Machine learning analysis of topic modeling Reranking of clinical records

V. Kakulapati[1], S. Mahender[2], B.S.S.Deepthi[3], João Manuel R.S. Tavares[4]
[1,2] Sreenidhi Institute of Science and Technology, [3]Mamatha Medical College
[1,2] Yamnam pet, Ghatkesar, Hyderabad, Telangana-501301.
[3]Khammam, Telangana-507002
[4]Instituto de Ciência e Inovação em Engenharia Mecânica e Engenharia Industrial, Departamento de Engenharia Mecânica, Faculdade de Engenharia, Universidade do Porto, Porto, PORTUGAL
[1]vldms@yahoo.com, [2]mahendersheri@gmail.com, [3]deepthisharma.3421@gmail.com, [4]tavares@fe.up.pt

Technologies in Big data have improved the analysis of clinical information for better understanding diseases in order to provide more efficient diagnoses. An online healthcare system has created huge data by record maintaining, taking into account acceptable requirements and the patient's care. These clinical records are in files that pose a challenge for data processing and finding relevant documents. In this work, we used a method that combines Statistical Topic Models, Language Models and Natural Language Processing, in order to retrieve clinical records. On the other hand, for analysing large clinical records in the form of documents, Topic models are used to finding related clusters of disease patterns. Here, it is explored the decomposition of clinical record summaries into topics which enables the effective clustering of relevant documents based on the topic under study. Clinical documents selected in a Topic-based approach give proper information to the users for better understanding and derive insights from the related data. In our proposed method, it is used clustering-based semantic similarity topic modelling in order to summarizing the clinical reports based on Latent Dirichlet Allocation (LDA) in a MapReduce framework. Automated unsupervised analysis of LDA models are used to identify different disease patterns and to rank topic significance. In this, topic and keyword re-ranking methods which assist physicians to get improved information through the LDA-obtained topics. The experimental assessment confirmed the value of the used methods in clinical documents summarization.

Keyword: Big Data, framework, Reranking, LDA, clinical records.

INTRODUCTION

Today Electronic medical records (EMRs) are created by many technologies like sensors, wearable and devices. The integration of wearable devices and medical records can be used to study a variety of physical conditions, for example, in fitness. The combination of EHR with big data, for instance, current medical records into the EMRs together with inherent data is very promising. This type of combination of electronic health records can supply important, complete and consistent basis of data for medical studies [1]. The main purpose of wearable devices is to generates data without human intervention by zero attempts need from enduring [2]. On the other hand, the manual creation of health records of patients visiting a hospital is much more time consuming, especially when testing different parameters, like level of glucose in blood and heartbeat rate. Whereas wearable devices are efficient and precise in time as, usually, such as devices can verify all the acquired features and analyse them directly, unlike manual methods. In this chapter, the discharge sheets are generated by taken into account electronic health record details and analysing them based on topic modelling technology. The topic models are used to detect behavior patterns in electronic medical records, which are produced by a health monitoring system.

The medical reports of patients provide information regarding growing data for managing patient information and predicting trends in diseases. Healthcare service providers facing a major challenge regarding patients that are suffering from multiple problems with inefficient diagnosis due, which leads to frequent and increasing visits to hospitals. Therefore, the discover of symptoms related to health conditions is increasing interesting since it can help to obtain improved predictions for hospitalization, disease or death. In this regard, Moumita Bhattacharya et al. [3], proposed the Electronic Medical Record (EMRs) topic modeling to analyze the symptoms and patient behavior.

Summarization in text mining is one of the major challenges. Because of this summarization, researchers give information to stakeholders and developed several real-time applications. The huge number of documents is converted into a decreased and compacted in summarization indicates the summary of the document collections. The document summarization gives better knowledge about the overall content of the dataset. This summarization reducing the physician time consuming without reading of the entire patient report. Generally, a function of converting entire document information to small chunks is calling as document summarization. These chunks of information hold the entire description of the document collection as shown (1), Here D represents the entire document collection and d is documented summarization, and the size of D is better than the size of d.

Text summarizer carries out by the algorithm is derived from the text summarization task. These are classifying in two ways, one is single-document, and another one is multi-document. The first type, a single document is summarized in the summary of the document, whereas as in second one, a collection documents are summarizing in the summary of the document, which gives the total knowledge of the various documents.

One of the popular Statistical approaches is LDA Topic modeling to allocation of items in large corpus into subsets into semantically-meaningful and used on textual corpus. Documents are arbitrary combinations over topics in the dataset, which makes logic between topics which is exceptional as regards a particular topic. For example, a news article on the President of the USA moves towards healthcare. The topics in the news would be reasonable to allocate like President, the USA, health assurance, and political opinions, though it is to confer the medicinal service.

The dataset contains documents which are a collection of a related number of topics; these topics are associating with a diversity of phrases which shows every document is the consequence of a combination of probabilistic samples: possible topics distribution and selected topic possible word listing — one of the main advantages of LDA than PLSA and LSI topic modelling techniques. LDA is a generative model which employs to split the text into the topic to documents on the outside the dataset. For instance, LDA group news articles into classes like Sports, Entertainment, and Politics, potential use of the fitted model to facilitate classification recently-circulated news. This facility is away from the scope of approaches like LSI. The number of parameters to an approximation for LDA model dimensions with the number of topics is much lower number, which makes LDA is apt to effective with huge data sets. LDA is to model documents as occurring from several topics, where each topic is described to be an allocation over a fixed glossary of words. Each document is a collection of topics and shows these topics with diverse parts because documents in a dataset are apt to be heterogeneous, merging a subset of main themes that filter through the group as a whole.

Now day researchers are concentrating on summarization techniques for the document. Numerous methods are developing to digest by retrieving the significant topics from particular corpora. For analyzing the unstructured text is utilizing probabilistic topic models, provides the latent to be incorporating into patient medical record summarization. A patient medical record contains metadata about the patient's diagnosis history and multiple topic concepts that can be precious for exactly understanding the document. The unified model [4] is utilizing for free-text medical reports which incorporate appropriate patient and data at document-level and identifies multiword in medical documents.

Clinical reports contain information regarding the patient is accumulating as the free text in the general practitioner's medical documents. These reports give medical description can be a computationally challenging task to make understandable by the inconsistency in physicians writing styles, disparities in their observation, and the intrinsic linguistic. However, a clinical report provides case-based reasoning [5] and automatic summarization [6] data for medical applications. Topic modeling of documents provides indexing large, unstructured data with conditional semantics [7]. These methods show potential results due to these basic methods that have not integrated further progression in the field of topic modeling. Improvement of advances in topic modeling methods shows varied medical data and potential structure to release the data in medical documents. Medical document processing and summarization of a database is a difficult undertaking and in the Big Data era where data is more and more, which requires algorithms for summarizing the large clinical reports.

2. RELATED WORK

In [8], topic models explain about to produce the concept from a prescription combination. In the same way, traditional Chinese medicine utilized the interactions between herbs to retrieve symptoms and analyzed [9] variants of LDA. Though, topic modeling using in clinical documents analyzing is a promising field. Topic modeling of unstructured clinical documents is classified and represents clinical reports. By utilizing topic models, the content has been exploring for an association between symptoms and topic adaptations are Topic-Concept models [10,11]. Similarly, the investigation of entertaining drug conversations [12] and, pertinent to clinical practice, clinical case repossession [13]. In this work we focus on the patient discharge summary report and the involvement of different patient-related information.

Text data contains Bag-of-Words (BoW) require to be changing to an appropriate format for computerized processing. For BoW, each report develops into a token/word vector. Patient clinical information is retrieved by analyzing Electronic health records (EHRs) [14]. These EHR data contains empty spaces and needs to be preprocessing to utilize in computer-based methods. By using this data could be efficient and effective for the speed and quality of health care. In our implementation, we utilize discharge summary reports. Generally for regular text classification, topic modeling is implemented on the entire dataset in diverse methods. Clinical reports topic modeling provides understandable topics which exist in medical reports. This type of representing reports based on their topic allocations is additional dense than representation of bag-of-words and can be improving in-process documents than raw text in successive computerized processes.

