Short Text Topic Modelling using Non-negative Matrix Factorization with Neighbourhood-based Assistance

W.S. Athukorala* and W.A. Mohotti

Department of Computer Science, University of Ruhuna, Matara
**Corresponding Author E-mail: shalani7007@gmail.com, TP: +94710835972*

A massive number of short texts are generated every day in the forms of tweets, news headlines, questions, and answers. Analyzing short texts is an effective method to acquire valuable insights from these online archives that show diverse applications in community detection, trend analysis, classification, and summarization. Topic modeling is a widely used technique for this purpose as it is capable of latent topic discovery, and finding relationships among terms, topics, and text documents. In discovering thematic structure in collections of texts, a higher number of terms appear in the document × term matrix representation and associated sparseness creates issues for distance-based and density-based document similarities calculations. This phenomenon is known as distance concentration where the distance differences between points become negligible due to sparseness in high dimensions. Additionally, the short text shows a shorter length compared to conventional documents. This leads short texts to create extremely sparse, high-dimensional text and challenge finding documents that share the same topic structure within them. Non-negative Matrix Factorization (NMF) which is aligned with the natural non-negativity of text data is proposed as an effective technique that handles high dimensional representation with lower-dimensional projection. However, this higher-to-lower dimensional projection results in an information loss. This paper proposes Neighbourhood-based assistance to compensate for this loss. Neighborhood information within documents is captured using Jaccard similarity considering term sets included in the documents. We coupled a symmetric document × document matrix that carries this neighborhood information with the document × term matrix using NMF to identify the lower order topic × document matrix. This unsupervised method learns a dense lower-order topic presentation by minimizing the encoding error of factor matrices. We empirically evaluate the effectiveness of the method against the state-of-the-art short text topic modeling methods belongs to probabilistic and matrix factorization categories. Experimental results using three Twitter datasets show that the proposed approach is able to deal with information loss attached with higher dimensional matrix factorization of short-text and attain high accuracy compared to relevant benchmarking methods.

Keywords: Topic Modelling; Short Text; Non-negative Matrix Factorization; Neighbourhood-based Assistance