kea Documentation

Release 0.2

Florian Boudin

October 27, 2011

CONTENTS

Contents:

Name kea

Authors Florian Boudin (florian.boudin@univ-nantes.fr)

Version 0.2-dev

Date

- 26 oct. 2011

Description kea is a tokenizer for French. The tokenization process is decomposed in two steps:

1. A rule-based tokenization approach is employed using the punctuation as an indication of token boundaries.
2. A large-coverage lexicon is used to merge over-tokenized units (e.g. fixed contractions such as *aujourd'hui* are considered as one token)

History

- 0.2 (26 oct. 2011), adding a large lexicon constructed from the lefff.
- 0.1 (20 oct. 2011), first released version.

Usage A typical usage of this module is sample:

```
>>> import kea
>>> sentence = "Le Kea est le seul perroquet alpin au monde."
>>> keatokenizer = kea.tokenizer()
>>> tokens = keatokenizer.tokenize(sentence)
['Le', 'Kea', 'est', 'le', 'seul', 'perroquet', 'alpin', 'au', 'monde',
'.']
```

CHAPTER

ONE

TOKENIZER CLASS

class `kea.`**`tokenizer`**

The Kea Tokenizer is a rule-based tokenizer for french.

`lexicon`

The dictionary containing the lexicon

`loadlist`(*path*)

Load a resource list and generate the corresponding regexp part

`resources`

The path of the resources folder

`tokenize`(*text*)

Tokenize the sentence given in parameter and return a list of tokens. This is a two-steps process: 1. tokenize text using punctuation marks, 2. merge over-tokenized units using the lexicon.

CHAPTER

TWO

INDICES AND TABLES

- *genindex*
- *modindex*
- *search*

PYTHON MODULE INDEX

INDEX