AN EFFICIENT NOVEL APPROACH FOR SOCIETAL COMMUNICATION OF Q&A COMMUNITY SYSTEM USING TOPIC MODELING TECHNIQUES

Venkateswara Rao P[1],
[1]Research Scholar of Computer Science and Engineering, Jawaharlal Nehru Technological University, Anantapur A.P
*Email: pvenkat2004@gmail.com , *Corresponding author*

A.P Siva kumar[2],
[2]Assistant Professor of Computer Science and Engineering, Jawaharlal Nehru Technological University, Anantapur, A.P
Email: sivakumar.ap@gmail.com

Abstract: *The emerging trend in technical research is to use customer-generated data collected by community media to probe community opinion and scientific communication on employment and care issues. This review of the collected data, the launch of a question-and-answer social website, is a separate stack for exploring the key factors that influence public preferences for technical knowledge and opinions. by means of a web search engine, topic modeling, and regression data modeling, this study quantified the effect of the response textual and auxiliary functions on the number of votes received with the response. Compared to previous studies based on open estimates, the model results show that Quora users are more likely to only talk about technology. It can fail when the keywords in the query do not match the text content of large documents that contain relevant questions of existing methods, ie. CNNMF and NMF, as well as some restrictions are not enough. Also, users are often not experts and provide ambiguous queries leading to mixed results and encountering problems with existing methods. To address this problem, in this article we propose a Hadoop model, distributed using semantics, non-negative matrix factorization (HDiSANNMF), to find topics for short texts. It effectively incorporates the semantic correlations of the word context into the model, where the semantic connections between words and their context are learned by omitting the grammatical view of the corpus. The researchers are trying to reorganize the main results and present modern techniques for modeling distributed themes to address technologies and platforms with increasing attributes, as well as how much time and space it takes to generate the model. This document briefly describes the structure of public questions and answers around the world and tracks the development of the main topics Housing and employment opportunities for next generation technologies in the world in real time.*

Keywords: *LDA, NNMF, HdiSANNMF, Hadoop, Topic models, quora, stack overflow, Twitter API, NLTK.*

1 Introduction

Large amounts of short text are generated every day, such as tweets, search queries, questions, image tags, keywords, headlines, and more. They have played an important role in our daily life. Discovering the knowledge of topic becomes an interesting but challenging research task that has garnered a lot of attention. Consequently, research and improvement of big data processing frameworks is increasing rapidly [16]. This comment looks at one of the promising open source software frameworks [17]. Hama is a distributed computing framework based on group synchronous equivalent computing for performing a variety of massive daily computational tasks in graphs, matrices, deep learning, machine learning, and network algorithms. It is written in Java and is on the Hadoop Distributed File System (HDFS), which makes it fully compatible with Hadoop clusters. Software engineers and programmers often search for answers to questions using websites like Stack Overflow. Analysis of this data [1] can provide insight into which aspects of programming and APIs are most difficult to understand. In this comment to classify

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

redundant stack problems that are of interest to two overlapping views, which is the programming concepts and the type of information sought. The flood attack contains a large amount of information related to a wide range of computer programming topics.

2 Related Works:

The open question and answer software systems have a Start system developed by the Info Lab group of the Laboratory for Computer Science and Artificial Intelligence at MIT. However, the question-and-answer systems used to solve assignments for a particular course are very rare. As a result, an intelligent question-and-answer system has been developed that returns answers to users' questions according to the principles of a course based course[3][19].

In this article, we introduce how open vocabulary fits in natural language processing along with machine learning can be used to infer a person's personality from their language use in a hiring interview. Using data from more than 46,000 people responded to an open interview questions [20]

The use of online reviews may be biased in relation to the topics being discussed. We expect biases in selecting the types of customers who post reviews, especially regarding the extreme of criticism. For example, customers with particularly good or exceptional Bad experiences are more likely to post an online review to recommend or warn others about accommodation. While this would likely help differentiate between topics of interest, this may be a deviation knowledge of what issues are presented in terms of a trend toward extremes [21][22].

However, it does not provide any good intuition for topic modeling. In addition, we cannot get the document representation from SymNMF directly. Therefore, the proposed method in this paper is the first work that considers to build a standard NMF-based topic model for the short texts [23][24].

3 Proposed Architecture:

The topic models range from NLP (Natural Language Processing) to the acquisition of immeasurable knowledge in the fields of technical, which are highly motivated to analyze the models.. Analysis of topics in TM is a major problem and facilitates the unreliable number of topics in TM, Which are related to poor results in technical field for placement assistantship. In this sense, the necessary visualizations are a vital part of cutting the information to determine the direction of the cluster for job assistantship. So, to believe and contribute to the proposed Topic models of Hadoop, which is distributed with Hadoop non-negative matrix factorization (NNMF) and Hadoop distribution with latent Dirichlet Allocation (DLDA) are correct approaches to balance and trim towards the direction of question and answer pairs or topics or terms or to from different perspective data sources in the technical data pool for placement. These proposed models are derived in below.

