# Module Discussion

## HTML Parser

## Tokenizer

The tokenizer utilizes the Tokenizer module from OpenNLP. It uses a pre-trained model which is freely available for download from their website [1]. This gives it an advantage over a more naïve approach to tokenization, such as using a regular expression; for instance, it can properly distinguish acronyms as a single token rather than separating the letters and dots as individual tokens. On the other hand, the training data used by the Apache Foundation to train the model has its own short-comings, since it doesn’t properly recognize entities such as email address (which it splits by its special characters such as the ‘@’ or ‘.’), and showed some trouble separating phrases like “Poesio/Kruschwitz” into three tokens, since it doesn’t seem to recognize the forward slash as a stop character. However, it still proves fairly useful at distinguishing periods which should belong within a token (e.g. “1.8 m” or “Y.M.C.A.”).

## Part-of-Speech Tagger

## Phrase Detector

## Ranker

## Entity Extracter

# References

|  |  |
| --- | --- |
| [1] | The Apache Software Foundation, "OpenNLP Tools Models," [Online]. Available: http://opennlp.sourceforge.net/models-1.5/. [Accessed 9 December 2015]. |