Stopwords

Stopwords are any word in a list of words which are filtered out (i.e., stopped) before or after processing of natural language data (text). There is no single universal list of stop words used by all NLP tools, nor any agreed upon rules for identifying stop words. Typically, pronouns, determiners, auxiliaries, adverbs are included in the list. But, any group of words can be chosen as the stop words for a given purpose. As Manning et al. (2008: 27) write: "[The] general trend in [information retrieval] systems over time has been from standard use of quite large stop lists (200–300 terms) to very small stop lists (7–12 terms) to no stop list whatsoever."

Sometimes, stopwords are also called **junk words** or **function words**, as opposed to **content words** that provide the bulk of meaning in a text.

Caveat. Pennebaker's work on function words show the importance these words have in detecting people's writing style and psychological state (Pennebaker and King 1999; Pennebaker et al. 2003; Chung and Pennebaker 2007).

For more information, history, and scholarly references, see the good **Wikipedia** page https://en.wikipedia.org/wiki/Stop_word#cite_note-2.

You will find a couple of lists of common stopwords under the lib\wordLists subdirectories: stopwords.txt and sw.txt.

References

Chung, Cindy and James Pennebaker. 2007. "The Psychological Functions of Function Words." In: pp. 343-359, Klaus Fiedler (Ed.), Social Communication, New York: Psychology Press.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

Pennebaker, James W., Matthias R. Mehl, and Kate G. Niederhoffer. 2003. "Psychological Aspects of Natural Language Use: Our Words, Our Selves." Annual Review of Psychology, Vol. 54, pp. 547–77.

Pennebaker, James W. and Laura A. King. 1999. "Linguistic Styles Language Use as an Individual Difference," *Journal of Personality and Social Psychology*, Vol. 77, No.6, pp. 1296-1312.