For generating topic models of discharge summary reports, we utilize LDA due to the probabilistic system for clinical documents and its toughness to overfitting. LDA believes that medical reports contain underlying topics and every topic classified by an allocation transversely words [15]. LDA is utilizing for a large variety of healthiness and clinical applications for predicting textual data [16], learning appropriate medical models and arrangements in clinical records [17], identifying prototypes of medical events in brain cancer patients [18], and examining the results [19]. The pattern contains information about enlightening the formation, semantics, and dynamics. These patterns give physicians with precise information which is utilized to guide better treatment actions of each patient. For finding treatment behavior of patients, LDA utilized these patterns [20], to predict medical classifying patterns, and to form diverse diagnosis activities [21] and pattern timestamps [22]. To determine enduring transience customized by LDA [23] and also identifying the knowledge based on characteristics of the patient and modeling disease [24]. Better performance than LDA for managing issues related to redundancy in clinical report using Redundancy aware LDA [25].

Generate summaries from the huge collection of documents, a MapReduce construction based summarization technique intended. Implementation results evaluation time for summarizing the huge collection of documents is significantly decrease utilizing this framework and also offers scalability for accepting huge document assortments for summarizing, which is a trendier programming model for processing huge data. By using Mapreduce, this provides several benefits in maintaining a huge amount of data, for example, scalability, flexibility, fault tolerance, and several benefits. Now a day's many researchers [26–32] are presented in several works in the aspect of Big Data and processing of the huge amount of data. It is extensively utilized for processing and handling the huge amount of data in a disseminated cluster, which has been utilize for several domains, for example, text clustering, access log investigation, creating search catalogs and diverse data analytical functions. The MapReduce framework [33] is to execute clustering on the huge amount of data by utilizing customized K-means clustering algorithm.

The MapReduce framework is effectively employe for several document processing tasks for dealing with large text are the complicated task in the knowledge discovery process. In-Text analytics, summarizing the huge amount of text set is a motivating and challenging crisis. Many researchers propose for dealing large text for automatic text summarization [34, 35]. Utilizing prosodic elements and enhance lexical element technique is proposed [36] for gathering summarization. An unsupervised technique [37] use for the regular summarization of source code text, which is employed for code folding and allocates one to discriminating conceal chunks of code.

Parametric shortest path algorithm utilizing phrase graphs is a multi-sentence compression technique [38] presents for multi-sentence compression. For creating the required summary, a parametric method of edge weights is utilizing. The execution is carried out by utilizing the MPI and framework of MapReduce, which is exhibited by Parallel implementation of Latent Dirichlet Allocation (PLDA) [39] to it can be useful to huge, real-world applications and accomplishes superior scalability.

3. HEALTH AND MEDICAL TOPIC MODELING

In-text mining, two leaning approaches are there: classification, which is known as supervised and clustering, which is known as unsupervised. In the first approach, to make the unknown formation in labeled datasets, whereas the second one is to identify the patterns in unlabeled data collection. The supervised learning method is the classification, and the unsupervised learning method is clustering. In the first approach is to prepare data with labels are predefined and assign to a new record [40]. Unsupervised learning allocates a set of every record in a data collection based on clustering similarity functions. Topic modeling is the most acceptable clustering techniques for a broad category of applications. Topic modeling deriving every topic is distribution of probability words and reports as probability distribution over topics. In clinical reports mining latent Dirichlet allocation gives more relevant information than other models.

In large corpus, topic modeling is unsupervised learning, which discovers the contents of a text collection. Techniques utilized Latent Semantic Analysis [41], probabilistic Latent Semantic Analysis [42] and LDA [43]. The unknown semantic arrangement of a word-text matrix where the text is rows and words are columns [44] depends on Singular Value Decomposition. The main disadvantage of Latent Semantic Analysis is every word is delighted as the similar meaning; word polysemes cannot distinguish. The result of this analysis consists of axes in Euclidean space is not understandable [45].

CBR (Case-Based Reasoning) is a technique implemented from knowledge-based classification in diverse provinces, which utilizes occurrences from prior related cases to resolve the latest crisis. The reason behind CBR is the hypothesis that related cases have analogous solutions [46]. By using CBR in different research problems, including similarity estimation algorithms, catalog methods to enhance the effectiveness of retrieval methods, case depiction techniques, and techniques to add the latest cases [47]. CBR main the history of past cases before the individual determined in rules, every case includes a depiction of the case, solution, which is the implicit solution.

CBR used to solve the latest case is that the case matched beside the cases in the case base, and analogous cases are repossessed, which is utilized to imply a solution reprocessed and examined for accomplishment. At last, the most recent case and its solution saved as the segment of a most recent case.

For creating a short, precise, and assured summary of a longer text document is known as text summarization. ATS (Automatic text summarization) techniques are required to address a large amount of text data accessible online to assist relevant information retrieval and reducing the user retrieval process. ATS is a text summarization, which is the procedure of generating a small and logical description of a large document.

LDA and LSI are statistical methods; whereas the former one is used complex probability and later used for simple. LSI is less complex than LDA, and LDA is a considerable extension of LSI. The major weakness of LSI is ambiguity. In LDA, words grouped into topics, which can exist in more than one topic. LDA deal with ambiguity by evaluating a document to two topics and resolving which topic is nearer to the document, transversely all permutations of topics. LDA assists the search engine to establish which documents are most significant to which topics. Probabilistic latent semantic analysis (pLSA) is identical to LDA except that the topic allocation is supposed to have SDP (sparse Dirichlet prior). SDPs determine the perception that documents cover only a few topics; these topics utilize only a few words frequently. The results are disambiguation of words and the precise task of documents to topics. The generalization process of pLSA model is LDA.

The probabilistic version of LSA is pLSA where an unseen variable is related to every incidence of a topic in a specific record. Topics are then contingent from the participation in clinical reports. The polysemis problem is solving by PLSA; but it is not considering a completely generative model of reports which is calling as overfitting. Multiple factors produce linearly with numerous documents. Topic distribution describing in LDA over a fixed language and every document can display topics with diverse sections. LDA creates the topics in a 2-step process for every medical report:

1. Topics are arbitrarily choosing in an allocation.
2. for every topic in the report:

(i) Arbitrarily select a word from the allocation over words.

(ii) Arbitrarily select a topic from the consequent language distribution.

The possibility of creating the topic tj from report ri can be defined as follows:

$$P\ (T_j \mid r_i;\Theta,\varphi)=\sum_{k=1}^{K} P(Tj \mid zi;\varphi)P(zk \mid di;\Theta d)$$

Where θ is a model from the distribution of Dirichlet for every text di and Θ model from the distribution of Dirichlet for each word zk. Using different sampling methods like Gibbs Sampling [48] and optimization methods [49] to prepare a topic model in LDA. The efficiency of LDA better than PLSA for simple data collection as it avoids overfitting and polysemy support. In dissimilarity of PLSA, LDA has also considered a completely creative method for text.

LDA is an extension of PLSA where the topic and word allocations have Dirichlet priors [50]. PLSA supposes that have consistent prior. The term allocations in LDA p(w|z) have Dirichlet preceding with parameter α, and the topic allocations p (z|d) have Dirichlet preceding with parameter β. Empirical experiments in LDA shows to better PLSA in cases where the number of parameters largely evaluated to the size of the data [51].

The LDA [52] is an effort to get better pLSA by establishing a Dirichlet prior on document-topic allocation. Multinomial distributions of prior association [53] of Dirichlet prior simplify the statistical inference problem. The LDA [54], successfully applied in diverse applications for recognizing topics. Performance of the LDA compared with other models, such as unigram, mixture unigram, and the pLSA in terms of perplexity. In this, they addressed that the LDA demonstrated superior performance and also LDA is not experiencing the severe overfitting crisis, whic related with the pLSA.

MapReduce [55] is a programme representation and a related implementation for doing out and creating big corpus with an equivalent, disseminated cluster algorithm. We utilized a novel structure, which is based on MapReduce tools for summarizing the huge document collection. This method is determining to by means of clustering semantic similarity and topic modeling utilizing LDA for the document collection summarization. The main advantage of the proposed framework is observable from the testing and also affords a faster execution of summarizing the huge collection of documents and is an influential tool in analysis of big data.

Conversely, the results retrieved by LDA [56] may not be initiative for understandable format and use. In our proposed model we implement various topic and keyword re-ranking approaches which helps stakeholder's healthier knowledge and utilize the words derived by LDA in the analysis of records. We utilized techniques to process the LDA results depends on a set of conditions that will provide required information for the patient. Our experiment analysis exhibits the effectiveness of the techniques in summarizing patient discharge summary reports.

