International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 03 | Mar 2020 www.irjet.net p-ISSN: 2395-0072

Conversion of Unsupervised Data to Supervised Data using Topic Modelling

Dhamodharan R[1], Chedhella Sai Goutham[2], Kavin kumar B[3], R.M.Shiny[4]

[1,2,3] Department of Computer Science and Engineering, Agni College of Technology
[4]Assistant Professor, Computer Science and Engineering Department, Agni college of Technology

---

Abstract - *Over the past five years, topic models have been applied to research as an efficient tool for discovering latent and potentially useful content. The combination of topic modeling algorithms and unsupervised learning has generated new challenges of interpret and understanding the outcome of topic modeling. Motivated by these new challenges, this paper proposes a systematic methodology for an automatic topic assignment for an unsupervised dataset. Relations among the clustered words for each topic are found by word similarities to each other. Clustered words are generated by NMF. To demonstrate feasibility and effectiveness of our methodology, we present Amazon Product Review. Possible application of the methodology in telling good stories of a target corpus is also explored to facilitate further research management and opportunity discovery. In addition to this we have perform Sentimental analysis and Wordcloud to get a deep insight into the data.*

***Key Words*: Topic Modeling, Methods of Topic Modeling, Latent Dirichlet Allocation (LDA), Non- Negative Matrix Factorization(NMF), Sentimental Analysis, Word Cloud.**

1. INTRODUCTION

Analytics Industry is all about obtaining the "Information" from the info . With the growing amount of knowledge in recent years, that too mostly unstructured, it's difficult to get the relevant and desired information.But, technology has developed some powerful methods which may be wont to mine through the info and fetch the knowledge that we are trying to find .

One such technique within the field of text mining is Topic Modelling. As the name suggests, it's a process to automatically identify topics and to derive hidden patterns exhibited by a text corpus. Thus, assisting better decision making.

Topic modeling is an unsupervised approach used for finding and observing the bunch of words (called "topics") in large clusters of texts.

Topic Modeling is extremely useful for document clustering, organizing large blocks of textual data and information retrieval from unstructured data. for instance – NY Times news are using topic models to boost their user – article recommendation engines. they're going to arrange large datasets of emails, customer reviews, and user social media profiles.

2. RELATED WORK

For the proposed system we have studied some papers which are related to topic modeling. Some of the related papers have been described below.

Kedar.S et al, [2] proposed a "Augmented Latent Dirichlet Allocation Topic Model With Gaussian Mixture Topics". In this work the LDA topic model that can handle data over a continuous domain, but discrete approximations to continuous data can lead to loss of information.

S. Sendhil Kumar et al,[3] worked on "Generations of Word Clouds Using Document Topic models". In this work Document topic modeling approach had proposed to generate topics and word cloud, but very difficult to access the data from corpus.

Mehdi Allahyari et al,[4]analyzed on "Discovering coherent topics with entity topic models". In this work the EntLDA with a regularization framework used to integrate the probabilistic topic models with the knowledge graph of the ontology, but ignores the rich information carried by entities.

Halima Banu et al, [5] presented "Trending Topic Analysis Using Novel Sub Topic Detection Model". In this work trending topic analysis system has been developed that is able to analyse twitter hot topics in a constructive manner, but does not scale up to user's expectation as it does not provide any analyzed summary.

3. LATENT DIRICHLET ALLOCATION

LDA assumes documents are produced from a mix of topics. Those topics then generate words depending on their probability distribution. Given a corpus, the Latent Dirichlet Allocation backtracks and tries to find out what topics would create those documents within the first place.

LDA is a matrix factorization technique. In vector space, any corpus (set of documents) are often represented as a document-term matrix. the subsequent matrix shows a corpus of N documents D1, D2, D3 … Dn and vocabulary size of N words W1,W2 .. Wn. The final outcome of i,j cell gives the frequency count of word Wj in Document Di.

It Iterates through each word "w" for each & every document "d" tries to regulate this topic – word assignment with a replacement assignment. A replacement topic "k" may

be assigned to word "w" with a probability P which is a product of two probabilities p1 and p2.

	W1	W2	W3	Wn
D1	0	2	1	3
D2	1	4	0	0
D3	0	2	3	1
Dn	1	1	3	0

Where P1 – p(topic t/document d) = the proportion of words in document d that are currently assigned to topic t and P2–p(word w/topic t)=Assignments to topic t over all documents that come from this word 'w'.

Below there is a python code for Topic modelling using the Latent Dirichlet Allocation Algorithm and the sample output presented in spyder application by giving the input of amazon product reviews.

