

Use regular expression to split multiple-sentence titles

Context and Problem Statement

Some entry titles are composed of multiple sentences, for example: “Whose Music? A Sociology of Musical Language”, therefore, it is necessary to first split the title into sentences and process them individually to ensure proper formatting using ‘[Sentence Case](#)’ or ‘[Title Case](#)’

Considered Options

- [Regular expression](#)
- [OpenNLP](#)
- [ICU4J](#)

Decision Outcome

Chosen option: “Regular expression”, because we can use Java internal classes (Pattern, Matcher) instead of adding additional dependencies

Positive Consequences

- Less dependencies on third party libraries
- Smaller project size (ICU4J is very large)
- No need for model data (OpenNLP is a machine learning based toolkit and needs a trained model to work properly)

Negative Consequences

- Regular expressions can never cover every case, therefore, splitting may not be accurate for every title
-