konoha

himkt

Jan 07, 2023

CONTENTS

CHAPTER

ONE

QUICK START

Welcome to Konoha! Konoha is a library for text processing in Japanese. In Japanese, we have to split a sentence into a sequence of words, called `tokenization`. There are many tools available for tokenizing sentences, and usages of them are not the same.

Konoha provides a unified interface to use these tools. You can try Konoha only by running `docker run` on your computer.

CHAPTER

TWO

INSTALLATION

Konoha supports Python 3.5 or newer.

We recommend to install via pip:

```
$ pip install konoha[all]
```

If you want to install Konoha with AllenNLP integration, please run:

```
$ pip install konoha[all_with_integrations]
```

You can also install Konoha with specific tokenizer, please run:

```
$ pip install konoha[janome,kytea,mecab,sentencepiece,sudachi,nagisa]  # specify one or␣
↪more of them
```

If you run *pip install konoha*, Konoha will be installed only with sentence splitter.

You can also install the development version of Konoha from the main branch of Git repository:

```
$ pip install git+https://github.com/himkt/konoha.git
```

CHAPTER

THREE

API REFERENCE

3.1 Word Level Tokenizer Interface

class konoha.word_tokenizer.**WordTokenizer**(*tokenizer: str = 'MeCab'*, *user_dictionary_path: Optional[str] = None*, *system_dictionary_path: Optional[str] = None*, *model_path: Optional[str] = None*, *mode: Optional[str] = None*, *dictionary_format: Optional[str] = None*, *endpoint: Optional[str] = None*, *ssl: Optional[bool] = None*, *port: Optional[int] = None*)

batch_tokenize(*texts: List[str]*) → List[List[*Token*]]

Tokenize input texts

tokenize(*text: str*) → List[*Token*]

Tokenize input text

3.2 Sentence Level Tokenizer Interface

class konoha.sentence_tokenizer.**SentenceTokenizer**(*period: Optional[str] = None*, *patterns: Optional[List[Pattern[str]]] = None*)

3.3 Word Tokenizer Implementations

3.3.1 Base Word Tokenizer

class konoha.word_tokenizers.tokenizer.**BaseTokenizer**(*name: str*)

Base class for word levelkonoha.tokenizer

property name: str

Return name ofkonoha.tokenizer

abstract tokenize(*text: str*) → List[*Token*]

Abstract method forkonoha.tokenization

3.3.2 Character Tokenizer

class konoha.word_tokenizers.character_tokenizer.**CharacterTokenizer**

tokenize(*text: str*)

Abstract method forkonoha.tokenization

3.3.3 MeCab Tokenizer

class konoha.word_tokenizers.mecab_tokenizer.**MeCabTokenizer**(*user_dictionary_path: Optional[str] = None*, *system_dictionary_path: Optional[str] = None*, *dictionary_format: Optional[str] = None*)

tokenize(*text: str*) → List[*Token*]

Abstract method forkonoha.tokenization

3.3.4 KyTea Tokenizer

class konoha.word_tokenizers.kytea_tokenizer.**KyTeaTokenizer**(*model_path: Optional[str] = None*)

tokenize(*text: str*) → List[*Token*]

Abstract method forkonoha.tokenization

3.3.5 Sentencepiece Tokenizer

class konoha.word_tokenizers.sentencepiece_tokenizer.**SentencepieceTokenizer**(*model_path: str*)

tokenize(*text: str*) → List[*Token*]

Abstract method forkonoha.tokenization

3.3.6 Sudachi Tokenizer

class konoha.word_tokenizers.sudachi_tokenizer.**SudachiTokenizer**(*mode: str*)

tokenize(*text: str*) → List[*Token*]

Abstract method forkonoha.tokenization

3.3.7 Janome Tokenizer

class konoha.word_tokenizers.janome_tokenizer.**JanomeTokenizer**(*user_dictionary_path: Optional[str] = None*)

tokenize(*text: str*) → List[*Token*]

Abstract method forkonoha.tokenization

3.3.8 nagisa Tokenizer

class konoha.word_tokenizers.nagisa_tokenizer.**NagisaTokenizer**

tokenize(*text: str*) → List[*Token*]

Abstract method forkonoha.tokenization

3.3.9 Whitespace Tokenizer

class konoha.word_tokenizers.whitespace_tokenizer.**WhitespaceTokenizer**

Simple rule-based word tokenizer.

tokenize(*text: str*) → List[*Token*]

Abstract method forkonoha.tokenization

3.4 Token

(Deprecated)

3.5 Data classes

3.5.1 Token

class konoha.data.token.**Token**(*surface: str*, *postag: Optional[str] = None*, *postag2: Optional[str] = None*, *postag3: Optional[str] = None*, *postag4: Optional[str] = None*, *inflection: Optional[str] = None*, *conjugation: Optional[str] = None*, *base_form: Optional[str] = None*, *yomi: Optional[str] = None*, *pron: Optional[str] = None*, *normalized_form: Optional[str] = None*)

Token class for konoha.

3.5.2 Resource

class konoha.data.resource.**Resource**(*path: Optional[str]*)

download_from_s3(*path: str*) → str

Download file(s) from Amazon S3.

3.6 Server

TBD

CHAPTER

FOUR

INDICES AND TABLES

- genindex
- modindex
- search

PYTHON MODULE INDEX

k

INDEX

N

R

S

T

W