Drug Dote And Healthcare Analysis Using Topic Modeling

Sunil Bhutada

Abstract:The health of an individual is quite a substantial factor in the growing world. We can witness numerous cases where people are found to take drugs devouring their lives. As opposed to obtaining information from regular methods like probabilistic analysis, latent semantic, LDA etc, this paper aims at introducing a new method of procuring the information of an individual's health data where akin symptoms are put together to predict how much extent an individual is suffering from the drug which helps in taking preventive measures. The system which is being proposed by us works really fine without any third party intervention.

Keywords: LDA, Gibbs sampling, Topic modeling, generalized clusters, etc..

I. INTRODUCTION

Technology in today's health care is rapidly advancing."DRUG DOTE" intents at dealing with drug addicts, where we give their symptoms as inputs in order to predict the after effect of taking particular drugs, providing preventive measures. In this algorithm, we used a supervised learning technique so that it can obtain hidden information from the prearranged data. It also gives a new method of obtaining results of an individual using text corpus, cluster analysis, and probabilistic methods to acquire data from latent semantic, Latent Dirichlet Allocation(LDA) and Gibbs sampling using LDA. The Output is in form of a graph(comparative results) which is having both concealed and observed results. This algorithm can give effective information to a troop of people and if necessary it also uses algorithms like linear regression and multi-regression. It is also tied up with a medical organization to give good medical inputs without the intervention of the third party.

For a world expecting changes every moment there is a need of meeting expectations especially in the field of Information Technology where everything is digitalized. Requirement of the software systems is a dire need for all the Firms / Organizations. For the retrieval of information, techniques have been drastically changing for the past ages. Today every individual expects easy and well-organized result for any requirement.

Data Retrieval is not a modern advancement the techniques may be changed but the concept is from times. There are several techniques for extracting the information from vast amount of data where the data may be in any format. The final outcome should be the desired information to suit their requirements. Here we use unsupervised learning practices where we pull out the information from the unordered data.

Topic Modeling is the above kind of technique use for the text data where there are several sub modules in order to dig out the preferred topics. In this method we take intact text corpus and cluster the similar kind of topics based on the probability of occurrence i.e. each topic is a combination of various words distributed over with probabilities. Therefore we select a topic which is alike to the cluster and the words for the topic are obtained based on the probability.

Supervised learning techniques can be used to obtain the hidden information from prearranged data where training and testing phases are separate. In this method we obtain the topics based on several machine learning techniques to dig out the hidden data.

II. LITERATURE SURVEY

Topic modeling has a step wise solution process of managing, organizing, and annotating huge archival text. The annotations provide us the solutions for retrieving information, classifying data and corpus description with good sort of exploring the things [1] gives theoretical details of probabilistic topic modeling and gives practical steps on implementing topic models.Topic modeling where relative similarities among documents are given is foreseen and formulation of the constraints as a loss function and proposal of a general probabilistic model that combines LDA with such constraints is made[6].A family of probabilistic time series models is developed to analyze the time evolution of topics in large document collections[11].The combination of data-driven clustering and theory-driven classification allowed for complex analysis workflows on very large text collections, thus making qualitative aspects of diachronic discourses quantifiable[7].

Supervised learning can be used when there is a requirement for assignment of the data from the labelled files (data). But in case of searching clusters based on the text corpus unsupervised algorithms can be used. A class of CNNs called deep convolution which evolved have certain architectural constraints, and demonstrate that they are a strong candidate for unsupervised learning [13].

A callow collapsed Gibbs sampling method for the widely used latent Dirichlet allocation (LDA) model is introduced resulting in speedups of real world text corpus[8].LDA estimates how much each topic contributes to each document and estimate of how much an each word contributes. Regression technique helps in predicting the possibility of a topic based on a dependent variable. This forms another pillar for the

topics to mention how accurate they are other than the various topic modelling algorithms. The use of generalised linear models for regression analysis of cost data and criteria for choosing an appropriate model are presented.[10]

Text extraction is the proper knowledge of text mining, topic building , predicting terms of the particular text. A system called Auto Slog-TS is developed that creates dictionaries of extraction patterns using only untagged text[12].Clustering takes dataset which further group the words into clusters which helps in finding hidden topic. Clustering techniques like pattern clustering methods is discussed [9].

