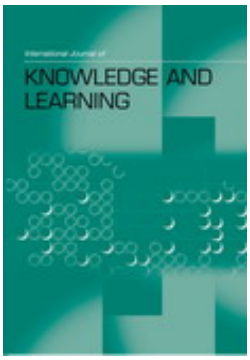


International Journal of Knowledge and Learning

ISSN online: 1741-1017 - ISSN print: 1741-1009
https://www.inderscience.com/ijkl

Three level weight for latent semantic analysis: an efficient approach to find enhanced semantic themes

Pooja Kherwa, Poonam Bansal

DOI: 10.1504/IJKL.2022.10047822

Article History:
Received: 18 May 2021
Accepted: 04 April 2022
Published online: 30 November 2022

Three level weight for latent semantic analysis: an efficient approach to find enhanced semantic themes

Pooja Kherwa*

Maharaja Surajmal Institute of Technology,
New Delhi, 110058, India
and
Affiliated to: GGSIPU, India
Email: Poona281280@gmail.com
*Corresponding author

Poonam Bansal

Indira Gandhi Delhi Technical University for Women,
Opp. St., Kashmere Gate, New Delhi, Delhi 110006, India
Email: pbansal89@gmail.com

Abstract: Latent semantic analysis is a prominent semantic themes detection and topic modelling technique. In this paper, we have designed a three-level weight for latent semantic analysis for creating an optimised semantic space for large collection of documents. Using this novel approach, an efficient latent semantic space is created, in which terms in documents comes closer to each other, which appear far away in actual document collection. In this approach, authors used two dataset: first is a synthetic dataset consists of small stories collected by the authors; second is benchmark BBC-news dataset used in text mining applications. These proposed three level weight models assign weight at term level, document level, and at a corpus level. These weight models are known as: 1) NPC; 2) NTC; 3) APC; 4) ATC. These weight models are tested on both the dataset, compared with state of the art term frequency and it has shown significant improved performances in term set correlation, document set correlation and has also shown highest correlation in semantic similarity of terms in semantic space generated through these three level weights. Our approach also shows automatic context clustering generated in dataset through three level weights.

Keywords: single value decomposition; SVD; latent semantic analysis; LSA; context clustering; semantic space.

Reference to this paper should be made as follows: Kherwa, P. and Bansal, P. (2023) 'Three level weight for latent semantic analysis: an efficient approach to find enhanced semantic themes', *Int. J. Knowledge and Learning*, Vol. 16, No. 1, pp.56–72.

Biographical notes: Pooja Kherwa is an Assistant Professor of the Maharaja Surajmal Institute of Technology, New Delhi. She received her MTech in Information Technology from the Guru Govind Singh Indraprastha University, Dwarka, New Delhi, in 2010. Currently, she is pursuing her PhD from the Guru Govind Singh Indraprastha University, Dwarka, New Delhi. Her research interest includes topic modelling, sentiment analysis and machine learning.

Poonam Bansal is a Professor in the Indira Gandhi Delhi Technical University for Women, New Delhi. She has received her PhD from the Guru Govind Singh Indraprastha University, Dwarka, New Delhi, in 2010. Her area of interest includes speech recognition, data mining and machine learning.

1 Introduction

Latent semantic analysis (LSA) is a dimension reduction technique, used to detect hidden semantic themes in data through an algebraic matrix decomposition method called single value decomposition (SVD) (Deerwester et al., 1990). LSA is a fully automatic in true sense and does not use any semantic network, knowledge base, conceptual hierarchy syntactic analyser to reveal set of latent dependence between the words in text or their context. Both the latent concept and their meaning is represented in same semantic space. LSA's power to derive this extraordinary interrelated kind of meaning depends on an aspect of it strong mathematical structure. The most important criteria that influence the performance of text mining techniques are:

a preprocessing of data (like stop-word, stemming)

b weighting of term and document in dataset.

In this paper, we mainly concentrate only weighting of term and documents, in the LSA, to keep the experiment to narrow down mainly for semantic themes detection with high relevance score. In text mining applications vector space model for documents is the most widely used approach. The TF-IDF function includes both the component of vector space, i.e., each word in the vocabulary and also individual document in the corpus (Wu and Salton, 1981). First, it incorporates word frequency in the document using term frequency, and if a word appears many times in the document its TF will be high. In addition, IDF, measures how infrequent a word is in the document. IDF is calculated on the whole corpus or document collection. TF-IDF has been used in many information retrieval, classification, and clustering models (Truica et al., 2016). Weighting is a methodology adopted to make the vector representation of data more appropriate as per their application requirement. This is very useful for efficient results in information retrieval, classification, clustering modelling (Leopold and Kindermann, 2002). In another level, appropriate weighting also present vector representation so beautifully that the burden on information retrieval model, classification algorithm and clustering algorithms can be minimised.

LSA aims to capture the latent concept structure in collection of documents through the co-occurrence analysis using SVD. Our work adds more to this latent concept extraction using three level weight for LSA. It reveals relationship between words and documents and more deeply contextual ideas through words.

The proposed work with four models: NTC, NPC, ATC, APC using three level weight for LSA can be a very powerful tool for text mining application domain specifically in information retrieval.

With the proposed models for LSA, the potential of LSA is increased at term level, document level. These enhanced semantic spaces generated through three level weight,

has produced contextual clusters in corpus spaces without any external help like dictionary WorldNet, SentiNet, etc.

The approach also reduces the computational complexity involved in this three level weight model for LSA. In general, vector space model in which only term frequency weight is used, for N documents the complexity is O(N2).Three level weight model for LSA, has reduce the complexity of organising term information from O(N2) to O(N).

So in totality this research includes:

1 A novel approach based on three level weight produced four weight model known as NTC, NPC, ATC, and APC for LSA in document collection is presented.

2 These four weight model has shown significant improvement in term correlation and document correlation in semantic spaces as compare to traditional term frequency weight approach.

3 These enhanced semantic spaces generated through three level weight, has produced contextual clusters in corpus spaces without any external help like dictionary WorldNet, SentiNet, etc.

