Stoplist ReadMe

A stoplist is a file containing a string of tokens, including words, numbers, paratextual information, and symbols that are considered burdensome to the process of machine reading. Words that do not provide a meaningful interpretive capacity to a text are removed in order to make the outputs of algorithmic processes clear and interpretable by humans. Examples include articles (*a*, *an*, *the*), proper names, and stage directions.

This folder contains the stoplists used for the NovelTM: Text Mining the Novel Project in the process of topic modeling the Eighteenth-Century Collections Online Text Creation Partnership files (ECCO-TCP).

All R scripts located in the NovelTM Google Drive use the ECCOstoplist file by default.

Stoplist Files:

ECCOstoplist:

This stoplist is the combination of both the JockersExpandedStoplist and taporstoplist below and the includes several tokens common in 18th and 17th century texts, as well as several Classical, Medieval, or literary names common to literary texts. This stoplist was developed over the course of two years specifically for topic modeling tragedies and comedies, and includes some tokens particular to both the ECCO and EEBO corpora, including OCR and unicode errors. Caution is advised for using this stoplist with other datasets or methods of text analysis.

Both .rtf and .txt versions of this list are included. The .rtf file lists the stoplist tokens without line breaks. The .txt version includes line breaks. If you want to load the stoplist into R as a data frame (if you are using the sLDA script, for example), using the .txt version of the file is recommended.

JockersExpandedStoplist:

This stoplist is pulled from a Matt Jockers blog post. It contains several common tokens, but also several thousand personal names for the purpose of helping to clean literary datasets of common character names. This list is primarily designed for 19th-century datasets.

taporstoplist:

This is a stoplist pulled from the Text Analysis Portal for Research (TAPoR). It features a small amount of commonly used tokens that are excluded in the process of using most TAPoR-born tools.