4. FRAMEWORK:

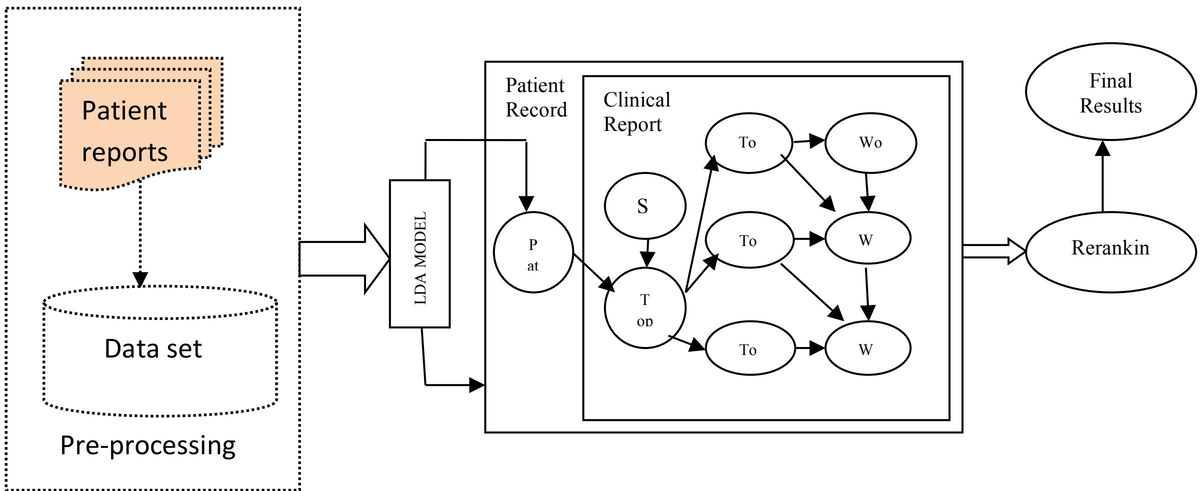

Fig 4.1. Reranking frame work for patient clinical records

4.1. Patient reports: Deidentifying discharge summary reports using in this investigation are provided by the i2b2 National Center for Biomedical Computing funded by U54LM008748 and data set is preparing for the sharing of Challenges in NLP for Clinical Data organizing by Dr. Ozlem Uzuner, i2b2 and SUNY." The dataset contains 390 discharge summaries for different patients. These reports are all the patients which contain details of the patient like patient history, symptoms, patient id, etc. In this dataset, the data set collected from the homogeneous set of patients from a medical perspective. For implementation purpose, the dataset categorized as the training set and test set.

The patient discharge summaries including patient name and all patient related information, prescriptions and conditions illustrating the enduring (*e.g.,* “heart pain”). The patient’s discharge summary perceptions in this assignment consist of things linked to a long-suffering, which are frequent in medical reports and determines co-references are crucial for the receipt of an overall description of the clinical situation. The discharge summary consists of different topic, such as the “patient History” and “prescription” describes the patient data in diverse situations. Additionally, the text format is not particular, so numerous names can be present in each entity. For instance, physician names, clinician, doctor, etc., can refer to the similar person in a medical report summary.

4.2. Dataset: This study used i2b2 patient discharge summary report and each report contains the patient related information and this data set contains 380 discharge summaries. The summary contains above 4000 topics indicate that patient relevant information that is patient id, name, patient history, symptoms, medication, etc. We classified this data set into training set about 100 records and 200 discharge summaries are test data set.

The clinical discharge summary dataset also included metadata for each report concerning data about the therapeutic history, with the date, the name of clinician and prescription status. In the same way, demographic data was related with every discharge summary report, as well as the age of patient, gender, family history and course. These data was through existing to the representation to utilize as earlier in generating topics precise to the present report.

4.3. Preprocessing: The clinical discharge summary data set cleaned every medical report to remove irrelevant inconsistencies in the data collection. Subsequent to cleaning, the discharge summary collection consists of 4000 topics in total.

In topic modelling the input data is a document-term matrix, where the tuples equal to text and the attributes to the words. The total number of tuples is corresponding to the data set size and the total number of attributes to the magnitude of the language. Document mapping to the term occurrences of vector includes tokenizing the text and then handling the tokens, for instance, by translating tokens to lower-case, eliminating punctuations, eliminating numbers, stemming, eliminating stop words and the missing terms with a length under a certain minimum.

4. 4. LDA model:

LDA defines the topic as a distribution of language, where each report demonstrates with diverse proportions. LDA utilizes probabilities and characterize documents as the combination of topics that categorize words with certain probabilities. Discharge summaries are produced in the subsequent approach

- According to Poisson distribution, the number of terms in the text.
- According to the distribution of Dirichlet distribution more than a predetermined set of K topics, select a topic combination for the document.

- Create every topic Ti in the report by:
    - A topic selection
    - According to the topic's multinomial distribution, the topic of creating the word itself.

Reports generative representation, LDA tries to back off the reports to discover a collection of topics created the set.

4.5. The distribution of topics:

Patient discharge summary reports topic modeling generates a distribution of topics for every summary report. These topics can utilized as topic vectors, which correspond to another approach for Bag of Words. In these topic vectors, terms are swapping in every summary document shows the probability of a precise topic with topics and entries for that report. The topic vector concept is further precise than Bag of Words as the languages for a report generally has thousands of entries, while a topic model usually constructed with a limit of topics.

5. EXPERIMENT ANALYSIS:

For evaluating the accuracy of the model, which is utilizing in machine learning algorithm has been attaining is a significant measure by interpreting the outcomes. Supervised classification is the best in this step is straightforward, for instance, as a class label known in supervised learning of the data classified, which can evaluate performance as simple as calculates the number of faults. In the topic modeling the condition is not so simple, with LDA utilizing an algorithm to identify logical subgroupings in data. In Evaluation, should continue with an assessment of homogeneity of the words consist of the documents in every grouping is often done. In Topic modeling [57], it is possible to calculate the topic model from a statistical perception utilizing hold-out investigating document assortment.

Implementing LDA on document data set, observe the topmost frequent words that can originate in every group. Every document can allocated to a topic, based on the combination of topics. LDA will allocate every document is a set of possibilities analogous to every probable topic.

5. 1. Extracting and Visualizing Topics

5.1.1. Extraction of topics by using Latent Dirichlet allocation

The most popular topic modeling technique is Latent Dirichlet Allocation, which forms clinical discharge summaries as the combination of hidden topics; these topics are main models existed in the report. Clinical reports in the topic model is a probability model, whereas every report congaing a grouping of topics and these topics correspond to the collection of words that be inclined to happen mutually. Φk is every topic represented as the distribution of probability over lexical words. Every topic is representing as a word's vector with the probability. A clinical discharge summary characterized as an allocation of probability topics.

The LDA Topic modelling process depends on a combined distribution of probability between topics unknown and the words observed to collect the words with the probability elevated in every topic by utilizing the posterior distribution. In LDA, the popularly accepted method collapsed Gibbs sampling used in analyzing the results. These methods require several repetitions lead to the cost of computational linearly with multiple clinical reports.

In our clinical discharge summaries, the resulting topics and patterns originate from associating with suitable topics of medical reports. Table 5.1 shows various symptoms obtained by the model and Figure 1 shows the frequency of symptoms, and their probability is showing in Table 5.2 from the clinical discharge summary dataset. Topics obtained by learning across all patients, generally patients exhibit a subset of all potential topics. Medical report data set where similar words are employe across summary records, which leads too complex because there are several unique words connected to the total number of words.

Our dataset contains above 4000 topics. Some of them are as follows

```
‟Discharge”,”histori”,”medic”,”admiss”,”hospit”,”date”,”pain”,”status,
”normal”,”blood",”time",”show",”follow",”report",”diagnosi",”present",
‟cours", ”examin", ”admit", ”year", ‟summari", ‟diseas", ‟past",
”sign", ‟care", ”bilater" and many more.
```

5.1.2. Categorization of Disease symptom categorization:

5.1.2. a. Without symptom-interdependency models: In this category, each disease treats different independent symptoms. This type is generally using in vector space models by the orthogonality hypothesis of symptom vectors by an independency assumption of symptom variables.

5.1.2.b. With immanent symptom interdependency models: This type of representation allows interdependencies between symptoms, whereas the degree of the interdependency among two Symptoms are defining the model itself. These models are straightforward or not directly derived from the co-occurrence of symptoms in the entire set of clinical reports.