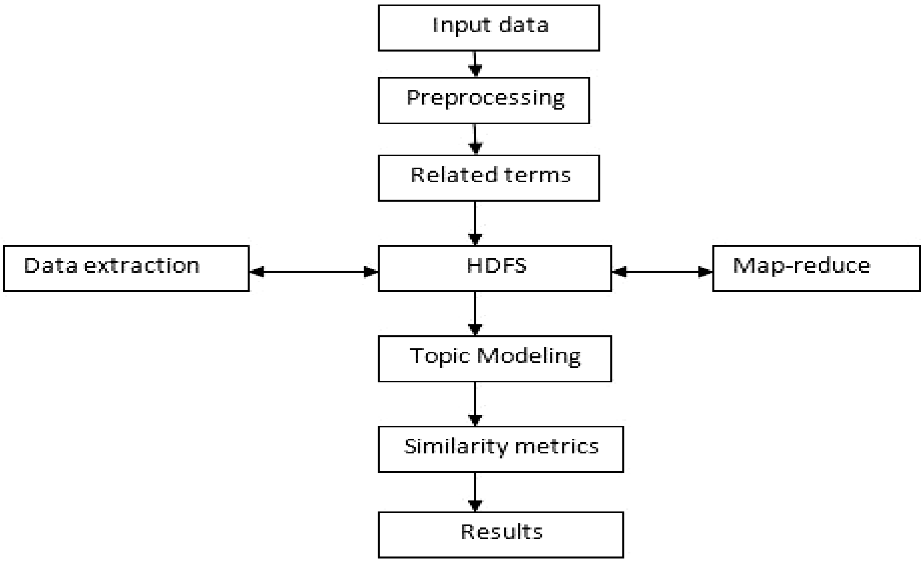


Fig 1: Proposed Architecture

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

The functions are distinct as exclusive and measurable attributes / properties for every surveillance in the dataset. These characteristics are the context / category / marker that correctly and uniquely defines the capacity of the question text. The same function, while introduced into a machine learning algorithm, find learning patterns with the intention can be applied to a opportunity text of the question. conversely, these functions must be vectorized so that the algorithm can optimize and decrease loss when demanding to define models. There are numerous models to apply characteristic recovery techniques such as Word2vec, Bag of Words and TF-IDF[2] and initiate with

The vector space model, wherever text data is represented as numeric vectors of explicit terms for vector dimensions and subsequently experienced the Word Bag model, which converts the text into a vector that represents the incidence of all the individual words current in the text vector space for that scrupulous text. consequently, the heaviness of every word is equivalent to its regularity in this text. The representation can also be twisted for n-grams. The trouble we originate with the word bag is that the vectors are based on the unconditional frequencies of the word encounters, so if the word appears a lot throughout the text of the question, it has a higher occurrence and eclipses other words they may not be so common We then tested the term document frequency and inverse frequency (TF-IDF)[2].Use the TF-IDF[2] weight to assemble the vector. Topic modeling uses a supplementary statistical and topical modeling approach to take out basic concepts from a body of documents, which in our container was our question of abundant a Python stack. The Topic models are statistical probabilistic models with the intention of use single festering decomposition (SVD) and latent Dirichlet mapping (LDA) to find secreted semantic structures in the text that give topics. The major attribute in which we complete the function extraction is the body of the question. In this structure, the hadoop distributed topic modeling techniques are addressed for getting better results and the methods shows in section 6.

3 .1 Methodologies

Our survey responded to the user uncertainty by primary assigning the appropriate tag or objective to the question and subsequently matching the user question mark with the most related question in the case of stack overflow. Finally, the first 10 most important questions were accessible. In this section, we will first provide some preliminaries along with the block coordinate descent method and its applications in NMF for topic modeling. Then, we will propose our SeaNMF model, and a block-coordinate descent algorithm to estimate latent representations of terms and short documents.

3.2 Data Collection

The data link is the official Stack Overflow blog view and social sentiment analysis. The command contains approximately 2.98,000 user data. Stack Overflow conducts a survey every year to find out the interests of its users, and based on these responses, you can perform various analyzes using the results. Initially, I intend to download the Stack Overflow file data commencing the records However, the collection file is 12 GB with zip. Outstanding to the limitations of our restricted PC configuration and the impossibility of working through such a bulky data set, we decided to use the data set provide by Kaggle in this regard. Kaggle has a specific data set for Python problems. Hadoop is open source and supports the process of processing an extremely large body of data in a HDFS. Map reduction is also a model processing and programming for Java, Python, R, and Spark-based computational calculations. The Map Reduce algorithm contains two important fields, including Map and Map Reduct. The target accepts one data conjunction and converts another data conjunction, where individual elements are divided into tuples (key / value keys).[19].

Table 1: Hadoop Network Distributed with Memory Allocation

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

Node	Transferring Address	last contact	Configure_ capacity (GB)	used (GB)	non _DFS used (GB)	Remaining (GB)	Used (%)	Remaining (%)	blocks	Block pool used(GB)
1	172.16.7.01:50010	3	9112.91	14.80	71.01	8157.09	0.53	83.99	78.00	5.71
2	172.16.7.02:50010	2	1154.19	14.68	77.66	1421.85	1.03	92.77	76.00	5.98
3	172.16.7.03:50010	1	1450.26	14.84	87.56	1417.88	1.08	89.63	79.00	5.36
4	172.16.7.04:50010	4	1450.26	14.84	87.56	1417.88	1.08	89.63	79.00	5.36
5	172.16.7.05:50010	3	1450.26	14.84	87.56	1417.88	1.08	89.63	79.00	5.36
6	172.16.7.06:50010	3	9112.91	14.80	71.01	8157.09	0.53	83.99	78.00	5.71
7	172.16.7.07:50010	2	1154.19	14.68	77.66	1421.85	1.03	92.77	76.00	5.98
8	172.16.7.03:50010	1	1450.26	14.84	87.56	1417.88	1.08	89.63	79.00	5.36
9	172.16.7.04:50010	4	1450.26	14.84	87.56	1417.88	1.08	89.63	79.00	5.36
10	172.16.7.05:50010	3	1450.26	14.84	87.56	1417.88	1.08	89.63	79.00	5.36

Table 2: Hadoop file Distributed Memory Segregation:

Data set(DS)	Size of the fill	Replication®	Block_ size(BS)	Modification_ Time(MT)	Ownr (O)
1	15.96 MB	5	128 MB	2/5/2020 14:44	pvrao
2	25.56 MB	5	128 MB	2/4/2020 14:44	pvrao
3	346.53 MB	5	128 MB	2/4/2020 14:44	pvrao
4	346.53 MB	5	128 MB	2/4/2020 14:54	pvrao
5	246.53 MB	5	128 MB	2/4/2020 14:24	pvrao
6	9.86 MB	5	128 MB	2/5/2020 14:44	pvrao
7	9.86 MB	5	128 MB	2/4/2020 14:44	pvrao
8	146.53 MB	5	128 MB	2/4/2020 14:44	pvrao
9	246.53 MB	5	128 MB	2/4/2020 14:54	pvrao
10	246.53 MB	5	128 MB	2/4/2020 14:24	pvrao