1.1 Motivation

The most popular weighting mechanism in vector space models term frequency unable to find the relevance of a term in respect to the query in an information retrieval application. For example, in document collection related to a topic like 'car', the term 'car' will appear almost in all documents. To attenuate the effect of a frequent term, it is important to scale down the term weight with high collection frequency. For this, it is better idea to consider document frequency instead of term frequency, document frequency defined as number of documents in the collection that contain a term *t*. Document frequency of a term used to scale its weights are inverse document frequency (IDF).

$$\text{Inverse document frequency} - idf_t = \log \frac{N}{df_t} \tag{1}$$

Si IDF of a rare term is higher, whereas the *idf* of a frequent term will be low.

$$\text{Probabilistic inverse document frequency} = p - idf = \max\left\{0, \log \frac{N - df_t}{df_t}\right\} \tag{2}$$

In distributional semantics, each document is represented as vector with one component to each term in the vocabulary with a weight for each component given by

$$tf - idf_{t,d} = tf_{t,d} * idf_t \tag{3}$$

This form of document representation is very fundamental in text mining applications including information retrieval model, classification, clustering and in many statistical language modelling techniques. The motivation for multilevel weight comes from the fact that this fundamental document representation carry useless occurrence of one term multiple times if term present multiple times in document collection.so going beyond term frequency a common modification is to use logarithm of the term frequency, which assign a weight given by

$$wf_{t,d} = \begin{Bmatrix} 1+\log tf_{t,d}, IF - tf_{t,d} > 0 \\ 0 \end{Bmatrix} \quad (4)$$

- Sub-linear term frequency scaling: It is called sub-linear term frequency scaling, in this instead of counting just occurrence of terms in document, use logarithm of term frequency which assign weight

$$wf - idf_{t,d} = wf_{t,d} * idf_t \quad (5)$$

- Term frequency normalisation: Maximum term frequency (tf) Normalisation in which the *tf* weights of all terms in a document are normalised by maximum *tf* in that document.

1.2 Paper organisation

In this paper, we are presenting a three level weight mechanism in LSA that is applicable at both term and document level and corpus level to the entire corpus. In Section 2, the prevalent weighting methods adopted in text mining research are discussed. In Section 3, methodology of our novel three level weight for LSA is presented, in Section 4, the proposed approach is evaluated through empirical experiments using two dataset and finally, in Section 5, the paper is concluded with future work.

2 Related work

In text mining application, both in classification and clustering, it is believed that representation of text is most important factor that dominates the text analysis techniques rather than tuning the classification and clustering models. So, motivated by this concept in text mining literature, it is quite useful to analyse the terms discriminating power in the document collection, as a result many supervised kinds of term weighting method are investigated (Leopold and Kindermann, 2002).

Tf-rf is a method which shows experimental evidences for the effectiveness of method (Lan et al., 2005b). These methods are based on local and global weighting at term and document level.

Another major area of work based on collection frequency factor which describes weighting of term using some statistical threshold in document collection These statistics are used both at term level as feature selection known as information gain (Deng et al., 2004), gain ratio factor or odd ratio factor at document level (Debole and Sebastiani, 2003). Another approach in this category is based on confidence interval (Soucy and Mineau, 2005).

In the literature of semantic themes detection and intelligent information retrieval, clustering, and classification model, many papers exist, that uses both supervised and unsupervised weighting approaches Some work also used different variants local and global weighting like term frequency (TF), logarithm, and entropy as local weight at term level and IDF, GF-IDF-global frequency of term and number of documents in which it appears and entropy at global level. Weighting in non-negative matrix factorisation (NNMF) and latent Dirichlet allocation (LDA) (Soucy and Mineau, 2005; Lan et al., 2005a).

An extended LDA model with term weighting in sampling is used to improve and balance topic distribution. This approach adds wonderful power to feature generation for classification (Lee et al., 2015). Tag weighted topic model is another framework for enhanced topic detection in documents (Li et al., 2013). A graph oriented approach to capture the relationship between words and number of contexts of co-occurrence as alternative term weight in LSA and LDA used (Bekoulis and Rousseau, 2016). Local and global term weighting for efficient clustering is also used (Domeniconi et al., 2015). Smart weight have many application in natural language processing application including IOT-based smart parking system (Sant et al., 2021).

An efficient approach by researchers based on artificial intelligence provides multi-criteria node selection for efficient scheduling (Jorge-Martinez et al., 2021). Another technique known as EI-Stream based on the drift detection for optimal classification used majority voting technique (Abbasi et al., 2021).

3 Three level weight for latent semantic analysis (three level weight-LSA)

3.1 Three level weight

Three level weight for LSA got inspiration from Galton (Salton and Buckley, 1988; Buckley et al., 1994). In this weights for term relevancy defined at multiple level in the corpus. These factors are:

- Term frequency: It represents the total number of terms in the corpus.
- Collection frequency: This factor considers or separates relevant document from irrelevant documents.
- Document length: Third factor for text analysis, a cosine normalisation is used to normalise the length of documents in the corpus.

We defined our three level weight as collection of three above said factors.

3.2 Three level – LSA

1 Text files are converted in .CSV format.

2 Data is cleaned by removing all noise like removing missing values and then preprocessed by removing all punctuation symbols, converting all capital letters to small letters, removing stop word symbols, stemming and all whitespaces, etc.

