IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, November 2022
Impact Factor: 6.252

Legal Document Classification using TF-IDF and KNN

Mr. Nikhil Wani, Ms. Gayatri Mangire, Mr. Aman Kumar, Ms. Nandini Solse, Mrs. P. S. Gaikwad
Department of Computer Engineering
AISSMS-Institute of Information Technology, Pune, Maharashtra, India

Abstract: *The increase in use of NLP, throughout various domains has travelled its way upto legal environment in the recent years. The collection and analysis of data required in lawsuits and topic segmentation of the filed petitions into the respective categories electronically can reduce a significant amount of time and cost compared to human efforts. Our aim is to come up with the process of automating the Text Classification of documents. We plan to implement NLP techniques such as count Vectorizer along with TF-IDF and KNN to categorize the legal documents in a supervised environment. We have proposed a model which works on high dimensional data.*

Keywords: Text Classification, Countvectorizer, TFIDF (Term Frequency–Inverse Document Frequency), KNN.

REFERENCES

[1]. Herbert L Roitblat, Anne Kershaw and Patrick Oot, "Document Categorization in Legal Electronic Discovery: Computer Classification vs. Manual Review", Journal of the American Society for Information Science and Technology Volume 61Issue 1January 2010 pp 70–80.

[2]. R. Chhatwal, P. Gronvall, N. Huber-Fliflet, R. Keeling, J. Zhang and H. Zhao, "Explainable Text Classification in Legal Document Review A Case Study of Explainable Predictive Coding," 2018 IEEE International Conference on Big Data (Big Data), 2018, pp. 1905-1911, doi: 10.1109/BigData.2018.8622073

[3]. R. Beale M. Y. Noguti, E. Vellasques and L. S. Oliveira, "Legal Document Classification: An Application to Law Area Prediction of Petitions to Public Prosecution Service," 2020 International Joint Conference on Neural Networks (IJCNN), 2020, pp. 1-8, doi: 10.1109/IJCNN48605.2020.9207211.

[4]. J. R. Brzezinski and G. J. Knafl, "Logistic regression modeling for context-based classification," Proceedings. Tenth International Workshop on Database and Expert Systems Applications. DEXA 99, 1999, pp. 755-759, doi: 10.1109/DEXA.1999.795279.

[5]. C. Liu, Y. Sheng, Z. Wei and Y. Yang, "Research of Text Classification Based on Improved TF-IDF Algorithm," 2018 IEEE International Conference of Intelligent Robotic and Control Engineering (IRCE), 2018, pp. 218-222, doi: 10.1109/IRCE.2018.8492945.

[6]. P. Soucy and G. W. Mineau, "A simple KNN algorithm for text categorization," Proceedings 2001 IEEE International Conference on Data Mining, 2001, pp. 647-648, doi: 10.1109/ICDM.2001.989592.

[7]. H. Li, H. Jiang, D. Wang and B. Han, "An Improved KNN Algorithm for Text Classification," 2018 Eighth International Conference on Instrumentation & Measurement, Computer, Communication and Control (IMCCC), 2018, pp. 1081-1085, doi: 10.1109/IMCCC.2018.00225.

[8]. A. Moldagulova and R. B. Sulaiman, "Using KNN algorithm for classification of textual documents," 2017 8th International Conference on Information Technology (ICIT), 2017, pp. 665-671, doi: 10.1109/ICITECH.2017.8079924.