CODE:

```
def topicassign():
    stoplist = stopwords.words('english')
    stoplist.extend(['omitted', 'voice','missed','call','video','deleted','media','messag
                     'like','love','well','bought','get','better','much','would','good',
    cv = CountVectorizer(max_df=0.95, min_df=2, stop_words='english')# Assign the value f
    dtm = cv.fit_transform(dataset['text'])# Transforming cummerlist to tfidf formula
    LDA = LatentDirichletAllocation(n_components=5,random_state=42)#Assigning number of t
    LDA.fit(dtm)
    print('**********************************************')
    ttopic=[]
    for index,topic in enumerate(LDA.components_):
        print(f'THE TOP 15 WORDS FOR TOPIC #{index}')
        result=([cv.get_feature_names()[i] for i in topic.argsort()[-10:]])
        print(result)
                    #ff=[]
                    #ttopic[]
        indexs=tt2(result)
        print(result[indexs])
                    #indexs=tt2(result)
        ta=result[indexs]
        print(indexs)
        results.append([ta,result])

        ttopic.append(result[indexs])
        #results.append([filename,result])
        print('\n')
```

OUTPUT:

4. NON-NEGATIVE MATRIX FACTORIZATION

In the proposed system, Non-negative Matrix Factorization(NMF) has been applied to a matrix of Term Frequency-Inverse Document Frequency (TF-IDF) which is used to extract topics from large collections of text.

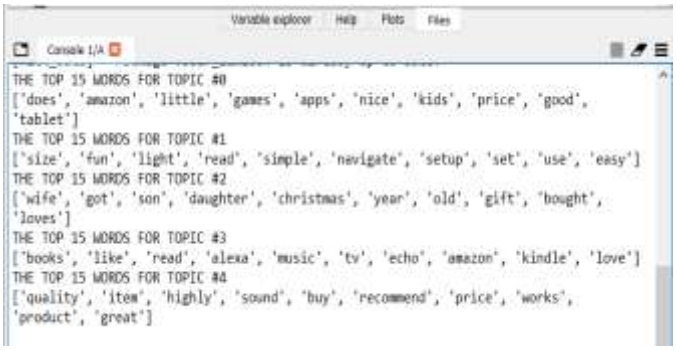
In many applications,The NLP go through the crucial step is to transforming the words into machine- readable numerical vectors. Where TF-IDF fulfills this role with an extra feature: it also gives us a measure of how important a word is to a document in the corpus

Below there is a proposed Formula for Automatic Topic Modelling for Unsupervised data based on the Non-Negative Matrix Factorization with the Term Frequency – Inverse Document Frequency and a sample output for the product reviews which is an Unsupervised data.

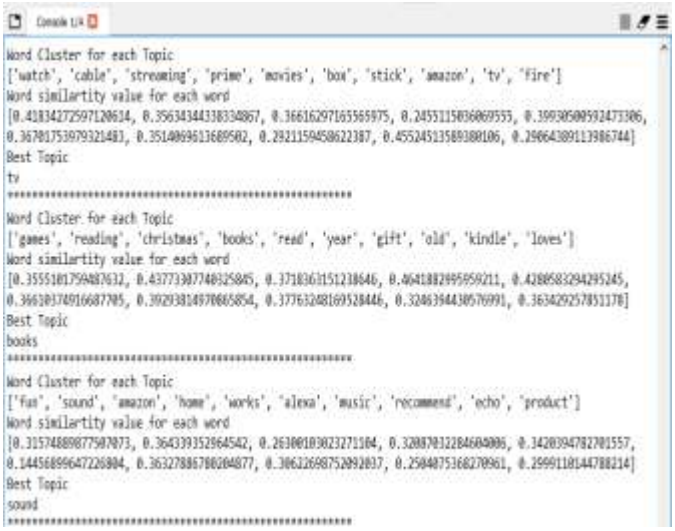
FORMULA:

$$Cluster\ Word = Series(Fw_i)$$
$$where, Fw_i = Feature\ Word$$
$$where\ i = 0\ to\ 10$$
$$Sim_j = \sum_{i=1}^{10}(Fw_i . \varsigma\ Fw_{1\ to\ 10})/total\ length$$
$$where\ j = Each\ cluster\ count$$
$$where\ \varsigma = Word\ Similarity$$
$$Tw = MaxValue[Sim_j]$$
$$Where,\ Tw = Best\ Topic\ for\ each\ cluster$$

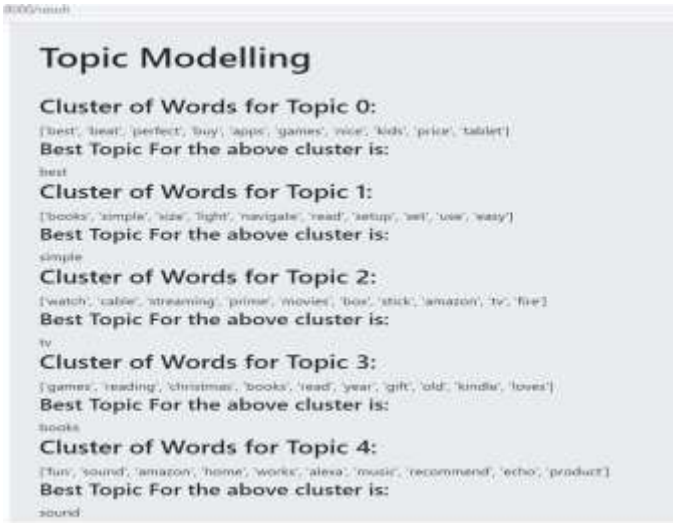
OUTPUT:



Below there is an example of Word similarity value for each word in the cluster.



