Generating an NLP Corpus from Java Source Code: The SSL Javadoc Doclet

Submitted by ninus [1] on Wed, 2011-03-16 12:53

- nlp [2]
- NLP Components [3]
- software engineering [4]

Title	Generating an NLP Corpus from Java Source Code: The SSL Javadoc Doclet
Publication Type	Conference Paper
Year of Publication	2010
Refereed Designation	Refereed
Authors	Khamis, N. [5], R. Witte [6], and J. Rilling [7]
Conference Name	New Challenges for NLP Frameworks
Pagination	41–45
Date Published	May 22
Publisher	ELRA
Conference Location	Valletta, Malta
Keywords	NLP [8], NLP Components [9], Software Engineering [10]
Abstract	Source code contains a large amount of natural language text, particularly in the form of comments, which makes it an emerging target of text analysis techniques. Due to the mix with program code, it is difficult to process source code comments directly within NLP frameworks such as GATE. Within this work we present an effective means for generating a corpus using information found in source code and in-line documentation, by developing a custom doclet for the Javadoc tool. The generated corpus uses a schema that is easily processed by NLP applications, which allows language engineers to focus their efforts on text analysis tasks, like automatic quality control of source code comments. The SSLDoclet is available as open source software.
URL	http://www.lrec-conf.org/proceedings/lrec2010/workshops/W10.pdf [11]
Copyright	

Attachment	Size
nlpf2010-ssldoclet.pdf [12]	436.13 KB

Source URL (retrieved on *2025-12-15 07:27*):
https://www.semanticsoftware.info/biblio/generating-nlp-corpus-java-source-code-ssl-javadoc-doclet

Links:

[1] https://www.semanticsoftware.info/users/ninus
[2] https://www.semanticsoftware.info/category/blog-tags/nlp
[3] https://www.semanticsoftware.info/category/blog-tags/nlp-components
[4] https://www.semanticsoftware.info/category/blog-tags/software-engineering
[5] https://www.semanticsoftware.info/biblio/author/9
[6] https://www.semanticsoftware.info/biblio/author/1
[7] https://www.semanticsoftware.info/biblio/author/10
[8] https://www.semanticsoftware.info/biblio/keyword/3
[9] https://www.semanticsoftware.info/biblio/keyword/6
[10] https://www.semanticsoftware.info/biblio/keyword/5
[11] http://www.lrec-conf.org/proceedings/lrec2010/workshops/W10.pdf
[12] https://www.semanticsoftware.info/system/files/nlpf2010-ssldoclet.pdf