SYMPTOM BASED CLINICAL DOCUMENT CLUSTERING THROUGH NON-NEGATIVE MATRIX FACTORIZATION

Dr.K.Kasturi,
Associate Professor,
Department of Information Technology,
School of Computing Sciences,
VISTAS, Chennai.
kasturi.scs@velsuniv.ac.in

Dr.K.Rohini,
Professor,
Department of Information Technology,
School of Computing Sciences,
VISTAS, Chennai.
rohini.scs@velsuniv.ac.in

Abstract

In the proposed a Doctor's clinic management kind of system, where patients will visit the clinic and if the patient is new to clinic, then receptionist will feed his/her details into the system else if the patient is already registered then the receptionist will search for the patient's name and add into the queue. When the patients turn appear, doctor will be able to see his/her details and also can check details about previous conditions if any. After doctor sees the patient, he will make entry of medicines patient needs to take. Receptionist will get those details, and there will be a section to add symptom. So at the end we will have list of symptom and medication provided for that. On this list we will apply NMF and retrieve our best possible symptom and its medication.

Keywords: *Clustering*, ***Non-Negative Matrix Factorization, Symptom-Based Clinical Document, Clinical Data.***

1. Introduction

Clustering of clinical documents is the major research area in the field of machine learning and artificial intelligence which aims to acquaint some type of association with the information that helps to highlight huge examples and patterns. The rich corpus of clinical notes consists of several unprocessed data which needs to be mined with appropriate technique to improvise the existing healthcare system. Biomedical information mining is a research strategy to recover, break down and analyze clinical information from a collection of medicinal records.Various researches have been made in this domain that have doubtlessly accomplishedincredible pace in the locale of restorative exploration and clinical practice. Clinical data mining is the way toward applying the information mining strategies on the acquired literary clinical reports. Rich content information of clinical reports contains data about drugs and syndromes. Extricating this data has proven useful to refine the medical system. Document clustering helps gather them into pertinent groups. This is done basically to find critical patterns to put a collection of the comparable articles into groups. By utilizing this method, we can perform grouping on literary information corpus that is in unstructured and semi organized configuration. These sorts of important data separated from biomedical prescriptions are very useful to build a consolidated summary for patients. Pharmaceutical data is an integration of prescription names and other medication information, for example drug name, dosage, course, etc.

There are many applications of data mining in medical field, as it has wide spread use in medical

area. It is getting great pace in medical research as well as in clinical practice. Clinical data mining is nothing but mining and make clustering of clinical data, so as to get essential data based on our requirement. Clinical documents contain textual data. By applying data mining technique on these data, we can fetch key information like medication names andsymptom names from clinical narratives. Extraction of these essential data helps health care provider to advance health care system. Clinical document plays a vital role in analysis and diagnosis of disease.

Non-Negative Matrix Factorization (NMF) is getting a lot of attention in document clusteringas well as in information retrieval. Document clustering is a fundamental technique for grouping data based on similarity and dissimilarity and then dividing theses data into subsets. Each subset is having similar kind of data or we can say that each item present in subset is having same characteristic.

Document clustering provides an intelligible summary of the collection, which can be used to provide random vision and to locate important patterns. Document clustering is a fundamental technique for content summarization, cluster-based information retrieval and automatic topic extraction. Clustering has shown remarkable progress in past decade. This paper presents a novel approach that utilizes Non-Negative Matrix Factorization Clustering approach to mine the medication names based on age and diagnosis of the patients.

This proposed System design concentrates on three types of users

1. Admin
2. Receptionist
3. Doctor

Here we proposed a Doctor's clinic management kind of system, where patients will visit the clinic and if the patient is new to clinic, then receptionist will feed his/her details into the system else if the patient is already registered then the receptionist will search for the patient's name and add into the queue. When the patients turn appear, doctor will be able to see his/her details and also can check details about previous conditions if any. After doctor sees the patient, he will make entry of medicines patient needs to take. Receptionist will get those details, and there will be a section to add symptom. So at the end we will have list of symptom and medication provided for that. On this list we will apply NMF and retrieve our best possible symptom and its medication.

