SANTI-POS v20210817 DOCUMENTATION

The first version of Indonesian POS tagger implemented using Nooj. POStags are shallow. Improvements are in progress. Drop your suggestions, questions, criticisms, here: prihantoro@live.undip.ac.id. Thank you!

--Prihantoro—

POS Tagset (adapted from Wicaksono & Purwarianti, 2010)

Label	Description	Example
JJ	Adjective	*besar, manis*
RB	Adverb	*lebih, nanti*
NN	Noun	*buku, mobil*
VB	Verb	*makan, tidur*
IN	Preposition	*di, ke, dari*
MD	Modal	*bisa, akan*
CC	Conjunction	*dan, jika*
DT	Determiner	*para, si*
UH	Interjection	*ah, aduh*
CD	Numeral	*satu, pertama*
PR	Pronoun	*dia, mereka*
NEG	Negation	*tidak, bukan*

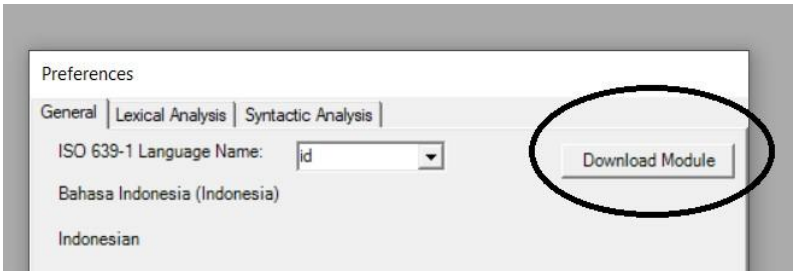
Requirement

Nooj (http://nooj4nlp.org/)

How to use

1. Download Indonesian language resources via Nooj Preferences

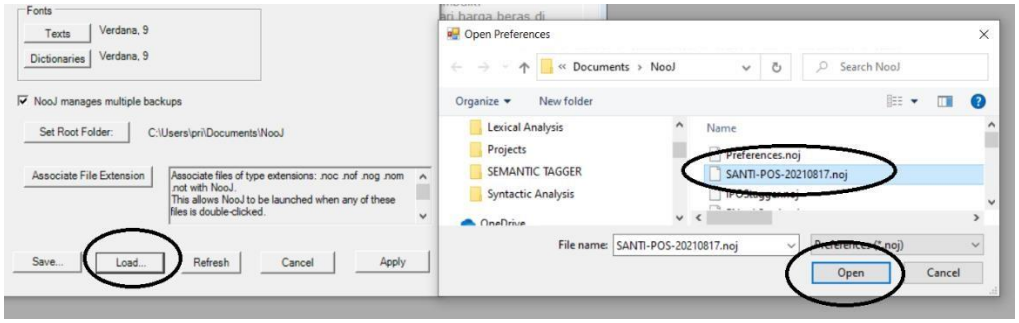
Info > Preferences > Next to language name, choose 'id' > Download Module



Or manually download them from here
https://drive.google.com/drive/folders/10XtOlLyW3tgX5SWaVLdFU_8inejbKi2G?usp=sharing

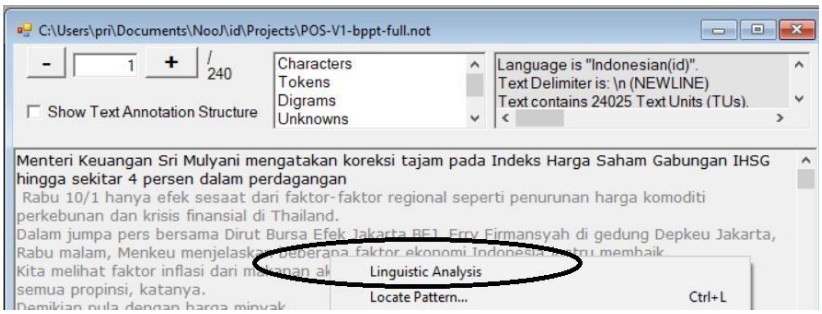
2. Load SANTI-POS configuration file

Info > Preferences > Load > id > Choose SANTI-POS-20210817.noj > Open



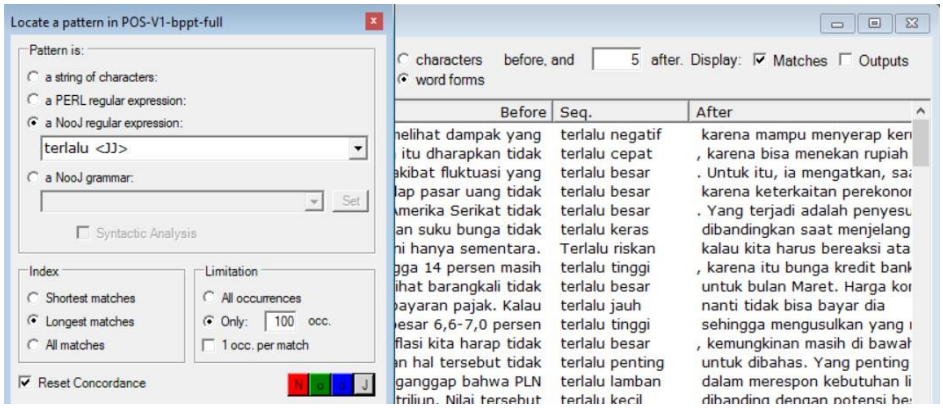
3. Annotate your corpus

File > Open > Text/Corpus > choose text/corpus file(s) > Open > right click on Text/Corpus > Linguistic Analysis



4. Use <POS Label> to search with pos tags

Right click on text > Locate Pattern > supply query > Enter



Corpus

The corpus that comes with the resources is BPPT-PAN Corpus (POS-V1-bppt-full.not). But you can use your own corpus if you want.

Exporting to .txt document

Open your Text/Corpus > Right Click > Export annotated text as XML format > see nameofyourfile.not.xml.txt

Citation

Wicaksono, A. F., & Purwarianti, A. (2010, August). HMM based part-of-speech tagger for Bahasa Indonesia. In *Fourth International MALINDO Workshop, Jakarta*.

Riza, H., & Hakim, C. (2009, August). Resource report: building parallel text corpora for multidomain translation system. In *Proceedings of the 7th Workshop on Asian Language Resources (ALR7)* (pp. 92-95).