5.1.2.c. With transcendent term interdependency Models: This type of representation allows interdependencies between symptoms. These models do not assert how the interdependency between the two symptoms is derived.

De-identifying discharge summary reports using in this investigation are granting by the i2b2 National Center for Biomedical Computing funded by U54LM008748, and these are preparing for the Sharing for Challenges in NLP for Clinical Data organizing by Dr. Ozlem Uzuner, i2b2, and SUNY.

The generative model in LDA is summarized as follows:

1. For every topic: choose what words are probable.
2. For every clinical discharge summary report,

a) Choose what percentage of topics supposed to be in the report,

b) For every term,

i. Selecting a topic

ii. Specified this topic, decide a likely word (created in step 1).

The probabilistic generative process described as:

1. For every topic k, illustrate an allocation over terms
2. For every report d,

a) Illustrate a vector of topic percentages

b) For every term

i. Illustrate a topic assignment

ii. Illustrate a term

5.3. Re-ranking of Keyword

An association among a set of items, for any two items with probable relations, item one is either 'ranked higher than', 'ranked lower than' or 'ranked equal to' the second item, which called as a weak order or total pre-order of items. None of the items can have a similar ranking. For instance, Google search engine can rank the pages it locates according to relevance information, which is making possible for the user quickly to select the pages according to their wish. Re-ranking is to enhance the precision of retrieval documents. The reranking provides more relevant information with higher ranking to the users. After ranking consequences are returning; the user can prefer information of importance as the seed information and apply the re-ranking by which documents re-rank based on similarity measures.

In this work, topic and keyword Reranking techniques to improve the LDA amount produced for more efficient human consumption. First, illustrate re-rank topic keywords derived from LDA because these keywords order directly influences the semantics and as a result the topic importance. The topic keywords order by LDA cannot be the model for stakeholders to be aware of the topic semantics. For instance, when LDA applied to a clinical discharge summary records, common diseases such as diabetes, cancer, heart issues, fever, etc., are generally ranked elevated in numerous topics due to their relevance in all topics. These words are not made use of patients identify knowledgeable topics as all of these are not relate to them. For providing better information; the topic keywords derived from the LDA to filter the topic definitions by implementing reranking technique.

5.4. Re-ranking of topics

The randomly ordered derived topic by the LDA; those may not be equally important to the patient. The order of topics, those are more useful and important shown first. Generally, the meaning of importance may be different from one patient to another. For instance, a patient may desire to see the most important symptoms, which covers several summary reports. In this situation, the rank of a symptom would be elevated, because it refers summary report content in the dataset. On the contrary, a patient may be concern with a group of distinct symptoms that contain the smallest content be related to one another. Such situation, rank symptom depends on their uniqueness in content. Subsequently, illustrate a small number of independent application symptoms re-ranking techniques that divide the topic based ranks on diverse ranking conditions.

5.5. Clinical reports reranking

The clinical report reranking, the rule [58] is about to rank the symptoms of the patient with the highest probability in the medical reports. Which is completing by replacement ranking to redefine topics in discharge summary reports.

1. Algorithm: Ranking (**clinical reports result set CRRS*****)***

Input: **clinical reports result set CRRS**.

Output: Arranging the Result List with Ranking r.

```
do
        if (CRRS i >CRRS j) then
                Swap (Ii,Ij)
        else
                Return CRRS I with ranking Order
Until (no more Items in CRRS)
```

Table 5.1. List of symptoms

S.No.	Symptom 1	Symptom 2	Symptom 3	Symptom 4	Symptom 5	Symptom 6	Symptom 7	Symptom 8	Symptom 9	Symptom 10
1	endometri	aortic	unit	arteri	histori	confirm	date	blood	hemorrhag	overrid
2	recent	cord	per	diseas	medic	ultim	follow	status	magnet	amiodaron
3	gallbladd	cathet	time	coronari	left	around	summari	cell	tomographi	elev
4	pelvic	aneurysm	hospit	cardiac	admiss	lenni	diagnosi	white	reson	interact
5	duct	spinal	given	underw	normal	breutzoln	report	chest	side	hcl

6	nonfoc	everi	care	left	hospit	degen	procedur	increas	gait	start

In table 5.1, list of symptoms of discharge summary sheets. All these symptoms described in section 5.11.

Table 5.2. Probability of symptoms for disease by using LDA and ranking.

Patient id	symptom1	symptom2	symptom3	symptom4	symptom5	symptom6	symptom7	symptom8	symptom9	symptom10
1	0.0093	0.0148	0.0391	0.0670	0.0335	0.0111	0.2290	0.0689	0.0130	0.5139
2	0.0760	0.0869	0.0652	0.0543	0.0543	0.0543	0.4347	0.0652	0.0543	0.0543
3	0.0467	0.0607	0.0560	0.0747	0.2476	0.0747	0.0607	0.1915	0.1168	0.0700
4	0.0311	0.0249	0.0623	0.3956	0.2585	0.0373	0.0716	0.0436	0.0467	0.0280
5	0.0701	0.0491	0.0596	0.0877	0.3929	0.0421	0.1017	0.0491	0.1228	0.0245
6	0.0436	0.0187	0.0769	0.0852	0.1559	0.0727	0.0415	0.4137	0.0602	0.0311
7	0.0331	0.0165	0.0900	0.2559	0.2417	0.1042	0.0687	0.0781	0.0781	0.0331
8	0.0537	0.0950	0.0619	0.0785	0.3181	0.0289	0.2107	0.0619	0.0702	0.0206
9	0.1710	0.0427	0.0690	0.0230	0.3223	0.0263	0.1282	0.1085	0.0789	0.0296
10	0.0370	0.0679	0.0370	0.0370	0.1358	0.0370	0.4814	0.0617	0.0679	0.0370
11	0.0914	0.0242	0.0471	0.06	0.2428	0.0171	0.0328	0.4571	0.0128	0.0142
12	0.0921	0.0401	0.0141	0.0330	0.3120	0.0543	0.0378	0.3617	0.0236	0.0307
13	0.0555	0.0666	0.1	0.0666	0.0777	0.0888	0.3333	0.0555	0.0888	0.0666
14	0.0438	0.0350	0.0438	0.0657	0.5	0.0438	0.0877	0.0701	0.0526	0.0570
15	0.0193	0.0502	0.0618	0.0541	0.1934	0.2978	0.1005	0.1934	0.0174	0.0116
16	0.0156	0.0254	0.0627	0.4980	0.1941	0.0137	0.0411	0.0980	0.0235	0.0274
17	0.0652	0.0380	0.0326	0.0434	0.2934	0.0380	0.0815	0.2771	0.0652	0.0652
18	0.0628	0.0571	0.0457	0.0342	0.04	0.04	0.6	0.04	0.04	0.04
19	0.0707	0.0530	0.0619	0.0442	0.0796	0.0442	0.4778	0.0442	0.0619	0.0619
20	0.0330	0.0301	0.1149	0.2011	0.1997	0.0186	0.091	0.2068	0.0833	0.0201
21	0.0341	0.0255	0.0447	0.0234	0.3411	0.0362	0.2409	0.1194	0.0298	0.1044
22	0.1351	0.0210	0.0510	0.0750	0.2822	0.0690	0.1351	0.1921	0.0210	0.0180
23	0.0725	0.0483	0.0403	0.0483	0.0645	0.0887	0.4838	0.0564	0.0483	0.0483
24	0.1209	0.0132	0.0491	0.0132	0.3686	0.0453	0.0567	0.2608	0.0378	0.0340
25	0.0512	0.0427	0.0512	0.0512	0.0598	0.0427	0.4786	0.0427	0.1025	0.0769
26	0.0292	0.0133	0.0937	0.1717	0.3922	0.0389	0.0085	0.1644	0.0499	0.0377
27	0.1975	0.0362	0.0443	0.0443	0.125	0.0322	0.3830	0.0564	0.0403	0.0403
28	0.0211	0.0246	0.0563	0.0387	0.2676	0.0774	0.1021	0.2992	0.0774	0.0352
29	0.0884	0.0619	0.0707	0.0442	0.0619	0.0530	0.4867	0.0442	0.0442	0.0442
30	0.0466	0.0333	0.0433	0.0266	0.49	0.0233	0.04	0.1066	0.1733	0.0166
31	0.0182	0.2145	0.0771	0.2061	0.1725	0.0196	0.0897	0.1556	0.0168	0.0294
32	0.0560	0.0560	0.0467	0.0467	0.0467	0.0654	0.4766	0.0467	0.0654	0.0934
33	0.0221	0.0202	0.5378	0.0904	0.1660	0.0249	0.0784	0.0452	0.0073	0.0073
34	0.0292	0.0439	0.0390	0.5723	0.1707	0.0260	0.0341	0.0292	0.0325	0.0227
35	0.1609	0.0218	0.0300	0.0627	0.3997	0.0163	0.0354	0.2482	0.0095	0.0150
36	0.0820	0.0597	0.0522	0.0522	0.1119	0.0522	0.4104	0.0597	0.0597	0.0597
37	0.0217	0.1959	0.1010	0.1306	0.2363	0.0233	0.0217	0.2270	0.0233	0.0186
38	0.0659	0.0494	0.0549	0.0329	0.1703	0.0714	0.3736	0.0604	0.0604	0.0604
39	0.0555	0.0833	0.0925	0.0462	0.0833	0.0740	0.3981	0.0648	0.0555	0.0462
40	0.0160	0.0140	0.0722	0.1124	0.3293	0.0381	0.0381	0.1726	0.1847	0.0220

