

Praktikum 2 IF4072

Pemrosesan Bahasa

Alami

Sequential Labeling

Practicum Rules !

1. Practicum is carried out independently (1 person / group)
2. All of the practicum will be done by using Python and Google Colab
3. All documentations are allowed (e.g. HF documentation, Keras documentation, StackOverflow)
4. QnA:
https://docs.google.com/spreadsheets/d/1FiMUwK5KW7GK94mVGXizZs6Lpc0q_NAv1bfQ_1WtQ80/edit?usp=sharing

Practicum 2

1. Fine-tune a sequential labeling model: **attention-based** (deadline: 24/09/23, 23.59 WIB) using [*id_nergrit_corpus](https://huggingface.co/datasets/id_nergrit_corpus), an Indonesian NER dataset. Model must provide metric values (from validation/train phase) and predictions from several test inputs (or test set).
2. Complete your source code with comments that explain the source code.
3. The submission in Edunex will be closed at 23.59 WIB. Convert your .ipynb file to PDF and submit it with the name format is “**Prak2_<NIM>.pdf**”


*https://huggingface.co/datasets/id_nergrit_corpus


Practicum 2


Hint:


1. Try using Indonesian / multilingual pre-trained models (e.g. IndoBERT, IndoROBERTa, mBERT).
2. To load the dataset:

```
✓ 18s ▶ from datasets import load_dataset  
  
nergrit = load_dataset("id_nergrit_corpus", "ner")
```

Download data: 100%  15.0M/15.0M [00:00<00:00, 57.9MB/s]

Generating train split: 100%  12532/12532 [00:02<00:00, 5400.48 examples/s]

Generating test split: 100%  2399/2399 [00:00<00:00, 4315.91 examples/s]

Generating validation split: 100%  2521/2521 [00:00<00:00, 4826.48 examples/s]

Practicum 2

Hint:

3. Helpful documentation:

https://github.com/huggingface/notebooks/blob/main/transformers_doc/en/token_classification.ipynb