Package ‘finnsurveytext’

February 16, 2024

Type Package

Title Analyse Open-Ended Survey Responses in Finnish

Version 1.0.0


License MIT + file LICENSE

Depends R (>= 3.5.0)

Imports dplyr, ggplot2, ggrepudr, ggraph, gridExtra, igraph, magrittr, RColorBrewer, stopwords, stringr, textrank, tibble, tidywr, upipe, wordcloud

Suggests knitr, rmarkdown

VignetteBuilder knitr

Encoding UTF-8

LazyData true

RoxygenNote 7.3.1


NeedsCompilation no

Author Adeline Clarke [cre, aut],
Krista Lagus [aut],
Katja Laine [aut],
Maria Litova [aut],
Matti Nelimarca [aut],
Joni Oksanen [aut],
Jaakko Peltonen [aut],
Tuukka Oikarinin [aut],
Jani-Matti Tirkkonen [aut],
Ida Toivanen [aut],
Maria Valaste [aut],
Shannon Emilia Carson [ctb],
Sirpa Lappalainen [ctb],
Tuukka Puonti [ctb],
Kimmo Vehkalahti [ctb],
DARIAH-FI [cph, fnd]

Maintainer  Adeline Clarke <adelinepclarke@gmail.com>
Repository  CRAN
Date/Publication  2024-02-16 14:50:02 UTC

R topics documented:

child_barometer_data ........................................... 3
conllu_cb_bullying .............................................. 4
conllu_cb_bullying_iso .......................................... 5
conllu_dev_q11_1 ................................................. 6
conllu_dev_q11_1_f ................................................ 7
conllu_dev_q11_1_f_nltk .......................................... 8
conllu_dev_q11_1_m ............................................... 9
conllu_dev_q11_1_m_nltk .......................................... 10
conllu_dev_q11_1_na ............................................. 11
conllu_dev_q11_1_na_nltk ......................................... 12
conllu_dev_q11_1_nltk ........................................... 13
conllu_dev_q11_1_snow ........................................... 14
conllu_dev_q11_2 ................................................ 15
conllu_dev_q11_2_nltk ........................................... 16
conllu_dev_q11_3 ................................................ 17
conllu_dev_q11_3_nltk ........................................... 18
dev_data .......................................................... 19
dev_data_f ........................................................ 19
dev_data_m ....................................................... 20
dev_data_na ...................................................... 21
fst_cn_compare_plot .............................................. 21
fst_cn_edges ....................................................... 23
fst_cn_get_unique ................................................ 24
fst_cn_nodes ....................................................... 24
fst_cn_plot ........................................................ 25
fst_cn_search ...................................................... 26
fst_comparison_cloud ............................................. 26
fst_concept_network .............................................. 28
fst_concept_network_compare .................................... 29
fst_find_stopwords ............................................... 30
fst_format_conllu ............................................... 31
fst_freq .......................................................... 31
fst_freq_compare ................................................ 32
fst_freq_plot ...................................................... 34
fst_get_top_ngrams ............................................... 35
**Description**

This data contains the responses to q7 "Kertoisitko, mitä sinun mielestäsi kiusaaminen on? (Avokysymys)" in the FSD3134 Lapsibarometri 2016 dataset.

**Usage**

`child_barometer_data`

**Format**

```r
# 'child_barometer_data' A dataframe with 414 rows and 2 columns:

**fsd_id** FSD case id

**q7** response text
```

**Source**

Description

This data contains the responses to q7 "Kertoisitko, mitä sinun mielestäsi kiusaaminen on? (Avokysymys)" in the FSD3134 Lapsibarometri 2016 dataset in CoNLL-U format using 'finnish-tdt' model from [udpipe] package.

Usage

conllu_cb_bullying

Format

```
## 'conllu_cb_bullying' A dataframe with 2722 rows and 14 columns:

doc_id  the identifier of the document
paragraph_id  the identifier of the paragraph
sentence_id  the identifier of the sentence
sentence  the text of the sentence for which this token is part of
token_id  Word index, integer starting at 1 for each new sentence; may be a range for multi-word tokens; may be a decimal number for empty nodes.
token  Word form or punctuation symbol.
lemma  Lemma or stem of word form.
upos  Universal part-of-speech tag.
xpos  Language-specific part-of-speech tag; underscore if not available.
feats  List of morphological features from the universal feature inventory or from a defined language-specific extension; underscore if not available.
head_token_id  Head of the current word, which is either a value of token_id or zero (0).
dep_rel  Universal dependency relation to the HEAD (root iff HEAD = 0) or a defined language-specific subtype of one.
deps  Enhanced dependency graph in the form of a list of head-deprel pairs.
misc  Any other annotation.
```

Source

**conllu_cb_bullying_iso**

*Child Barometer 2016 Bullying response data in CoNLL-U format with ISO stopwords removed*

---

**Description**

This data contains the responses to q7 "Kertoisitko, mitä sinun mielestäsi kiusaaminen on? (Avokysymys)" in the FSD3134 Lapsibarometri 2016 dataset in CoNLL-U format with ISO stopwords and punctuation removed.

**Usage**

`conllu_cb_bullying_iso`

**Format**

`# 'conllu_cb_bullying_iso' A dataframe with 1240 rows and 14 columns:

doc_id the identifier of the document
paragraph_id the identifier of the paragraph
sentence_id the identifier of the sentence
sentence the text of the sentence for which this token is part of
token_id Word index, integer starting at 1 for each new sentence; may be a range for multi-word tokens; may be a decimal number for empty nodes.
token Word form or punctuation symbol.
lemma Lemma or stem of word form.
upos Universal part-of-speech tag.
xpos Language-specific part-of-speech tag; underscore if not available.
feats List of morphological features from the universal feature inventory or from a defined language-specific extension; underscore if not available.
head_token_id Head of the current word, which is either a value of token_id or zero (0).
dep_rel Universal dependency relation to the HEAD (root iff HEAD = 0) or a defined language-specific subtype of one.
deps Enhanced dependency graph in the form of a list of head-deprel pairs.
misc Any other annotation.

**Source**

### `conllu_dev_q11_1` A dataframe with 6782 rows and 14 columns:

<table>
<thead>
<tr>
<th>doc_id</th>
<th>the identifier of the document</th>
</tr>
</thead>
<tbody>
<tr>
<td>paragraph_id</td>
<td>the identifier of the paragraph</td>
</tr>
<tr>
<td>sentence_id</td>
<td>the identifier of the sentence</td>
</tr>
<tr>
<td>sentence</td>
<td>the text of the sentence for which this token is part of</td>
</tr>
<tr>
<td>token_id</td>
<td>Word index, integer starting at 1 for each new sentence; may be a range for multi-word tokens; may be a decimal number for empty nodes.</td>
</tr>
<tr>
<td>token</td>
<td>Word form or punctuation symbol.</td>
</tr>
<tr>
<td>lemma</td>
<td>Lemma or stem of word form.</td>
</tr>
<tr>
<td>upos</td>
<td>Universal part-of-speech tag.</td>
</tr>
<tr>
<td>xpos</td>
<td>Language-specific part-of-speech tag; underscore if not available.</td>
</tr>
<tr>
<td>feats</td>
<td>List of morphological features from the universal feature inventory or from a defined language-specific extension; underscore if not available.</td>
</tr>
<tr>
<td>head_token_id</td>
<td>Head of the current word, which is either a value of token_id or zero (0).</td>
</tr>
<tr>
<td>dep_rel</td>
<td>Universal dependency relation to the HEAD (root iff HEAD = 0) or a defined language-specific subtype of one.</td>
</tr>
<tr>
<td>deps</td>
<td>Enhanced dependency graph in the form of a list of head-deprel pairs.</td>
</tr>
<tr>
<td>misc</td>
<td>Any other annotation.</td>
</tr>
</tbody>
</table>

Source

<https://urn.fi/urn:nbn:fi:fsd:T-FSD2821>
conllu_dev_q11_1_f

### Description

This data contains the female responses to q11.1 ‘Jatka lausetta: Kehitysmaa on maa, jossa... (Avokysymys)’ in CoNLL-U format using ‘finnish-ftb’ model from [udpipe] package.

### Usage

conllu_dev_q11_1_f

### Format

```plaintext
## 'conllu_dev_q11_1_f' A dataframe with 5251 rows and 14 columns:

**doc_id** the identifier of the document

**paragraph_id** the identifier of the paragraph

**sentence_id** the identifier of the sentence

**sentence** the text of the sentence for which this token is part of

**token_id** Word index, integer starting at 1 for each new sentence; may be a range for multi-word tokens; may be a decimal number for empty nodes.

**token** Word form or punctuation symbol.

**lemma** Lemma or stem of word form.

**upos** Universal part-of-speech tag.

**xpos** Language-specific part-of-speech tag; underscore if not available.

**feats** List of morphological features from the universal feature inventory or from a defined language-specific extension; underscore if not available.

**head_token_id** Head of the current word, which is either a value of token_id or zero (0).

**dep_rel** Universal dependency relation to the HEAD (root iff HEAD = 0) or a defined language-specific subtype of one.

**deps** Enhanced dependency graph in the form of a list of head-deprel pairs.

**misc** Any other annotation.
```

### Source

<https://urn.fi/urn:nbn:fi:fsd:T-FSD2821>
Young People’s Views on Development Cooperation 2012 Female q11_1 response data in ConLL-U format with NTLK stopwords removed

Description
This data contains the female responses to Development Cooperation q11_1 dataset in ConLL-U format with ISO stopwords and punctuation removed.

