International Journal of Engineering Sciences & Research Technology

(A Peer Reviewed Online Journal)
Impact Factor: 5.164

IJESRT

RESEARCHERID
THOMSON REUTERS

[Mahajabin* *et al.*, 9(2): February, 2020]
IC™ Value: 3.00

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

DERIVATION OF PERTINENT ABILITIES FOR PILOTS USING LSA & SVD: A TEXT MINING APPROACH

Mistu Mahajabin, Scientist
Defence Institute of Psychological Research (DIPR), DRDO, India

DOI: 10.5281/zenodo.3692921

ABSTRACT

This paper discusses an experimental study on 'abilities required for the pilots' data using the Latent Semantic Analysis (LSA)/ Singular Value Decomposition (SVD) technique for text extraction. LSA has been used for structuring the documents as Term Document Matrix (TDM). Term-Frequency-Inverse Document Frequency (Tf-idf) has been used for weighting the terms in TDM. Then Singular Value Decomposition (SVD) has been applied for dimension reduction and calculated U dot S to get the importance of the terms in the whole documents corpus. Finally, words are grouped using cosine similarity method. 14 abilities have been derived as a result of the study.

KEYWORDS: Text analysis, LSA/SVD, NLP, TDM, Tf-idf.

1. INTRODUCTION

Documents text extraction is one of the important topics in Natural Language Processing (NLP). This research aims to determine the internal meaning or semantic of the terms/words used in the documents. The internal meaning of a word changes depending on the context the word is being used at the same time completely different set of words can be used to convey the same meaning. For example, let's take two sentences `Human management is a very important quality for a pilot ' and `Crew resource management is a required quality to be a pilot' express the same meaning. Semantic analysis is used to bring these two sentences into the same concept space. In this research, the data are in the form of written documents. The documents have been written by the pilots. They have written the qualities, a pilot should possess. Every word has a different purpose in the documents but few stand out as principal purposes. The objective of this research is to find out the principal purpose of a document and to find similar purposes in other documents.

2. THEORETICAL REVIEW

Text analysis

Text analysis is a process of extracting valuable information from a document. Text analysis can be done manually but it is an inefficient method for a huge amount of data. The process of text analysis involves a few sub-processes like structuring the input text, deriving the pattern in the structure, and finally stemming and interpreting the output. This technology is being widely used in various sectors like data mining, search engine, social media monitoring, and semantics analysis. In this investigation, text analysis has been used as semantic analysis.

Latent semantic analysis (LSA)

LSA is a mathematical model for automatic semantic analysis [3]. It is neither traditional NLP nor an artificial program and does not use any resources like a dictionary, semantic network or any kind [2]. It takes raw text documents as input and follows some steps before producing output without any human intervention; parse it into words then stem those to concentrate more on the concept. Term document Matrix (TDM) is built considering the stop words and the special characters don't need to be put into the matrix. In TDM each row is represented as terms, each column is represented as documents and each cell contains the number of times the words appear in the corresponding document. The number of times a specific word appears in a document is put

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

in the corresponding TDM cell, apply Tf-idf to weighting the terms on TDM then SVD for dimension reduction finally interpretation.

Term-Frequency-Invers Document Frequency (Tf-idf)
Tf-idf is one of the most popularly used weighting techniques for information retrieving [10]. It calculates how important a word is, in a document among the whole documents corpus. It gives more weight to the words that appear less than the most commonly used words. In Tf all words are weighted as equal importance but in a document, few words appear often but have less importance. So Inverse Document Frequency (idf) is calculated to give less weight to the more frequent words and more weight to the rare and meaningful words. Tf-idf is calculated as below.

$$\text{Tf}(w) = \frac{\textit{The number of times word w appear in a document}}{\textit{Total number of words in that document}}$$

$$\text{Idf}(w) = \log_e \frac{\textit{Total number of documents}}{\textit{The number of documents where w appears}}$$

$$\text{Tf-idf}(w) = \text{Tf}(w) \,.\, \text{Idf}(w)$$

Singular Value Decomposition (SVD)
SVD is a method, use in linear algebra to factorize a complex matrix. It breaks a rectangular matrix M into the product of three matrices. The left singular matrix U where the columns of U are the orthonormal eigenvector of MMT, The right singular matrix V where the columns of V are the orthonormal eigenvector of MTM and ∑ is a diagonal matrix with the square roots of eigenvalues of U or V in descending order.
SVD of matrix M can be written as below where m and n are dimensions of matrices [6].

$$M_{m\,x\,n} = U_{m\,x\,m} \sum_{m\,x\,n} V^{t}_{n\,x\,n}$$

As the values of the S matrix are in decreasing order so for dimensionality reduction *m* and *n* can be selected as required dimension.SVD has a wide range of application like solving homogeneous linear equations, Total least-squares minimization, and matrix dimensionality reduction; has been used for this investigation.

Co-sine similarity
Cosine similarity is the most popular metric to measure the similarity between two term-vectors [7]. It is often used with Tf-idf for information retrieval. It is the value of cos Θ where Θ is the angle between the two vectors. The value of cos Θ lies between 1 and 0. If cos Θ =1 which happens only when Θ=0, the distance between two vectors is 0. So they are similar vector. For all other values of Θ, cos Θ<1. Two term-vectors are more similar when the cosine similarity value of those two vectors is near to 1 and if near to 0 they are dissimilar. Cosine similarity between two vectors can be shown, mathematically as below [8][9]

$$\text{Cosine similarity} = \cos\Theta = \frac{\vec{A}\times\vec{B}}{|\vec{A}||\vec{B}|}$$

3. METHODOLOGY

3.1 Data collection and pre-processing
Data have been collected from pilots in hard document form. The documents have been converted into soft forms. These data contain the qualities a pilot should possess and should not possess. It was optional for pilots to write the qualities required and not required in a pilot. This research is done with only the qualities required for a pilot. Total 608 pilot's data have been collected, among which only 142 pilots have written the qualities pilot should possess.
All the special charters have been removed from the whole data corpus.