3.3 Preprocessing and Data Preparation:

The Bag of Words (BoW) Algorithm:

The Bag of Words (BoW) algorithm is the simplest way to present text in numbers. Like the term itself, which is a sentence like a bag of words (a string of numbers). Recall the three types of technical inspections (R) for data preparation, ie. A1: "This JAVA is very scary and long." R2: "This JAVA is not scary and slow." A3: "This JAVA is scary and good" and will first build a dictionary of all the unique words in the previous three revisions. The dictionary consists of these 11 words = {'This',' java ',' is', 'very', 'scary', 'and', 'long', 'no', 'slow', 'creepy', ' good '} = {W1, W2, W3, W4, W5, W6, W7, W8, W9, W10, W11} FROM Reviews for R1, R2 and R3. Now take each of these words and mark their occurrence in the previous three technical reviews with 1 and 0. This will give us 3 vectors for 3 reviews:

Table: 3word and review interfacing

	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	W11	Weight of the review(W)
R1	1	1	1	1	1	1	1	0	0	0	0	7
R2	1	1	2	0	0	1	1	0	1	0	0	8
R3	1	1	1	0	0	0	1	0	0	1	1	6

Vector space of review one(R1) is {1,1, 1, 1 ,1, 1, 1, 0, 0, 0, 0}==length of the review =7, Vector space of review two(R2) is {1 ,1, 2, 0, 0, 1, 1, 0, 1, 0, 0} length of the review ==8 and Vector space of review three(R3) is {1 ,1 ,1 ,0 ,0 ,0 ,1 ,0 ,0 ,1 ,1} == length of the review =6

Term Frequency-Inverse Document Frequency (TF-IDF)

Term frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.”

Term Frequency (TF) : Let’s first understand Term Frequent (TF). It is a measure of how frequently a term, t, appears in a document, d:

$$tf_{t,d} = \frac{n_{t,d}}{Number\ of\ terms\ in\ the\ document}$$

Here, in the numerator, n is the number of times the term “t” appears in the document “d”. Thus, each document and term would have its own TF value.

Table 4:Term Frequency on words and Review

Word/Reviews	R1	R2	R3	TF(R1)	TF(R2)	TF(R3)
W1	**1**	**1**	**1**	**1/7**	**1/8**	**1/6**
W2	**1**	**1**	**1**	**1/7**	**1/8**	**1/6**
W3	**1**	**2**	**1**	**1/7**	**1/4**	**1/6**
W4	**1**	**0**	**0**	**1/7**	**0**	**0**
W5	**1**	**1**	**0**	**1/7**	**1/8**	**0**
W6	**1**	**1**	**1**	**1/7**	**1/8**	**1/6**
W7	**1**	**0**	**0**	**1/7**	**0**	**0**
W8	**0**	**1**	**0**	**0**	**1/8**	**0**
W9	**0**	**1**	**0**	**0**	**1/8**	**0**
W10	**0**	**0**	**1**	**0**	**0**	**1/6**
W11	**0**	**0**	**1**	**0**	**0**	**1/6**

Inverse Document Frequency (IDF)

IDF is a measure of how important a term is. We need the IDF value because computing just the TF alone is not sufficient to understand the importance of words:

$$idf_t = \log \frac{number\ of\ documents}{number\ of\ documents\ with\ term\ 't'}$$

Here, evaluate the I.D.F values from the review two from data corpus

$$\text{I.d.f}_{\#\text{this}}=\log_2 2\left(\frac{\text{No.of Documents}}{\text{No of Documents containing word}}\right)$$
$$=\log_2 2\left(\frac{3}{3}\right)$$
$$=\log_2 2(1)$$
$$=0$$

Here , computing the I.D.F values on behalf of each word from the given data corpus

Table 5: Inverse Document Frequency on words and Review

Word/Reviews	R1	R2	R3	IDF
W1	1	1	1	0.00
W2	1	1	1	0.00
W3	1	2	1	0.00
W4	1	0	0	0.48
W5	1	1	0	0.18
W6	1	1	1	0.00
W7	1	0	0	0.48
W8	0	1	0	0.48
W9	0	1	0	0.48
W10	0	0	1	0.48
W11	0	0	1	0.48

for this reason, the words(W) like “W1”, “W2”, “W3”, etc., are concentrated to zero(0) and have small significance of the words similar to “W4”, “W5”, “W7”, etc. are words with further importance with high value. Now ,Determine T.F-I.D.F values for every word(W) from the review corpus data .

$$\prod(\boldsymbol{tf}-\boldsymbol{idf})_{\boldsymbol{t,d}}=\prod_{\boldsymbol{t}}^{\boldsymbol{d}}(\boldsymbol{tf}*\boldsymbol{idf})$$

Here, to compute T.F-I.D.F values on behalf of every word(W) from the data corpus of review two(R2) in the same way ,to determine the the

$$\prod(\boldsymbol{tf}-\boldsymbol{idf})_{\boldsymbol{t,d}}=\prod_{\boldsymbol{t}}^{\boldsymbol{d}}(\boldsymbol{tf}*\boldsymbol{idf})$$

values on behalf of every words(W) from the data corpus of the all the reviews

Table 6: Relation between TF and IDF on words and Review

Word/Reviews	R1	R2	R3	IDF	$\prod(\boldsymbol{tf}-\boldsymbol{idf})_{\boldsymbol{R1,d}}$	$\prod(\boldsymbol{tf}-\boldsymbol{idf})_{\boldsymbol{R2,d}}$	$\prod(\boldsymbol{tf}-\boldsymbol{idf})_{\boldsymbol{R3,d}}$
W1	1	1	1	0.00	0.000	0.000	0.000
W2	1	1	1	0.00	0.000	0.000	0.000
W3	1	2	1	0.00	0.000	0.000	0.000
W4	1	0	0	0.48	0.068	0.000	0.000
W5	1	1	0	0.18	0.025	0.022	0.000
W6	1	1	1	0.00	0.000	0.000	0.000
W7	1	0	0	0.48	0.068	0.000	0.000
W8	0	1	0	0.48	0.000	0.060	0.000
W9	0	1	0	0.48	0.000	0.060	0.000
W10	0	0	1	0.48	0.000	0.000	0.080
W11	0	0	1	0.48	0.000	0.000	0.080