Usage
conllu_dev_q11_1_f_nltk

Format
```r
## 'conllu_dev_q11_1_f_nltk' A dataframe with 3268 rows and 14 columns:

  doc_id  the identifier of the document
  paragraph_id  the identifier of the paragraph
  sentence_id  the identifier of the sentence
  sentence  the text of the sentence for which this token is part of
  token_id  Word index, integer starting at 1 for each new sentence; may be a range for multi-word tokens; may be a decimal number for empty nodes.
  token  Word form or punctuation symbol.
  lemma  Lemma or stem of word form.
  upos  Universal part-of-speech tag.
  xpos  Language-specific part-of-speech tag; underscore if not available.
  feats  List of morphological features from the universal feature inventory or from a defined language-specific extension; underscore if not available.
  head_token_id  Head of the current word, which is either a value of token_id or zero (0).
  dep_rel  Universal dependency relation to the HEAD (root iff HEAD = 0) or a defined language-specific subtype of one.
  deps  Enhanced dependency graph in the form of a list of head-deprel pairs.
  misc  Any other annotation.
```

Source
<https://urn.fi/urn:nbn:fi:fsd:T-FSD2821>
conllu_dev_q11_1_m

conllu_dev_q11_1_m  Young People’s Views on Development Cooperation 2012 Male q11_1
response data in CoNLL-U format

Description
This data contains the male responses to q11_1 ‘Jatka lausetta: Kehitysmaa on maa, jossa... (Avokysymys)’

Usage
conllu_dev_q11_1_m

Format
## ‘conllu_dev_q11_1_m’ A dataframe with 1006 rows and 14 columns:

doc_id  the identifier of the document
paragraph_id  the identifier of the paragraph
sentence_id  the identifier of the sentence
sentence  the text of the sentence for which this token is part of
token_id  Word index, integer starting at 1 for each new sentence; may be a range for multi-word
tokens; may be a decimal number for empty nodes.
token  Word form or punctuation symbol.
lemma  Lemma or stem of word form.
upos  Universal part-of-speech tag.
xpos  Language-specific part-of-speech tag; underscore if not available.
feats  List of morphological features from the universal feature inventory or from a defined language-specific extension; underscore if not available.
head_token_id  Head of the current word, which is either a value of token_id or zero (0).
dep_rel  Universal dependency relation to the HEAD (root iff HEAD = 0) or a defined language-specific subtype of one.
deps  Enhanced dependency graph in the form of a list of head-deprel pairs.
misc  Any other annotation.

Source
<https://urn.fi/urn:nbn:fi:fsd:T-FSD2821>
Description

This data contains the male responses to Development Cooperation q11_1 dataset in CoNLL-U format with ISO stopwords and punctuation removed.

Usage

`conllu_dev_q11_1_m_nltk`

Format

```markdown
## `conllu_dev_q11_1_m_nltk` A dataframe with 651 rows and 14 columns:

<table>
<thead>
<tr>
<th>doc_id</th>
<th>the identifier of the document</th>
</tr>
</thead>
<tbody>
<tr>
<td>paragraph_id</td>
<td>the identifier of the paragraph</td>
</tr>
<tr>
<td>sentence_id</td>
<td>the identifier of the sentence</td>
</tr>
<tr>
<td>sentence</td>
<td>the text of the sentence for which this token is part of</td>
</tr>
<tr>
<td>token_id</td>
<td>Word index, integer starting at 1 for each new sentence; may be a range for multi-word tokens; may be a decimal number for empty nodes.</td>
</tr>
<tr>
<td>token</td>
<td>Word form or punctuation symbol.</td>
</tr>
<tr>
<td>lemma</td>
<td>Lemma or stem of word form.</td>
</tr>
<tr>
<td>upos</td>
<td>Universal part-of-speech tag.</td>
</tr>
<tr>
<td>xpos</td>
<td>Language-specific part-of-speech tag; underscore if not available.</td>
</tr>
<tr>
<td>feats</td>
<td>List of morphological features from the universal feature inventory or from a defined language-specific extension; underscore if not available.</td>
</tr>
<tr>
<td>head_token_id</td>
<td>Head of the current word, which is either a value of token_id or zero (0).</td>
</tr>
<tr>
<td>dep_rel</td>
<td>Universal dependency relation to the HEAD (root iff HEAD = 0) or a defined language-specific subtype of one.</td>
</tr>
<tr>
<td>deps</td>
<td>Enhanced dependency graph in the form of a list of head-deprel pairs.</td>
</tr>
<tr>
<td>misc</td>
<td>Any other annotation.</td>
</tr>
</tbody>
</table>
```

Source

<https://urn.fi/urn:nbn:fi:fsd:T-FSD2821>
Description

This data contains the gender not specified responses to q11_1 `Jatka lausetta: Kehitysmaa on maa, jossa... (Avokysymys)` in CoNLL-U format using `finnish-ftb` model from [udpipe] package.

Usage

conllu_dev_q11_1_na

Format

## `conllu_dev_q11_1_na` A dataframe with 525 rows and 14 columns:

- **doc_id**  the identifier of the document
- **paragraph_id** the identifier of the paragraph
- **sentence_id** the identifier of the sentence
- **sentence** the text of the sentence for which this token is part of
- **token_id** Word index, integer starting at 1 for each new sentence; may be a range for multi-word tokens; may be a decimal number for empty nodes.
- **token** Word form or punctuation symbol.
- **lemma** Lemma or stem of word form.
- **upos** Universal part-of-speech tag.
- **xpos** Language-specific part-of-speech tag; underscore if not available.
- **feats** List of morphological features from the universal feature inventory or from a defined language-specific extension; underscore if not available.
- **head_token_id** Head of the current word, which is either a value of token_id or zero (0).
- **dep_rel** Universal dependency relation to the HEAD (root iff HEAD = 0) or a defined language-specific subtype of one.
- **deps** Enhanced dependency graph in the form of a list of head-deprel pairs.
- **misc** Any other annotation.

Source

<https://urn.fi/urn:nbn:fi:fsd:T-FSD2821>
Description

This data contains the gender not specified responses to Development Cooperation q11_1 dataset in CoNLL-U format with ISO stopwords and punctuation removed.

Usage

conllu_dev_q11_1_na_nltk

Format

## `conllu_dev_q11_1_na_nltk` A dataframe with 338 rows and 14 columns:

- **doc_id** the identifier of the document
- **paragraph_id** the identifier of the paragraph
- **sentence_id** the identifier of the sentence
- **sentence** the text of the sentence for which this token is part of
- **token_id** Word index, integer starting at 1 for each new sentence; may be a range for multi-word tokens; may be a decimal number for empty nodes.
- **token** Word form or punctuation symbol.
- **lemma** Lemma or stem of word form.
- **upos** Universal part-of-speech tag.
- **xpos** Language-specific part-of-speech tag; underscore if not available.
- **feats** List of morphological features from the universal feature inventory or from a defined language-specific extension; underscore if not available.
- **head_token_id** Head of the current word, which is either a value of token_id or zero (0).
- **dep_rel** Universal dependency relation to the HEAD (root iff HEAD = 0) or a defined language-specific subtype of one.
- **deps** Enhanced dependency graph in the form of a list of head-deprel pairs.
- **misc** Any other annotation.

Source

<https://urn.fi/urn:nbn:fi:fsd:T-FSD2821>
Description

This data contains the responses to Development Cooperation q11_1 dataset in CoNLL-U format with ISO stopwords and punctuation removed.

Usage

conllu_dev_q11_1_nltk

Format

### 'conllu_dev_q11_1_nltk' A dataframe with 4257 rows and 14 columns:

**doc_id** the identifier of the document  
**paragraph_id** the identifier of the paragraph  
**sentence_id** the identifier of the sentence  
**sentence** the text of the sentence for which this token is part of  
**token_id** Word index, integer starting at 1 for each new sentence; may be a range for multi-word tokens; may be a decimal number for empty nodes.  
**token** Word form or punctuation symbol.  
**lemma** Lemma or stem of word form.  
**upos** Universal part-of-speech tag.  
**xpos** Language-specific part-of-speech tag; underscore if not available.  
**feats** List of morphological features from the universal feature inventory or from a defined language-specific extension; underscore if not available.  
**head_token_id** Head of the current word, which is either a value of token_id or zero (0).  
**dep_rel** Universal dependency relation to the HEAD (root iff HEAD = 0) or a defined language-specific subtype of one.  
**deps** Enhanced dependency graph in the form of a list of head-deprel pairs.  
**misc** Any other annotation.

Source

<https://urn.fi/urn:nbn:fi:fsd:T-FSD2821>
### Description

This data contains the responses to Development Cooperation q11_1 dataset in CoNLL-U format with ISO stopwords and punctuation removed.