Bioinformatics Tool helps to analyze the data and interpret the results in a biologically meaningful manner. Organization of data is done in such a way that allows researchers to access existing information and to submit new entries as they are produced[2]. In order to estimate, we use the concept of a "genome type." Genome type refers to the genomes in the population having a specific level of genetic risk for a specified disease [3]. Using this concept, an estimation of the maximum capacity of whole-genome sequencing to identify individuals at clinically significant risk for 24 different diseases is made[4]. Craving drugs can be assessed with striking reliability using instruments that are relatively easy to deploy with little burden for either drug users or clinician and craving has considerable clinical significance across multiple domains [5].

III. RELATED WORK

As the intelligence of humans is increasing in context of machine learning, artificial intelligence, deep learning there are several techniques that exist such as unsupervised and supervised learning. Each of this may relay with many algorithms based on the type of the requirements for the assignment of data from the labeled files (data) we can use the supervised learning. But in case of searching the clusters based on the text corpus unsupervised algorithms can be used.

Topic modelling gives us a detailed solution process so that we can manage, organize, annotate huge archival text. The annotations will provide us the solutions for retrieving the information, classifying data and the corpus description in order to explore the things.

They provide us an easy way to analyze the extensive quantity of the unstructured textual data. A "topic" is something which consists of collection of words which are repeatedly seen or observed. By using some of the methods it will provide us the necessary details of the similar words and it also differentiates the words with different meanings that belong to the other clusters. To understand about the present topic we can have a glance at Probabilistic Topic Models by Steyvers and Griffiths (2007) .

In most the existing and trending technologies such as deep learning and natural language processing these models are capable of generating things which give us a probabilistic idea for the word frequencies in the given files of the dataset. In this type of models/frameworks we can evaluate the data provided by ourselves which will be coming from the generative process which consists of the concealed variables. This is the process which gives us the combined probability or prediction of both the concealed and the observed variables. We will be able to calculate the conditional distribution which is termed as the posterior distribution by the help of the combined distribution.

Latent Dirichlet (said like deer-ih-CLAY) allocation, which is more commonly shortened to LDA, is one of the unsupervised ways of doing the topic modelling.

IV. PROBLEM STATEMENT

The environment /circumstances in which we live today is generating huge amount of data from dawn to dusk in so many ways such as giving up a Phone call, Texting friends, Taking pictures etc. All such activities are done for their personal benefits where we can see the world generating the data from one side and used by other side has been changed to data generated by every Individual and data being used by other Individual. It is essential for us to provide the required digital information to each and every individual and let them notice that their daily life activities (food preferences) are being affected by them on hand or the other with a little percent as today's Diet is not entirely organic.

Health is the most important factor to each individual and there is a requirement of digitizing in the health field. There are many individuals who consume drugs whereas some may not but the effect may relay on the each person with various fractions. Here we propose a novel method similar to a digital doctor where we consider and note down the symptoms of the patient and provide them what they need based on the probability of the drugs a particular person is affected with.

V. OBJECTIVE AND SCOPE

Our system is well organized and efficient enough to provide all the necessary information to each and every individual person about their respective health so that they can take the preventive measures before as it is very complicated to take actions after the person gets affected. Whoever is been affected and has come up with symptoms may provide their data so that they can know to what extent they are healthy and also know on how deep is their problem. Our System is proficient enough to provide the detailed information without the intervention of any kind of the third party or a doctor.

Our system may be extended to good platform so that we can provide every individual to access and get the information by themselves,also there is a lot of scope to get it extended with good medical inputs from the doctors so that we can obtain a better result comparatively to the checkups at the hospitals by tying up with medical organizations.

VI. ALGORITHM

LDA algorithm is one of the text processing and extraction tool where it will be following the basic pre-processing step as of the

other algorithms and thereby it has its own specific steps.
LDA may be implemented in many formats but here we will follow the Gibbs sampling Technique. This is the algorithm taken from the paper of "A Theoretical and Practical Implementation Tutorial on Topic Modelling and Gibbs Sampling by William M. Darling School of Computer Science University of Guelph"[2]

Input : words wrd E documents doc
Outcome : topic assigning z and counts $n_{d,k}$, $n_{k,w}$, and n_k

Start

Randomly initializes z followed by increasing the counter values for every repeating iteration perform

for j = 0 → N− 1 perform

wrd ← wrd[j]

tpc ← z[j]

$n_{d,tpc}$-=1; $n_{wrd,tpc}$-=1; n_{tpc}-=1 for s = 0 → s − 1 perform

$p(z = s|\cdot) = (n_{d,s} + \alpha_s)n_{s,w}+\beta_w / n_s+\beta \times W$

end

tpc ← sample from $p(z|\cdot)$

z[j] ← tpc

$n_{d,tpc}$+=1; $n_{wrd,tpc}$+=1; n_{tpc}+=1 end

end

return z, $n_{d,s}$, $n_{s,w}$, n_s end

Where the inputs and the output are :
INPUT: The huge text corpus consisting of the text files will be the input to the algorithm and thereby we need to specify the desired number/quantity of topics to be extracted out for the required number of documents. We can also specify about the required number of words to be seen in each topic.

