On “Document clustering based on non-negative matrix factorization”

This paper proposes a matrix factorization method that finds a semantic space for text documents. This space has the property of having similar documents (in terms of their topics) near each other or in clusters.

The non-requirement of orthogonality of the latent vectors for clusters is essential to this method because it allows overlap that is naturally present in document topics, making this space more consistent with human reasoning. This is not the case with SVD, for example.

The paper uses word frequency to build the starting document vectors. It would certainly be interesting how using a different representation (such as Glove or Word2Vec) could affect performance and latent space. If I understood the study correctly, a positive representation of the documents is required.

Related to the above, this algorithm should not be limited to the domain in which the authors apply it. This is because it only requires a vector representation of *items* (documents in the study). However, I do not know if changing the domain affects performance. That and the variation of the initial representation method could have been tested by the authors or at least proposed as future work. Their conclusions seem to be too oriented to results and not to applicability, which is important as well and is often overlooked.