Links

* <https://www.gutenberg.org/>
* <https://github.com/ageitgey/Gutenberg>

Pipeline

1. Download files with gutenberg python package
2. Remove files not in the public domain
3. Separate files by language
4. Filter old books by comparing their vocabulary to average vocab
5. Extract dialogs
   1. ignore books which don’t have enough delimiters
   2. choose delimiter with highest count
   3. Split dialogs with dialog\_space, and utterances by paragraphs
   4. Join into a single utterance delimiters within a paragraph
   5. Remove utterances longer than a threshold
6. Build a vocab based on extracted dialogs and remove dialogs which have more than 20% unkown words (keep top 100k words)
7. Split dataset into train, dev, test

Parameters

* config.py