RESEACH ARTICLE

Effect Of Stop-Word Removal For Marathi Language Text Retrieval

by Harshali B. Patil, Ajay S. Patil

International Journal of Computer Applications

Foundation of Computer Science (FCS), NY, USA

Volume 185 - Number 47

Year of Publication: 2023

Authors: Harshali B. Patil, Ajay S. Patil

10.5120/ijca2023923289

Harshali B. Patil, Ajay S. Patil . Effect of Stop-Word Removal for Marathi Language Text Retrieval. International Journal of Computer Applications. 185, 47 (Dec 2023), 30-34. DOI=10.5120/ijca2023923289

Abstract

Automatic e-document processing systems have been one of the main fields of research and development over past decades. Preprocessing techniques are found to be useful for the process of organizing unstructured text while implementation of various web and data mining techniques like information retrieval, clustering, classifications, etc. Stop-word removal is one of the important preprocessing techniques used for removal of the tokens that do not have any linguistic meaning, and affects on the performance of text mining tasks. These words serve no purpose for Information Retrieval, but they are used very frequently in composing the documents. In modern Information Retrieval process, effective indexing can be achieved by removal of stop words. Many stop word lists have been developed for the major European languages that motivated researchers to work on Asian languages. In case of Indian languages, attempts could be found for Hindi, Bengali, etc. This paper discusses the procedures of two types of stop-word list construction for Marathi text retrieval systems and their impact on reduction in index size. The experimental results reveals that the proposed stop-word list achieves maximum reduction in index size over prior published lists.

References

1. F. Zou, F. L. Wang, X. Deng, S. Han, and L. S. Wang, "Automatic Construction of Chinese Stop Word List", in Proceedings of the 5th WSEAS International Confere nce on Applied Computer Science, pp. 1010-1015, 2006.
2. A. S. Patil and B. V. Pawar, "Analysis of Traditional Information Retrieval Techniques Applied to the World Wide Web", International Journal in Computer Science a nd Information Technology (IJCSIT), Vol. 1, No. 2, pp-63-73,2008.
3. L. Dolamic and Jacques Savoy, "Comparative Study of Indexing and Search Strategies for the Hindi, Marathi and Bengali Language", ACM Transactions on Asian Language Information Processing Vol. 9, No. 3, Article No.: 11, pp 1–24, 2010.
4. K. Zipf, "Human behaviour and the principle of least effort: an introduction to human ecology", Addison-Wesley Press, 1949.
5. C. Fox, "A stop list for general text." In ACM SIGIR Forum, vol. 24, no. 1-2, pp. 19-21, ACM 1989.
6. H. P. Luhn, "A statistical approach to mechanized encoding and searching of literary information." IBM Journal of research and development Vol. 1, no. 4, , pp. 3 09-317, 1957.
7. R. Al-Shalabi, G. Kanaan, J. M. Jaam, A. Hasnah and E. Hilat, "Stop-word removal algorithm for Arabic language," Proceedings of the International Conference o n Information and Communication Technologies: From Theory to Applications, pp. 545, 2004.
8. R. T.W. Lo, B. He and I. Ounis, "Automatically Building a Stopword List for an Information Retrieval System", Journal of Digital Information Management, Vol. 3 N o. 1 pp 03- 08, 2005.
9. F. Lazarinis, "Engineering and utilizing a stopword list in Greek Web retrieval", Journal of the American Society for Information Science and Technology, Vol. 58, No. 11 , pp. 1645-1652., 2007
10. M. Makrehchi and M. S. Kamel, "Automatic Extraction of Domain-Specific Stopwords from Labeled Documents", In European Conference on Information Retriev al pp. 222-233, 2008.
11. C. T. Yuang, R. E. Banchs, and C. E. Siong, "An empirical evaluation of stop-word removal in statistical machine translation", in Proc.13th conference of the Europe an chapter of the association for computational linguistics 2012, pp 30-37.
12. A. Alajmi, E. M. Saad, and R. R. Darwish. "Toward an ARABIC stop-words list generation." International Journal of Computer Applications Vol.46, no. 8, pp.8-13,2 012.
13. S. Hassan, F. Miriam and A. Harith, "On stopwords, filtering and data sparsity for sentiment analysis of twitter", in Proceedings of the 9th International Language Resources and Evaluation Conference, pp. 810-817, 2014
14. J. K. Raulji and J. R. Saini, "Stop-Word Removal Algorithm and its Implementation for Sanskrit Language", International Journal of Computer Applications, Vol. 15 0 – No.2, pp. 15-17, 2016.
15. J. Kaur and P. K. Buttar, "A Systematic Review on Stopword Removal Algorithms", International Journal on Future Revolution in Computer Science & Commu nication Engineering, Vol. 4, No. 4, pp. 207-210, 2018.
16. R. Rani, and D. K. Lobiyal, "Automatic Construction of Generic Stop Words List for Hindi Text", Procedia Computer Science, Vol. 132 pp. 362-370, 2018.
17. S. S. Sahu and S. Pal, "Effect of Stop-words in Indian Language IR", Sadhana, Vol .47, No. 1, 2022 .
18. Y. Zhou, and C. Ze-wen. "Research on the construction and filter method of stop-word list in text preprocessing." In Intelligent Computation Technology and Au tomation (ICICTA), vol. 1, IEEE, pp. 217-221, 2011.

19. J. Savoy, "A Stemming Procedure and Stopword List for General French Corpora", Journal of the American Society for Information Science, Vol. 50, No. 10, pp. 944–952, 1999.
20. A.K Pandey., and T.J. Siddiqui, " Evaluating Effect of Stemming and Stop-word Removal on Hindi Text Retrieval", In Proceedings of the First International Conference on Intelligent Human Computer Interaction, 2009.
21. A. Almeida and P. Bhattacharya, "Using Morphology to Improve Marathi Monolingual Information Retrieval", in proceeding of Forum for Information Retrieval Evaluation, 2010

Index Terms

Computer Science

Information Sciences

Keywords

Stop-word Marathi Fox guidelines Zipf's law.

The week's pick

Artificial Bee Colony (ABC) optimization Algorithm Based Automatic Segmentation and Detection of Suspicious Lesions in Lung CT Images

Divya A. Janaki Sathya D.

Random Articles

2016