The Bag of Words immediately creates a collection of vectors containing the count up the word occurrence in the data corpus .while the T.F-I.D.F replica contains information on the further significant words and the a smaller amount important ones as well. The Bag of Words vectors are trouble-free to read between the lines. $\prod(\boldsymbol{tf}-\boldsymbol{idf})_{t,d}$=$\prod_{t}^{d}(\boldsymbol{tf}*\boldsymbol{idf})$ more often than not performs improved in machine learning models. at the same time as mutually BOW and T.F-I.D.F include well-liked in their be the owner of regard and Detecting the correspondence between the words and efficiency of the data corpus and data preparation classified and addressed by The Matrix Representation of proximity analysis from table 4.and Reliability scale of Data Analysis from table 5.and apply the topic modeling methods in section 4.3.

3.4 Map Reduce Works

The Map Reduce architecture contains two main components, such as Daemon services, responsible for performing mapping and reduction tasks, monitoring, and re-failover tasks. In Hadoop 2, Resource Manager and Node Manager are daemon services. When the customer submits a job to reduce the card, these demons go into effect. They are also responsible for the parallel processing and fault tolerance characteristics of Map Reduce tasks. In this article, the Map Reduce ecosystem is well maintained for finding the number of words or topics in the data corpus in a system of question and answer pairs in real time.