2. Existing System

In the existing system, almost each and every clustering algorithms mainly deals with exploratory data analysis, where there was no prior knowledge regarding the documents to be clustered.

There was no proper method for documents clustering in multi view point. Almost all the clustering algorithms are efficient in clustering the data in single view point mechanism, but till

now there was no best approach for clustering in multi view point.

2.1 Disadvantage of Existing System

The following are the limitations of existing system that occur during clustering.

- For document having same similarities exactly, the existing system can give bestresults but fails in different similarities documents.
- No appropriate mechanisms are available for symptoms based clinical document clustering

3. Proposed System

Many clustering algorithms have been studied for decades, and the literature on the subject is very huge. Therefore decided to choose a Non-Negative Matrix Factorization algorithm in order to show the potential of the proposed approach. This algorithm was run with different combinations of their parameters, resulting in finding clusters.

3.1 Advantage of Proposed System

- Clustering is a technique of identifying all the useful data from both relevant and non-relevant data sets.
- Our proposed approach is suitable for symptom based clinical documents clustering which is used for retrieving best medications for patients as much as easy to develop the health care system.
- Our proposed system is always suitable for identifying similarities between each and every clinical document in an multi view point mechanism, whereas the existing failed in doing this.

3.2 Algorithmic Description

Non-Negative Matrix Factorization

NMF is a matrix factorization algorithm that finds the positive factorization of a given positive matrix [7, 8, 6]. Assume that the given document corpus consists of k document clusters. Here the goal is to factorize X into the non-negative m×k matrix U and the non- negative k ×n matrix VT that minimize the following objective function:

J =1/ 2X−UVT (1) where ·denotes the squared sum of all the elements in the matrix. The objective function J can be re-written as:

J =1/2 tr((X−UVT)(X−UVT)T) = 1/2 tr(XXT −2XVUT + UVT VUT) = ½ (tr(XXT)−2tr(XVUT) + tr(UVT VUT)) (2)

where the second step of derivation uses the matrix property tr(AB) = tr(BA). Let U =[uij], V =[vij], U = [U1,U2,...,Uk]. The above minimization problem can be restated as follows: minimize J with respect to U and V under the constraints of uij ≥ 0, vxy ≥ 0, where 0 ≤ i ≤ m, 0 ≤ j ≤ k,0≤ x ≤ n, and 0≤ y ≤ k. This is a typical constrainted optimization problem, and can be solved using the Lagrange multiplier method. Let αij and βij be the Lagrange multiplier for constraint uij ≥ 0 andvij ≥ 0, respectively, and α =[αij], β =[βij], the Lagrange L is,

$$L = J + tr(\alpha U^T) + tr(\beta V^T) \quad (3)$$

The derivatives of L with respect to U and V are:

$$\partial L / \partial U = -XV + UV^T V + \alpha \quad (4)$$

$$\partial L / \partial V = -X^T U + VU^T U + \beta \quad (5)$$

Using the Kuhn-Tucker condition $\alpha_{ij}u_{ij} = 0$ and$\beta_{ij}v_{ij} = 0$, we get the following equations foruij and vij: $(XV)_{ij}u_{ij} - (UV^TV)_{ij}u_{ij} = 0$ (6) $(X^TU)_{ij}v_{ij} - (VU^TU)_{ij}v_{ij} = 0$ (7)

These equations lead to the following updating formulas:uij
← uij(XV)ij (UVT V)ij (8)
vij ← vij(XT U)ij (VUT U)ij (9)

It is proven by Lee [8] that the objective function J is nonincreasing under the above iterative updating rules, and that the convergence of the iteration is guaranteed. Note that the solution to minimizing the criterion function J is not unique. If U and V are the solution to J, then,UD, VD−1 will also form a solution for any positive diagonal matrix D. To make the solutionunique, we further require that the Euclidean length of the column vector in matrix U is one.