3.2 Tokenize and stop word remove
Tokenization is breaking the sentences into words/token. Then match the words with the set of stop words. Stop words are those words that don't carry much information for concept building. The stop words for this study are
Stop words = [all, pilot, quality ,when, what, which, who , I , will, must, than, are, should, then, with, vice, versa, their ,the, a, an, be, new, he, for ,and, at, as, it, in, on, of, to, i.e., is, or, from, if, also, that, his,by]

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

After stop word removes 287 words/terms have been received; considered only the words which have been appeared more the once in the whole documents set.

3.3 Stemming

Stemming is the process of getting the main word from an extended word which generates concept. It is done by removing suffix and prefix from a word. As an example after stemming of the word 'writing' we get 'write' as the main word. After stemming a total of 271 words/terms have been considered for further study.

3.4 TDM creation and apply Tf-idf

TDM has been built in the next step. At first, we have parsed the document and build TMD. There are 271 rows (words) and 142 columns (documents) in TMD. Then Tf-idf has been applied as a weighing technique on TDM.

3.5 SVD calculation

After calculating SVD of the weighted TDM the dimension of the decomposed matrices are Term-term matrix (U) is 271x271, significance matrix (∑) is 271x142, and document- document matrix (Vt) is 142x142.

3.6 Dimension reduction

Dimension reduction is performed on the singular matrix to eliminate less significant or unwanted dimensions from the matrix. 90 % of the entropy in ∑ [5] has been taken dimension reduction. That is the sum of the squares of the present singular values should be at least 90% of the square of all singular values together. In this process, k largest singular values are kept and remainders are set to 0. The value of the k for our finding is 72.

3.7 Interpretation

Finally calculated the dot product of U with S to get the importance of the words.The words with co-sine similarity >0.5 [2] have been taken together and manually named the concepts as Inquisitiveness, Decision making, Information processing, Assertiveness, Emotional stability, Resilience/Hardiness, Analytical, Fearlessness, Disciplined, Substantively, Logical, Self-control, Mental stamina, Perceptual ability.

4. CONCLUSIONS

The purpose of this research was to find out the abilities required for the pilots from the open-ended question. 7 steps LSA model has been performed and got a successful result as 14. LSA provides results close to human evaluators' results [1] so the accuracy of this work has not been evaluated. Calibrations in the model may give a more efficient result. Particularly lemmatizers will give better results than stemmers.

5. ACKNOWLEDGEMENTS

I would like to express my deep gratitude to Dr. K Ramachandran, Director DIPR for allowing me to publish this paper. I wish to acknowledge the help provided by A K Dhamija, scientist DIPR for his valuable and constructive suggestions during the planning and development of this research work. I am particularly grateful for the assistance given by Dr. D. Ravi, scientist DIPR for sharing the data and naming the final abilities.

REFERENCES

[1] F Alves dos Santos, J.C. & Favero, E.L. J Braz ComputSoc (2015) 21: 21. doi:10.1186/s13173-015-0039-7
[2] Landauer, T. K., Foltz, P. W., &Laham, D. (1998).Introduction to Latent Semantic Analysis. Discourse Processes, 25, 259-284.
[3] April Kontostathis & William M. Pottenger, Ph.D. "Detecting Patterns in the LSI Term-Term Matrix" Lehigh University, Bethlehem, PAV. Chakravarthy, X. Li, Z. Wu, M. Temple, and F. Garber, "Novel overlay/underlay cognitive radio waveforms using SD-SMSE framework to enhance spectrum efficiency—Part I," IEEE Trans. Commun., vol. 57, no. 12, pp. 3794–3804, Dec. 2009.
[4] S. G. Chodhary1, Rajkumar S. Jagdale2, Sachin N. Deshmukh 3 "Semantic Analysis of Tweets using LSA and SVD" International Journal of Emerging Trends & Technology in Computer Science (IJETTCS) (2016)

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

[5] Russ Albright, Ph.D."Taming Text with the SVD " SAS Institute Inc., Cary, NC (January, 2004)
[6] Panagiotis Symeonidis and Ivaylo Kehayov and Yannis Manolopoulos "Text Classification by Aggregation of SVD Eigenvectors" Aristotle University, Department of Informatics, Greece
[7] Anna Huang "Similarity Measures for Text Document Clustering" Department of Computer Science The University of Waikato, Hamilton, New Zealand
[8] AlfirnaRizqiLahitani1), Adhistya Erna Permanasari2), Noor AkhmadSetiawan3) "Cosine Similarity to Determine Similarity Measure: Study Case in Online Essay Assessment" Department of Electrical Engineering and Information Technology, Faculty of Engineering , Universitas Gadjah Mada, Indonesia , 2016
[9] Faisal Rahutomo*, Teruaki Kitasuka, and Masayoshi Aritsugi "Semantic Cosine Similarity" Graduate School of Science and Technology, Kumamoto University 2012
[10] MingyongLiu1+ and Jiangang Yang "An improvement of TFIDF weighting in text categorization" Zhejiang University 2012.