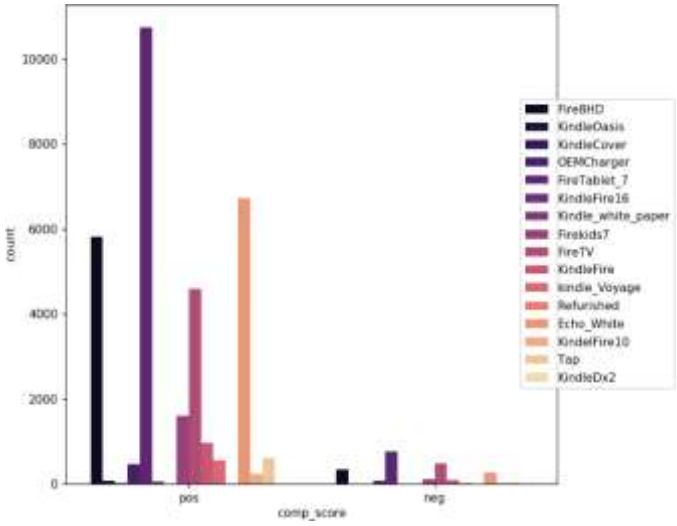
In the below picture represents the Result of the Web Development where the best topic for every 10 words of cluster is shown as the output based on the NMF with the TF-IDF algorithm



5. SENTIMENTAL ANALYSIS:

In the corpus or the product reviews there is high probability of finding the similar words which is results in producing the less accurate topics. So we implement the sentimental analysis to over come the problem of similar words.

In the proposed system will represent the sentimental analysis in the bar chart like X - axis represent positive and negative of the words and Y-axis represent the words count which are encounter in the corpus



6. WORD CLOUD:

word cloud is a visualization of word frequency in a given text as a weighted list. This technique has recently been popularly used to visualize the topical content of product reviews, political speeches. Where the big font size of the word represents the most frequently encountered in the unstructured data and the small font size word represent the less frequently encountered in corpus or unstructured data.

7. CONCLUSIONS

The proposed system gives a better result for the combination of TF-IDF Vectorization with Non-Matrix Factorization. Using this combination automatic labeling has achieved.

The proposed system successfully converted unsupervised data into supervised data using Automatic Labelling with Topic modeling. This creates a reasonable impact on Topical modeling domain.

In addition to this, sentimental analysis and word cloud also performed because of the product review dataset which creates data insight for business development.

REFERENCES

[1] Hongshu Chen, Ximeng Wang , Shirui Pan , and Fei Xiong had proposed "Identify Topic Relations in Scientific Literature Using Topic Modeling , " 2019.

[2] Kedar S. Prabhudesai, Boyla O. Mainsah, Leslie M. Collins, and Chandra S. Throckmorton , "Augmented Latent Dirichlet Allocation(LDA) Topic Model With Gaussian Mixture Topics " , Department of Electrical and Computer Engineering, Duke University, Durham, NC, USA,2018.

[3] S. Sendhilkumar, M. Srivani,G.S. Mahalakshmi, "Generation of Word Clouds Using Document Topic Models", 2017, Anna University Chennai, India.

[4] Mehdi Allahyari,Krys Kochut,"Discovering Coherent Topics with Entity Topic Models",2016,Computer Science Department University of Georgia, Athens, GA, USA.

[5] Halima Banu S and S Chitrakala,"Trending Topic Analysis Using Novel Sub Topic Detection Model", 2016, Department of Computer Science and Engineering, Anna University, Chennai, Tamil Nadu, India.

[6] I. Ketata, W. Sofka, and C. Grimpe, "The role of internal capabilities and firms' environment for sustainable innovation: Evidence for Germany," R&D Manage., vol. 45, no. 1, pp. 60–75, 2015.

[7] C. K. Yau, A. Porter, N. Newman, and A. Suominen, "Clustering scientific documents with topic modeling, "Scientometrics, vol.100, no.3, pp.767– 786, 2014.

[8] A.McAfee, E.Brynjolfsson, T.H.Davenport, D.Patil, and D. Barton, "Big data: The management revolution," Harvard Bus. Rev., vol. 90, no. 10, pp. 60–68, 2012.

[9] Y.-H. Tseng, C.-J. Lin and Y.-I. Lin, "Text mining techniques for patent analysis," Inf. Process. Manage., vol. 43, no. 5, pp. 1216–1247, 2007.

[10] S.W.Cunningham, A.L.Porter, and N.C.Newman, "Special issue on tech mining, "Technol Forecasting Social Change,vol.8,no.73,pp.915–922, 2006.

[11] A. L. Porter and S. W. Cunningham had proposed Tech Mining: Exploiting New Technologies for Competitive Advantage. Hoboken, NJ, USA: Wiley, 2004.

[12] R. N. Kostoff, D. R. Toothman, H. J. Eberhart, and J.A. Humenik had proposed "Text mining using data base tomography and bibliometrics: Areview, "Technol. Forecasting Social Change, vol. 68, no. 3, pp. 223–253, 2001.