This requirement of normalizing U can be achieved by 1:

vij ← vij i u2 ij (10)uij
← uij i u2 ij (11)
There is an analogy with the SVD in interpreting the meaning of the two non-negative matrices U and V. Each element uij of matrix U represents the degree to which term fi ∈Wbelongs to cluster j, while each element vij of matrix V indicates to which degreedocument i is associated with cluster j. If documenti solely belongs to cluster x, thenvix will take on a large value while rest of the elements in i'th row vector of V will take on a small value close to zero. In summary, our document clustering algorithm is composed of the following steps:

1. Given a document corpus, construct the termdocument matrix X in which column irepresents the weighted term-frequency vector of document di.

2. Perform the NMF on X to obtain the two non-negative matrices U and V using Eq.(8) and

Eq.(9).

3. Normalize U and V using Eq.(11) and Eq.(10).

4. Use matrix V to determine the cluster label of each data point. More precisely, examine each row i of matrix V. Assign document di to cluster x if x = argmax j vij.

The computation complexity for Eq.(11) and Eq.(10) is O(kn) and the total computation time is O(tkn), where is t is number of iterations performed.

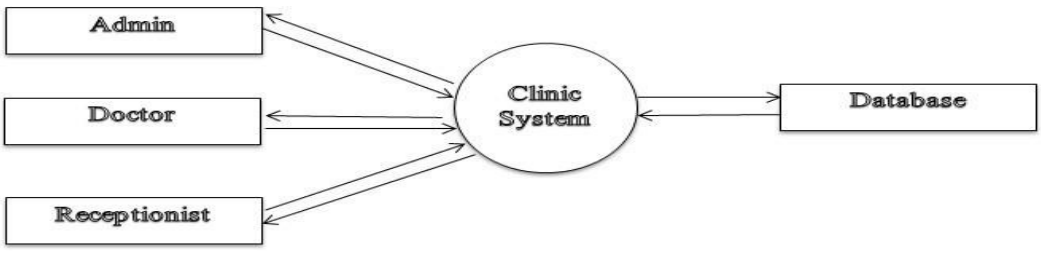
4. System Design

4.1 DFD/ ER DIAGRAM

The DFD takes an input-process-output view of a system i.e. data objects flow into the software, are transformed by processing elements, and resultant data objects flow out of the software.Data objects represented by labeled arrows and transformation are represented by circles also called as bubbles. DFD is presented in a hierarchical fashion i.e. the first data flow model represents the system as a whole. Subsequent DFD refine the context diagram (level 0 DFD), providing increasing details with each subsequent level.

The DFD enables the software engineer to develop models of the information domain & functional domain at the same time. As the DFD is refined into greater levels of details, the analyst performs an implicit functional decomposition of the system. At the same time, the DFD refinement results in a corresponding refinement of the data as it moves through the process that embodies the applications.