3 After preprocessing Term document matrix is generated from corpus object of dataset for further processing.

4 Four three level weight models (NTC, NPC, ATC, and APC) are applied to term document matrix created in Step 3.

5 SVD function is applied to all weighted matrix generated in step 4 and saved as semantic spaces.

Figure 1 Methodology for three level latest semantic analysis

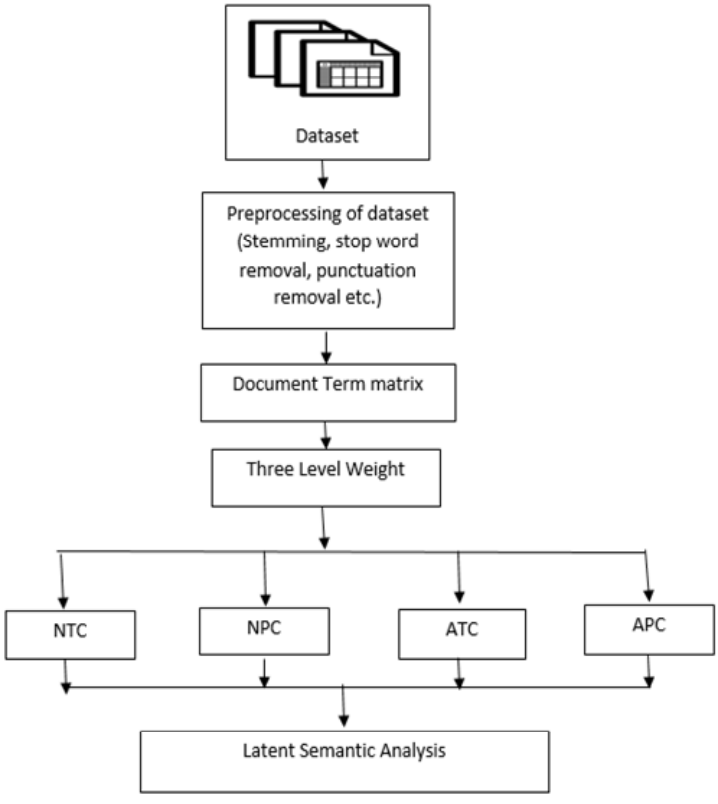
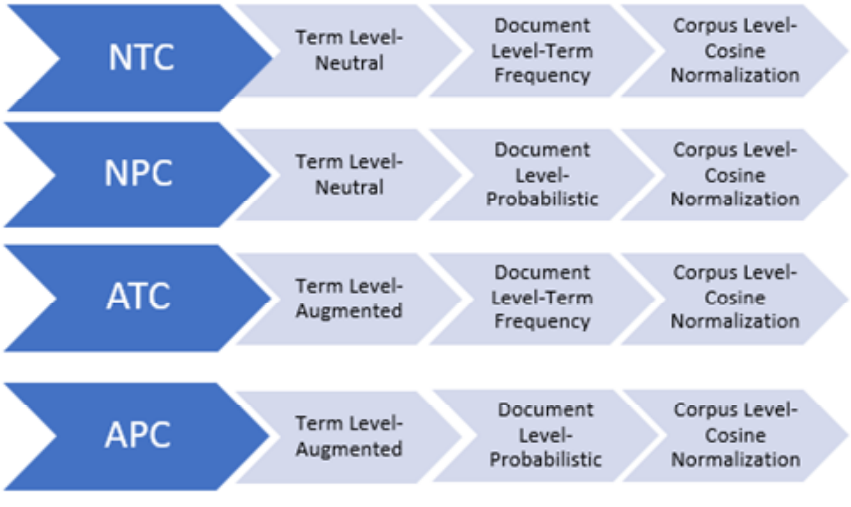


Figure 2 Three level weight model (see online version for colours)



4 Experimental setup

4.1 Dataset

- *Synthetic story dataset*: In this experiment a novel dataset is prepared from collection of some stories as text files. In this total 14 stories are stored, and a csv file is created from these text files. These text files are converted to a corpus object using text mining techniques.
- *BBC-news dataset*: The second dataset is a collection of BBC news stories collected in 2004–2005. The dataset consists of 2,225 documents of five categories – tech, sports, politics, entertainment, Business.
- *Data preprocessing*: The corpus vector of dataset is further processed for necessary preprocessing. Here, the main steps included are convert all capital letters to lowercase, all punctuation symbols removed, removed all stop words like 'the', 'to', 'is', 'was', 'were', etc. defined in English language as stop words, stemming is also performed and whitespaces are removed from corpus.
- *Parameter selection*: After preprocessing of both datasets, we get our data in vector form as term document matrix (M*N) consists of total M terms in the entire corpus and having N documents in the corpus. Term document matrix automatically generates only term frequency weighting. This is the default vector space model used in text mining and natural language processing application domain.

For both the dataset, we get term document matrix for synthetic dataset –Story consists of 4,654 ∗ 14, 4,654 terms in 14 documents with a sparsity of 84%, and for BBC News dataset the term document matrix consists of 2,226 documents with 20,473 terms with a sparsity of 89%.

So in next step we use our novel three level weighting: first model is NTC-defined as natural, i.e., default term frequency, at term level represented in NTC triplet at first level, T is term frequency of collection is taken at document level defined as dft, This consider number of documents containing term t, So in this way through this document frequency of term, scale down the term weight of terms with high collection frequency. Last third factor C is cosine normalisation is taken from cosine similarity between two documents is normalised by product of their Euclidean length. Another weighting models are NPC ATC, APC, Here, we modelled A-Augmented term weighting, P-probabilistic document frequency weighting, T as term frequency as document level instead of term frequency of collection, and third factor in weighting is C-cosine normalisation of document lengths, it is considered as most important factor in all four three level weighting models as shown in Table 1.

Table 1 Smart weight for LSA

Term weighting (tf)	*Document weighting (idf)*	*Normalisation*
$n(neutral) = tf_{t,d}$	$t = idf = \log \frac{N}{df_t}$	$c(cosine) = \frac{1}{\sqrt{w_1^2 + w_2^2 + ...w_m^2}}$
$a = \frac{0.5 + 0.5 * tf_{t,d}}{\max t(tf_{t,d})}$	$P(prob) = \max\left[0, \log \frac{N - df_t}{df_t}\right]$	$c(cosine) = \frac{1}{\sqrt{w_1^2 + w_2^2 + ...w_m^2}}$

After weighting the term document matrices using the proposed three level weight with four models, the SVD is applied, and semantic spaces are generated for each model. Semantic spaces are saved as text matrices. These semantic spaces contain three matrices -tk, dk, and sk known as the term matrix, the document matrix, and the singular value matrix. For term correlation analysis a value-weighted matrix of Terms is calculated by multiplying tk with sk matrix. For keeping our semantic space more revealing of co-occurrence relationship the value of dims = 3 taken in all the experiment in both the dataset.

5 Result and analysis

In this experiment, here we implemented four different version of three level of weighted for LSA. Four weight models are NTC NPC, ATC, APC, with cosine normalisation of document length in all four version. For analysing the impact of these four-three level weight on LSA, we need to evaluate the different semantic space generated by each weight model with SVD process. SVD generates three matrices. These three matrices are term matrix, document matrix and singular value matrix. The term correlation analysis and document correlation is done by utilising these matrices. For term correlation analysis, cosine similarity score between vectors ∑*U matrix (term matrix).is calculated. For document correlation cosine similarity between vectors in ∑*VT matrix (document matrix) is calculated. From this relatedness information, it is possible to organise term and document into groups, means both term and documents in same semantic space. So in this way, all the terms related to a semantic themes in a document clustered together called topic. LSA improve grouping of documents into cluster of interest with high correlation between documents. Semantic space generates three matrices as: term matrix Tk, document matrix Dk and singular matrix Sk.

