python
from typing import Any, Optional

from langchain.evaluation import StringEvaluator
from evaluate import load


class PerplexityEvaluator(StringEvaluator):
    """Evaluate the perplexity of a predicted string."""

    def __init__(self, model_id: str = "gpt2"):
        self.model_id = model_id
        self.metric_fn = load(
            "perplexity", module_type="metric", model_id=self.model_id, pad_token=0
        )

    def _evaluate_strings(
        self,
        *,
        prediction: str,
        reference: Optional[str] = None,
        input: Optional[str] = None,
        **kwargs: Any,
    ) -> dict:
        results = self.metric_fn.compute(
            predictions=[prediction], model_id=self.model_id
        )
```

```
        ppl = results["perplexities"][0]
        return {"score": ppl}
```

API Reference:

- StringEvaluator from `langchain.evaluation`

```
evaluator = PerplexityEvaluator()
```

```
evaluator.evaluate_strings(prediction="The rains in Spain fall mainly
on the plain.")
```

```
    Using pad_token, but it is not set yet.


    huggingface/tokenizers: The current process just got forked, after
parallelism has already been used. Disabling parallelism to avoid
deadlocks...
    To disable this warning, you can either:
        - Avoid using `tokenizers` before the fork if possible
        - Explicitly set the environment variable
TOKENIZERS_PARALLELISM=(true | false)




    0%|              | 0/1 [00:00<?, ?it/s]




    {'score': 190.3675537109375}
```

```
# The perplexity is much higher since LangChain was introduced after
'gpt-2' was released and because it is never used in the following
context.
evaluator.evaluate_strings(prediction="The rains in Spain fall mainly
on LangChain.")
```

```
Using pad_token, but it is not set yet.
```

```
0%|          | 0/1 [00:00<?, ?it/s]
```

```
{'score': 1982.0709228515625}
```