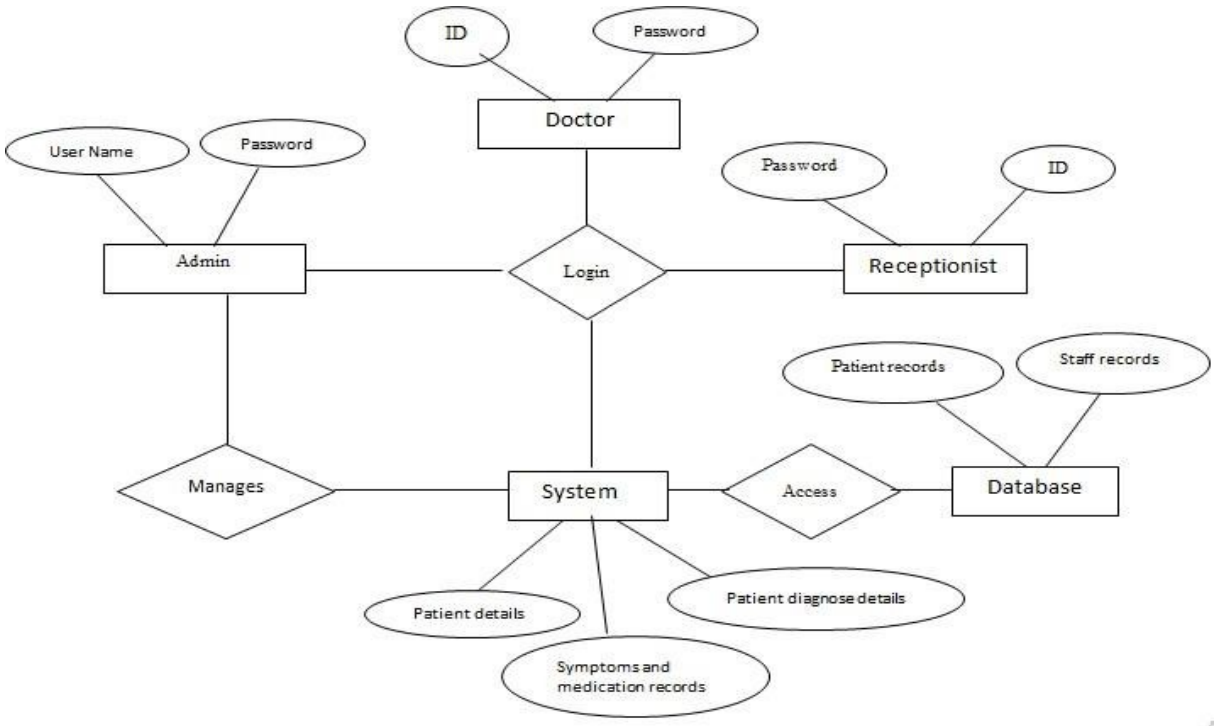
DFD Level 0



DFD Level 1

4.2 E-R Diagram

An entity relationship model, a graphical representation of entities and their relationships to each others, typically used in computing in regard to the organization of data within database or information systems .an entity is a piece of data –an objects or concept about which data is stored. A relationship is how the data is shared between entities.



5. System Implementation

Implementation is the stage in the project where the theoretical design is turned into a

working system and is giving confidence on the new system for the users that it will work efficiently and effectively. It involves careful planning, investigation of the current systemand its constraints on implementation, design of methods to achieve the change over, an evaluation of change over methods. Apart from planning major task of preparing the implementation are education and training of users. The implementation process begins with preparing a plan for the implementation of the system. According to this plan, the activities are to be carried out, discussions made regarding the equipment and resources and the additional equipment has to be acquired to implement the new system. In network backup system no additional resources are needed. Implementation is the final and the most important phase. The most critical stage in achieving a successful new system is giving the users confidence that the new system will work and be effective. The system can be implemented only after thorough testing is done and if it is found to be working according to the specification. This method also offers the greatest security since the old system can take over if the errors are found or inability to handle certain type of transactions while using the new system.

6. Performance and Limitations

Merits of the System

- Doctor can check the past health history of the patient.
- Sequential patient appointment on first come, first serve basis.