5.1 Term matrix evaluation

In Figures 3 and 4, term matrices in semantic space generated using four three level weight LSA models are shown in both the dataset, these figures demonstrate the terms in two dimensions. In this semantic space evaluation, it is clearly visible that term matrix plot of weight-ATC() is the best weight mechanism for generating semantic space where terms are equally correlated in both dimensions. In next level is weight-APC(), giving term correlation concentrated on both dimension, in weight NTC terms in both the dimension are correlated, and in NPC terms are less correlated with second dimension, and with default term frequency weighting, the correlation between two dimensions in semantic space become hard to find. Both are independent as shown in Figure 3(e). These highly correlated semantic space generated using proposed three level weight enhanced the semantic themes in the corpus. In this way more cohesive topics are identified.

In Figure 4, term matrix correlation in two dimensions for BBC-news dataset is shown. Here, weight-NTC has highest correlation as maximum terms in dimension one and two has same semantic score. Weight-NPC is at 2nd position in this correlation score and Weight APC has third position in this correlation, and weight APC has the worst correlation as very few terms are correlated, most of terms are scattered in entire semantic space in two dimension.

Figure 3 Term matrix in different semantic space generated by different three level weight, (a)–(e) for story dataset, (a) weight-NTC (b) weight-NPC (c) weight-ATC (d) weight-APC (e) basic-TF (see online version for colours)

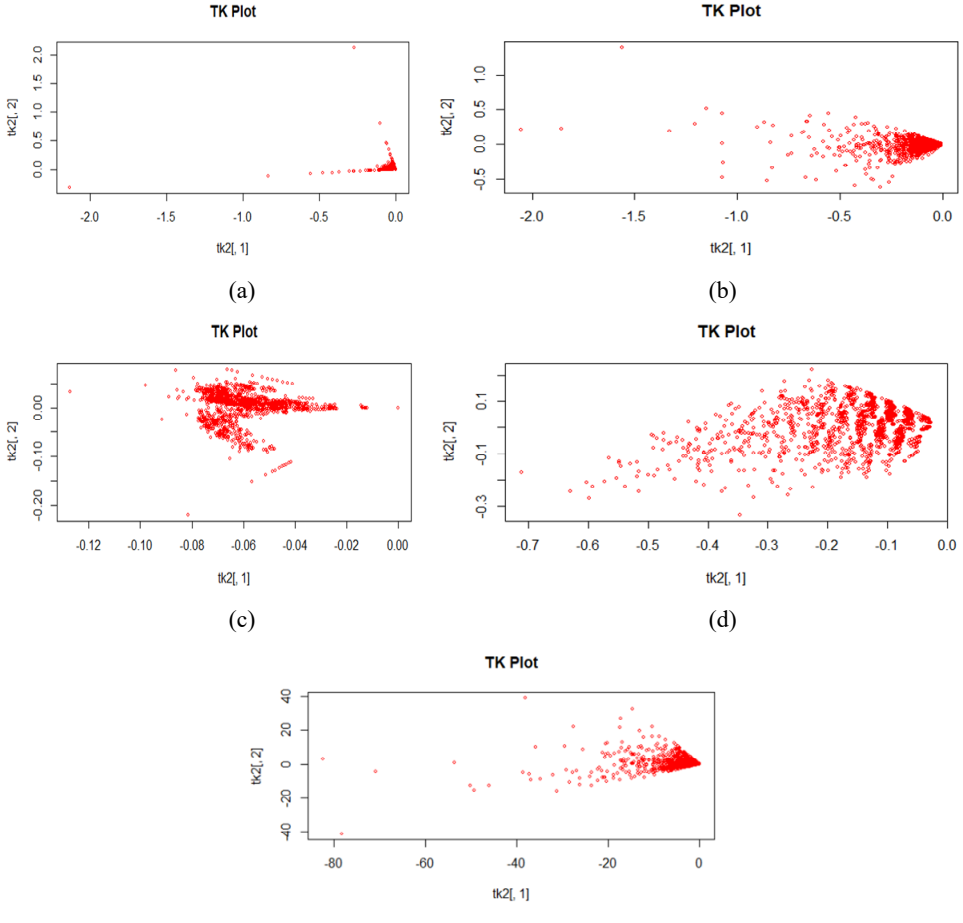


Figure 4 Term matrix in different semantic space generated by different three level weight, (a)–(e) (for BBC-news dataset), (a) weight-NTC (b) weight-NPC (c) weight-ATC (d) weight-APC (e) basic-TF (see online version for colours)

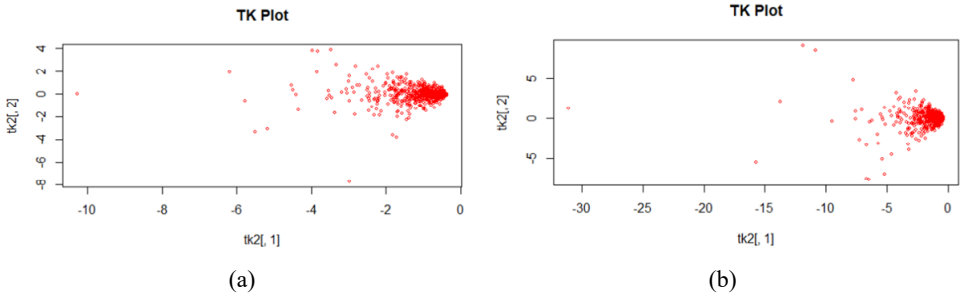
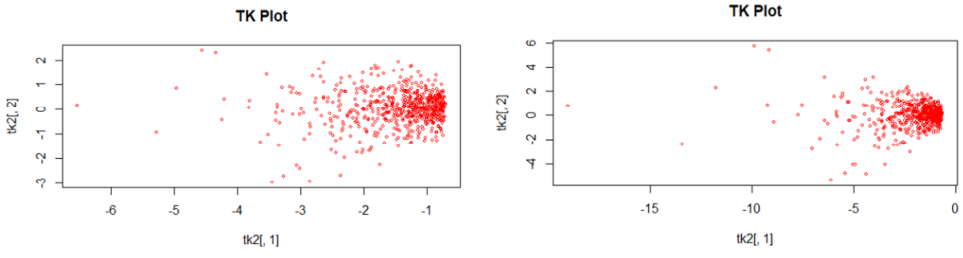
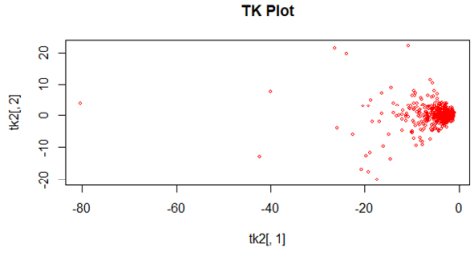