We can also adjust our perplexity and measures of the outputs by setting the proper inputs to the application.

OUTPUT: There are majorly 2 outputs to the system they are the topics which we are extracting and the topic probabilities of each topic in each document.

We also have some of the other outputs which give us the information to document probabilities and also the word frequencies of the documents and all these are obtained in the form of the csv files where we can easily visualize each and every detail clearly. There are many other graphs which will form the major outputs to visualize how the topics are varying and how specific are the values belonging to the document.

All these inputs and outputs are to the one end and on the other end we have our regression technique where we can predict the possibility of the topic based on one of the dependent variables. This will form another pillar for the topics to classify how accurate they are when compared to the other various topic modelling algorithms.

Overall we can say that the LDA algorithm using Gibbs sampling follows the unique procedure where we can easily combine the other algorithms to optimize the results such as genetic algorithm and we can add privacy using the greedy algorithm and also many other algorithms can be used accordingly.

VII. WORKING

The present working system is based on the LDA algorithm where we can use different probabilistic, hierarchical techniques. The working of LDA is shown in the Figure 1

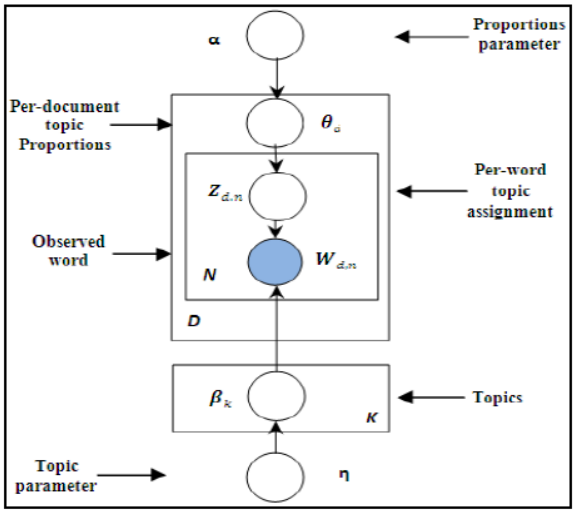


Fig.1. Graphical model representation of LDA. The boxes are "plates" representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document.

This algorithm is entirely based on the random selection that is:

First we need to select random topics from the random text files using text corpus. It consists of set of texts and used for statistic and hypothesis testing. Next we need to assign random words to the topics so, that we can obtain entire random file.

Now we have to reassign the work to the entire document based on the probability of occurrence and word probabilities.

The distortion and assigning of words to the topics is obtained because the occurrences are applied several times in order to optimize our results as we need to obtain the required number of words in each topic which are nearly related. They perform imprecise number of iterative classification in hundreds and thousands of times in order to get the final result.

Apart from this our algorithm consists of some basic notations i.e., they are

Alpha *= the topic distribution over the documents.*
Beta *= words distributed over the topics.*
K *= number of topics.*
Gamma *= the topic specificity to the document.*

W = ***specific word.***

N = ***number of words.***

These all are the basis for our project, a small case study can help us to know the exact condition and working of this algorithm :

Here we are going to apply our method to the below paragraph i.e.

"There lived a father and the son who went for a dinner on a specific occasion to a restaurant where they ordered biryani ,the restaurant was famous for its special pulav and meanwhile father has seen their cousin was sitting beside them eating roti they met accidentally had their day and left the restaurant after eating biryani"

Here if we consider there are two topics majorly they are

1) Family & 2) Food

If we consider the words included in the topics they are

Family - father, son, cousin.

Food - biryani, roti, pulav.