41	0.0933	0.03	0.07	0.0766	0.3666	0.11	0.1	0.09	0.0333	0.03
42	0.0342	0.0868	0.0473	0.1105	0.3289	0.0552	0.0789	0.1131	0.0947	0.05
43	0.0515	0.0515	0.0309	0.0360	0.0670	0.0618	0.5670	0.0515	0.0463	0.0360
44	0.0495	0.0965	0.0693	0.2202	0.299	0.0247	0.0643	0.1311	0.0198	0.0247
45	0.0512	0.1002	0.0645	0.1603	0.2516	0.0222	0.0489	0.2293	0.0289	0.0423
46	0.1648	0.0358	0.0609	0.0573	0.3405	0.0430	0.1362	0.0752	0.0322	0.0537
47	0.0124	0.0651	0.1914	0.1511	0.2371	0.0208	0.0166	0.2621	0.0249	0.0180
48	0.0427	0.0539	0.0408	0.0241	0.2918	0.0576	0.0464	0.3680	0.0390	0.0353
49	0.0351	0.0351	0.1022	0.3738	0.2108	0.0383	0.0702	0.0543	0.0255	0.0543
50	0.1076	0.0311	0.0708	0.0396	0.2407	0.0226	0.0368	0.3937	0.0226	0.0339
51	0.0195	0.0160	0.0409	0.0587	0.4750	0.0177	0.0284	0.0498	0.2704	0.0231

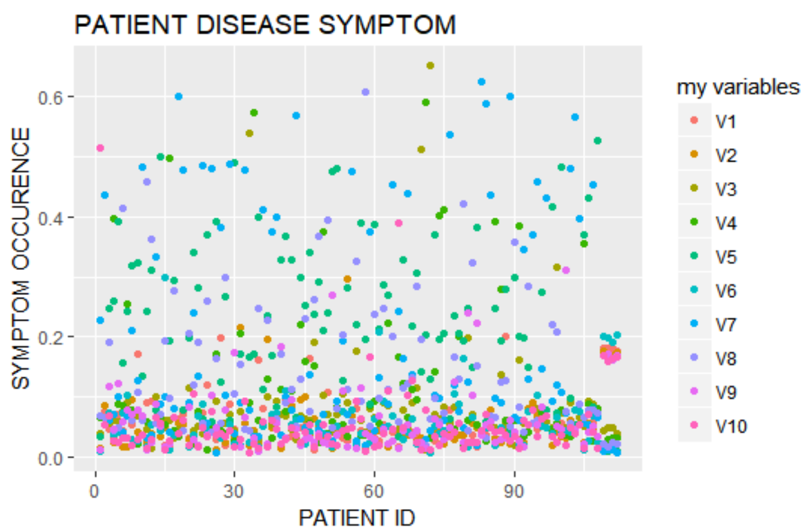

Fig: 5.1. Frequency Topic distribution

2. Algorithm: Re-ranking (Ranked* clinical reports result set RCRRS*)

Input: ***Ranked* clinical reports result set CRRS**

Output: Ordered Result List with Re-Ranking r.

CRD<--GetClinical Report data (q, r, s);

do

if (CRD=True && RCRRS i > RCRRS j) then

Swap (Ii, Ij)

else

Return RCCRS I with Re-ranking Order

Until (no more Items in RCRRS)

Table 5.3. List of Symptoms after reranking.

S.No.	Symptom 1	Symptom 2	Symptom 3	Symptom 4	Symptom 5	Symptom 6	Symptom 7	Symptom 8	Symptom 9	Symptom 10
1	duct	aneurysm	hospit	cardiac	admiss	lenni	diagnosi	white	reson	Interact
2	endometri	aortic	unit	arteri	histori	confirm	date	blood	hemorrhag	Overrid
3	gallbladd	cathet	time	coronari	left	around	summari	cell	tomographi	Elev
4	nonfoc	cord	per	diseas	medic	ultim	follow	status	magnet	amiodaron
5	pelvic	everi	care	left	hospit	degen	procedur	increas	gait	Start
6	recent	spinal	given	underw	normal	breutzoln	report	chest	side	Hcl

Table 5.4. Probability of symptoms for diseases after re-ranking.

Patient id	symptom 1	symptom 2	symptom 3	symptom 4	symptom 5	symptom 6	symptom 7	symptom 8	symptom 9	symptom 10
1	0.0149	0.0093	0.0391	0.0670	0.0335	0.0112	0.2291	0.0689	0.0130	0.5140
2	0.0870	0.0761	0.0652	0.0543	0.0543	0.0543	0.4348	0.0652	0.0543	0.0543
3	0.0607	0.0467	0.0561	0.0748	0.2477	0.0748	0.0607	0.1916	0.1168	0.0701
4	0.0312	0.0249	0.0623	0.3956	0.2586	0.0374	0.0717	0.0436	0.0467	0.0280
5	0.0702	0.0491	0.0596	0.0877	0.3930	0.0421	0.1018	0.0491	0.1228	0.0246
6	0.0437	0.0187	0.0769	0.0852	0.1559	0.0728	0.0416	0.4137	0.0603	0.0312
7	0.0332	0.0166	0.0900	0.2559	0.2417	0.1043	0.0687	0.0782	0.0782	0.0332
8	0.0950	0.0537	0.0620	0.0785	0.3182	0.0289	0.2107	0.0620	0.0702	0.0207
9	0.1711	0.0428	0.0691	0.0230	0.3224	0.0263	0.1283	0.1086	0.0789	0.0296
10	0.0679	0.0370	0.0370	0.0370	0.1358	0.0370	0.4815	0.0617	0.0679	0.0370
11	0.0914	0.0243	0.0471	0.0600	0.2429	0.0171	0.0329	0.4571	0.0129	0.0143
12	0.0922	0.0402	0.0142	0.0331	0.3121	0.0544	0.0378	0.3617	0.0236	0.0307
13	0.0667	0.0556	0.1000	0.0667	0.0778	0.0889	0.3333	0.0556	0.0889	0.0667
14	0.0439	0.0351	0.0439	0.0658	0.5000	0.0439	0.0877	0.0702	0.0526	0.0570
15	0.0503	0.0193	0.0619	0.0542	0.1934	0.2979	0.1006	0.1934	0.0174	0.0116
16	0.0255	0.0157	0.0627	0.4980	0.1941	0.0137	0.0412	0.0980	0.0235	0.0275
17	0.0652	0.0380	0.0326	0.0435	0.2935	0.0380	0.0815	0.2772	0.0652	0.0652
18	0.0629	0.0571	0.0457	0.0343	0.0400	0.0400	0.6000	0.0400	0.0400	0.0400
19	0.0708	0.0531	0.0619	0.0442	0.0796	0.0442	0.4779	0.0442	0.0619	0.0619
20	0.0330	0.0302	0.1149	0.2011	0.1997	0.0187	0.0920	0.2069	0.0833	0.0201
21	0.0341	0.0256	0.0448	0.0235	0.3412	0.0362	0.2409	0.1194	0.0299	0.1045

