In search of efficient tokenizers, I found 3 libraries to be best suited for the use case: keras, spacy and re. re and keras work in the same manner and have similar performance. spacy can get slower past a certain threshold because of the extra functionalities it brings. This can be seen from the following figure where notes can be replaced with number of words. Number of seconds consists of the time elapsed for a complete run.

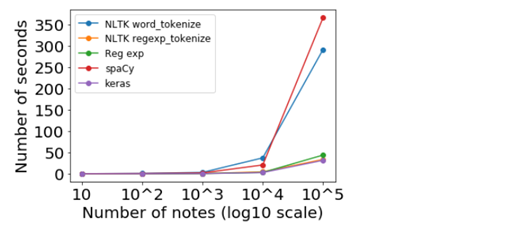


Figure Comparing tokenizers performance(<https://towardsdatascience.com/benchmarking-python-nlp-tokenizers-3ac4735100c5>)

An important thing to note is that even though performance varies, they all got the same results. Pointing to the fact that the methods are good, or the source is problematic. To ensure that there are minimal problems in the source, the python version, for transforming pdf in text, has been compared with files that have been manually converted from pdf to text. Both had similar enough content to ensure that the source is not the reason for the fact that each method has given the same results.

The extracted NPs are of type base noun phrase, or “NP chunk”, which is a noun phrase that does not permit other NPs to be nested within it – so no NP-level coordination, no prepositional phrases, and no relative clauses. I tried using my own patterns, but what was gained was little and what was lost was bigger.

For extracting VPs, I used a basic structure that doesn’t cover all possible verb phrases but does a better job than Pandey and Pandey. The code finds verb phrases with the pattern Optional Verb, 0 to many Adverbs and Verb.

Spacy uses different statistical models. I’ve listed below the different statistical models in spaCy along with their specifications:

en\_core\_web\_sm: English multi-task CNN trained on OntoNotes. Size – 11 MB

en\_core\_web\_md: English multi-task CNN trained on OntoNotes, with GloVe vectors trained on Common Crawl. Size – 91 MB

en\_core\_web\_lg: English multi-task CNN trained on OntoNotes, with GloVe vectors trained on Common Crawl. Size – 789 MB