IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

IJARSCT

Impact Factor: 7.301

Volume 3, Issue 1, December 2023

Text Language Identification and Translator

Tejas Pinge[1], Prajwal Patil[2] , Mayur Sherki[3], Aditya Nandurkar[4] , Prof. Ravindra Chilbule[5]

Students, Department of Computer Science Engineering[1,2,3,4]

Guide, Department of Computer Science Engineering[5]

Rajiv Gandhi College of Engineering Research and Technology, Chandrapur, Maharashtra, India

tejaspinge232@gmail.com, prajwalpatil2812@gmail.com, sherkimayur@gmail.com, nandurkaraditya8@gmail.com

Abstract: *Language Identification refers to the process of detecting the language(s) of the text in the document based on the script used for writing and observing the diacritics particular to a language. This research area has always fascinated researchers as early as 1970 and till now due to varied applications and increased demands of this field. In this work, I address the problem of detecting language of textual documents. I have introduced a method which is able to detect language of text more efficiently and accurately by determining their respective proportions and finding the greatest of them which represents the language of the text. I have demonstrated the performance comparison of three different approaches which are using n-gram approach (word-wise), using n-gram approach (character-wise) and using a combination of word search and stop words detection. My project currently contains language models for 4 languages. On an average the accuracy of my program is about 96.5%.*

Keywords: Language

REFERENCES

[1]. Cavnar, William B., and John M. Trenkle. "N-gram-based text categorization." Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval. 1994.

[2]. Dunning, Ted. "Accurate methods for the statistics of surprise and coincidence." Computational Linguistics 19.1 (1993): 61-74

[3]. Cortes, C., and Vapnik, V. "Support-vector networks." Machine Learning 20, 3 (1995): 273–297.

[4]. Baldwin, Timothy, and Su Nam Kim. "Multiword expressions." Language and Linguistics Compass 3.1 (2009): 870-894.

[5]. Lui, Marco, and Timothy Baldwin. "langid.py: An off-the-shelf language identification tool." Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 2012.

[6]. Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems. 2013.

[7]. Bojanowski, Piotr, et al. "Enriching word vectors with subword information." Transactions of the Association for Computational Linguistics 5 (2017): 135-146.

[8]. Devlin, Jacob, et al. "BERT: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

DOI: 10.48175/IJARSCT-14055

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH IN SCIENCE, COMMUNICATION AND TECHNOLOGY
ISSN
2581-9429
IJARSCT