### Usage

conllu_dev_q11_1_snow

### Format

```
`conllu_dev_q11_1_snow` A dataframe with 4259 rows and 14 columns:

- **doc_id** the identifier of the document
- **paragraph_id** the identifier of the paragraph
- **sentence_id** the identifier of the sentence
- **sentence** the text of the sentence for which this token is part of
- **token_id** Word index, integer starting at 1 for each new sentence; may be a range for multi-word tokens; may be a decimal number for empty nodes.
- **token** Word form or punctuation symbol.
- **lemma** Lemma or stem of word form.
- **upos** Universal part-of-speech tag.
- **xpos** Language-specific part-of-speech tag; underscore if not available.
- **feats** List of morphological features from the universal feature inventory or from a defined language-specific extension; underscore if not available.
- **head_token_id** Head of the current word, which is either a value of token_id or zero (0).
- **dep_rel** Universal dependency relation to the HEAD (root iff HEAD = 0) or a defined language-specific subtype of one.
- **deps** Enhanced dependency graph in the form of a list of head-deprel pairs.
- **misc** Any other annotation.
```

### Source

<https://urn.fi/urn:nbn:fi:fsd:T-FSD2821>
Description

This data contains the responses to q11_2 'Jatka lausetta: Kehitysyhteistyö on toimintaa, jossa... (Avokysymys)' in CoNLL-U format using 'finnish-ftb' model from [udpipe] package.

Usage

conllu_dev_q11_2

Format

```markdown
## 'conllu_dev_q11_2' A dataframe with 5495 rows and 14 columns:

doc_id the identifier of the document  
paragraph_id the identifier of the paragraph  
sentence_id the identifier of the sentence  
sentence the text of the sentence for which this token is part of  
token_id Word index, integer starting at 1 for each new sentence; may be a range for multi-word tokens; may be a decimal number for empty nodes.  
token Word form or punctuation symbol.  
lemma Lemma or stem of word form.  
upos Universal part-of-speech tag.  
xpos Language-specific part-of-speech tag; underscore if not available.  
_feats List of morphological features from the universal feature inventory or from a defined language-specific extension; underscore if not available.  
head_token_id Head of the current word, which is either a value of token_id or zero (0).  
dep_rel Universal dependency relation to the HEAD (root iff HEAD = 0) or a defined language-specific subtype of one.  
deps Enhanced dependency graph in the form of a list of head-deprel pairs.  
misc Any other annotation.
```

Source

<https://urn.fi/urn:nbn:fi:fsd:T-FSD2821>
conllu_dev_q11_2_nltk  Young People’s Views on Development Cooperation 2012 q11_2 response data in CoNLL-U format with NTLK stopwords removed

Description

This data contains the responses to Development Cooperation q11_2 dataset in CoNLL-U format with ISO stopwords and punctuation removed.

Usage

conllu_dev_q11_2_nltk

Format

```
## 'conllu_dev_q11_2_nltk' A dataframe with 4407 rows and 14 columns:

doc_id  the identifier of the document
paragraph_id  the identifier of the paragraph
sentence_id  the identifier of the sentence
sentence  the text of the sentence for which this token is part of
token_id  Word index, integer starting at 1 for each new sentence; may be a range for multi-word tokens; may be a decimal number for empty nodes.
token  Word form or punctuation symbol.
lemma  Lemma or stem of word form.
upos  Universal part-of-speech tag.
xpos  Language-specific part-of-speech tag; underscore if not available.
feats  List of morphological features from the universal feature inventory or from a defined language-specific extension; underscore if not available.
head_token_id  Head of the current word, which is either a value of token_id or zero (0).
dep_rel  Universal dependency relation to the HEAD (root iff HEAD = 0) or a defined language-specific subtype of one.
deps  Enhanced dependency graph in the form of a list of head-deprel pairs.
misc  Any other annotation.
```

Source

<https://urn.fi/urn:nbn:fi:fsd:T-FSD2821>
Description

This data contains the responses to 'q11_3' Jatka lausetta: Maailman kolme suurinta ongelmaa ovat... (Avokysymys)' in CoNLL-U format using ‘finnish-ftb’ model from [udpipe] package.

Usage

conllu_dev_q11_3

Format

## ‘conllu_dev_q11_3’ A dataframe with 6610 rows and 14 columns:

- **doc_id** the identifier of the document
- **paragraph_id** the identifier of the paragraph
- **sentence_id** the identifier of the sentence
- **sentence** the text of the sentence for which this token is part of
- **token_id** Word index, integer starting at 1 for each new sentence; may be a range for multi-word tokens; may be a decimal number for empty nodes.
- **token** Word form or punctuation symbol.
- **lemma** Lemma or stem of word form.
- **upos** Universal part-of-speech tag.
- **xpos** Language-specific part-of-speech tag; underscore if not available.
- **feats** List of morphological features from the universal feature inventory or from a defined language-specific extension; underscore if not available.
- **head_token_id** Head of the current word, which is either a value of token_id or zero (0).
- **dep_rel** Universal dependency relation to the HEAD (root iff HEAD = 0) or a defined language-specific subtype of one.
- **deps** Enhanced dependency graph in the form of a list of head-deprel pairs.
- **misc** Any other annotation.

Source

<https://urn.fi/urn:nbn:fi:fsd:T-FSD2821>
conllu_dev_q11_3_nltk  Young People’s Views on Development Cooperation 2012 q11_3 response data in CoNLL-U format with NTLK stopwords removed

Description
This data contains the responses to Development Cooperation q11_3 dataset in CoNLL-U format with ISO stopwords and punctuation removed.

Usage

conllu_dev_q11_3_nltk

Format

## 'conllu_dev_q11_3_nltk' A dataframe with 4192 rows and 14 columns:

doc_id  the identifier of the document
paragraph_id  the identifier of the paragraph
sentence_id  the identifier of the sentence
sentence  the text of the sentence for which this token is part of
token_id  Word index, integer starting at 1 for each new sentence; may be a range for multi-word tokens; may be a decimal number for empty nodes.
token  Word form or punctuation symbol.
lemma  Lemma or stem of word form.
upos  Universal part-of-speech tag.
xpos  Language-specific part-of-speech tag; underscore if not available.
feats  List of morphological features from the universal feature inventory or from a defined language-specific extension; underscore if not available.
head_token_id  Head of the current word, which is either a value of token_id or zero (0).
dep_reln  Universal dependency relation to the HEAD (root iff HEAD = 0) or a defined language-specific subtype of one.
deps  Enhanced dependency graph in the form of a list of head-deprel pairs.
misc  Any other annotation.

Source

<https://urn.fi/urn:nbn:fi:fsd:T-FSD2821>
Dev_data Young People’s Views on Development Cooperation 2012 response data

Description
This data contains the responses to q11_1 'Jatka lausetta: Kehitysmaa on maa, jossa... (Avokysymys)', q11_2 'Jatka lausetta: Kehitysyhteistyö on toimintaa, jossa... (Avokysymys)', q11_3' Jatka lausetta: Maailman kolme suurinta ongelma ovat... (Avokysymys)’ in the FSD2821 Nuorten ajatuksia kehitysyhteistyöstä 2012 dataset.

Usage
dev_data

Format
## 'dev_data' A dataframe with 925 rows and 4 columns:

fsd_id FSD case id
q11_1 response text for q11_1
q11_2 response text for q11_2
q11_3 response text for q11_3

Source
<https://urn.fi/urn:nbn:fi:fsd:T-FSD2821>

Dev_data_f Young People’s Views on Development Cooperation 2012 Female response data

Description
This data contains the female responses to q11_1 'Jatka lausetta: Kehitysmaa on maa, jossa... (Avokysymys)', q11_2 'Jatka lausetta: Kehitysyhteistyö on toimintaa, jossa... (Avokysymys)', q11_3' Jatka lausetta: Maailman kolme suurinta ongelma ovat... (Avokysymys)’ in the FSD2821 Nuorten ajatuksia kehitysyhteistyöstä 2012 dataset.

Usage
dev_data_f
Format

```r
dev_data_m
```

A dataframe with 673 rows and 4 columns:

- `fsd_id` FSD case id
- `q11_1` response text for q11_1
- `q11_2` response text for q11_2
- `q11_3` response text for q11_3

Source

<https://urn.fi/urn:nbn:fi:fsd:T-FSD2821>

---

Description

This data contains the male responses to q11_1 'Jatka lausetta: Kehitysmaa on maa, jossa... (Avokysymys)', q11_2 'Jatka lausetta: Kehitysyhteistyö on toimintaa, jossa... (Avokysymys)', q11_3' Jatka lausetta: Maailman kolme suurinta ongelma on ovat... (Avokysymys)' in the FSD2821 Nuorten ajatuksia kehitysyhteistyöstä 2012 dataset.

Usage

```r
dev_data_m
```

Format

```r
dev_data_m
```

A dataframe with 183 rows and 4 columns:

- `fsd_id` FSD case id
- `q11_1` response text for q11_1
- `q11_2` response text for q11_2
- `q11_3` response text for q11_3

Source

<https://urn.fi/urn:nbn:fi:fsd:T-FSD2821>
Description

This data contains the gender not specified responses to q11_1 'Jatka lausetta: Kehitysmaa on maa, jossa... (Avokysymys)', q11_2 'Jatka lausetta: Kehitysyhteistyö on toimintaa, jossa... (Avokysymys)', q11_3' Jatka lausetta: Maailman kolme suurinta ongelma ovat... (Avokysymys)' in the FSD2821 Nuorten ajatuksia kehitysyhteistyöstä 2012 dataset.