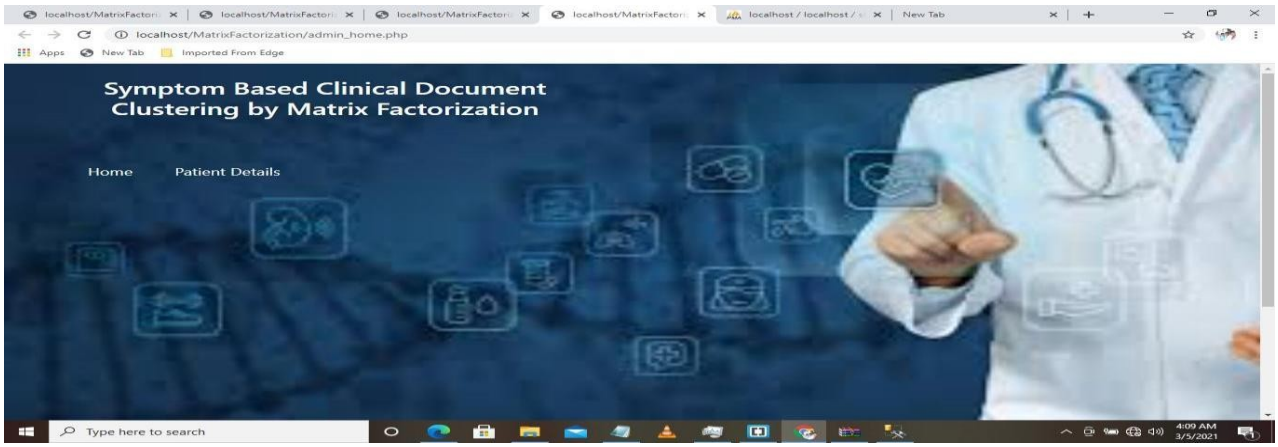
Limitations of the System

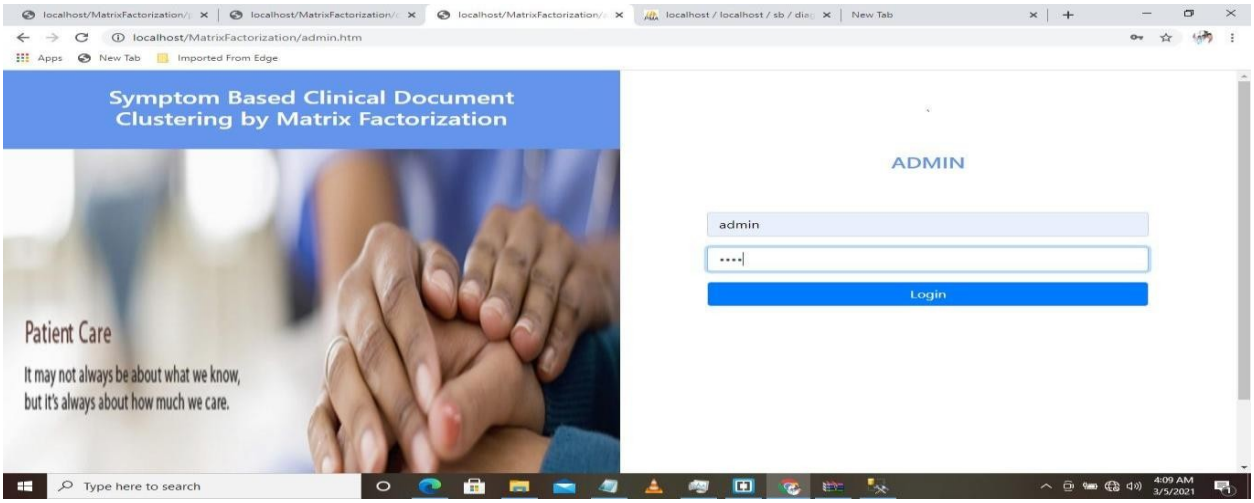
- If the data is entered incorrectly, the system will provide inaccurate results.

