4

1 Paper Details

Title Text mining using nonnegative matrix factorization and latent semantic analysis

Authors Ali Hassani, Amir Iranmanesh, Najme Mansouri

Venue Neural Computing and Applications, Volume 33

Year of Publication 2021

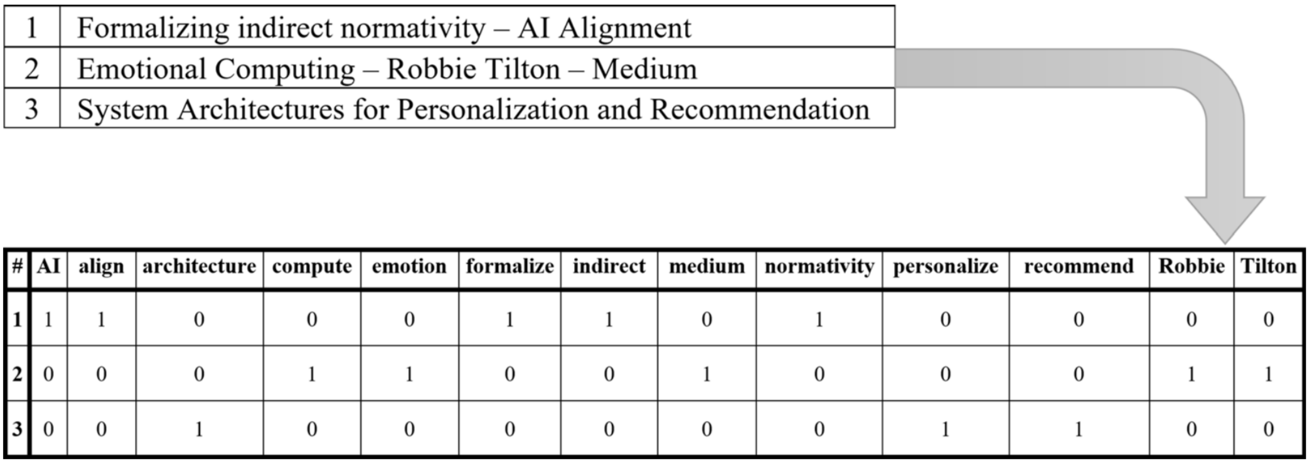
URL `https://link.springer.com/article/10.1007/s00521-021-06014-6`

Problem The paper proposes an algorithm for Text Clustering which is a Natural Language Processing problem involving grouping a collection of textual documents into clusters based on their content's similarity.

Subproblem involving NMF The paper uses NMF for **Feature Extraction**, Latent semantic analysis (LSA) for dimension reduction and deterministic K-means algorithm for clustering. Usage of NMF also decreases dimensionality of the data.

Terminology A *term* is a unit which could be a word or combination of words of the language.

A matrix of concern is the *term-document matrix* $X \in \mathbb{R}^{n \times m}$ which is obtained after processing the matrix which stores the frequency of a term (m different terms) of a particular document (n available documents) in that row as shown in the below figure from the same paper. The processing step involves stemming (reduces words to their stems) and weighting (applying weights to terms based on their frequency)



1	Formalizing indirect normativity – AI Alignment
2	Emotional Computing – Robbie Tilton – Medium
3	System Architectures for Personalization and Recommendation

#	AI	align	architecture	compute	emotion	formalize	indirect	medium	normativity	personalize	recommend	Robbie	Tilton
1	1	1	0	0	0	1	1	0	1	0	0	0	0
2	0	0	0	1	1	0	0	1	0	0	0	1	1
3	0	0	1	0	0	0	0	0	0	1	1	0	0

Solution of the subproblem involving NMF Feature Extraction involves creating a more suitable feature space for text clustering. This is done by separating the terms into groups and then combining each group's term vectors into new feature vectors. Mathematically, $\|X - WH\|_F$ is minimised such that $W \in \mathbb{R}^{n \times p}, R \in \mathbb{R}^{p \times m}$ and $W, H \geq 0$.

Significance of the Dictionary and its coefficients After the optimisation, W represents the dictionary and H gives the dictionary coefficients. H is used in order to group terms together into p categories and then represent each such category with a feature vector whereas an initial feature vector of each term is given by linear combination of columns of W according to the dictionary coefficients. Then for each group these initial feature vectors are taken and new feature vectors are generated.