Usage

dev_data_na

Format

## 'dev_data_na' A dataframe with 89 rows and 4 columns:

- fsd_id  FSD case id
- q11_1  response text for q11_1
- q11_2  response text for q11_2
- q11_3  response text for q11_3

Source

<https://urn.fi/urn:nbn:fi:fsd:T-FSD2821>

Description

Creates a Concept Network plot from a list of edges and nodes (and their respective weights) which indicates unique words in this plot in comparison to another Network.

Usage

fst_cn_compare_plot(
edges,
nodes,
concepts,
unique_lemmas,
name = NULL,
concept_colour = "#cd1719",
unique_colour = ":#4DAF4A",
min_edge = NULL,
max_edge = NULL,
min_node = NULL,
max_node = NULL
)

Arguments

edges Output of ‘fst_cn_edges()’, dataframe of ‘edges’ connecting two words.

nodes Output of ‘fst_cn_nodes()’, dataframe of relevant lemmas and their associated pagerank.

concepts List of terms which have been searched for, separated by commas.

unique_lemmas List of unique lemmas, output of ‘fst_cn_get_unique()’

name An optional “name” for the plot, default is ‘NULL’ and a generic title (“Textrank extracted keyword occurrences”) will be used.

color

unique_colour Colour to display concept words, default is “indianred”.

unique_colour Colour to display unique words, default is “darkgreen”.

min_edge A numeric value for the scale of the edges, the smallest co_occurrence value for an edge across all Networks to be plotted together.

max_edge A numeric value for the scale of the edges, the largest co_occurrence value for an edge across all Networks to be plotted together.

min_node A numeric value for the scale of the nodes, the smallest pagerank value for a node across all Networks to be plotted together.

max_node A numeric value for the scale of the nodes, the largest pagerank value for a node across all Networks to be plotted together.

Value

Plot of concept network with concept and unique words (nodes) highlighted.

Examples

cb <- conllu_ch_bullying_iso
pos_filter = c("NOUN", "VERB", "ADJ", "ADV")
e1 <- fst_cn_edges(cb, "lyödä", pos_filter = pos_filter)
e2 <- fst_cn_edges(cb, "lyöminen", pos_filter = pos_filter)
n1 <- fst_cn_nodes(cb, e1)
n2 <- fst_cn_nodes(cb, e2)
u <- fst_cn_get_unique(n1, n2)

fst_cn_compare_plot(e1, n1, "lyödä", unique_lemma = u)
fst_cn_compare_plot(e2, n2, "lyöminen", u, unique_colour = "purple")
**fst_cn_edges**

**Concept Network - Get textrank edges**

**Description**

This function takes a string of terms (separated by commas) or a single term and, using `fst_cn_search()` find words connected to these searched terms. Then, a dataframe is returned of edges between two words which are connected together in an frequently-occurring n-gram containing a concept term.

**Usage**

```
fst_cn_edges(
  data, 
  concepts, 
  threshold = NULL, 
  norm = "number_words", 
  pos_filter = NULL 
)
```

**Arguments**

- **data**  
  A dataframe of text in CoNLL-U format.

- **concepts**  
  List of terms to search for, separated by commas.

- **threshold**  
  A minimum number of occurrences threshold for 'edge' between searched term and other word, default is 'NULL'. Note, the threshold is applied before normalisation.

- **norm**  
  The method for normalising the data. Valid settings are "number_words" (the number of words in the responses, default), "number_resp": (the number of responses), or 'NULL' (raw count returned).

- **pos_filter**  
  List of UPOS tags for inclusion, default is 'NULL' to include all UPOS tags.

**Value**

Dataframe of co-occurrences between two connected words.

**Examples**

```
con <- "kiusata, lyöminen"
cb <- conllu_cb_bullying_iso
fst_cn_edges(cb, con, pos_filter = c("NOUN", "VERB", "ADJ", "ADV"))
fst_cn_edges(cb, "lyöminen", threshold = 2, norm = 'number_resp')
```
fst_cn_get_unique  

**Concept Network - Get Unique Nodes**

**Description**

Takes at least two tables of nodes and pagerank (output of ‘fst_cn_nodes()’) and finds nodes unique to one table.

**Usage**

```
fst_cn_get_unique(table1, table2, ...)
```

**Arguments**

- **table1**  
  The first table.
- **table2**  
  The second table.
- **...**  
  Any other tables you want to include.

**Value**

Dataframe of words and whether word is unique or not.

**Examples**

```r
cb <- conllu_cb_bullying_iso
pos_filter = c("NOUN", "VERB", "ADJ", "ADV")
e1 <- fst_cn_edges(cb, "lyödä", pos_filter = pos_filter)
e2 <- fst_cn_edges(cb, "lyöminen", pos_filter = pos_filter)
n1 <- fst_cn_nodes(cb, e1)
n2 <- fst_cn_nodes(cb, e2)
fst_cn_get_unique(n1, n2)
```

fst_cn_nodes  

**Concept Network - Get textrank nodes**

**Description**

This function takes a string of terms (separated by commas) or a single term and, using ‘textrank_keywords()’ from ‘textrank’ package, filters data based on ‘pos_filter’ ranks words which are the filtered for those connected to search terms.

**Usage**

```
fst_cn_nodes(data, edges, pos_filter = NULL)
```
**fist_cn_plot**

**Arguments**
- **data**: A dataframe of text in CoNLL-U format.
- **edges**: Output of `fist_cn_edges()`, dataframe of co-occurrences between two words.
- **pos_filter**: List of UPOS tags for inclusion, default is 'NULL' to include all UPOS tags.

**Value**
A dataframe containing relevant lemmas and their associated pagerank.

**Examples**
```r
con <- "kiusata, lyöminen"
cb <- conllu_cb_bullying_iso
dges <- fist_cn_edges(cb, con, pos_filter = c("NOUN", "VERB", "ADJ", "ADV"))
nodes <- fist_cn_nodes(cb, edges, c("NOUN", "VERB", "ADJ", "ADV"))
fist_cn_plot(edges = edges, nodes = nodes, concepts = con)
```

---

**Description**
Creates a Concept Network plot from a list of edges and nodes (and their respective weights).

**Usage**
`fist_cn_plot(edges, nodes, concepts, title = NULL)`

**Arguments**
- **edges**: Output of `fist_cn_edges()`, dataframe of 'edges' connecting two words.
- **nodes**: Output of `fist_cn_nodes()`, dataframe of relevant lemmas and their associated pagerank.
- **concepts**: List of terms which have been searched for, separated by commas.
- **title**: Optional title for plot, default is 'NULL' and a generic title ("Textrank extracted keyword occurrences") will be used.

**Value**
Plot of Concept Network.

**Examples**
```r
con <- "kiusata, lyöminen"
cb <- conllu_cb_bullying_iso
dges <- fist_cn_edges(cb, con, pos_filter = c("NOUN", "VERB", "ADJ", "ADV"))
nodes <- fist_cn_nodes(cb, edges, c("NOUN", "VERB", "ADJ", "ADV"))
fist_cn_plot(edges = edges, nodes = nodes, concepts = con)
```
**fst_cn_search**  
*Concept Network - Search textrank for concepts*

**Description**

This function takes a string of terms (separated by commas) or a single term and, using `textrank_keywords()` from `textrank` package, filters data based on `pos_filter` and finds words connected to search terms.

**Usage**

```r
define_usage(fst_cn_search)```

```r
fst_cn_search(data, concepts, pos_filter = NULL)
```

**Arguments**

- `data`  
  A dataframe of text in CoNLL-U format.
- `concepts`  
  String of terms to search for, separated by commas.
- `pos_filter`  
  List of UPOS tags for inclusion, default is `NULL` to include all UPOS tags.

**Value**

Dataframe of n-grams containing searched terms.

**Examples**

```r
c <- c("kiusata, lyöminen, lyödä, potkia")
pf <- c("NOUN", "VERB", "ADJ", "ADV")

data1 <- conllu_cb_bullying_iso

data2 <- conllu_cb_bullying_iso

data3 <- NULL

data4 <- NULL

fst_cn_search(data1, concepts = c, pos_filter = pf)
```

**fst_comparison_cloud**  
*Make comparison cloud*

**Description**

Creates a comparison wordcloud showing words that occur differently between each group.

**Usage**

```r
define_usage(fst_comparison_cloud)```

```r
fst_comparison_cloud(data1, data2, data3 = NULL, data4 = NULL, name1 = "Group 1", name2 = "Group 2")
```
Arguments

data1 A dataframe of text in CoNLL-U format for the first group.
data2 A dataframe of text in CoNLL-U format for the second group.
data3 An optional dataframe of text in CoNLL-U format for the third group, default is ‘NULL’.
data4 An optional dataframe of text in CoNLL-U format for the fourth group, default is ‘NULL’.
name1 A string describing data1, default is ‘Group 1’.
name2 A string describing data2, default is ‘Group 2’.
name3 A string describing data3, default is ‘Group 3’.
name4 A string describing data4, default is ‘Group 4’.
pos_filter List of UPOS tags for inclusion, default is ‘NULL’ which means all word types included.
max The maximum number of words to display, default is ‘100’.