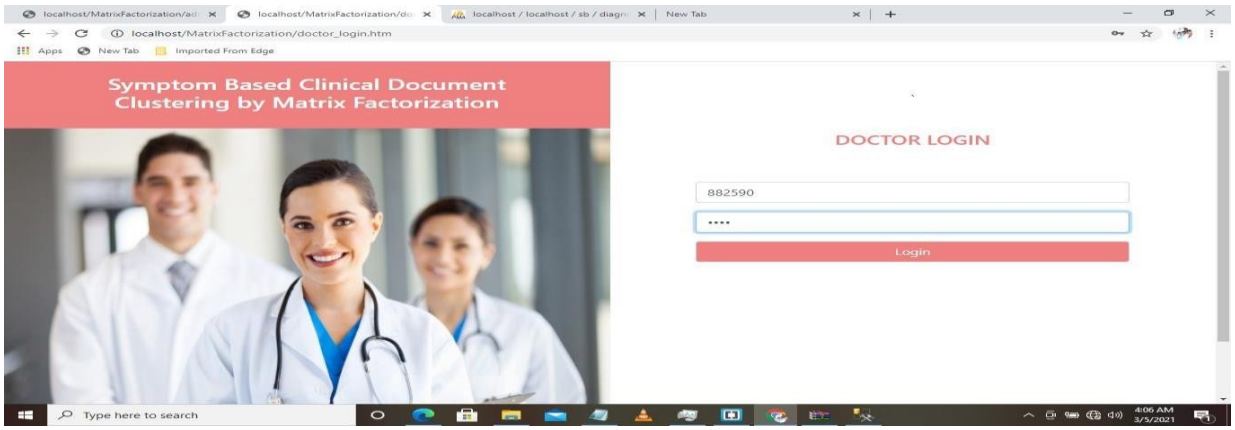
7. Conclusion

This proposed system extracts medication and symptom names from clinical documents. We have used matrix factorization to get different sections of clinical documents as well as medication names and symptom names. Proposed system also shows the accuracy of medication association with each symptom. Pre-processing before clustering provides an efficient result. Multi-view NMF is used to cluster clinical documents, which provides better performance than simple NMF. In future, we may do the clinical document clustering using the artificial intelligence. We willalso try to improve processing speed of our proposed system.