Word frequency is

Father = 2, son = 1, cousin = 1

Biryani =2, pulav = 1, roti = 1

If we consider alpha value to the topic it will be of the probability of topic family and word father it is 2/4 i.e. 0.5 and if we consider to the topics itself it will be 4/8 i.e. 0.5 only. In this way we can calculate the alpha and beta values and further we can use different terminology .

$$p(Z_{d,n} = k) \propto \frac{\alpha\beta}{C_k^{-n} + V\beta} + \frac{C_k^d \beta}{C_k^{-n} + V\beta} + \frac{C_k^w(\alpha + C_k^d)}{C_k^{-n} + V\beta}$$

Where z denotes the topic, k denotes number of topics ,n denotes numbers and so alpha and beta values.

Random Gibbs sampling is calculated using the following:

Where t is the topic i.e.. we can assign randomly words to the topics from 1 to n.

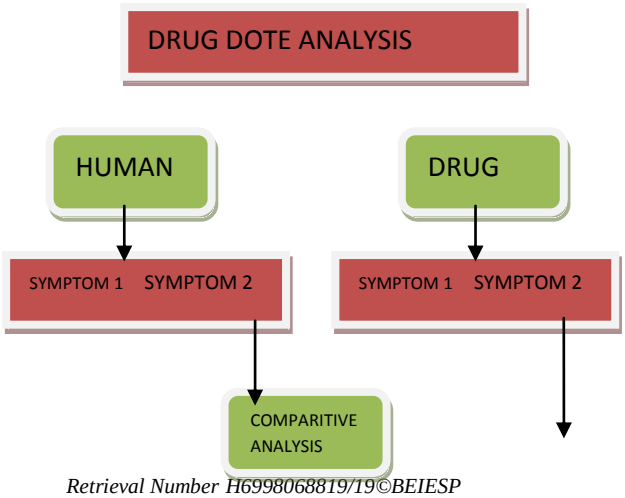
1. Randomly initialize each x_i
2. For $t = 1, ..., T$:

2.1 $x_1^{t+1} \sim p(x_1|x_2^{(t)}, x_3^{(t)}, ..., x_m^{(t)})$

2.2 $x_2^{t+1} \sim p(x_2|x_1^{(t+1)}, x_3^{(t)}, ..., x_m^{(t)})$

2.m $x_m^{t+1} \sim p(x_m|x_1^{(t+1)}, x_2^{(t+1)}, ..., x_{m-1}^{(t+1)})$

VIII. FRAMEWORK OF SYSTEM



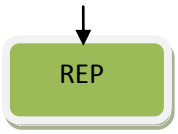


Fig1. SYSTEM FRAMEWORK

The program flows starting with two independent modules i.e, patient module and the Drug module where the files contain the consumed drug symptoms .

Step1: Preprocessing of data is used in order to remove unnecessary things such as: Delimiters, punctuations, numerics, stop words etc. After the input is provided to the text files.

Step 2: Topic distribution and word distribution are applied to the specific document where Gibbs sampling is followed.

Step 3: Here in this step we use topic and word probabilities by comparing them with other words and topics in the files.

Step 4: The number of words in the topic and topic selection is done in this step.

Step 5: In this step we should calculate the accuracy of the topics by using gamma value where it denotes the topic specification.

Step 6: The document term and word frequency matrix are build here and the final topics are retrieved.

Step 7: Here we can obtain the term probability matrix from topic and the topics probabilities from document .

Step 8: Here we can merge two topic modeling modules result i.e. of drugs module and the patient's module.

Step 9: Here comparative analysis is used in order to obtain the specific results i.e. here it compares the symptoms that are matching with the patients symptom details or not.If it does then it specifies the drug symptoms.

Step 10: Dirichlets parameters and particular graphs are generating this.

IX. RESULTS

In DRUGDOTE we can collect datasets from the database websites and the comparative analysis report of a patient with particular symptoms are calculated and observed .
Below is an analysis done and it is found that a patient with id A101 topic 7(symptoms) is of high probability which belongs to the topic 1 (symptoms) of drug database.

With this we can conclude that the occurrence of topic 1 in drug data prefers from the following drugs i.e..steroids that is calculated from gamma value table.

