Abstract

This paper presents the results of an experimental implementation of a document classifier leveraging contextual word embeddings clustered on a self-organizing map. The problem of document categorization is further compounded when there are no predefined categories, or conversely there are too many categories, that documents may be bucketed into. This paper proposes to address these problems by modelling the major themes contained in the document corpus into a cluster-map using a self-organizing neural network. The cluster-map provides a visual representation to explore the corpus, and a near-semantic search interface of the many concepts outlined across the corpus.

References

5. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. and Dean, J., “Distributed
representations of words and phrases and their compositionality,” Advances in Neural
retrieval,” Proceedings of the 14th Annual International ACM SIGIR Conference on Research

Index Terms

Computer Science
Pattern Recognition

Keywords

Clustering, knowledge retrieval, natural language processing, neural nets, self organizing map,
topic modelling, semantic search, unsupervised.