Figure 4 Term matrix in different semantic space generated by different three level weight, (a)–(e) (for BBC-news dataset), (a) weight-NTC (b) weight-NPC (c) weight-ATC (d) weight-APC (e) basic-TF (continued) (see online version for colours)



(c) (d)



(e)

5.2 Document matrix evaluation

In Figures 5 and 6, document matrices generated from multiple three level weight semantic space are shown in Figures 5(a)–5(e) and 6(a)–6(e). These matrices are also plotted in two dimensions. This document matrix shows document in the dataset and how much correlation exist between them. A three level weight LSA try to establish a semantic space where all documents put into a correlated space, this type of semantic analysis can help in many types of text analysis application domains also in bioinformatics. In Figure 5(a), weight NTC make the semantic space, where higher document correlation can be seen, and next weight-ATC, where correlation is good, next in correlation score is weight NPC and in last where least correlation is found in synthetic dataset-story is weight APC. Here as dataset consists of short stories the most common term frequency also shown better correlation in documents better than weight APC.5(b) and 5(d) has document 1 and document 2 are related and also document 3 and 5, and also document 8, document 10, and document 13 are correlated in dimension one as shown. So, our three level weighting has shown great correlation in three proposed three level weight-LSA models in synthetic dataset story.

In Figure 6, the documents correlation in two dimensions for BBC-dataset is shown. Here again weight NTC is best in document correlation in first dimension, at 2nd rank is weight-ATC and at third position is weight APC, then weight NPC and with least correlation is most basic term frequency

Figure 5 Document matrix in different semantic space generated by different three level weight (synthetic dataset-story), (a) weight-NTC (b) weight-NPC (c) weight-ATC (d) weight-APC (e) basic-TF (see online version for colours)

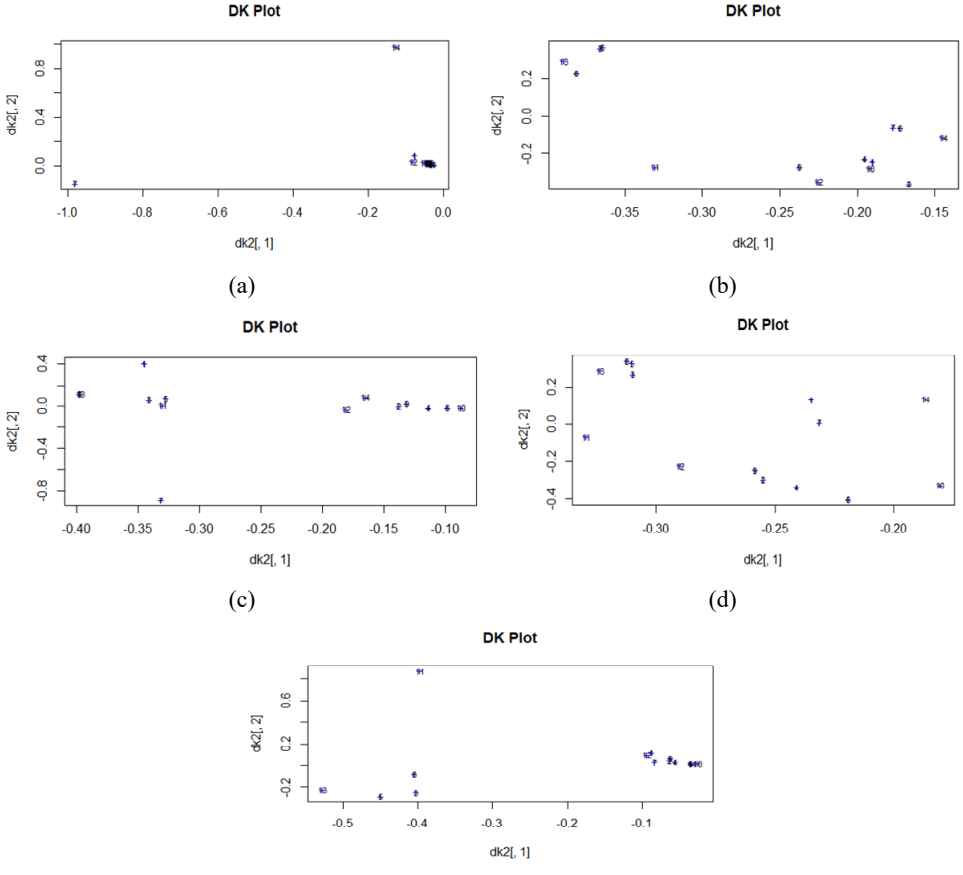


Figure 6 Document matrix in different semantic space generated by different three level weight (for BBC-news dataset), (a) smart weight-NTC (b) smart weight-NPC (c) smart weight-ATC (d) smart-weight-APC (e) basic TF (see online version for colours)

