corpus.tools

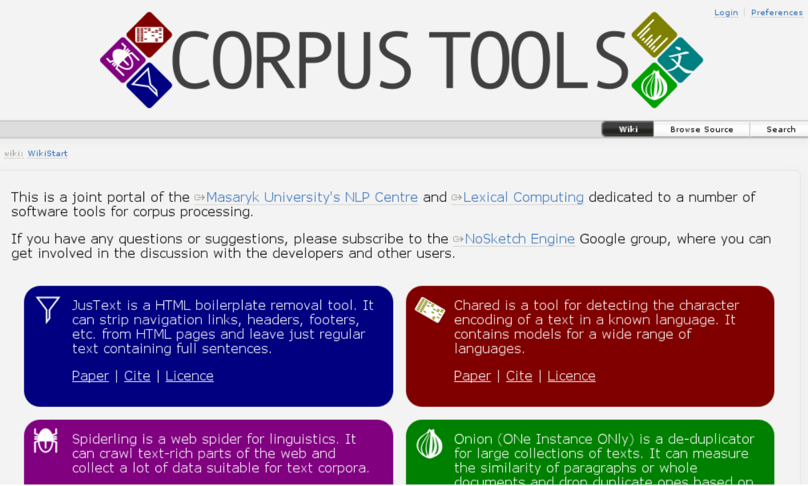
Jan Michelfeit

Lexical 文 Computing

jan.michelfeit@sketchengine.co.uk

6th Sketch Engine Workshop
Herstmonceux, August 10, 2015

corpus.tools



This is a joint portal of the Masaryk University's NLP Centre and Lexical Computing dedicated to a number of software tools for corpus processing.

If you have any questions or suggestions, please subscribe to the NoSketch Engine Google group, where you can get involved in the discussion with the developers and other users.

JusText is a HTML boilerplate removal tool. It can strip navigation links, headers, footers, etc. from HTML pages and leave just regular text containing full sentences.

Paper | Cite | Licence

Chared is a tool for detecting the character encoding of a text in a known language. It contains models for a wide range of languages.

Paper | Cite | Licence

Spiderling is a web spider for linguistics. It can crawl text-rich parts of the web and collect a lot of data suitable for text corpora.

Onion (ONe Instance ONly) is a de-duplicator for large collections of texts. It can measure the similarity of paragraphs or whole documents and drop duplicate ones based on

Our strategy

- academia a very fertile environment

Our strategy

- academia a very fertile environment
- but projects often end up abandoned

Our strategy

- academia a very fertile environment
- but projects often end up abandoned
- adopt useful tools and continue developing and supporting them

Our strategy

- academia a very fertile environment
- but projects often end up abandoned
- adopt useful tools and continue developing and supporting them
- business-friendly open source licences

JusText

- “boilerplate” content removal tool
- boilerplate = headers, footers, navigation etc. in web pages
- leaves only large contiguous chunks of text (whole paragraphs)

Chared

- character encoding detection tool
- webpage metadata often not reliable
- a small amount of noise can pollute keyword reports
- languge-aware detection more successful than language-agnostic

SpiderLing

- = web-SPIDER for LINGuistics
- utilizes both JusText and Chared
- prioritizes sites rich in text
- goal: maximize number of words per megabyte downloaded

Onion

- = ONe Instance ONly
- duplicate (and near-duplicate) removal tool
- based on word n-grams
- typically runs on paragraphs
- similarity threshold, structure and *n* configurable

Unitok

- = UNIversal TOKenizer
- based on regular expressions
- URLs, e-mail addresses
- abbreviations, clictics (do – n't, ca – n't)
- configuration files for a variety of languages
- versioning promotes replicability
- includes uninorm, UNIversal NORMalizer

NoSketch Engine

- like Sketch Engine, but without Word Sketches
- manatee – backend, storage, indexing and query evaluation
- bonito – user interface for querying
- corpus architect – user corpora – not included
- google group for support