Value

A comparison cloud from wordcloud package.

Examples

d1 <- conllu_dev_q11_1_nltk
d2 <- conllu_dev_q11_3_nltk
pf1 <- c("NOUN", "VERB", "ADJ", "ADV")
fst_comparison_cloud(d1, d2, pos_filter = pf1)

f <- conllu_dev_q11_1_f_nltk
m <- conllu_dev_q11_1_m_nltk
na <- conllu_dev_q11_1_na_nltk
n1 <- "Female"
n2 <- "Male"
n3 <- "NA"
fst_comparison_cloud(f, m, na, name1 = n1, name2 = n2, name3 = n3, max = 400)
fst_comparison_cloud(f, m, na, name1 = n1, name2 = n2, name3 = n3, max = 100)
Description

This function takes a string of terms (separated by commas) or a single term and, using ‘textrank_keywords()’ from ‘textrank’ package, filters data based on ‘pos_filter’ and finds words connected to search terms. Then it plots a Concept Network based on the calculated weights of these terms and the frequency of co-occurrences.

Usage

```r
fst_concept_network(
  data,
  concepts,
  threshold = NULL,
  norm = "number_words",
  pos_filter = NULL,
  title = NULL
)
```

Arguments

- `data` A dataframe of text in CoNLL-U format.
- `concepts` List of terms to search for, separated by commas.
- `threshold` A minimum number of occurrences threshold for ‘edge’ between searched term and other word, default is ‘NULL’. Note, the threshold is applied before normalisation.
- `norm` The method for normalising the data. Valid settings are “"number_words"” (the number of words in the responses, default), ""number_resp"" (the number of responses, or ‘NULL’ (raw count returned).
- `pos_filter` List of UPOS tags for inclusion, default is ‘NULL’ to include all UPOS tags.
- `title` Optional title for plot, default is ‘NULL’ and a generic title (“Textrank extracted keyword occurrences”) will be used.

Value

Plot of Concept Network.

Examples

```r
data <- conllu_cb_bullying_iso
c <- c("kiusata, lyöminen")
pf <- c("NOUN", "VERB", "ADJ", "ADV")
title <- "Bullying Concept Network"
fst_concept_network(data, concepts = c, pos_filter = pf, title = title)
```
**fst_concept_network_compare**

*Concept Network- Compare and plot Concept Network*

**Description**

This function takes a string of terms (separated by commas) or a single term and, using `textrank_keywords()` from `textrank` package, filters data based on `pos_filter` and finds words connected to search terms for each group. Then it plots a Concept Network for each group based on the calculated weights of these terms and the frequency of co-occurrences, indicating any words that are unique to each group’s Network plot.

**Usage**

```r
fst_concept_network_compare(
  data1, data2, data3 = NULL, data4 = NULL, pos_filter = NULL, name1 = "Group 1", name2 = "Group 2", name3 = "Group 3", name4 = "Group 4", concepts, norm = "number_words", threshold = NULL)
```

**Arguments**

- `data1`: A dataframe of text in CoNLL-U format for the first concept network.
- `data2`: A dataframe of text in CoNLL-U format for the second concept network.
- `data3`: An optional dataframe of text in CoNLL-U format for the third concept network, default is ‘NULL’.
- `data4`: An optional dataframe of text in CoNLL-U format for the fourth concept network, default is ‘NULL’.
- `pos_filter`: List of UPOS tags for inclusion, default is ‘NULL’ which means all word types included.
- `name1`: A string describing data1, default is “Group 1”.
- `name2`: A string describing data2, default is “Group 2”.
- `name3`: A string describing data3, default is “Group 3”.
- `name4`: A string describing data4, default is “Group 4”.
- `concepts`: List of terms to search for, separated by commas.
**norm**

The method for normalising the data. Valid settings are "number_words" (the number of words in the responses, default), "number_resp" (the number of responses), or "NULL" (raw count returned).

**threshold**

A minimum number of occurrences threshold for 'edge' between searched term and other word, default is 'NULL'.

**Value**

Between 2 and 4 concept network plots with concept and unique words highlighted.

**Examples**

```r
d1 <- conllu_ch_bullying
d2 <- conllu_ch_bullying_iso
con1 <- "lyödä, lyöminen"
fst_concept_network_compare(d1, d2, concepts = con1)
```

---

**fst_find_stopwords**

*Get available Finnish stopwords lists*

**Description**

Returns a tibble containing available Finnish stopword lists, their contents, and the size of the lists.

**Usage**

```r
fst_find_stopwords()
```

**Value**

A tibble containing the stopwords lists.

**Examples**

```r
fst_find_stopwords()
```
Annotate open-ended survey responses in Finnish into CoNLL-U format

Description

Creates a dataframe in CoNLL-U format from a list of strings of Finnish text using the [udpipe] package and a Finnish language model.

Usage

```r
fst_format_conllu(data, field, model = "ftb")
```

Arguments

- `data` A dataframe of survey responses which contains an open-ended question.
- `field` The field in the dataframe which contains the open-ended question.
- `model` A Finnish language model available for [udpipe], "ftb" (default) or "tdt".

Value

Dataframe of annotated text in CoNLL-U format.

Examples

```r
define_data <- function() {
  child_barometer_data <- data.frame(
    q7 = c("This is a survey response.\n    Other responses as well.\n    This one is too.
    "
  )
}
df <- define_data()
fst_format_conllu(data = df, field = "q7")
fst_format_conllu(data = df, field = "q7", model = "tdt")
unlink("finnish-ftb-ud-2.5-191206.udpipe")
unlink("finnish-tdt-ud-2.5-191206.udpipe")
```

Find and Plot Top Words

Description

Creates a plot of the most frequently-occurring words (unigrams) within the data.
Usage

```r
fst_freq(
  data,
  number = 10,
  norm = "number_words",
  pos_filter = NULL,
  strict = TRUE,
  name = NULL
)
```

Arguments

data A dataframe of text in CoNLL-U format.
number The number of top words to return, default is '10'.
norm The method for normalising the data. Valid settings are "number_words" (the number of words in the responses, default), "number Resp" (the number of responses), or 'NULL' (raw count returned).
pos_filter List of UPOS tags for inclusion, default is 'NULL' which means all word types included.
strict Whether to strictly cut-off at 'number' (ties are alphabetically ordered), default is 'TRUE'.
name An optional "name" for the plot to add to title, default is 'NULL'.

Value

Plot of top words.

Examples

```r
q11_1 <- conllu_dev_q11_1
n1 <- "number_resp"
fst_freq(q11_1, number = 12, norm = n1, strict = FALSE, name = "All")
fst_freq(q11_1, number = 15, name = "Not Spec")
```

Description

Find top and unique top words for between 2 and 4 sets of prepared data. Results will be shown within the plots pane. If 2 or 3 plots, they will be in a single row, if there are 4 plots, they will be in 2 rows of 2.
Usage

```r
def(FST_frequency_compare)

fst_freq_compare(
  data1,
  data2,
  data3 = NULL,
  data4 = NULL,
  number = 10,
  norm = "number_words",
  pos_filter = NULL,
  name1 = "Group 1",
  name2 = "Group 2",
  name3 = "Group 3",
  name4 = "Group 4",
  unique_colour = "indianred",
  strict = TRUE
)
```

Arguments

data1 A dataframe of text in CoNLL-U format for the first plot.
data2 A dataframe of text in CoNLL-U format for the second plot.
data3 An optional dataframe of text in CoNLL-U format for the third plot, default is ‘NULL’.
data4 An optional dataframe of text in CoNLL-U format for the fourth plot, default is ‘NULL’.
number The number of top words to return, default is ‘10’.
norm The method for normalising the data. Valid settings are “number_words” (the number of words in the responses, default), “number_resps” (the number of responses), or ‘NULL’ (raw count returned).
pos_filter List of UPOS tags for inclusion, default is ‘NULL’ which means all word types included.
name1 An optional "name" for the first plot, default is “Group 1”.
name2 An optional "name" for the second plot, default is “Group 2”.
name3 An optional "name" for the third plot, default is “Group 3”.
name4 An optional "name" for the fourth plot, default is “Group 4”.
unique_colour Colour to display unique words, default is “indianred”.
strict Whether to strictly cut-off at ‘number’ (ties are alphabetically ordered), default is ‘TRUE’.

Value

Between 2 and 4 plots of Top n-grams in the plots pane with unique n-grams highlighted.
Examples

```r
f <- conllu_dev_q11_1_f_nltk
m <- conllu_dev_q11_1_m_nltk
na <- conllu_dev_q11_1_na_nltk
fst_freq_compare(f, m, number = 10)
fst_freq_compare(f, m, na, number = 5, norm = "number_resp")
fst_freq_compare(f, m, na, name1 = "F", name2 = "M", name3 = "NA")
fst_freq_compare(f, m, na, strict = FALSE)
```

---

### fst_freq_plot

**Description**

Plots most common words.

