TenTen corpora

TenTen is a new generation of Web corpora. These corpora are created by Web crawling and processed with our latest boilerplate cleaning and de-duplication tools. The "TenTen" designates the target sizes of the corpora which is $10^{10}$ (10 billion) words.

Available corpora:

**Error: Macro SubWiki(Corpora/TenTen/,2) failed**

unsupported operand type(s) for +: 'NoneType' and 'str'

**Structural attributes**

A list of structural attributes shared by all TenTen corpora follows. Attributes specific to a corpus can be found on the information page of the corpus.

**Document**

- Crawl date = date of downloading of the document from the web
- Second level domain
- Top level domain
- url = URL of the source document
- Web domain
- wordcount

**Paragraph**

- heading = 1 if the paragraph is a heading, 0 otherwise

**Related papers**

?The TenTen Corpus Family

?Efficient Web Crawling for Large Text Corpora
by Jan Pomikálek, Vít Suchomel
at ?ACL SIGWAC Web as Corpus (at conference WWW), Lyon, April 2012.