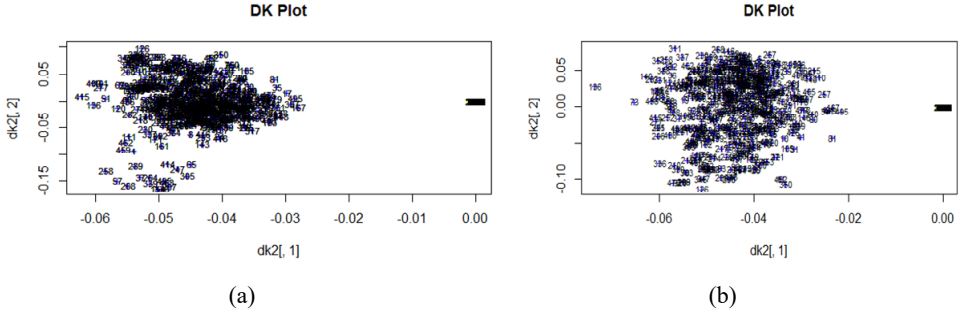
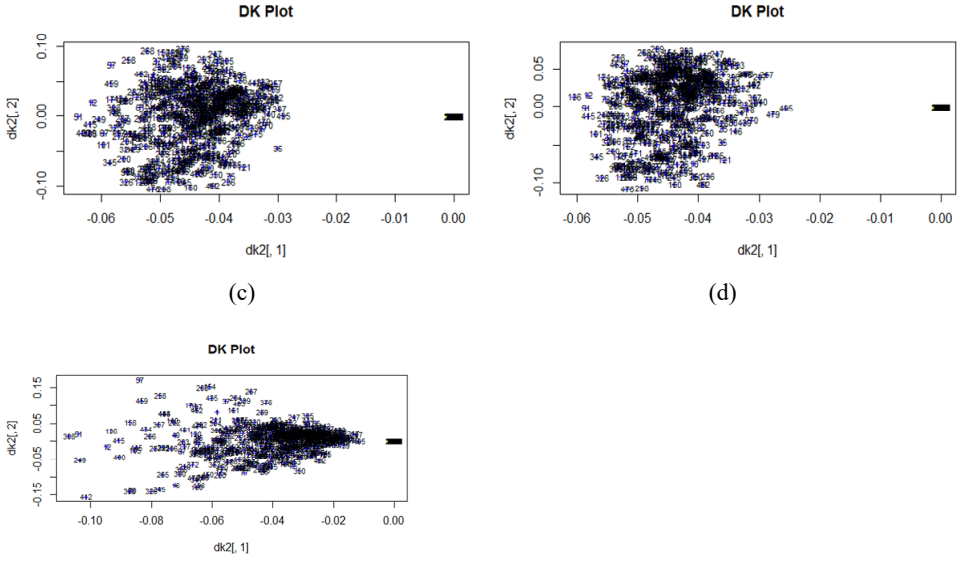


Figure 6 Document matrix in different semantic space generated by different three level weight (for BBC-news dataset), (a) smart weight-NTC (b) smart weight-NPC (c) smart weight-ATC (d) smart-weight-APC (e) basic TF (continued) (see online version for colours)



(c) (d)

(e)

In Figure 7, we have shown the most powerful semantic space given by our proposed three level weight LSA as compare to shown in Figure 7(e), this efficiency of highly semantic correlation is shown using heat maps of all four proposed -weight models. The heat maps demonstrates semantic themes generated from the 20 nearest term to 'trust' in the corpus. This semantic clustering of this term in heat map, shows that 'trust' has two distinct meaning in our corpus. Heat map arranges terms in hierarchical clustering based on the distance or similarity between them. Two colour schemes are used in these heat maps, each colour represents semantic relatedness score between terms related to 'trust', Here, yellow colour depicts highest semantic relatedness (0.8–1.0), orange (0.6–0.8), light red (0.4–0.6), red colour shows similarity score between (0–0.4), In all the four smart weight LSA, max yellow portion indicates that term chosen for 'trust' has high correlation score between(0.8–1.0), and here in Figure 3, the model weight NTC has highest semantic score, and at second rank is NPC and third in ranking ATC and fourth is weight-APC. And it is clearly visible with maximum red colour in weight-TF that normal LSA provides lowest semantic similarity in terms as shown in Figure 3(e).

In Figure 8, the similar behaviour is found for BBC-news dataset, here the term chosen for semantic relatedness is 'economist'. Here again in Figure 4, the model weight NTC has highest semantic score, and at second rank is NPC and third in ranking ATC and fourth is weight-APC.

Figure 7 Term correlation based heatmap as clustering in different semantic space generated by different three level weight for synthetic dataset story, (a) weight-NTC (b) weight-NPC (c) weight-ATC (d) weight-APC (e) basic TF (see online version for colours)