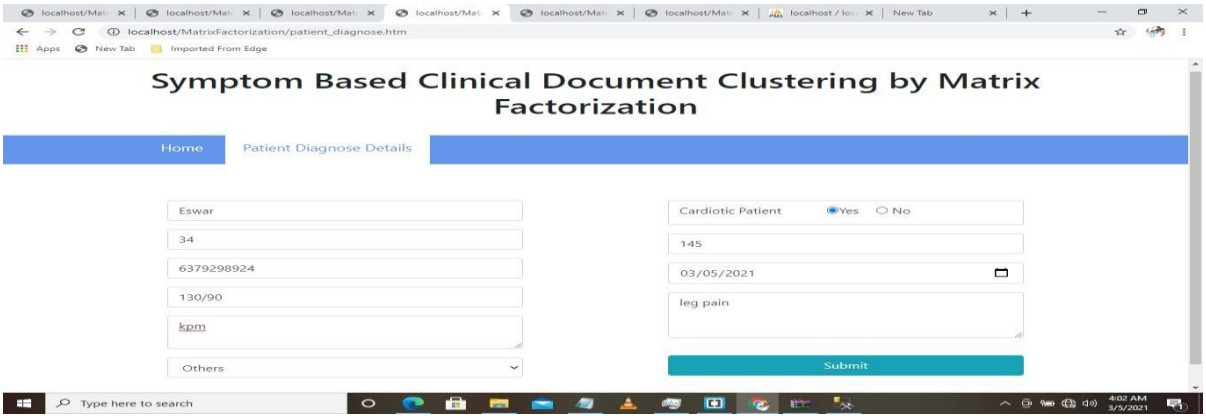
Appendices

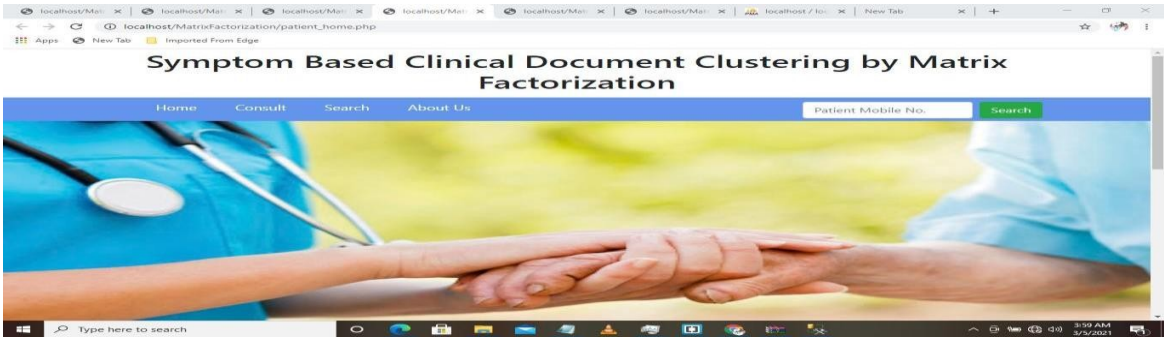

localhost/MatrixFactorization/admin_home.php
Symptom Based Clinical Document Clustering by Matrix Factorization
Home
Patient Details
Type here to search

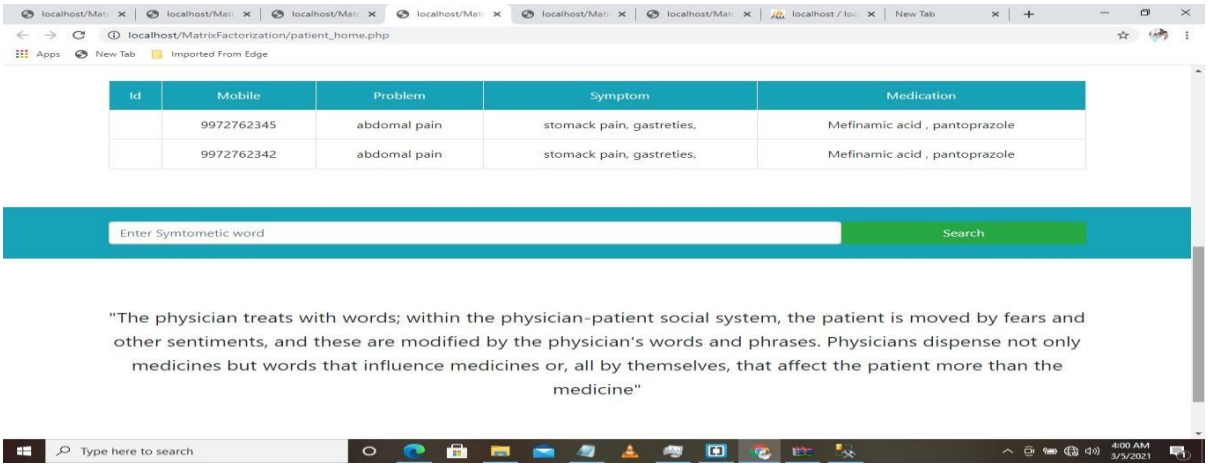

localhost/MatrixFactorization/admin.htm
Symptom Based Clinical Document Clustering by Matrix Factorization
Patient Care
It may not always be about what we know, but it's always about how much we care.
ADMIN
admin
Login
Type here to search