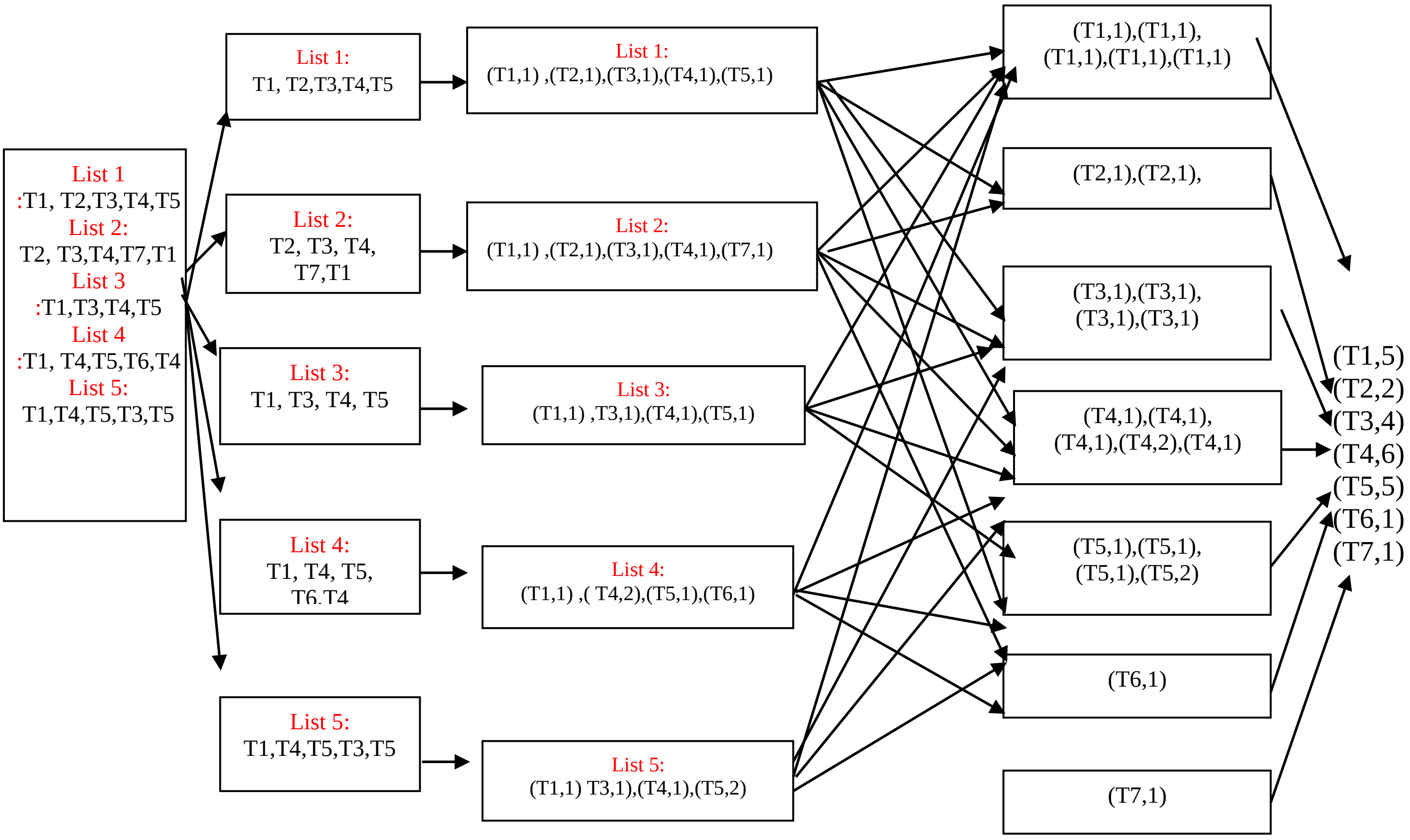


Fig 2: Working procedure for word count from question and Answer pair

Table 7: Mapreduce classification of proximity matrix correlation

Catego ry	Topics to Covered in the corpus	Count

1	Rupy,Go,C,C++,Php,TypeScript,C#,Bash &Shell,JAVA,SQL,CSS,Java Script,HTML,R,scala,Python	2.1738
2	Go,C,C++,Php,TypeScript,C#,Bash &Shell,JAVA,SQL,CSS,Java Script,HTML,R,scala,Python	1.4106
3	C,C++,Php,TypeScript,C#,Bash &Shell,JAVA,SQL,CSS,Java Script,HTML,R,scala,Python	9.5694
4	C++,Php,TypeScript,C#,Bash &Shell,JAVA,SQL,CSS,Java Script,HTML,R,scala,Python	0.5981
5	Php,TypeScript,C#,Bash &Shell,JAVA,SQL,CSS,Java Script,HTML,R,scala,Python	5.5904
6	TypeScript,C#,Bash &Shell,JAVA,SQL,CSS,Java Script,HTML,R,scala,Python	0.1271
7	C#,Bash &Shell,JAVA,SQL,CSS,Java Script,HTML,R,scala,Python	3.2747
8	Bash &Shell,JAVA,SQL,CSS,Java Script,HTML,R,scala,Python	1.4157
9	JAVA,SQL,CSS,Java Script,HTML,R,scala,Python	9.838
10	SQL,CSS,Java Script,HTML,R,scala,Python	9.5847
11	CSS,Java Script,HTML,R,scala,Python	8.5622
12	Java Script,HTML,R,scala,Python	0.9193
13	HTML,R,scala,Python	1.1131
14	R,scala,Python	2.604
15	scala,Python	4.0775
16	Python	9.1667

In table 7, The Mapreduce classification of proximity matrix correlation addressed the each topic and word representation from the data corpus in real time stream data analysis. And each topic covered by data corpus .

Table 8: Data Association:

	lhs	rhs	support	confidence	lift	count
[1]	{}	{python}	0.500	0.5000000	1.000000	4
[2]	{}	{R}	0.500	0.5000000	1.000000	4
[3]	{}	{hadoop}	0.500	0.5000000	1.000000	4
[4]	{}	{java}	0.750	0.7500000	1.000000	6
[5]	{python}	{java}	0.375	0.7500000	1.000000	3
[6]	{java}	{python}	0.375	0.5000000	1.000000	3
[7]	{R}	{java}	0.375	0.7500000	1.000000	3
[8]	{java}	{R}	0.375	0.5000000	1.000000	3
[9]	hadoop}	{java}	0.500	1.0000000	1.333333	4
[10]	{java}	{hadoop}	0.500	0.6666667	1.333333	4

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

Topic models:

Distributed Latent Dirichlet Allocation (DLDA)-The Dirichlet distributed latent distribution or allocation to big data is one of the most valuable models for a topics or terms. The latent Dirichlet distribution or allocation is a very complex application of the Bayesian technique and here, the latent word means to capture the meaning of the text to find the hidden terms or themes of the words in the corpus of each document in the corpus [11].

NNNMF:

the non-negative matrix factorization is a house of linear algebra algorithms for obtaining the latent structure in the data designed as a non-negative matrix and its applied for topic models to hadoop environment for distributed the tweet corpus is called as distributed visual non-negative matrix factorization [9][10] .it is similar and equivalent to NMF where the input is TDM(terim-document-matrix),TF-IDF distributed available in table 6.. the input of matrix factorizations are TDM and Number of topics and illustrated two decompose matrices i.e. one for every topic and second one is every topic from tweet corpus.after decomposition will get two non-negative matrices of the original n words by **k** topics .so that the nmf model is going to allocate it into one of those topics. The first thing to note is that non-negative matrix factorization can be shown to be equivalent to optimizing the same objective function as the one from probabilistic latent semantic analysis. The following algorithm NMF [12][13] addresses for topic modeling on a term-document matrix, using scikit-learn

SANMF:

We propose a new model, supported by NMF semantics (HdiSANMF) to model topics with short text, which is described in fig. 3. In this figure, the documents, words and context are indicated as DI, WI and CI respectively. The proposed HDiSANNMF model can capture the semantics of short text body based on word-document and word-context correlations available in table 9 and table 10, and our target function combines the advantages of both the NMF topic modeling model and the grammar skip capture model in the context of word semantic correlations. Figures H, Wc and W are the vector images of documents, contexts and words in latent space. Each column of W represents a theme. We use a block algorithm for coordinated descent to solve the optimizations. For better interpretability, we also present a slim version of the HdiSANNMF.