22	0.1351	0.0210	0.0511	0.0751	0.2823	0.0691	0.1351	0.1922	0.0210	0.0180
23	0.0726	0.0484	0.0403	0.0484	0.0645	0.0887	0.4839	0.0565	0.0484	0.0484
24	0.1210	0.0132	0.0491	0.0132	0.3686	0.0454	0.0567	0.2609	0.0378	0.0340
25	0.0513	0.0427	0.0513	0.0513	0.0598	0.0427	0.4786	0.0427	0.1026	0.0769
26	0.0292	0.0134	0.0938	0.1717	0.3922	0.0390	0.0085	0.1644	0.0499	0.0378
27	0.1976	0.0363	0.0444	0.0444	0.1250	0.0323	0.3831	0.0565	0.0403	0.0403
28	0.0246	0.0211	0.0563	0.0387	0.2676	0.0775	0.1021	0.2993	0.0775	0.0352
29	0.0885	0.0619	0.0708	0.0442	0.0619	0.0531	0.4867	0.0442	0.0442	0.0442
30	0.0467	0.0333	0.0433	0.0267	0.4900	0.0233	0.0400	0.1067	0.1733	0.0167
31	0.2146	0.0182	0.0771	0.2062	0.1725	0.0196	0.0898	0.1557	0.0168	0.0295
32	0.0561	0.0561	0.0467	0.0467	0.0467	0.0654	0.4766	0.0467	0.0654	0.0935
33	0.0221	0.0203	0.5378	0.0904	0.1661	0.0249	0.0784	0.0452	0.0074	0.0074
34	0.0439	0.0293	0.0390	0.5724	0.1707	0.0260	0.0341	0.0293	0.0325	0.0228
35	0.1610	0.0218	0.0300	0.0628	0.3997	0.0164	0.0355	0.2483	0.0095	0.0150
36	0.0821	0.0597	0.0522	0.0522	0.1119	0.0522	0.4104	0.0597	0.0597	0.0597
37	0.1960	0.0218	0.1011	0.1306	0.2364	0.0233	0.0218	0.2271	0.0233	0.0187
38	0.0659	0.0495	0.0549	0.0330	0.1703	0.0714	0.3736	0.0604	0.0604	0.0604
39	0.0833	0.0556	0.0926	0.0463	0.0833	0.0741	0.3981	0.0648	0.0556	0.0463
40	0.0161	0.0141	0.0723	0.1124	0.3293	0.0382	0.0382	0.1727	0.1847	0.0221
41	0.0933	0.0300	0.0700	0.0767	0.3667	0.1100	0.1000	0.0900	0.0333	0.0300
42	0.0868	0.0342	0.0474	0.1105	0.3289	0.0553	0.0789	0.1132	0.0947	0.0500
43	0.0515	0.0515	0.0309	0.0361	0.0670	0.0619	0.5670	0.0515	0.0464	0.0361
44	0.0965	0.0495	0.0693	0.2203	0.2995	0.0248	0.0644	0.1312	0.0198	0.0248
45	0.1002	0.0512	0.0646	0.1604	0.2517	0.0223	0.0490	0.2294	0.0290	0.0423
46	0.1649	0.0358	0.0609	0.0573	0.3405	0.0430	0.1362	0.0753	0.0323	0.0538
47	0.0652	0.0125	0.1914	0.1512	0.2372	0.0208	0.0166	0.2621	0.0250	0.0180
48	0.0539	0.0428	0.0409	0.0242	0.2918	0.0576	0.0465	0.3680	0.0390	0.0353
49	0.0351	0.0351	0.1022	0.3738	0.2109	0.0383	0.0703	0.0543	0.0256	0.0543
50	0.1076	0.0312	0.0708	0.0397	0.2408	0.0227	0.0368	0.3938	0.0227	0.0340
51	0.0196	0.0160	0.0409	0.0587	0.4751	0.0178	0.0285	0.0498	0.2705	0.0231
52	0.0436	0.0381	0.0708	0.0845	0.4796	0.0518	0.0926	0.0817	0.0381	0.0191
53	0.0497	0.0331	0.1050	0.0552	0.1271	0.0773	0.1934	0.2044	0.1105	0.0442
54	0.2961	0.0269	0.0911	0.0352	0.2816	0.0166	0.0725	0.1139	0.0186	0.0476
55	0.0680	0.0680	0.0485	0.0485	0.0583	0.0777	0.4757	0.0583	0.0485	0.0485
56	0.0804	0.0285	0.1759	0.0151	0.2211	0.0201	0.0268	0.3250	0.0151	0.0921
57	0.0614	0.0433	0.0614	0.0361	0.3899	0.0397	0.0939	0.0939	0.1155	0.0650
58	0.0194	0.0166	0.0416	0.0374	0.1953	0.0111	0.0208	0.6080	0.0402	0.0097
59	0.0468	0.0468	0.0809	0.0851	0.0894	0.0255	0.3745	0.0511	0.0340	0.1660
60	0.0333	0.0250	0.0667	0.0944	0.3861	0.0639	0.0444	0.2389	0.0306	0.0167
61	0.0455	0.0265	0.1326	0.0568	0.2121	0.2083	0.1326	0.0985	0.0682	0.0189

62	0.0537	0.0488	0.0902	0.0463	0.2878	0.0927	0.0488	0.2463	0.0463	0.0390
63	0.0199	0.0179	0.0857	0.2231	0.2689	0.0518	0.0876	0.1135	0.0219	0.1096
64	0.0411	0.0274	0.0457	0.0457	0.0457	0.0685	0.4521	0.2009	0.0274	0.0457
65	0.0172	0.0153	0.0460	0.1667	0.0345	0.0153	0.2414	0.0594	0.0153	0.3889
66	0.0409	0.0297	0.0781	0.1227	0.3271	0.0297	0.1636	0.1152	0.0446	0.0483
67	0.0427	0.0366	0.0488	0.0671	0.0427	0.0427	0.4390	0.1220	0.0549	0.1037
68	0.0426	0.0372	0.0904	0.1117	0.2181	0.0638	0.1383	0.1330	0.0372	0.1277
69	0.0359	0.0265	0.1153	0.0964	0.3062	0.0227	0.0435	0.2836	0.0454	0.0246
70	0.0311	0.0115	0.5121	0.0230	0.1415	0.0219	0.0230	0.1968	0.0253	0.0138
71	0.0191	0.0172	0.0134	0.5897	0.2023	0.0153	0.0305	0.0191	0.0153	0.0782
72	0.0315	0.0102	0.6517	0.0331	0.1308	0.0079	0.0449	0.0646	0.0126	0.0126
73	0.0394	0.0276	0.0709	0.1417	0.3701	0.0512	0.1024	0.0984	0.0630	0.0354
74	0.0433	0.0236	0.0630	0.4016	0.1969	0.0394	0.0984	0.0591	0.0551	0.0197
75	0.0293	0.0220	0.0842	0.4103	0.2051	0.0440	0.0623	0.0733	0.0293	0.0403
76	0.0786	0.0429	0.0571	0.0500	0.0643	0.0500	0.5357	0.0357	0.0500	0.0357
77	0.0460	0.0293	0.0837	0.0460	0.2343	0.1046	0.1213	0.1674	0.1255	0.0418
78	0.0750	0.0375	0.0750	0.2000	0.2063	0.0875	0.0750	0.1313	0.0563	0.0563
79	0.0379	0.0238	0.0498	0.0195	0.1928	0.1647	0.0498	0.4204	0.0260	0.0152
80	0.0218	0.0218	0.1987	0.0611	0.2467	0.0175	0.1245	0.0437	0.2402	0.0240
81	0.0564	0.0359	0.0718	0.0667	0.1487	0.0667	0.1026	0.3231	0.1026	0.0256
82	0.0503	0.0186	0.0521	0.0540	0.3818	0.0242	0.0205	0.1508	0.2235	0.0242
83	0.0451	0.0451	0.0451	0.0451	0.0376	0.0376	0.6241	0.0376	0.0376	0.0451
84	0.0380	0.0326	0.0489	0.0272	0.0598	0.0489	0.5870	0.0326	0.0543	0.0707
85	0.0777	0.0583	0.0583	0.0680	0.0680	0.0583	0.4369	0.0680	0.0583	0.0485
86	0.0315	0.0280	0.0490	0.3916	0.2483	0.0210	0.0979	0.0664	0.0245	0.0420
87	0.0247	0.0247	0.1370	0.2795	0.1945	0.0493	0.0904	0.1260	0.0301	0.0438
88	0.2000	0.0483	0.0552	0.0621	0.2793	0.0310	0.1276	0.1276	0.0310	0.0379
89	0.0565	0.0217	0.0478	0.0522	0.0478	0.0348	0.6000	0.0522	0.0609	0.0261
90	0.0386	0.0386	0.0193	0.0611	0.2990	0.0418	0.0675	0.3569	0.0418	0.0354
91	0.0300	0.0158	0.1609	0.3833	0.2019	0.0205	0.0615	0.0836	0.0284	0.0142
92	0.1986	0.0244	0.0557	0.0348	0.1986	0.0209	0.3449	0.0418	0.0174	0.0627
93	0.0806	0.0538	0.0753	0.0430	0.1505	0.0753	0.1290	0.2849	0.0538	0.0538
94	0.0899	0.0562	0.0562	0.0562	0.0787	0.0787	0.3708	0.0562	0.0787	0.0787
95	0.0685	0.0479	0.0822	0.0342	0.0548	0.1027	0.4589	0.0616	0.0548	0.0342
96	0.0737	0.0737	0.0737	0.0737	0.2737	0.0526	0.1474	0.0842	0.0842	0.0632
97	0.0588	0.0490	0.0588	0.0686	0.0784	0.0490	0.4314	0.0686	0.0686	0.0686
98	0.0463	0.0193	0.0502	0.0463	0.4170	0.0386	0.0927	0.2201	0.0347	0.0347
99	0.0410	0.0410	0.3169	0.0574	0.0738	0.0656	0.1202	0.2077	0.0519	0.0246
100	0.0558	0.0340	0.0728	0.0777	0.4830	0.0388	0.1092	0.0461	0.0461	0.0364
101	0.0463	0.0327	0.0490	0.0381	0.3106	0.0436	0.0736	0.0763	0.3106	0.0191

