New Javadoc Doclet for NLP Analysis on Java Source Code

- Tools & Resources
- Corpora
- Semantic Computing
- Software Engineering

For those interested in performing NLP on source code, in particular Javadoc comments, we just released a Doclet at the NLP Frameworks workshop last week.

Its main feature is that it creates an XML corpus from Java source code that is optimised for processing in an NLP Framework (GATE in our case, but it should work for any framework that takes XML as input).

For more information and the download, have a look at the Web page. And for details, background, and an application example at our paper [1].

We currently use it for automatic quality assessment of source code comments, but obviously there are many other use cases as well.

References

1. Khamis, N., R. Witte, and J. Rilling, "Generating an NLP Corpus from Java Source Code: The SSL Javadoc Doclet", *New Challenges for NLP Frameworks*, Valletta, Malta : ELRA, pp. 41–45, May 22, 2010.

Source URL (retrieved on *2026-01-06 20:31*):
https://www.semanticsoftware.info/content/new-javadoc-doclet-nlp-analysis-java-source-code