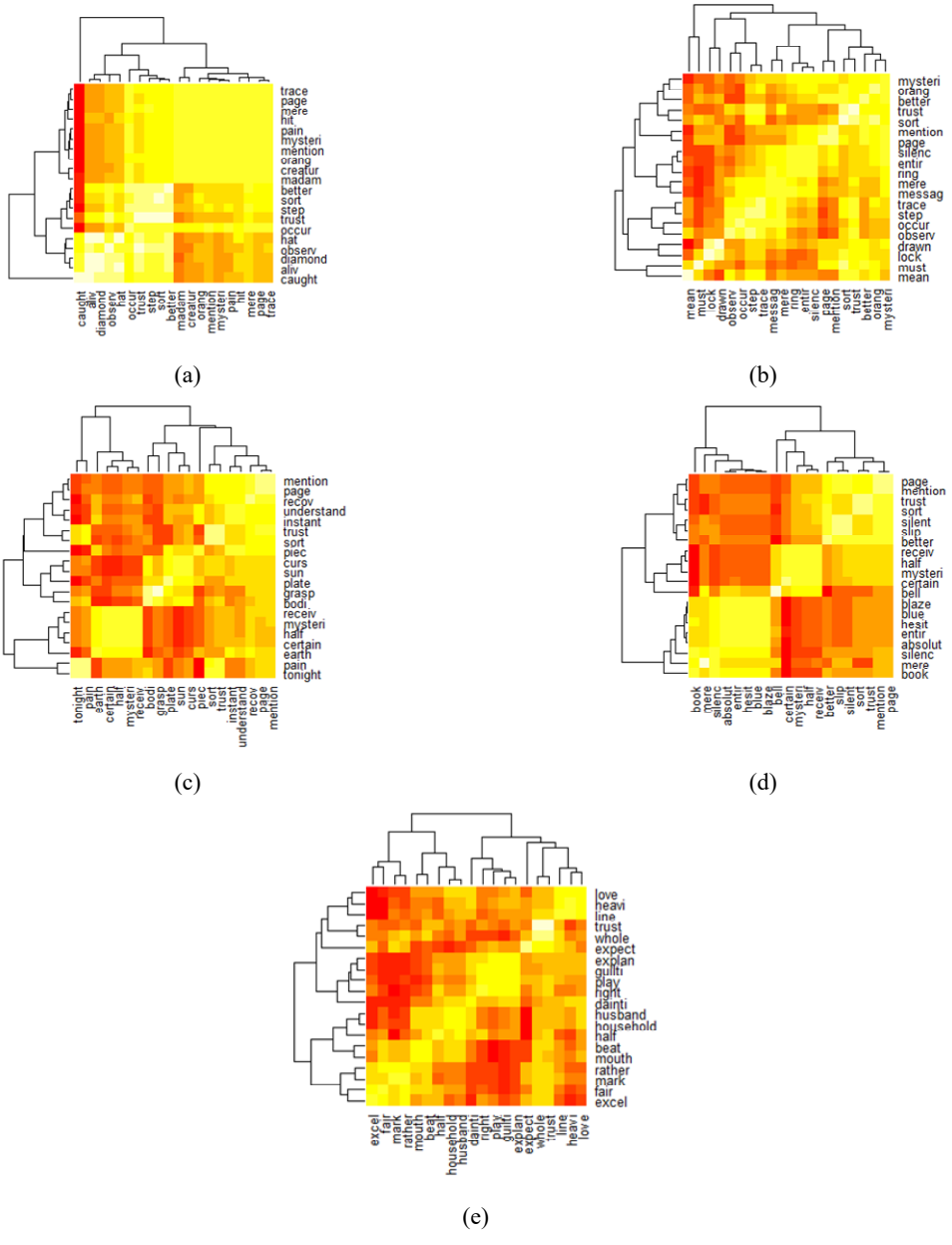
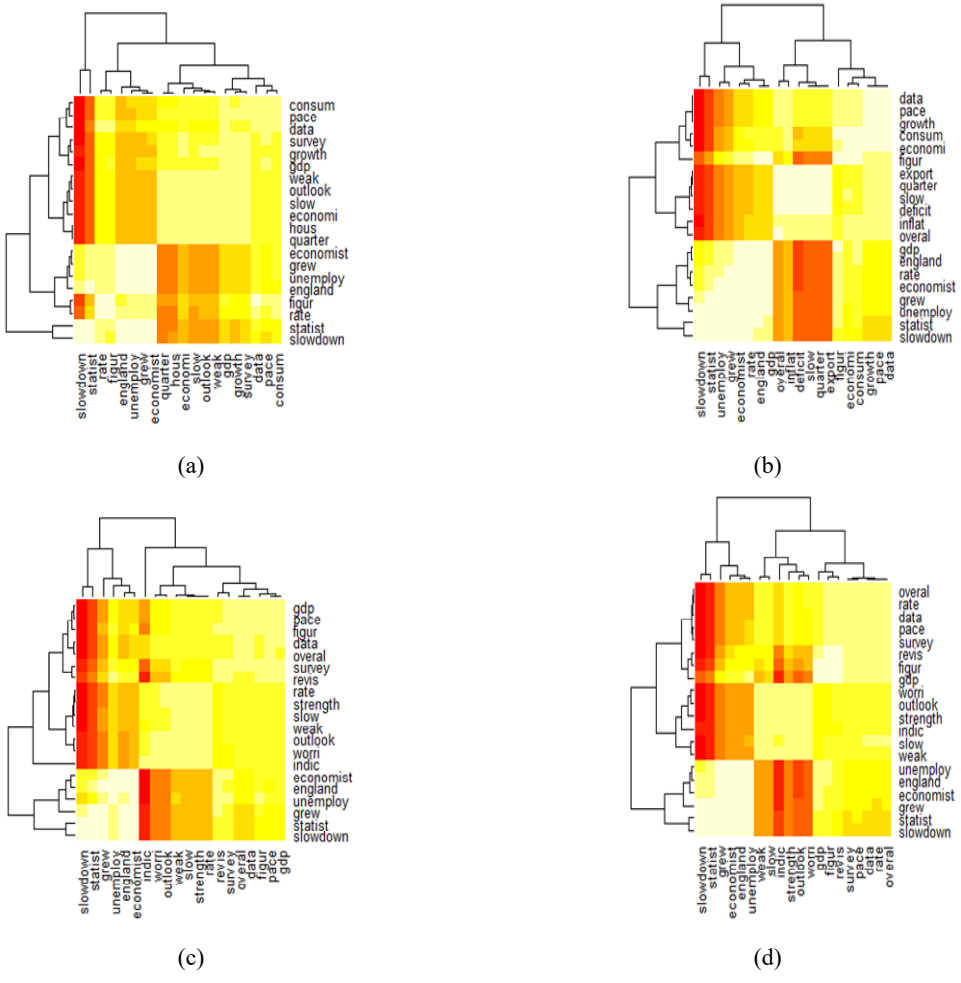


Figure 8 Term correlation based heatmap as clustering in different semantic space generated by different three level weight for BBC-news dataset, (a) weight NTC (b) weight NPC (c) weight AP (d) weight ATC (see online version for colours)



5.3 Term correlation using cosine similarity

In Table 2, the semantic correlation score between terms is shown in different three level weight semantic spaces generated through three level weight LSA models. In these models, we take set of two terms and calculate their similarity score, the chosen similarity measure is cosine, Here, it is clearly visible that in all three term set considered-habit-harm, wisdom-height, chain-chair shown high semantic similarity in all smart weight LSA models, with having highest similarity shown by -weight NTC, and in one case chain-chair, both weight ATC, weight APC shown high correlation. This semantic correlation between terms is in reference to the context of their meaning in dataset, not in their dictionary meaning. The benefit of proposed three level weight for latent sematic analysis, present the terms in sematic spaces contextually so closer, although in dataset the terms are very far from each other. Also each new document

added to the dataset can be projected into the existing semantic space and can be recalibrated with previously generated weights efficiently.

Table 2 Similarity score of terms in all four smart weight LSA models for synthetic dataset story

Chosen term	*Semantic space chosen*				
	Normal (NNN)	*Smart Weight-NTC*	*Smart Weight-NPC*	*Smart Weight-ATC*	*Smart Weight-APC*
Habit, Harm	0.66	*0.99*	*0.94*	*0.72*	*0.96*
Hard work, Tough	0.065	*0.095*	0.015	0.044	*0.084*
Wisdom, height	0.57	*0.97*	*0.82*	*0.77*	*0.83*
Chain, chair	0.53	*0.65*	*0.54*	*0.71*	*0.71*

Table 3 Similarity score of terms in all four smart weight LSA models for BBC-news dataset

Chosen term	*Semantic space chosen*				
	Normal (NNN)	*Smart weight-NTC*	*Smart weight-NPC*	*Smart weight-ATC*	*Smart weight-APC*
Scandal, campaign	0.89	*0.84*	*0.76*	*0.81*	0.18
bank, bankruptci	0.37	*0.56*	*0.55*	*0.52*	0.24
bank, revenu	*0.91*	*0.73*	*0.46*	*0.75*	0.20
Strong, scandal	*0.41*	*0.27*	*0.24*	*0.25*	0.17
economist, analyst	*0.66*	*0.62*	*0.57*	*0.63*	0.39
Agreement, boss	*0.93*	*0.95*	*0.97*	*0.97*	0.23
Prime, condit	*0.86*	*0.91*	*0.77*	*0.72*	0.65
Period, quarter	*0.97*	*0.96*	*0.96*	*0.97*	0.41
Boom, strong	*0.88*	*0.87*	*0.89*	*0.83*	0.17