102	0.0417	0.0625	0.0903	0.0347	0.0972	0.0417	0.4792	0.0556	0.0486	0.0486
103	0.0465	0.0465	0.0388	0.0388	0.0543	0.0775	0.5659	0.0465	0.0465	0.0388
104	0.0659	0.0549	0.0659	0.0549	0.0879	0.0769	0.3956	0.0879	0.0549	0.0549
105	0.0243	0.0243	0.0512	0.3558	0.3693	0.0135	0.0566	0.0674	0.0216	0.0162
106	0.0569	0.0569	0.0925	0.0890	0.4306	0.0285	0.0534	0.0783	0.0427	0.0712
107	0.0602	0.0463	0.0880	0.0278	0.0833	0.0694	0.4537	0.0787	0.0509	0.0417
108	0.0691	0.0259	0.0821	0.0799	0.5270	0.0259	0.0670	0.0734	0.0346	0.0151
109	0.1807	0.1772	0.0432	0.0176	0.0102	0.2015	0.0102	0.0219	0.1681	0.1694
110	0.1809	0.1718	0.0489	0.0304	0.0100	0.1978	0.0130	0.0174	0.1709	0.1590
111	0.1820	0.1701	0.0481	0.0319	0.0114	0.1915	0.0132	0.0255	0.1625	0.1638
112	0.1763	0.1735	0.0393	0.0319	0.0091	0.2033	0.0089	0.0221	0.1687	0.1668

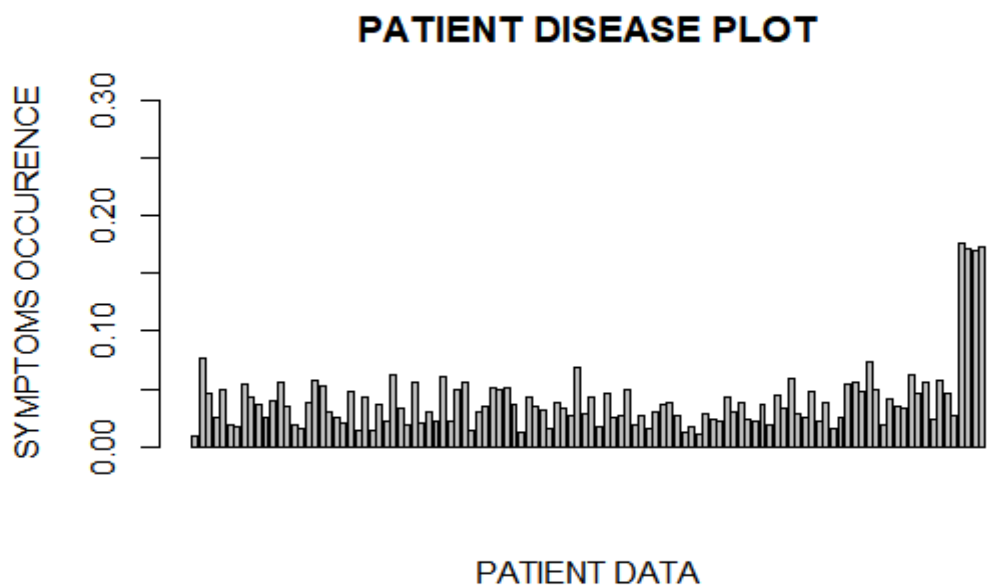

Fig 5.2. Frequency Topic distribution after Reranking.

6. CONCLUSION

By incorporating patient discharge summary metadata, over and above, in order to capturing topics in the clinical document, the topic representation of medical reports is improved. The integrate topic modeling of LDA, allows the concept, test and disease studies using discriminating words which are unclear using the Bag of words (BoW) method. Common unsupervised methods for topic modeling can determine hidden formation in huge datasets of unstructured medical records.

The integration of patient and medical report data generates more knowledge about the prior topics included in a text. Our implementation results of reranking technique indicated conditions grouped as topics. The performance achieved by our technique in exhibiting the recognized topics is promising and can be useful in more reliable clinical decision making, since all the available data is used to identifying related symptoms that can be used for facilitating clinical diagnosis with the patient's condition.

In the future, a hierarchical topic model is going to be developed using fuzzy concepts and dynamic application which will automatically summarize patient medical records. The approach will include the topic identification, concept and time-oriented views, providing support for multilingual text summarization with the help of MapReduce framework to smooth the progress of different medical records.

REFERENCES

[1].Shameer K, Badgeley MA, Miotto R, Glicksberg BS, Morgan JW, et al. Translational bioinformatics in the era of real-time biomedical, health care, and wellness data streams. Briefings in Bioinformatics 2016; 1-20.

[2].Chiauzzi E, Rodarte C, DasMahapatra P. Patientcentered activity monitoring in the self-management of chronic health conditions. BMC Medicine 2015; 1: 1-6.

[3] Moumita Bhattacharya1et al., "Identifying Patterns of Associated-Conditions through Topic Models of Electronic Medical Records", 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)

[4] William Speier et al ., "Using phrases and document metadata to improve topic modeling of clinical reports", Journal of Biomedical Informatics, 2016

[5] C.W. Arnold, S.M. El-Saden, A.A.T. Bui, R. Taira, Clinical case-based retrieval using latent topic analysis, AMIA Annu. Symp. Proc. 2010 (2010) 26–30.

[6] J.C. Feblowitz, A. Wright, H. Singh, L. Samal, D.F. Sittig, Summarization of clinical information: a conceptual model, J. Biomed. Inform. 44 (2011) 688-699, http://dx.doi.org/10.1016/j.jbi.2011.03.008.

[7] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (2012) 993–1022, http://dx.doi.org/10.1162/jmlr.2003.3.4-5.993.

[8]. Dean J, Ghemawat S (2010) MapReduce: A flexible data processing tool. Commun ACM 53(1):7277.

[9]. Borthakur, D. (2007) The hadoop distributed file system: Architecture and design. Hadoop Project Website (Available online at - https://hadoop.apache.org/docs/r1.2.1/hdfs_design.pdf). p 1-14, Accessed in 15 April 2014.

[10] J. Wiens, J. V. Guttag, and E. Horvitz, "On the promise of topic models for abstracting complex medical data: A study of patients and their medications," in NIPS Workshop on Personalized Medicine, 2011.

[11] Z. Jiang, X. Zhou, X. Zhang, and S. Chen, "Using link topic model to analyze traditional Chinese medicine clinical symptom-herb regularities," in e-Health Networking, Applications and Services (Healthcom), 2012 IEEE 14th International Conference on. IEEE, 2012, pp. 15-18.

[12] E. Sarioglu, K. Yadav, and H.-A. Choi, "Topic modeling-based classification of clinical reports," ACL 2013, p. 67, 2013.