**Usage**

```r
fst_freq_plot(table, number = NULL, name = NULL)
```

**Arguments**

- **table**: Output of ‘fst_get_top_words()’ or ‘fst_get_top_ngrams()’.
- **number**: Optional number of n-grams for the title, default is ‘NULL’.
- **name**: An optional "name" for the plot to add to title, default is ‘NULL’.

**Value**

Plot of top words.

**Examples**

```r
cb <- conllu.cb.bullying
pf <- c("NOUN", "VERB", "ADJ", "ADV")
top_bullying_words <- fst_get_top_words(cb, number = 15, pos_filter = pf)
fst_freq_plot(top_bullying_words, number = 5, name = "Bullying")

q11_1 <- conllu.dev.q11.1.nlkt
q11_1_ngrams <- fst_get_top_ngrams(q11_1, number = 10, ngrams = 1)
fst_freq_plot(q11_1_ngrams)
```
fst_get_top_ngrams

Make Top N-grams Table

**Description**

Creates a table of the most frequently-occurring n-grams within the data.

**Usage**

```r
defrost_get_top_ngrams(
  data,
  number = 10,
  ngrams = 1,
  norm = "number_words",
  pos_filter = NULL,
  strict = TRUE
)
```

**Arguments**

- **data**: A dataframe of text in CoNLL-U format.
- **number**: The number of n-grams to return, default is ‘10’.
- **ngrams**: The type of n-grams to return, default is ‘1’.
- **norm**: The method for normalising the data. Valid settings are “number_words” (the number of words in the responses, default), “number_resp” (the number of responses), or ‘NULL’ (raw count returned).
- **pos_filter**: List of UPOS tags for inclusion, default is ‘NULL’ which means all word types included.
- **strict**: Whether to strictly cut-off at ‘number’ (ties are alphabetically ordered), default is ‘TRUE’.

**Value**

A table of the most frequently occurring n-grams in the data.

**Examples**

```r
q11_1 <- conllu_dev_q11_1_nltk
defrost_get_top_ngrams(q11_1, norm = NULL)
defrost_get_top_ngrams(q11_1, number = 10, ngrams = 1, norm = "number_resp")
```
fst_get_top_ngrams2  

*Make Top N-grams Table 2*

**Description**

Creates a table of the most frequently-occurring ngrams within the data. Equivalent to ‘fst_get_top_ngrams()’ but does not print message.

**Usage**

```r
fst_get_top_ngrams2(
  data,
  number = 10,
  ngrams = 1,
  norm = "number_words",
  pos_filter = NULL,
  strict = TRUE
)
```

**Arguments**

- **data**: A dataframe of text in CoNLL-U format.
- **number**: The number of n-grams to return, default is ‘10’.
- **ngrams**: The type of n-grams to return, default is ‘1’.
- **norm**: The method for normalising the data. Valid settings are ‘"number_words"’ (the number of words in the responses, default), ‘"number_resp"’ (the number of responses), or ‘NULL’ (raw count returned).
- **pos_filter**: List of UPOS tags for inclusion, default is ‘NULL’ which means all word types included.
- **strict**: Whether to strictly cut-off at ‘number’ (ties are alphabetically ordered), default is ‘TRUE’.

**Value**

A table of the most frequently occurring n-grams in the data.

**Examples**

```r
fst_get_top_ngrams2(conllu_dev_q11_l_nltk)
fst_get_top_ngrams2(conllu_dev_q11_l_nltk, number = 10, ngrams = 1)
```
**fst_get_top_words**

Make Top Words Table

**Description**

Creates a table of the most frequently-occurring words (unigrams) within the data.

**Usage**

```r
fst_get_top_words(
  data,
  number = 10,
  norm = "number_words",
  pos_filter = NULL,
  strict = TRUE
)
```

**Arguments**

- **data**: A dataframe of text in CoNLL-U format.
- **number**: The number of top words to return, default is ‘10’.
- **norm**: The method for normalising the data. Valid settings are “"number_words"” (the number of words in the responses, default), “"number_resp"” (the number of responses), or ‘NULL’ (raw count returned).
- **pos_filter**: List of UPOS tags for inclusion, default is ‘NULL’ which means all word types included.
- **strict**: Whether to strictly cut-off at ‘number’ (ties are alphabetically ordered), default is ‘TRUE’.

**Value**

A table of the most frequently occurring words in the data.

**Examples**

```r
fst_get_top_words(conllu_dev_q11_1_nltk, number = 15, strict = FALSE)
cb <- conllu_cb_bullying
pf <- c("NOUN", "VERB", "ADJ", "ADV")
fst_get_top_words(cb, number = 5, norm = "number_resp", pos_filter = pf)
```
**fst_get_unique_ngrams**  
*Get unique n-grams*

### Description

Takes at least two tables of n-grams and frequencies (either output of ‘fst_get_top_words()’ or ‘fst_get_top_ngrams()’) and finds n-grams unique to one table.

### Usage

```r
goodnessistic::fst_get_unique_ngrams(table1, table2, ...)
```

### Arguments

- **table1**: The first table.
- **table2**: The second table.
- **...**: Any other tables you want to include.

### Value

Dataframe of words and whether word is unique or not.

### Examples

```r
top_f <- goodnessistic::fst_get_top_words(conllu_dev_q11_l_f_nltk)
top_m <- goodnessistic::fst_get_top_words(conllu_dev_q11_l_m_nltk)
top_na <- goodnessistic::fst_get_top_words(conllu_dev_q11_l_na_nltk)
topn_f <- goodnessistic::fst_get_top_ngrams(conllu_dev_q11_l_f_nltk)
topn_m <- goodnessistic::fst_get_top_ngrams(conllu_dev_q11_l_m_nltk)
topn_na <- goodnessistic::fst_get_top_ngrams(conllu_dev_q11_l_na_nltk)
goodnessistic::fst_get_unique_ngrams(top_f, top_m, top_na)
goodnessistic::fst_get_unique_ngrams(topn_f, topn_m, topn_n)
```

---

**fst_join_unique**  
*Merge N-grams table with unique words*

### Description

Merges list of unique words from ‘fst_get_unique_ngrams()’ with output of ‘fst_get_top_ngrams()’ or ‘fst_get_top_words()’ so that unique words can be displayed on comparison plots.

### Usage

```r
goodnessistic::fst_join_unique(table, unique_table)
```
Arguments

<table>
<thead>
<tr>
<th>table</th>
<th>Output of <code>fst_get_top_words()</code> or <code>fst_get_top_ngrams()</code>.</th>
</tr>
</thead>
<tbody>
<tr>
<td>unique_table</td>
<td>Output of <code>fst_get_unique_ngrams()</code>.</td>
</tr>
</tbody>
</table>

Value

A table of top n-grams, frequency, and whether the n-gram is "unique".

Examples

```r
top_f <- fst_get_top_words(conllu_dev_q11_1_f_nltk)
top_m <- fst_get_top_words(conllu_dev_q11_1_m_nltk)
top_na <- fst_get_top_words(conllu_dev_q11_1_na_nltk)
topn_f <- fst_get_top_ngrams(conllu_dev_q11_1_f_nltk)
topn_m <- fst_get_top_ngrams(conllu_dev_q11_1_m_nltk)
topn_na <- fst_get_top_ngrams(conllu_dev_q11_1_na_nltk)
unique_words <- fst_get_unique_ngrams(top_f, top_m, top_na)
unique_ngrams <- fst_get_unique_ngrams(topn_f, topn_m, topn_na)
fst_join_unique(top_f, unique_words)
fst_join_unique(topn_m, unique_ngrams)
```

Description

Compare length of text responses for between 2 and 4 sets of prepared data.

Usage

```r
fst_length_compare(  
  data1,  
  data2,  
  data3 = NULL,  
  data4 = NULL,  
  name1 = "Group 1",  
  name2 = "Group 2",  
  name3 = "Group 3",  
  name4 = "Group 4",  
  incl_sentences = TRUE)  
)
```

Arguments

<table>
<thead>
<tr>
<th>data1</th>
<th>A dataframe of text in CoNLL-U format for the first group.</th>
</tr>
</thead>
<tbody>
<tr>
<td>data2</td>
<td>A dataframe of text in CoNLL-U format for the second group.</td>
</tr>
</tbody>
</table>
data3  An optional dataframe of text in CoNLL-U format for the third group, default is ‘NULL’.
data4  An optional dataframe of text in CoNLL-U format for the fourth group, default is ‘NULL’.
name1  A string describing data1, default is “Group 1”.
name2  A string describing data2, default is “Group 2”.
name3  A string describing data3, default is “Group 3”.
name4  A string describing data4, default is “Group 4”.
incl_sentences  Whether to include sentence data in table, default is ‘TRUE’.

Value

Dataframe summarising response lengths.

Examples

```r
f <- conllu_dev_q11_1_f_nltk
m <- conllu_dev_q11_1_m_nltk
na <- conllu_dev_q11_1_na_nltk
all <- conllu_dev_q11_1_nltk
fst_length_compare(f, m, na, all, "Female", "Male", "Not Spec", "All")
fst_length_compare(f, m, name1 = "F", name2 = "M", incl_sentences = FALSE)
```

fst_length_summary  Make Length Summary Table

Description

Create a table summarising distribution of the length of responses.

