
FinnTK Documentation

Frankie Robertson

Nov 01, 2020

CONTENTS

1	finntk	1
1.1	finntk.omor.extract	1
1.2	finntk.omor.inst	2
1.3	finntk.omor.tok	2
1.4	finntk.omor.seg	2
1.5	finntk.wordnet	3
2	Indices and tables	5
	Python Module Index	7
	Index	9

`finntk.analysis_to_subword_dicts(ana)`

Returns a list of list of dicts. Each list element is an analysis. For each analysis, there is a list of subwords. Each dict contains an Omorfi analysis

`finntk.extract_lemmas(word_form)`

Extract lemmas specifically mentioned by OMorFi.

`finntk.extract_lemmas_combs(word_form)`

Works like *extract_lemmas*, but also tries to combine adjacent subwords to make lemmas which may be out of volcaburary for OMorFi.

Note that this will over generate (by design). For example: *voileipäkakku* will generate *voi*, *voileipä* and *voileipäkakku* as desired, but will also spuriously generate *leipäkakku*.

`finntk.extract_lemmas_recurs(word_form)`

Works like *extract_lemmas*, but also tries to expand each lemma into more lemmas. This helps in some cases (but can overgenerate even more). For example, it will mean that *synnyinkaupunkini* will generate *synty*, *kaupunki*, *synnyinkaupunki*, *synnyin* and *syntyä*.

`finntk.get_omorfi()`

Gets an Omorfi instance with everything possible enabled. Reuses the existing instance if already called once.

`finntk.get_token_positions(tokenised, text)`

Returns the start positions of a series of tokens produced by `Omorfi.tokenise(...)`

<code>finntk.omor.extract</code>	Functions for extracting lemmas from OMorFi analyses.
<code>finntk.omor.inst</code>	Function to get ahold of an OMorFi instance.
<code>finntk.omor.tok</code>	Functions for basic processing of OMorFi tokens.
<code>finntk.omor.seg</code>	Functions for basic processing of OMorFi segment labelling style analyses.
<code>finntk.wordnet</code>	Utilities for working with FinnWordNet

1.1 finntk.omor.extract

Functions for extracting lemmas from OMorFi analyses.

`finntk.omor.extract.extract_lemmas(word_form)`

Extract lemmas specifically mentioned by OMorFi.

`finntk.omor.extract.extract_lemmas_combs(word_form)`

Works like *extract_lemmas*, but also tries to combine adjacent subwords to make lemmas which may be out of volcaburary for OMorFi.

Note that this will over generate (by design). For example: *voileipäkakku* will generate *voi*, *voileipä* and *voileipäkakku* as desired, but will also spuriously generate *leipäkakku*.

`finntk.omor.extract.extract_lemmas_recurs(word_form)`

Works like *extract_lemmas*, but also tries to expand each lemma into more lemmas. This helps in some cases (but can overgenerate even more). For example, it will mean that *synnyinkaupunkini* will generate *synty*, *kaupunki*, *synnyinkaupunki*, *synnyin* and *syntyä*.

`finntk.omor.extract.extract_lemmas_span(word_form)`

Works like *extract_lemmas*, but doesn't extract individual subwords. However, if a word is only recognised by as a compound word by OMorFi it will glue the parts together, lemmatising only the last subword. This means it extracts only lemmas which span the whole word form.

`finntk.omor.extract.extract_true_lemmas_span(word_form, norm_func=<function
iden_func>, return_pos=False)`

Works like *extract_lemmas_span*, but uses *true_lemmatise*. It also returns some of the features associated with each lemma.

`finntk.omor.extract.lemma_intersect(toks1, toks2)`

Given two iterables of tokens, return the intersection of their lemmas. This can work as a simple, high recall, method of matching for example, two inflected noun phrases.

1.2 finntk.omor.inst

Function to get ahold of an OMorFi instance.

`finntk.omor.inst.get_omorfi()`

Gets an Omorfi instance with everything possible enabled. Reuses the existing instance if already called once.

1.3 finntk.omor.tok

Functions for basic processing of OMorFi tokens.

`finntk.omor.tok.get_token_positions(tokenised, text)`

Returns the start positions of a series of tokens produced by *Omorfi.tokenise(...)*

1.4 finntk.omor.seg

Functions for basic processing of OMorFi segment labelling style analyses.

`finntk.omor.seg.labelsegment_to_subword_tokens(labelsegmented)`

Returns an iterator of segments specified as (type, value) tuple. Type is one of "seg", "tag" or "surf".

`finntk.omor.seg.tokens_to_surf(it)`

Given an iterator of segments as (type, value) tuples, reconstruct the surface string.

1.5 finntk.wordnet

Utilities for working with FinnWordNet

`finntk.wordnet.has_abbrev(lemma)`

Given a FinnWordNet formatted lemma, e.g. `saada_tehdä_jtak` return whether it contains a placeholder abbreviation.

INDICES AND TABLES

- `genindex`
- `modindex`
- `search`

PYTHON MODULE INDEX

f

- `finntk`, [1](#)
- `finntk.omor.extract`, [1](#)
- `finntk.omor.inst`, [2](#)
- `finntk.omor.seg`, [2](#)
- `finntk.omor.tok`, [2](#)
- `finntk.wordnet`, [3](#)

A

`analysis_to_subword_dicts()` (in module *finntk*), 1

E

`extract_lemmas()` (in module *finntk*), 1
`extract_lemmas()` (in module *finntk.omor.extract*), 1

`extract_lemmas_combs()` (in module *finntk*), 1
`extract_lemmas_combs()` (in module *finntk.omor.extract*), 1

`extract_lemmas_rekurs()` (in module *finntk*), 1
`extract_lemmas_rekurs()` (in module *finntk.omor.extract*), 2

`extract_lemmas_span()` (in module *finntk.omor.extract*), 2

`extract_true_lemmas_span()` (in module *finntk.omor.extract*), 2

F

finntk
 module, 1
finntk.omor.extract
 module, 1
finntk.omor.inst
 module, 2
finntk.omor.seg
 module, 2
finntk.omor.tok
 module, 2
finntk.wordnet
 module, 3

G

`get_omorfi()` (in module *finntk*), 1
`get_omorfi()` (in module *finntk.omor.inst*), 2
`get_token_positions()` (in module *finntk*), 1
`get_token_positions()` (in module *finntk.omor.tok*), 2

H

`has_abbrev()` (in module *finntk.wordnet*), 3

L

`labelsegment_to_subword_tokens()` (in module *finntk.omor.seg*), 2

`lemma_intersect()` (in module *finntk.omor.extract*), 2

M

module
finntk, 1
finntk.omor.extract, 1
finntk.omor.inst, 2
finntk.omor.seg, 2
finntk.omor.tok, 2
finntk.wordnet, 3

T

`tokens_to_surf()` (in module *finntk.omor.seg*), 2