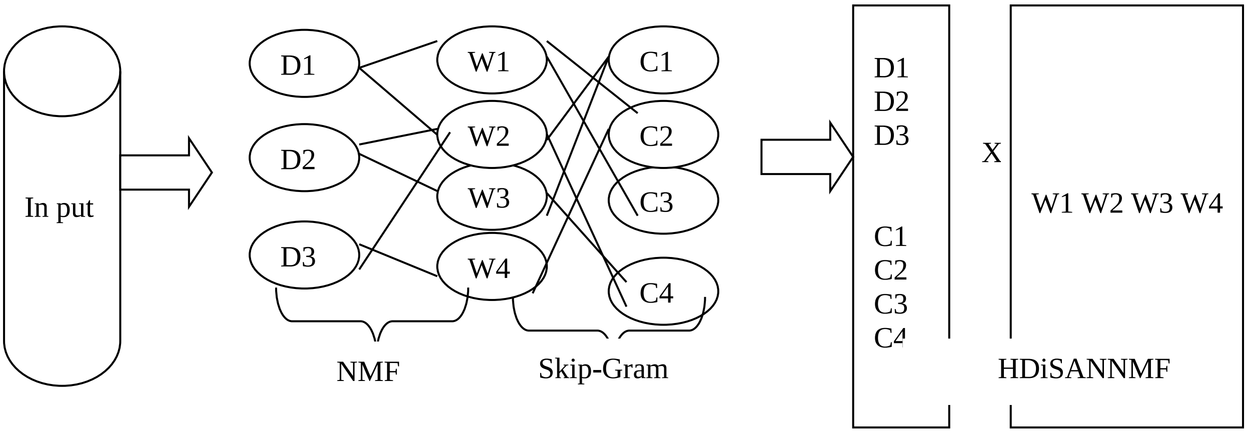
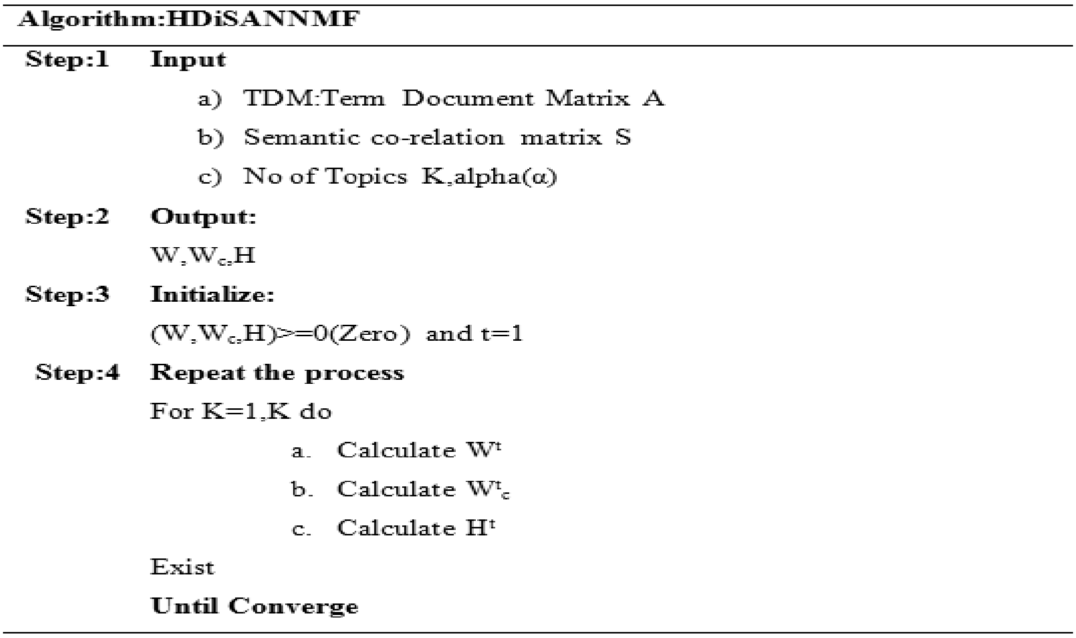


Fig3: working framework on Topic modeling

HDiSANMF is summarized into an algorithm and the document term matrix "A" is first constructed using a representation of a "bag of words". Then we calculate the semantic correlation matrix S, the latent factor matrices W, Wc

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

and H are arbitrarily initialized with non-negative real numbers. Then within each iteration your coordinates will be updated by columns.



```
Algorithm:HDiSANNMF
Step:1  Input
            a) TDM:Term Document Matrix A
            b) Semantic co-relation matrix S
            c) No of Topics K,alpha(α)
Step:2  Output:
        W,W_c,H
Step:3  Initialize:
        (W,W_c,H)>=0(Zero) and t=1
Step:4  Repeat the process
        For K=1,K do
                a. Calculate $W^t$
                b. Calculate $W^t_c$
                c. Calculate $H^t$
        Exist
        Until Converge
```

Experimental Results and Discussion:

In this study, to compared a series of resemblance factorization models used to find similar questions using topic modeling methods, and then created the most relevant questions that were successfully derived. To measured the relevance of the similarity of the questions by means of a call and use cosine similarity for k topics, topic modeling, and ensemble models to extract the 10 most relevant questions for each user query

Table 9: WTCM: Matrix for word-topic-count

Word-topic count matrix										
Topics/ Q&A pairs	**T1**	**T2**	**T3**	**T4**	**T5**	**T6**	**T7**	**T8**	**T9**	**T10**
1	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**1**	**0**
2	**0**	**0**	**1**	**0**	**1**	**2**	**0**	**0**	**0**	**0**
3	**0**	**0**	**3**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
4	**1**	**0**	**2**	**0**	**0**	**1**	**1**	**0**	**0**	**0**
5	**0**	**1**	**0**	**0**	**2**	**1**	**0**	**0**	**0**	**0**
6	**1**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
7	**0**	**0**	**0**	**0**	**1**	**1**	**0**	**0**	**0**	**0**
8	**1**	**1**	**1**	**1**	**0**	**1**	**0**	**0**	**0**	**0**
9	**0**	**1**	**1**	**1**	**0**	**0**	**0**	**0**	**0**	**0**
10	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**0**	**1**	**0**

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

Table 10: DTCM: Matrix for document-topic-count

Document-topic count matrix										
Topics/ Q&A pairs	**T1**	**T2**	**T3**	**T4**	**T5**	**T6**	**T7**	**T8**	**T9**	**T10**
1	**0**	**1**	**0**	**1**	**1**	**0**	**0**	**2**	**2**	**0**
2	**2**	**1**	**0**	**0**	**0**	**1**	**0**	**2**	**0**	**1**
3	**0**	**1**	**0**	**2**	**2**	**0**	**0**	**0**	**0**	**1**
4	**0**	**0**	**0**	**2**	**1**	**0**	**2**	**1**	**0**	**1**
5	**1**	**1**	**1**	**2**	**1**	**1**	**1**	**1**	**1**	**1**
6	**0**	**1**	**0**	**0**	**1**	**0**	**1**	**1**	**0**	**1**
7	**3**	**0**	**1**	**1**	**0**	**0**	**0**	**1**	**0**	**0**
8	**1**	**0**	**2**	**1**	**2**	**0**	**1**	**2**	**1**	**1**
9	**0**	**0**	**2**	**0**	**0**	**0**	**0**	**2**	**0**	**3**

From Fig 4 and 5 , the precision ,Recall, Accuracy and F-1 Score of the Topic Modeling technique applied for similarity metrics and its enable the hadoop Distributed Semantic Assisted Non Negative Matrix Factorization for better accuracy of the model from real time corpus data compare with other topic models . Here the comparison metrics computed by TP,TN,FP and FN of precision and Recall in topic modeling models.

