bnlm Documentation

Release latest

Jun 09, 2020

Contents

1 Installation **3**

2 Evaluation Result **5**
- 2.1 Language Model . . . 5
- 2.2 Training . . . 5

3 Features and API **7**

4 Download pretrained Model **9**

5 Predict N Words **11**

6 Get Sentence Encoding **13**

7 Get Embedding Vectors **15**

8 Sentence Similarity **17**

9 Get Simillar Sentences **19**

10 Classification **21**

Bengali language model is build with fastai's ULMFit and ready for `prediction` and `classfication` task.

NB:

- This tool mostly followed inltk
- We separated `Bengali` part with better evaluation results

CHAPTER 1

Installation

```
pip install bnlm
```

CHAPTER 2

Evaluation Result

2.1 Language Model

- Accuracy 48.26% on validation dataset
- Perplexity: ~22.79

2.2 Training

To train with your own corpus follow this repository

CHAPTER 3

Features and API

CHAPTER 4

Download pretrained Model

To start, first download pretrained Language Model and Sentencepiece model

```
from bnlm.bnlm import download_models

download_models()
```

CHAPTER 5

Predict N Words

```
from bnlm.bnlm import BengaliTokenizer
from bnlm.bnlm import predict_n_words
model_path = 'model'
input_sen = " "
output = predict_n_words(input_sen, 3, model_path)
print("Word Prediction: ", output)
```

CHAPTER 6

Get Sentence Encoding

```
from bnlm.bnlm import BengaliTokenizer
from bnlm.bnlm import get_sentence_encoding
model_path = 'model'
sp_model = "model/bn_spm.model"
input_sentence = "  "
encoding = get_sentence_encoding(input_sentence, model_path, sp_model)
print("sentence encoding is: ", encoding)
```

CHAPTER 7

Get Embedding Vectors

```
from bnlm.bnlm import BengaliTokenizer
from bnlm.bnlm import get_embedding_vectors
model_path = 'model'
sp_model = "model/bn_spm.model"
input_sentence = "  "
embed = get_embedding_vectors(input_sentence, model_path, sp_model)
print("sentence embedding is : ", embed)
```

CHAPTER 8

Sentence Similarity

```
from bnlm.bnlm import BengaliTokenizer
from bnlm.bnlm import get_sentence_encoding
from bnlm.bnlm import get_sentence_similarity
model_path = 'model'
sp_model = "model/bn_spm.model"
sentence_1 = "  "
sentence_2 = "  "
sim = get_sentence_similarity(sentence_1, sentence_2, model_path, sp_model)
print("similarity is: ", sim)
```

CHAPTER 9

Get Simillar Sentences

```
from bnlm.bnlm import BengaliTokenizer
from bnlm.bnlm import get_embedding_vectors
from bnlm.bnlm import get_similar_sentences

model_path = 'model'
sp_model = "model/bn_spm.model"

input_sentence = "  "
sen_pred = get_similar_sentences(input_sentence, 3, model_path, sp_model)
print(sen_pred)
```

CHAPTER 10

Classification

upcomming