Sentence Segmentation

Background

Pragmatic Segmenter is a Rule Based technique for Sentence Segmentation. The fact that it does not require any training data and that it can work even when the format and domain of the input text is unknown allows the user to implement it across many languages. The algorithm used is conservative in nature- it ignores any ambiguous sentence boundary. It does not split sentences within the parentheses and quotations.

NLTK Punct is an Unsupervised Machine Learning technique for Sentence Segmentation. The algorithm used for it is described in the paper by Tibor Kiss and Jan Strunk (2006) - Unsupervised Multilingual Sentence Boundary Detection. It is trained on Newspaper Corpora, BROwn and Poe Corpus. It can learn parameters such as a list of abbreviations.

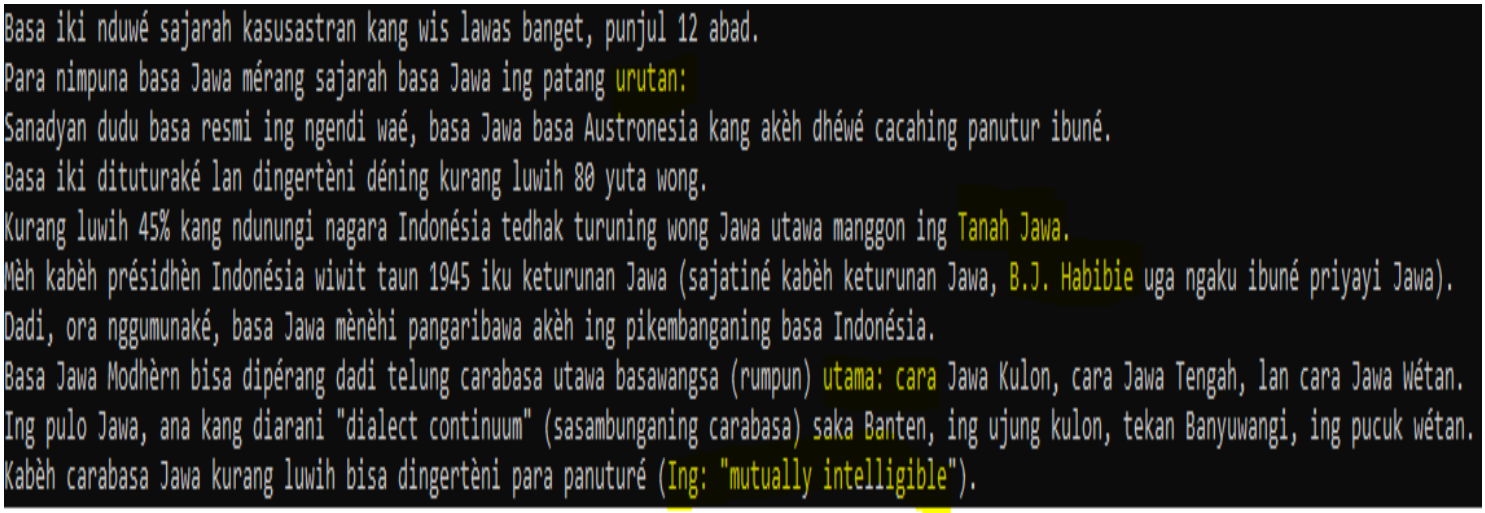
Quantitative evaluation:

NLTK Punct Accuracy = 96.67%
Pragmatic Segmenter Accuracy = 100%

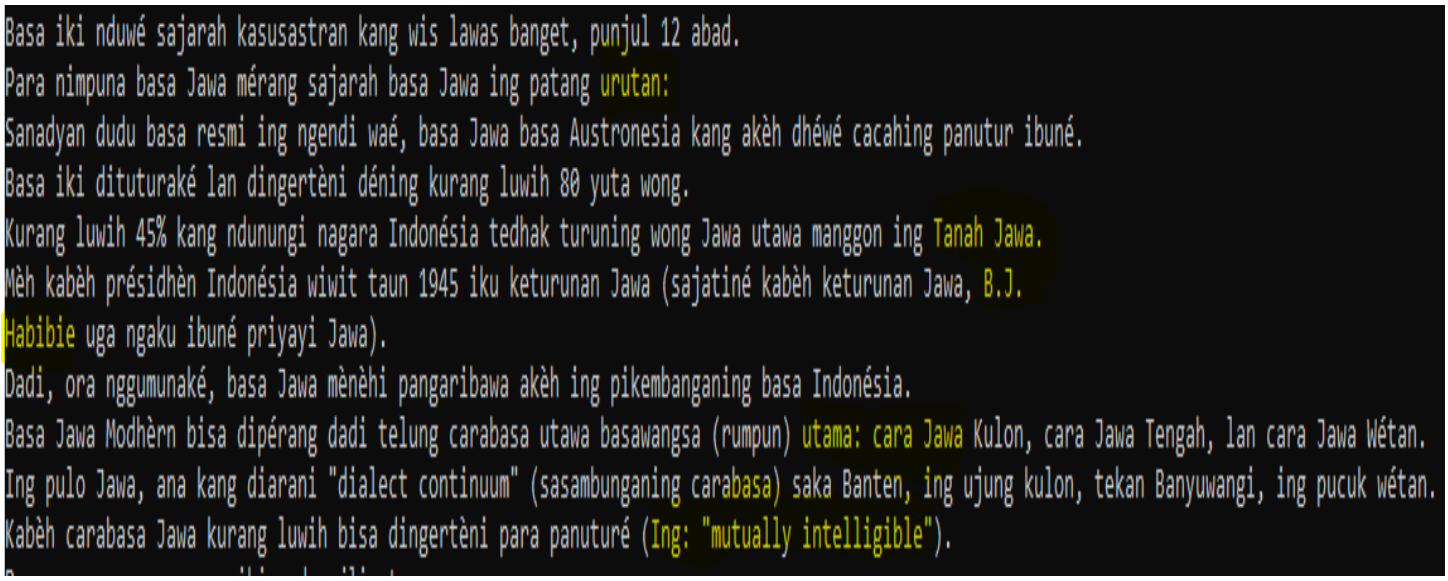
Qualitative evaluation:

- Both the segmenter correctly identifies the Sentence boundary when marked by the colon. It splits the sentences if colon is followed by a capital letter indicating new sentence and does not split the sentence when colon is followed by lower case letter indicating continuation of the sentence.
- Pragmatic Segmenter correctly identifies and distinguishes between the abbreviation like B.J. Habibie and the sentence boundary, but NLTK Punct considered the dot after B.J. as a full stop instead of an indication of abbreviation.

1. Pragmatic Segmenter



2. NLTK Punct



Note: One strange Example that i came across for NLTK Punct is that it Splitted B.J. and Habibie into separate sentences but did not do the same for E.M. Uhlenbeck