Usage

```r
fst_length_summary(data, desc = "All respondents", incl_sentences = TRUE)
```

Arguments

data  dataframe of text in CoNLL-U format.
desc  An optional string describing respondents, default is “All respondents”.
incl_sentences  Whether to include sentence data in table, default is ‘TRUE’.

Value

Table summarising distribution of lengths of responses.

Examples

```r
fst_length_summary(conllu_dev_q11_1, incl_sentences = FALSE)
fst_length_summary(conllu_dev_q11_1, desc = "Female")
```
**fst_ngrams**

**Find and Plot Top N-grams**

**Description**

Creates a plot of the most frequently-occurring n-grams within the data.

**Usage**

```r
fst_ngrams(
  data,  
  number = 10,  
  ngrams = 1,  
  norm = "number_words",  
  pos_filter = NULL,  
  strict = TRUE,  
  name = NULL
)
```

**Arguments**

- `data`: A dataframe of text in CoNLL-U format.
- `number`: The number of top words to return, default is ‘10’.
- `ngrams`: The type of n-grams, default is ‘1’.
- `norm`: The method for normalising the data. Valid settings are ‘"number_words"’ (the number of words in the responses, default), ‘"number_resp"’ (the number of responses), or ‘NULL’ (raw count returned).
- `pos_filter`: List of UPOS tags for inclusion, default is ‘NULL’ which means all word types included.
- `strict`: Whether to strictly cut-off at ‘number’ (ties are alphabetically ordered), default is ‘TRUE’.
- `name`: An optional "name" for the plot to add to title, default is ‘NULL’.

**Value**

Plot of top n-grams

**Examples**

```r
q11_1 <- conllu_dev_q11_1
fst_ngrams(q11_1, 12, ngrams = 2, norm = NULL, strict = FALSE, name = "All")
fst_ngrams(conllu_dev_q11_1_na, number = 15, ngrams = 3, name = "Not Spec")
```
**fst_ngrams_compare**  
*Compare and plot top n-grams*

**Description**

Find top and unique top n-grams for between 2 and 4 sets of prepared data. Results will be shown within the plots pane. If 2 or 3 plots, they will be in a single row, if there are 4 plots, they will be in 2 rows of 2.

**Usage**

```r
fst_ngrams_compare(
  data1,
  data2,
  data3 = NULL,
  data4 = NULL,
  number = 10,
  ngrams = 1,
  norm = "number_words",
  pos_filter = NULL,
  name1 = "Group 1",
  name2 = "Group 2",
  name3 = "Group 3",
  name4 = "Group 4",
  unique_colour = "indianred",
  strict = TRUE
)
```

**Arguments**

- `data1`: A dataframe of text in CoNLL-U format for the first plot.
- `data2`: A dataframe of text in CoNLL-U format for the second plot.
- `data3`: An optional dataframe of text in CoNLL-U format for the third plot, default is 'NULL'.
- `data4`: An optional dataframe of text in CoNLL-U format for the fourth plot, default is 'NULL'.
- `number`: The number of n-grams to return, default is '10'.
- `ngrams`: The type of n-grams to return, default is '1'.
- `norm`: The method for normalising the data. Valid settings are "number_words" (the number of words in the responses, default), "number_resp" (the number of responses), or 'NULL' (raw count returned).
- `pos_filter`: List of UPOS tags for inclusion, default is 'NULL' which means all word types included.
- `name1`: An optional "name" for the first plot, default is "Group 1".
name2  An optional "name" for the second plot, default is “Group 2”.
name3  An optional "name" for the third plot, default is “Group 3”.
name4  An optional "name" for the fourth plot, default is “Group 4”.
unique_colour  Colour to display unique words, default is “indianred”.
strict  Whether to strictly cut-off at ‘number’ (ties are alphabetically ordered), default is ‘TRUE’.

Value

Between 2 and 4 plots of Top n-grams in the plots pane with unique n-grams highlighted.

Examples

```r
f <- conllu_dev_q11_1_f_nltk
m <- conllu_dev_q11_1_m_nltk
na <- conllu_dev_q11_1_na_nltk
all <- conllu_dev_q11_1_nltk

fst_ngrams_compare(f, m, na, all, number = 10, strict = FALSE)
fst_ngrams_compare(f, m, ngrams = 2, number = 10, norm = "number_resp")
fst_ngrams_compare(f, m, ngrams = 2, number = 10, strict = FALSE)
fst_ngrams_compare(f, m, number = 5, ngrams = 3, name1 = "M", name2 = "F")
fst_ngrams_compare(f, m, na, number = 20, unique_colour = "slateblue", )
```

fst_ngrams_compare_plot

Plot comparison n-grams

Description

Plots frequency n-grams with unique n-grams highlighted.

Usage

```r
fst_ngrams_compare_plot(
  table,
  number = 10,
  ngrams = 1,
  unique_colour = "indianred",
  name = NULL,
  override_title = NULL
)
```
fst_ngrams_plot

Make N-grams plot

Description

Plots frequency n-grams.

Usage

fst_ngrams_plot(table, number = NULL, ngrams = 1, name = NULL)

Arguments

table Output of ‘fst_get_top_words()’ or ‘fst_get_top_ngrams()’.
number Optional number of n-grams for title, default is ‘NULL’.
ngrams The type of n-grams, default is ‘1’.
name An optional "name" for the plot to add to title, default is ‘NULL’.

Examples

top_f <- fst_get_top_words(conllu_dev_q11_1_f_nltk)
top_m <- fst_get_top_words(conllu_dev_q11_1_m_nltk)
top_na <- fst_get_top_words(conllu_dev_q11_1_na_nltk)
topn_f <- fst_get_top_ngrams(conllu_dev_q11_1_f_nltk)
topn_m <- fst_get_top_ngrams(conllu_dev_q11_1_m_nltk)
topn_na <- fst_get_top_ngrams(conllu_dev_q11_1_na_nltk)
unique_words <- fst_get_unique_ngrams(top_f, top_m, top_na)
unique_ngrams <- fst_get_unique_ngrams(topn_f, topn_m, topn_na)
top_fu <- fst_join_unique(top_f, unique_words)
topn_mu <- fst_join_unique(topn_m, unique_ngrams)
fst_ngrams_compare_plot(top_fu, ngrams = 1, name = "Female")
fst_ngrams_compare_plot(topn_mu, ngrams = 2, name = "Male")
Value

Plot of top n-grams.

Examples

topn_f <- fst_get_top_ngrams(conllu_dev_q11_1_f_nltk)
topn_m <- fst_get_top_ngrams(conllu_dev_q11_1_m_nltk)
topn_na <- fst_get_top_ngrams(conllu_dev_q11_1_na_nltk)
fst_ngrams_plot(topn_f, ngrams = 2, name = "Female")
fst_ngrams_plot(topn_f, ngrams = 1, number = 15)
fst_ngrams_plot(topn_m, ngrams = 2, number = 15)
fst_ngrams_plot(topn_na, ngrams = 2)

fst_plot_multiple

Display comparison plots

Description

Display between 2 and 4 plots within the plots pane. If 2 or 3 plots, they will be in a single row, if there are 4 plots, they will be in 2 rows of 2.

Usage

fst_plot_multiple(plot1, plot2, plot3 = NULL, plot4 = NULL, main_title = NULL)

Arguments

plot1 First plot to display.
plot2 Second plot to display.
plot3 Optional third plot to display, default is ‘NULL’.
plot4 Optional fourth plot to display, default is ‘NULL’.
main_title An optional title for the set of plots. The default is ‘NULL’ and no main title will be included.

Value

Up to 4 plots within the plots pane.

Examples

top_f <- fst_get_top_words(conllu_dev_q11_1_f_nltk)
top_m <- fst_get_top_words(conllu_dev_q11_1_m_nltk)
top_na <- fst_get_top_words(conllu_dev_q11_1_na_nltk)
topn_f <- fst_get_top_ngrams(conllu_dev_q11_1_f_nltk)
topn_m <- fst_get_top_ngrams(conllu_dev_q11_1_m_nltk)
topn_na <- fst_get_top_ngrams(conllu_dev_q11_1_na_nltk)
unique_words <- fst_get_unique_ngrams(top_f, top_m, top_na)
unique_ngrams <- fst_get_unique_ngrams(topn_f, topn_m, topn_na)
top_fu <- fst_join_unique(top_f, unique_words)
top_mu <- fst_join_unique(top_m, unique_words)
top_nau <- fst_join_unique(top_na, unique_words)
p1 <- fst_ngrams_compare_plot(top_fu, ngrams = 1, name = "Female")
p2 <- fst_ngrams_compare_plot(top_mu, ngrams = 1, name = "Male")
p3 <- fst_ngrams_compare_plot(top_nau, ngrams = 1, name = "Not Spec")
fst_plot_multiple(p1, p2, p3, main_title = "Comparison Plots")
fst_plot_multiple(p1, p1)

---

**fst_pos**

*Make POS Summary Table*

**Description**

Creates a summary table for the input CoNLL-U data which counts the number of words of each part-of-speech tag within the data.