Table 11: Hadoop Distributed semantics-assisted NNMF

Q&A Corpus	**Precision (P)**	**Recall ®**	**Accuracy**	**F-1 Score**
1	1.00	1.00	1.00	1.00
2	0.98	0.98	0.99	0.98
3	0.96	0.96	0.97	0.96
4	0.94	0.94	0.96	0.94
5	0.91	0.91	0.95	0.91
6	0.89	0.89	0.93	0.89
7	0.86	0.86	0.92	0.86
8	0.84	0.84	0.91	0.84
9	0.81	0.81	0.89	0.81
10	0.78	0.78	0.88	0.78

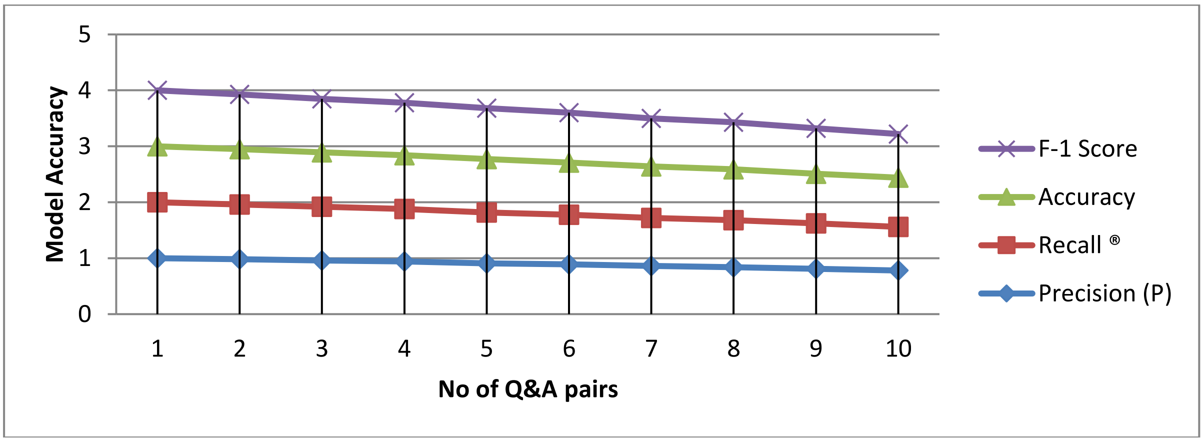


Fig 4: precison, Recall, Accuracy and F-1 Score of Distributed hadoop SANNMF

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

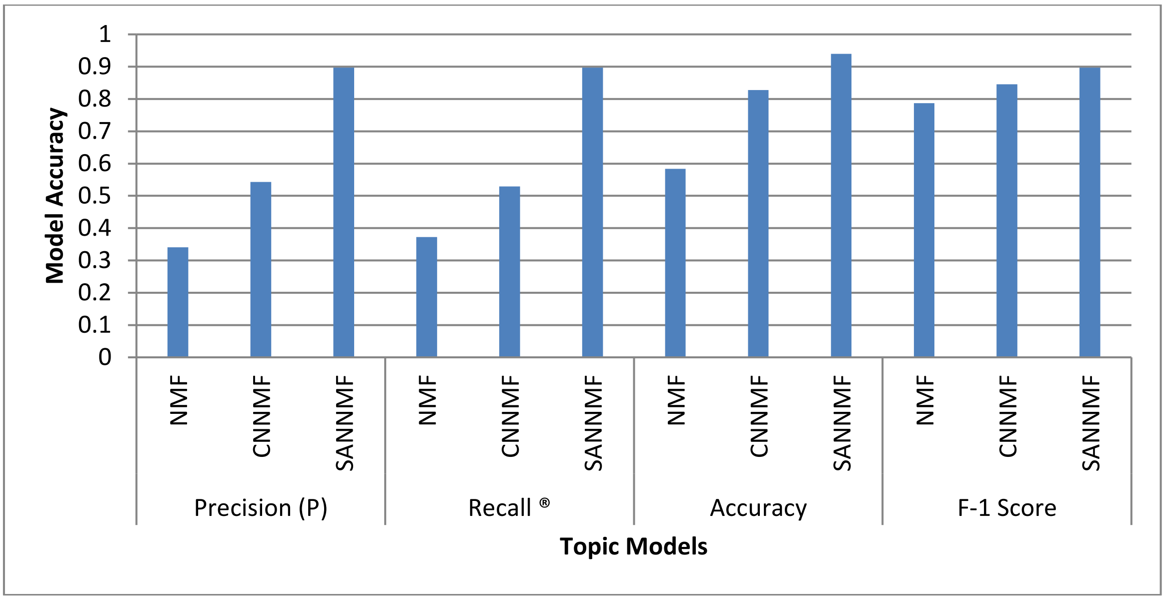


Fig: 5 Comparison results on Topic modeling techniques

From Fig 4 and , the precision ,Recall, Accuracy and F-1 Score of the Topic Modeling technique applied for similarity metrics and its enable the hadoop Distributed Semantic Assisted Non Negative Matrix Factorization for better accuracy of the model from real time corpus data compare with other topic models . here the comparison metrics computed by TP,TN,FP and FN of precision and Recall in topic modeling models.

The NMF, CNNMF and HDiSANNMF of topic models for analyzing data corpus and balanced in Topic Modeling Techniques. From Fig 5, the comparision analysis balanced by Hadoop Distributed Semantic assisted Non-Negative Matrix Factorization. Finally, this work briefly describes the public question-and-answer structure around the world and follows the development of the main themes Housing and employment opportunities for next-generation technologies worldwide in real time scrolling.

