Tokenizers (PyTorch)

File Edit View Insert Runtime Tools Help

+ Code + Text Copy to Drive

Tokenizers (PyTorch)

Install the Transformers, Datasets, and Evaluate libraries to run this notebook.

```
[ ] !pip install datasets evaluate transformers[sentencepiece]
```

```
[ ] tokenized_text = "Jim Henson was a puppeteer".split()
    print(tokenized_text)
```

```
['Jim', 'Henson', 'was', 'a', 'puppeteer']
```

```
[ ] from transformers import BertTokenizer

    tokenizer = BertTokenizer.from_pretrained("bert-base-cased")
```

```
[ ] from transformers import AutoTokenizer

    tokenizer = AutoTokenizer.from_pretrained("bert-base-cased")
```

```
[ ] tokenizer("Using a Transformer network is simple")
```

```
{'input_ids': [101, 7993, 170, 11303, 1200, 2443, 1110, 3014, 102],
```

```
[ ] tokenizer.save_pretrained("directory_on_my_computer")
```

```
[ ] from transformers import AutoTokenizer

    tokenizer = AutoTokenizer.from_pretrained("bert-base-cased")

    sequence = "Using a Transformer network is simple"
    tokens = tokenizer.tokenize(sequence)

    print(tokens)
```

```
['Using', 'a', 'transform', '##er', 'network', 'is', 'simple']
```

```
[ ] ids = tokenizer.convert_tokens_to_ids(tokens)

    print(ids)
```

```
[7993, 170, 11303, 1200, 2443, 1110, 3014]
```

```
[ ] decoded_string = tokenizer.decode([7993, 170, 11303, 1200, 2443, 1110, 3014])
    print(decoded_string)
```

```
'Using a Transformer network is simple'
```