[13] M. Bundschus, M. Dejori, S. Yu, V. Tresp, and H.-P. Kriegel, "Statistical modeling of medical indexing processes for biomedical knowledge information discovery from text," in Proceedings of the 8th International Workshop on Data Mining in Bioinformatics (BIOKDD'08), 2008.

[14] M. J. Paul and M. Dredze, "Experimenting with drugs (and topic models): Multi-dimensional exploration of recreational drug discussions," in AAAI 2012 Fall Symposium on Information Retrieval and Knowledge Discovery in Biomedical Text, 2012.

[15] C. W. Arnold, S. M. El-Saden, A. A. Bui, and R. Taira, "Clinical case-based retrieval using latent topic analysis," in AMIA Annual Symposium Proceedings, vol. 2010. American Medical Informatics Association, 2010, p. 26.

[16] Efsun Sarioglu et al. "Topic Modeling Based Classification of Clinical Reports", Proceedings of the ACL Student Research Workshop, pp. 67-3, Sofia, Bulgaria, August 4-9 2013. c 2013 Association for Computational Linguistics.

[17] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. Journal of machine Learning research, 3:993 1022, 2003.

[18]. T. Asou and K. Eguchi. Predicting protein-protein relationships from literature using collapsed variational latent Dirichlet allocation. In Proceedings of the 2nd international workshop on Data and text mining in bioinformatics, pp. 77-80. ACM, 2008.

[19]. C. W. Arnold, S. M. El-Saden, A. A. Bui, and R. Taira. Clinical case-based retrieval using latent topic analysis. In AMIA Annual Symposium Proceedings, volume 2010, page 26. American Medical Informatics Association, 2010.
[20] C. Arnold and W. Speier. A topic model of clinical reports. In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, pages 1031{1032. ACM, 2012.
[21]. J. A. Dawson and C. Kendziorski. Survival supervised latent Dirichlet allocation models for genomic analysis of time-to-event outcomes. arXiv preprint arXiv:1202.5999, 2012.
[22] Z. Huang, W. Dong, H. Duan, and H. Li. Similarity measure between patient traces for clinical pathway analysis: problem, method, and applications. IEEE Journal of biomedical and health informatics, 18(1): 4-14, 2014.
[23] J. H. Chen, M. K. Goldstein, S. M. Asch, L. Mackey, and R. B. Altman. Predicting inpatient clinical order patterns with probabilistic topic models vs. conventional order sets. Journal of the American Medical Informatics Association, page 136, 2016.
[24] G. Defossez, A. Rollet, O. Dameron, and P. Ingrand. Temporal representation of care trajectories of cancer patients using data from a regional information system: an application in breast cancer. BMC medical informatics and decision making, 14 (1):24, 2014.
[25] M. Ghassemi, T. Naumann, F. Doshi-Velez, N. Brimmer, R. Joshi, A. Rumshisky, and P. Szolovits. Unfolding physiological state: mortality modelling in intensive care units. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 75- 84. ACM, 2014.
[26] R. Pivovarov, A. J. Perotte, E. Grave, J. Angiolillo, C. H. Wiggins, and N. Elhadad. Learning probabilistic phenotypes from heterogeneous her data. Journal of biomedical informatics, 58:156-165, 2015.
[27] R. Cohen, I. Aviram, M. Elhadad, and N. Elhadad. Redundancy-aware topic modeling for patient record notes. PloS one, 9(2): e87555, 2014.
[28]. Steve L (2012) The Age of Big Data. Big Data's Impact in the World, New York, USA, pp 1-5
[29]. Russom P (2011) Big Data Analytics. TDWI Research Report, US, pp 1-38
[30]. McAfee A, Brynjolfsson E (2012) Big Data: The Management Revolution. Harv Bus Rev 90(10):60–68

[31]. Li F, Ooi BC, Özsu MT, Wu S (2013) Distributed Data Management Using MapReduce. ACM Computing Surveys 46:1-41

[32]. Shim K (2013) MapReduce Algorithms for Big Data Analysis. Databases in Networked Information Systems, Springer, Berlin, Heidelberg, Germany, pp 44-48.

[33]. Shim K (2012) MapReduce Algorithms for Big Data Analysis, Framework. Proceedings of the VLDB Endowment 5(12):2016–2017

[34]. Lee K-H, Lee Y-J, Choi H, Chung YD, Moon B (2011) Parallel Data Processing with MapReduce: A Survey. ACM SIGMOD Record 40(4):11–20

[35]. Li HG, Wu GQ, Hu XG, Zhang J, Li L, Wu X (2011) K-means clustering with bagging and mapreduce. Proc. 2011 44th Hawaii International Conference on IEEE System Sciences (HICSS). Kauai/Hawaii, US, pp 1-8.

[36]. Galgani F, Compton P, Hoffmann A (2012) Citation based summarisation of legal texts. Proc. of 12th Pacific Rim International Conference on Artificial Intelligence. Kuching, Malaysia, pp 40-52.

[37]. Hassel M (2004) Evaluation of Automatic Text Summarization. Licentiate Thesis, Stockholm, Sweden, pp 1-75

[38]. Hu Q, Zou X (2011) Design and implementation of multi-document automatic summarization using MapReduce. Computer Engineering and Applications 47(35):67-70

[39]. Lai C, Renals S (2014) Incorporating Lexical and Prosodic Information at Different Levels for Meeting Summarization, Proceedings of the 15th Annual Conference of the International Speech Communication Association, INTERSPEECH 2014. ISCA, Singapore, pp 1875-1879

[40]. Fowkes J, Ranca R, Allamanis M, Lapata M, Sutton C (2014) Autofolding for Source Code Summarization. Computing Research Repository 1403(4503):1-12

[41]. Tzouridis E, Nasir JA, Lahore LUMS, Brefeld U (2014) Learning to Summarise Related Sentences. The 25th International Conference on Computational Linguistics (COLING'14), Dublin, Ireland, pp 1-12, ACL

[42]. T. M. Mitchell, Topic modeling for medical data 38. Machine learning. Web, 1997.

[43] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6):391-407.

[44]. Thomas Hofmann. 1999. Probabilistic latent semantic analysis. In UAI.

[45] David M. Blei, Andrew Y. Ng and Michael I. Jordan. 2003. Latent Dirichlet Allocation. J. Mach. Learn.Res., 3:993-1022.

[46] Thomas Hofmann. 2001. Unsupervised learning by probabilistic latent semantic analysis. Mach. Learn., 42(1-2):177-196.

[47] Thomas L. Griffiths, Mark Steyvers, David M. Blei, and Joshua B. Tenenbaum. 2004. Integrating Topics and Syntax. In NIPS, pages 537–544.

[48] A. Aamold, "Case-based reasoning: Foundation issues." AICOM 7, pp 39-59, 1994

[49] A. Aamold and E. Plaza, "Case-based Reasoning: foundation issues, methodological variation, and System approach," AI Communication 7(1), pp. 39-59, 1994.

[51] Arthur Asuncion, Max Welling, Padhraic Smyth, and Yee-Whye Teh. 2009. On smoothing and inference for topic models.

[52] N K Nagwani, "Summarizing large text collection using topic modeling and clustering based on MapReduce framework" Journal of Big Data, 2015, DOI 10.1186/s40537-015-0020-5.

[53]. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. Journal of Machine Learning Research 3, 993–1022 (2003).

[54]. Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., Blei, D.: Reading tea leaves: How humans interpret topic models. In: Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C.K.I., Culotta, A. (eds.) Advances in Neural Information Processing Systems 22, pp. 288-296 (2009).

[55] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," Journal of Machine Learning Research, vol. 3, pp. 993-1022, 2003.

[56] T. L. Griffiths and M. Steyvers, "Finding scientific topics." Proc Natl Acad Sci USA, vol. 101 Suppl 1, pp. 5228-5235, April 2004.

[57] Yangqiu Song et al. "Topic and Keyword Re-ranking for LDA-based Topic Modeling." *CIKM'09,* November 2–6, 2009, Hong Kong, China. ACM 978-1-60558-512-3/09/11.

[58]. V Kakulapati et al. "A Re-Ranking Approach Personalized Web Search results by using Privacy Protection" Advances in Intelligent Systems and Computing, ISBN 978-81-322-2750-2 \ ISBN 978-81-322-2752-6 (e-book), DOI: 10.1007/978-81-322-2752-6, pp. 77-88.