Text before preprocessing

```
> writeLines(as.character(docs[[36]]))
c("Zolpidem exerts its effects by dampening signals in the brain. As a result, many o
f the side effects of abuse are related to this slowing of neuronal activity, and can
 be troublesome for the user.", "", "Amnesia is an unfortunate potential side effect
of taking zolpidem that can lead to troublesome results.", "", "According to an artic
le in the Huffington Post, a Washington man woke from an Ambien sleep to hallucinatio
ns of being in a plane crash, even going so far as to call 911.", "", "Some signs of
zolpidem abuse include 1,4:",
"", "    Fatigue.", "    Headache.", "    Dizziness.", "    Struggle to maintain bala
nce.", "    Nausea.", "    Gastrointestinal issues.", "    Changes in appetite.", "
```

In middle of preprocessing after converting It into lower case and removing punctuations, delimiters, stop words etc..

```
> docs <- tm_map(docs, removeWords, stopwords("english"))
> writeLines(as.character(docs[[36]]))
czolpidem exerts effects dampening signals brain result many side effects a
use related slowing neuronal activity can troublesome user amnesia unfort
nate potential side effect taking zolpidem can lead troublesome results accordir
   article huffington post washington man woke ambien sleep hallucinations
plane crash even going far call signs zolpidem abuse include
    fatigue headache dizziness struggle maintain balance nausea
 gastrointestinal issues changes appetite body shakes redness eyes
>
```

```
> dtm
<<DocumentTermMatrix (documents: 36, terms: 1452)>>
Non-/sparse entries: 2980/49292
Sparsity           : 94%
Maximal term length: 32
Weighting          : term frequency (tf)
>
```

Beta values for some topic terms

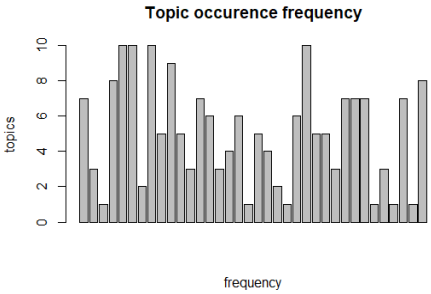
A table: 55 x 3

TOPIC	TERM	BETA
<int>	<chr>	<dbl>
1	xanax	0.0313
1	sign	0.0282
1	problem	0.0251
1	abus	0.0219
1	slow	0.0188
2	time	0.0215
2	mxe	0.0214
2	amount	0.0212

Gamma values for topics of drugs for specificity

title	topic	gamma
<chr>	*<int>*	*<dbl>*
1 Psilocybin mushroom.txt	8	0.675
2 LSD.txt	6	0.761
3 Hash drug.txt	3	0.767
4 links.txt	10	0.881
5 Streroids.txt	1	0.990
6 Actiq.txt	10	0.992
7 Soma.txt	7	0.992
8 Valium.txt	7	0.995
9 Ultrum.txt	10	0.995
10 Vicodin.txt	10	0.997

Word cloud of drug symptoms with word frequency 1



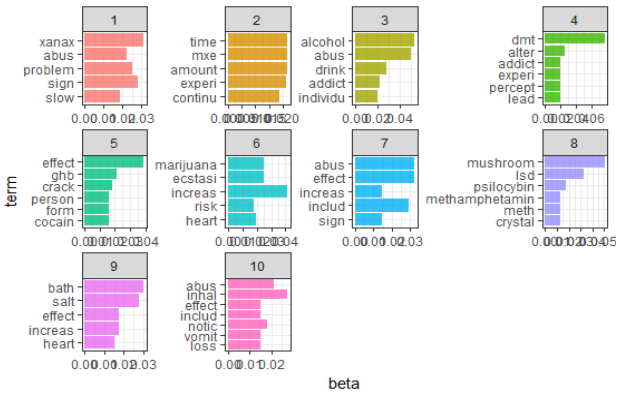
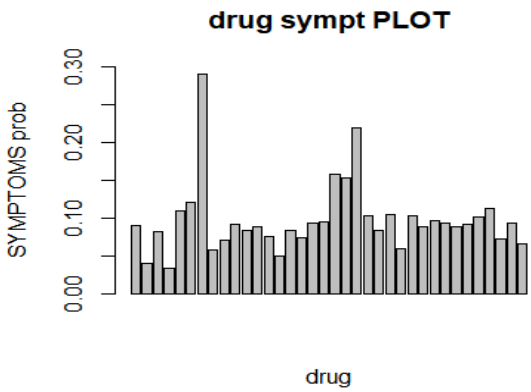
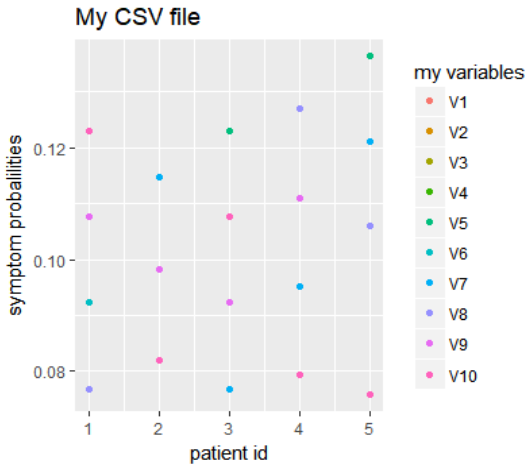