5 Conclusions:

As a new study with a data set and methodologies examines how topic Models (TM) provide information on technical terms instance of the problems faced by professionals. Quora and stackoverflow users, the ones who follow the questions and vote on the answers to the technical conditions or topics can only be the ones who take the issues seriously. Therefore, the result of this survey reflects only a small part of public opinion. But as Quora and Stack Overflow gain more and more users and a comprehensive look at the technical issues and Stack Overflow is important in the field of sociological research and communication advisers. This can cause a possible loss of important information. Therefore, combine cluster modeling with topic modeling to create an overall presentation and get better results while modeling a topic or terms using the HDiNNMF algorithm with precision, recovery, and F-1 indicators compared to existing results. Believe that the proposed methodology, which includes valuable question-and-answer data and a quantitative analytical and qualitative results process, shows that the topics found by HDiSNNNMF with their best keywords are more semantically correlated. Therefore, we conclude that the proposed method is an effective topic model for this work.

in future research, scientific statements and community input to support the job linking with RPA PEGA Robotics and AWS Eco-system for automate the manual work with streaming questions and answers. This study briefly describes the social structure of questions and answers around the world and follows the development of the main topics of hosting opportunities and use of next generation technologies in the world in real time.

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

References:

1. D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet Allocation,” the Journal of machine Learning research, vol. 3, pp. 993–1022, 2003.
2. LI Hua-Meng,LI Hai-Rui,XUE Liang. TFIDF Algorithm Based on Information Gain and Informati Entropy[J]. Computer Engineering, 2012, 38(08): 37-40.
3. Hanchen Jiang, Maoshan Qiang, Dongcheng Zhang, Qi Wen, Bingqing Xia, Nan An. "Climate Change Communication in an Online Q&A Community: A Case Study of Quora", Sustainability, 2018
4. Campbell, J.C., Hindle, A. and Stroulia, E., 2014. Latent Dirichlet allocation: extracting topics from software engineering data. In The art and science of analyzing software data (pp. 139-159). Morgan Kaufmann
5. Rainer Lienhart, Stefan Romberg, and Eva H¨orster. Multilayer pLSA for multimodal image retrieval. In Proceeding of the ACM International Conference on Image and Video Retrieval, CIVR ’09, pages 9:1–9:8, New York, NY, USA, 2009. ACM.
6. S. Arora, R. Ge, R. Kannan, and A. Moitra. Computing a nonnegative matrix factorization provably. In Proc. the 44th Symposium on Theory of Computing (STOC), pages 145–162,2012.
7. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: Annual Conference on Neural Information Processing Systems, pp. 556–562 (2000)
8. Yan X, Guo J Learning topics in short text using ncut-weighted non-negative matrix factorization on term correlation matrix,2013
9. Huang L, Ma J, Chen C (2017) Topic detection from microblogs using T-LDA and perplexity. In: 24th Asia-Pacific software engi- neering conference workshops,2018
10. W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In Proc. the 26th Annual International ACM SIGIR conference on Research and Development in Information Retrieval (SIGIR), pages 267–273, 2003
11. Peng Zhang ,Department of Mathematics, Zhejiang University, Hangzhou, 310027 China ; Wanhua Su Statistical inference on recall, precision and average precision under random selection,2012 7,Print on Demand(PoD) ISBN: 978-1-7281-4715-4,2019
12. Gaujoux R, Seoighe C. A flexible R package for nonnegative matrix factorization. BMC Bioinformatics 2010;11:367.
13. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999;286:531–7.
14. Dai L, Gao X, Guo Y, Xiao J, Zhang Z. Bioinformatics clouds for big data manipulation. Biol Direct 2012;7:43.
15. V. K. Vavilapalli, A. C. Murthy, C. Douglas, S. Agarwal, M. Konar, R. Evans, T. Graves, J. Lowe, H. Shah, S. Seth, B. Saha, C. Curino, O. O’Malley, S. Radia, B. Reed, and E. Baldeschwieler, “Apache Hadoop YARN: Yet Another Resource Negotiator,” In Proc. ACM
16. F. H. Gebara, H. P. Hofstee and K. J. Nowka, “Second-Generation Big Data Systems,” IEEE Computer, vol. 48, no. 1, pp. 36-41, 2015.
17. Apache Hama, accessed on June 16, 2016. [Online]. Available: https://hama.apache.org/
18. Yeung, K. Quora Now Has 100 Million Monthly Visitors up from 80 Million in January. Available online: http://venturebeat.com/2016/03/17/quora-now-has-100-million-monthly-visitors-up-from-80-million-in-january/ (accessed on 28 March 2016).
19. AlexaWebpage: Web Traffic Statistics of Quora. Available online: http://www.alexa.com/siteinfo/www.
20. quora.com (accessed on 15 April 2016).
21. M. Jayaratne, B. Jayatilleke:

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

Predicting Personality Using Answers to Open-Ended Interview Questions, Digital Object Identifier 10.1109/ACCESS.2020.3004002. July 2, 2020.

22. Ian Sutherland , Youngseok Sim, Seul Ki Lee, Jaemun Byun and Kiattipoom Kiatkawsin , Topic Modeling of Online Accommodation Reviews via Latent Dirichlet Allocation, Sustainability 2020, 12, 1821; doi:10.3390/su12051821
23. Ashesh Iqbal ,Sumi Khatun ,Mohammad Shamsul Arefin , andM. Ali Akber Dewan ,ERF: An Empirical Recommender Framework for Ascertaining Appropriate Learning Materials from Stack Overflow Discussions, Computers 2020, 9(3), 57; https://doi.org/10.3390/computers9030057
24. Tian Shi, Kyeongpil Kang, Jaegul Choo, Chandan K. ReddyShort-Text Topic Modeling via Non-negative Matrix Factorization Enriched with Local Word-Context Correlations, DOI: https://doi.org/10.1145/3178876.3186009, WWW '18: Proceedings of The Web Conference 2018, Lyon, France, April 2018
25. Chenliang Li, Haoran Wang, Zhiqian Zhang, Aixin Sun, and Zongyang Ma. 2016. Topic Modeling for Short Texts with Auxiliary Word Embeddings. In Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval . ACM, 165–174.

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731