**Usage**

`fst_pos(data)`

**Arguments**

- `data` A dataframe of text in CoNLL-U format.

**Value**

A dataframe with a count and proportion of each UPOS tag in the data and the full name of the tag.

**Examples**

`fst_pos(conllu_cb_bullying_iso)`
`fst_pos(conllu_dev_q11_3_nltk)`

---

**fst_pos_compare**

*Compare parts-of-speech*

**Description**

Compare words in responses based on part-of-speech tagging for between 2 and 4 sets of prepared data.
Usage

```
fst_pos_compare(
    data1,  
    data2,  
    data3 = NULL,  
    data4 = NULL,  
    name1 = "Group 1",  
    name2 = "Group 2",  
    name3 = "Group 3",  
    name4 = "Group 4"
)
```

Arguments

- `data1`: A dataframe of text in CoNLL-U format for the first group.
- `data2`: A dataframe of text in CoNLL-U format for the second group.
- `data3`: An optional dataframe of text in CoNLL-U format for the third group, default is `NULL`.
- `data4`: An optional dataframe of text in CoNLL-U format for the fourth group, default is `NULL`.
- `name1`: An optional "name" for the first group, default is "Group 1".
- `name2`: An optional "name" for the second group, default is "Group 2".
- `name3`: An optional "name" for the third group, default is "Group 3".
- `name4`: An optional "name" for the fourth group, default is "Group 4".

Value

Table of POS tag counts for the groups.

Examples

```
f <- conllu_dev_q11_1_f_nltk
m <- conllu_dev_q11_1_m_nltk
na <- conllu_dev_q11_1_na_nltk
all <- conllu_dev_q11_1_nltk
fst_pos_compare(f, m, na, all, "Female", "Male", "Not Spec.", "All")
fst_pos_compare(f, m, name1 = "Female", name2 = "Male")
```

Description

`fst_prepare_conllu()` produces a dataframe (and saves as csv) containing Finnish survey text responses in CoNLL-U format with stopwords removed.
Usage

```
fst_prepare_conllu(data, field, model = "ftb", stopword_list = "nltk")
```

Arguments

- `data`: A dataframe of survey responses which contains an open-ended question.
- `field`: The field in the dataframe which contains the open-ended question.
- `model`: A Finnish language model available for [udpipe], "ftb" (default) or "tdt".
- `stopword_list`: A valid Finnish stopword list, default is "nltk", or "none".

Value

A dataframe of Finnish text in CoNLL-U format.

Examples

```
cb <- child_barometer_data
fst_prepare_conllu(data = cb, field = "q7", stopword_list = "stopwords-iso")
unlink("finnish-ftb-ud-2.5-191206.udpipe")
unlink("finnish-tdt-ud-2.5-191206.udpipe")
```

---

**fst_rm_stop_punct**

Remove Finnish stopwords and punctuation from CoNLL-U dataframe

Description

Removes stopwords and punctuation from a dataframe containing Finnish survey text data which is already in CoNLL-U format.

Usage

```
fst_rm_stop_punct(data, stopword_list = "nltk")
```

Arguments

- `data`: A dataframe of Finnish text in CoNLL-U format.
- `stopword_list`: A valid Finnish stopword list, default is "nltk".

Value

A dataframe of Finnish text in CoNLL-U format without stopwords and punctuation.

Examples

```
fst_rm_stop_punct(conllu_dev_q11_3)
fst_rm_stop_punct(conllu_dev_q11_1, stopword_list <- "snowball")
fst_rm_stop_punct(conllu_cb_bullying, "stopwords-iso")
```
 fst_summarise  

**Make Summary Table**

**Description**

Creates a summary table for the input CoNLL-U data which provides the response count and proportion, total number of words, the number of unique words, and the number of unique lemmas.

**Usage**

```r
fst_summarise(data, desc = "All respondents")
```

**Arguments**

- `data` A dataframe of text in CoNLL-U format.
- `desc` A string describing respondents, default is "All respondents".

**Value**

A dataframe with summary information for the data including response rate and word counts.

**Examples**

```r
fst_summarise(conllu_dev_q11_1)
fst_summarise(conllu_dev_q11_2_nltk, "Q11_2")
```

-------------------

**fst_summarise_compare  Make comparison summary**

**Description**

Compare text responses for between 2 and 4 sets of prepared data.

**Usage**

```r
fst_summarise_compare(
    data1,
    data2,
    data3 = NULL,
    data4 = NULL,
    name1 = "Group 1",
    name2 = "Group 2",
    name3 = "Group 3",
    name4 = "Group 4"
)
```
fst_summarise_short

Arguments

- **data1**: A dataframe of text in CoNLL-U format for the first group.
- **data2**: A dataframe of text in CoNLL-U format for the second group.
- **data3**: An optional dataframe of text in CoNLL-U format for the third group, default is ‘NULL’.
- **data4**: An optional dataframe of text in CoNLL-U format for the fourth group, default is ‘NULL’.
- **name1**: A string describing data1, default is “Group 1”.
- **name2**: A string describing data2, default is “Group 2”.
- **name3**: A string describing data3, default is “Group 3”.
- **name4**: A string describing data4, default is “Group 4”.

Value

Summary table of responses between groups.

Examples

```r
f <- conllu_dev_q11_1_f_nltk
m <- conllu_dev_q11_1_m_nltk
na <- conllu_dev_q11_1_na_nltk
all <- conllu_dev_q11_1_nltk
fst_summarise_compare(m, f, na, all, "Male", "Female", "Not Spec.", "All")
fst_summarise_compare(m, f, name1 = "Male", name2 = "Female")
```

---------------------

fst_summarise_short  
Make Simple Summary Table

Description

Creates a summary table for the input CoNLL-U data which provides the total number of words, the number of unique words, and the number of unique lemmas.

Usage

```r
fst_summarise_short(data)
```

Arguments

- **data**: A dataframe of text in CoNLL-U format.

Value

A dataframe with summary information on word counts for the data.
**fst_wordcloud**

*Make Wordcloud*

**Description**

Creates a wordcloud from CoNLL-U data of frequently-occurring words.

**Usage**

```
fst_wordcloud(data, pos_filter = NULL, max = 100)
```

**Arguments**

- **data**: A dataframe of text in CoNLL-U format.
- **pos_filter**: List of UPOS tags for inclusion, default is ‘NULL’ which means all word types included.
- **max**: The maximum number of words to display, default is ‘100’

**Value**

A wordcloud from the data.

**Examples**

```r
cb <- conllu_cb_bullying_iso
fst_wordcloud(cb)
```

```r
cb <- conllu_cb_bullying_iso
fst_wordcloud(cb, pos_filter = c("NOUN", "VERB", "ADJ", "ADV"))
```

```r
cb <- conllu_dev_q11_1_nltk
fst_wordcloud(cb)
```

```r
cb <- conllu_dev_q11_1_nltk
fst_wordcloud(cb, pos_filter = "VERB", max = 50)
```
Index

* datasets
  - child_barometer_data, 3
  - conllu_cb_bullying, 4
  - conllu_cb_bullying_iso, 5
  - conllu_dev_q11_1, 6
  - conllu_dev_q11_1_f, 7
  - conllu_dev_q11_1_f_nltk, 8
  - conllu_dev_q11_1_m, 9
  - conllu_dev_q11_1_m_nltk, 10
  - conllu_dev_q11_1_na, 11
  - conllu_dev_q11_1_na_nltk, 12
  - conllu_dev_q11_2, 13
  - conllu_dev_q11_2_snow, 14
  - conllu_dev_q11_2_nltk, 15
  - conllu_dev_q11_3, 16
  - conllu_dev_q11_3_nltk, 17
  - dev_data, 19
  - dev_data_f, 19
  - dev_data_m, 20
  - dev_data_na, 21
  - dev_data_f, 19
  - dev_data_m, 20
  - dev_data_na, 21

  - fst_cn_compare_plot, 21
  - fst_cn_edges, 23
  - fst_cn_get_unique, 24
  - fst_cn_nodes, 24
  - fst_cn_plot, 25
  - fst_cn_search, 26
  - fst_comparison_cloud, 26
  - fst_concept_network, 28
  - fst_concept_network_compare, 29
  - fst_find_stopwords, 30
  - fst_format_conllu, 31
  - fst_freq, 31
  - fst_freq_compare, 32
  - fst_freq_plot, 34
  - fst_get_top_ngrams, 35
  - fst_get_top_ngrams2, 36
  - fst_get_top_words, 37
  - fst_get_unique_ngrams, 38
  - fst_join_unique, 38
  - fst_length_compare, 39
  - fst_length_summary, 40
  - fst_ngrams, 41
  - fst_ngrams_compare, 42
  - fst_ngrams_compare_plot, 43
  - fst_ngrams_plot, 44
  - fst_plot_multiple, 45
  - fst_pos, 46
  - fst_pos_compare, 46
  - fst_prepare_conllu, 47
  - fst_rm_stop_punct, 48
  - fst_summarise, 49
  - fst_summarise_compare, 49
  - fst_summarise_short, 50
  - fst_wordcloud, 51

dev_data, 19