
Kadot Documentation

Release 1.0dev

the_new_sky

Jan 24, 2020

Contents

1	What's in 1.0 ?	3
2	License	5

You are reading the documentation of Kadot 1.0 which is under development.

- [genindex](#)
- [modindex](#)
- [search](#)

Kadot is an open-source library to easily process text documents. It relies on vector representations of documents or words in order to solve NLP tasks such as **summarization**, **spellchecking** or **classification**.

```
# How to get n-grams using kadot.
>>> from kadot.tokenizers import regex_tokenizer
>>> hello_tokens = regex_tokenizer("Kadot just lets you process a text easily.")
>>> hello_tokens.ngrams(n=2)

[('Kadot', 'just'), ('just', 'lets'), ('lets', 'you'), ('you', 'process'), ('process',
↪ 'a'), ('a', 'text'), ('text', 'easily')]
```


CHAPTER 1

What's in 1.0 ?

All these new features may not yet be available on Github.

- **Vectorizers** : We are now offering Word2Vec, the state-of-the-art Fasttext and Doc2Vec algorithms using [Gensim](#) 's powerful backend.
- **Performances** : Using a much more efficient algorithm, the new word vectorizer is up to 95% faster and sparse vectors now take up to 94% less memory.
- **Models** : Kadot now includes an *automatic text summarizer* and an *entity labeler* which can be useful in many projects.
- **Bot Engine** ?
- **Dependencies** : In order to guarantee good performance without reinventing the wheel, we are adding [Gensim](#) and [Pytorch](#) to our list of dependencies. Although installed by default, these libraries (with scikit-learn) will be optional and only Numpy and Scipy are strictly required to use Kadot.

CHAPTER 2

License

Kadot is under [MIT license](#) .

I am not a native English speaker, if you see any language mistakes in the documentation or in the code, please open an issue on Github.