Stopword Removal: Why Bother? A Case Study on Verbose Queries

ABSTRACT

Stopword removal has traditionally been an integral step in information retrieval preprocessing. In this paper we question the utility of this step for verbose queries on standard datasets. We show in seven FIRE test collections in four languages – Bangla, Hindi, Gujarati and Marathi, that stopword removal does not lead to noticeable difference in retrieval performance as opposed to not removing stopwords. However, for European languages like English (TREC678 Ad Hoc) and French (CLEF 2005 to 2007), stopword removal leads to a statistically significant drop in performance.

CCS CONCEPTS

•Information systems →Information retrieval;

KEYWORDS

Stopword removal, retrieval, test collections

ACM Reference format:
. 2017. Stopword Removal: Why Bother? A Case Study on Verbose Queries. In *Proceedings of ACM CIKM conference, Pan Pacific, Singapore, November 2017 (CIKM'17)*, 2 pages.
DOI:

CIKM'17, Pan Pacific, Singapore

DOI:

Dataset	Number of documents	Number of topics
FIRE 2010 Ad Hoc Bangla	123047	50
FIRE 2011 Ad Hoc Bangla	377104	50
FIRE 2012 Ad Hoc Bangla	377111	50
FIRE 2010 Ad Hoc Hindi	149482	50
FIRE 2011 Ad Hoc Hindi	331599	50
FIRE 2011 Ad Hoc Gujarati	313163	50
FIRE 2010 Ad Hoc Marathi	99275	50
TREC678 Ad Hoc English	528155	150
CLEF 2005 to 2007 Ad Hoc French	177452	148

Table 1: Datasets.

Language	Source
Bangla	http://www.isical.ac.in/~fire/data/stopwords_list_ben.txt
Hindi	http://www.isical.ac.in/~fire/data/stopwords_list_hin.txt
Gujarati	http://irlab.daiict.ac.in/downloads/gujarati_stop_words.zip
Marathi	http://members.unine.ch/jacques.savoy/clef/marathiST.txt
English	http://www.lemurproject.org/stopwords/stoplist.dft
French	http://members.unine.ch/jacques.savoy/clef/frenchST.txt

Table 2: Stopword sources.

Dataset	Language	With stopwords	Stopwords removed
FIRE 2010	Bangla	**0.4337**	0.4334
FIRE 2011	Bangla	**0.2910**	0.2898
FIRE 2012	Bangla	0.2465	**0.2472**
FIRE 2010	Hindi	**0.4695**	0.4569
FIRE 2011	Hindi	0.1639	**0.1642**
FIRE 2011	Gujarati	**0.2818**	0.2753
FIRE 2010	Marathi	**0.2587**	0.2552
TREC678	English	**0.2150***	0.1971
CLEF 2005 to 2007	French	**0.2739***	0.2581

Table 3: This table reports the retrieval performance in terms of Mean Average Precision. For each dataset the better performance value is shown in bold font. * indicates that the performance difference is statistically significant ($p < 0.05$) at 95% confidence interval.