inputs for spellchecking process

augmented with likely proper nouns

*Gazetteer.txt*

only most common

*PrecisionDictionary.txt*

existing dictionaries of multiple languages

Augment dictionary with common words that tend to appear in Titlecase, and in clean documents.

Smaller “recall” dictionary (things we try *to* correct if misspelled.)

/filtered/\*.txt

Main list of token types for spellchecking

if **yes,** augment dictionary with information about #vols and ratio of titlecase / homogenouscase

*ExtractTitlecase1.py*

*ExtractTitlecase2.py*

*FilterTypes.py*

if **no,** send them to a filtered list of types where cased variants are compressed *unless* variant is *neither* Titlecase nor homogenous; e.g.,

“wiUful.”

in dictionary?

List of all types that occur in documents, with info about counts, #vols, and the average OCR quality of the documents where they occur.

Collection of documents

*typeindexer*