Fig 2 (Beta values for topics)



Published By:
Blue Eyes Intelligence Engineering &
Sciences Publication



X. TEXT ADVANCEMENT

Textual data inarguably plays a vital role as it is the easiest format for humans to both comprehend and express. Given any particular data, we already have a huge chunk of data on the internet. So the question here is can we obtain a summary report of the entire corpus? The recent text advancement says yes to the posed question. Sentiment analysis is applied by the organization to know the pulse of the employees or their feeling towards the work. After taking the score to the consideration they arrive at conclusion be it positive or negative feedback.

As the time progressed, topic modelling was used to obtain results which involved the details of the feedback with keywords, as they thought merely knowing whether the feedback negative or positive is insufficient.

On moving further there came another challenge, where keywords feedback is also no longer adequate. So we made a shift towards text summarization using deep learning. This also uses Natural Language Processing (NLP) which yields results that are indistinguishable whether the feedback is given by human or machine.

XI. CONCLUSION AND FUTURE SCOPE

This project steers towards the proper utilization of the topic modelling in new ways to analyze the data and comes up with the proper results of digging out the hidden topics. Thus we may conclude that this project not only deals with the addicted individuals but also with the normal humans which assist them in taking proper precautions.

In the future, as we expect more advancements in textual data where this project might facilitate in precisely predicting the details of diseases/side effects. With the proper guidelines, this can be a real-time implementation in medical centres and hospitals rendering facilities to the common man.

REFERENCES

1. Darling, W. M. “A Theoretical and Practical Implementation Tutorial on Topic Modeling and Gibbs Sampling”. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, (Portland, Oregon, USA, 2011)
2. Mabrouk M. S., Hamdy M., Mamdouh M., Aboelfotoh M., Kadah Y. M. " BIOINFTool: Bioinformatics and sequence data analysis in molecular biology using Matlab". Proceeding Cairo International Biomedical Engineering Conference, 2006
3. M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, “Cluster analysis and display of genomewide expression patterns,” Proceedings of the National Academy of Sciences, vol. 95, no. 25, pp. 14 863–14 868, 1998.
4. Nicholas John Roberts, Joshua T. Vogelstein, Giovanni Parmigani et al., “The predictive capacity of personal genome sequencing,” Science Translational Medicine 09 May 2012: Vol. 4, Issue 133, pp. 133ra58
DOI: 10.1126/scitranslmed.3003380
5. Stephen T Tiffany, “A cognitive model of drug urges and drug-use behaviour : Role of automatic and nonautomatic processes,” Psychological Review, Vol 97(2), Apr 1990, 147-168
6. Jianguang Du et al, "Topic Modelling with Document Relative Similarities ". Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)
7. David M. Blei, “Probabilistic topic models”, Communications of the ACM, v.55 n.4, April 2012. doi>10.1145 /2z33806.2133826
8. Ian Porteous et al, “Fast collapsed gibbs sampling for latent dirichlet allocation” , Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, August 24-27, 2008, Las Vegas, Nevada, USA [doi>10.1145/1401890.1401960]
9. Anil K Jain, M Narasimha Murty, and Patrick J Flynn,” Data clustering: a review”, ACM computing surveys (CSUR), 31(3):264–323, 1999.
10. Barber J, Thompson S,” Multiple Regression of cost data : use of generalized linear models”, J Health Serv Res Policy. 2004 Oct;9(4):197-204
11. David M. Blei , John D. Lafferty, “Dynamic topic models”, Proceedings of the 23rd international conference on Machine learning, p.113-120, June 25-29, 2006, Pittsburgh, Pennsylvania, USA [doi>10.1145/1143844.1143859]
12. Ellen Riloff. “Automatically generating extraction patterns from untagged text” In AAAI, 1996. 1.1, 2.1.1
13. A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” arXiv preprint arXiv:1511.06434, 2015. 13, 14, 16