Table 4 Classification performance of three level weight on LSA

Weight model for LSA	*Dataset*	*Precision*	*Recall*	*F-measure*
NNN (neutral at three level) weight.	Story dataset	0.732	0.741	0.736
	BBC news dataset	0.725	0.720	0.725
NTC	Story dataset	0.76	0.77	0.76
	BBC news dataset	0.83	0.82	0.81
NPC	Story dataset	0.76	0.78	0.77
	BBC news dataset	0.82	0.83	0.82
ATC	Story dataset	0.77	0.76	0.77
	BBC news dataset	0.84	0.83	0.82
APC	Story dataset	0.79	0.79	0.78
	BBC news dataset	0.83	0.81	0.81

5.4 Document classification accuracy

In this evaluation of our three level weight model for LSA, naïve Bayes classifier is used to check the effects of weight models on classification results. Classification results are evaluated using the state of the art measure precision, recall and F-measure. For story dataset, data has two class labels, fiction and non-fiction. For BBC news dataset classified into five categories: tech, sports, politics, entertainment, and business. It is found in this experiment that three level weight model for LSA has great potential in improving classification for multiclass dataset like BBC news in comparison with binary class dataset like story dataset. Although result are improved in all the four model of three level weight in story dataset in comparison with neutral at three level.

6 Conclusions and future scope

In this paper, we explored the concept of three level weights first time in LSA. We proposed four three level weight models for LSA known as:

1 NTC

2 NPC

3 ATC

4 APC.

In these weight models for LSA models, we use three level of weighting, i.e., at term level, document level, and third is document length normalisation. In these four models, we use term frequency, probability based IDF, and augmented term frequency, and term frequency based IDF. It is clearly demonstrated in this experiment that the proposed models have improved results in multiple dimensions in term and document matrix generated by LSA. Highly semantic rich semantic spaces were created by these three level weights for LSA's and enhanced term correlation has found for term pairs existing in documents. This experiment's result will be very useful for intelligent information retrieval application models in future. In the future the work can be extended to improve the performances of other topic modelling techniques like LDA and NNMF.

References

Abbasi, A. et al. (2021) 'ElStream: an ensemble learning approach for concept drift detection in dynamic social big data stream learning', *IEEE Access*, April, Vol. 9, pp.66408–66419.

Bekoulis, G. and Rousseau, F. (2016) 'Graph-based term weighting scheme for topic modeling', in *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, IEEE, December, pp.1039–1044.

Buckley, C., Salton, G., Allan, J. and Singhal, A. (1994) 'Automatic query expansion using SMART: TREC 3', in *Proc. of the Third Text Retrieval Conference*, pp.69–80.

Debole, F. and Sebastiani, F. (2003) 'Supervised term weighting for automated text categorization', in *SAC '03: Proceedings of the 2003 ACM Symposium on Applied Computing*, ACM Press, New York, NY, USA, pp.784–788.

Deerwester, S., Tumais, S.T., Landauer, T.K., Furnas, G.W. and Harshman, R.A. (1990) 'Indexing by latent semantic analysis', *J. Soc. Inform. Sci*., Vol. 41, No. 6, pp.391–407.

Deng, Z-H., Tang, S-W., Yang, D-Q., Zhang, M., Li, L-Y. and Xie, K.Q. (2004) 'A comparative study on feature weight in text categorization', in *AP Web*, Springer-Verlag, Heidelberg, March, Vol. 3007, pp.588–597.

Domeniconi, G., Moro, G., Pasolini, R. and Sartori, C. (2015) 'A comparison of term weighting schemes for text classification and sentiment analysis with a supervised variant of tf.idf', in *International Conference on Data Management Technologies and Applications*, Springer, Cham, July, pp.39–58.

Jorge-Martinez, D. et al. (2021) 'Artificial intelligence-based Kubernetes container for scheduling nodes of energy composition', *International Journal of System Assurance Engineering and Management*, July, pp.1–9.

Lan, M., Sung, S.Y., Low, H.B. and Tan, C.L. (2005a) 'A comparative study on term weighting schemes for text categorization', in *Proceedings of International Joint Conference on Neural Networks (IJCNN-05)*, Montreal, Canada, 31 July–4 August, Vol. 1, pp.546–551.

Lan, M., Tan, C.L., Low, H.B. and Sung, S.Y. (2005b) 'A comprehensive comparative study on term weighting schemes for text categorization with support vector machines', in *WWW '05: Special Interest Tracks and Posters of the 14th International Conference on World Wide Web*, ACM Press, New York, NY, USA, pp.1032–1033.

Lee, S., Kim, J. and Myaeng, S.H. (2015) 'An extension of topic models for text classification: a term weighting approach', in *2015 International Conference on Big Data and Smart Computing (BIGCOMP)*, IEEE, February, pp.217–224.

Leopold, E. and Kindermann, J. (2002) 'Text categorization with support vector machines. how to represent texts in input space?', *Machine Learning*, January–March, Vol. 46, Nos. 1–3, pp.423–444.

Li, S., Li, J. and Pan, R. (2013) 'Tag-weighted topic model for mining semi-structured documents', in *Twenty-Third International Joint Conference on Artificial Intelligence*, June.

Salton, G. and Buckley, C. (1988) 'Term-weighting approaches in automatic text retrieval', *Inf. Process. Manage.*, Vol. 24, No. 5, pp.513–523.

Sant, A. et al. (2021) 'A novel green IoT-based pay-as-you-go smart parking system', *CMC Comput. Mater. Cont.*, Vol. 67, No. 3, pp.3523–3544.

Soucy, P. and Mineau, G.W. (2005) 'Beyond tfidf weighting for text categorization in the vector space model', in *IJCAI*, July, Vol. 5, pp.1130–1135.

Truica, C.O., Radulescu, F. and Boicea, A. (2016) 'Comparing different term weighting schemas for topic modeling', in *2016 18th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, IEEE, September, pp.307–310.

Wu, H. and Salton, G. (1981) 'A comparison of search term weighting: term relevance vs. inverse document frequency', in *SIGIR'81: Proceedings of the 4th Annual International ACM SIGIR Conference on Information Storage and Retrieval*, ACM Press, New York, NY, USA, pp.30–39.