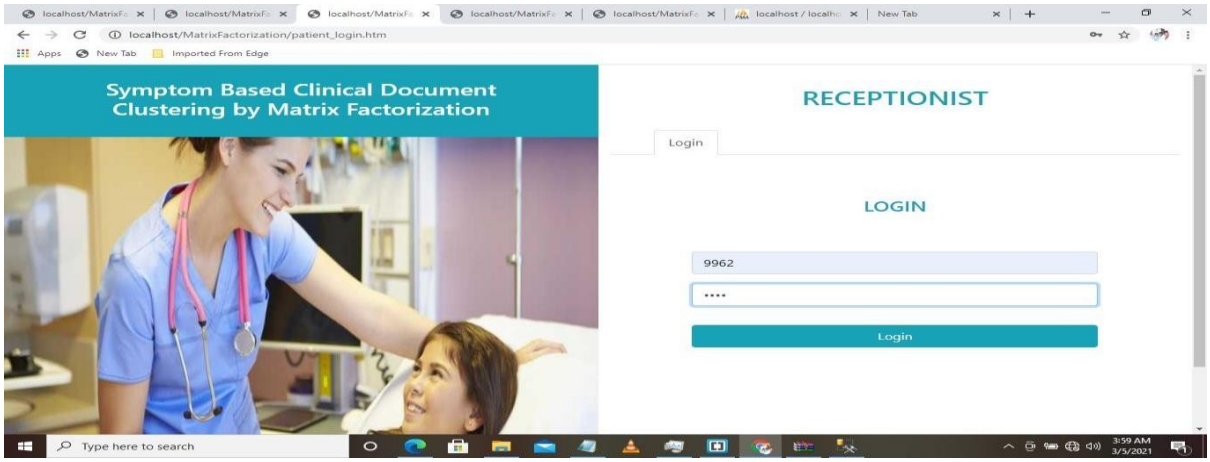


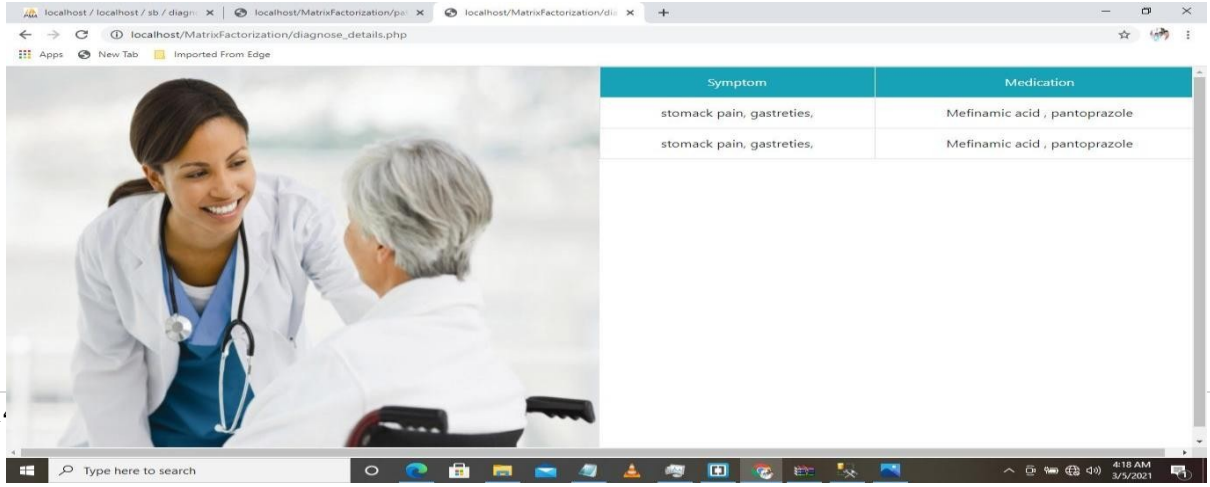
Name	Age	Mobile	BP	Address	Problem	Cardiotic	Sugar level	Others	Date
Anbu	23	9972762342	148/90	kpm	Others	Yes	156	sd	0000-00-00 00:00:00
Anbu	23	9972762341	148/90	kpm	Others	Yes	156	abdomal pain	0000-00-00 00:00:00
Anbu	34	9972762345	148/90	kpm	Others	Yes	156	abdoma pain	2021-03-04 00:00:00
Eswar	34	6379298924	130/90	kpm	Others	Yes	145	leg pain	2021-03-05 00:00:00

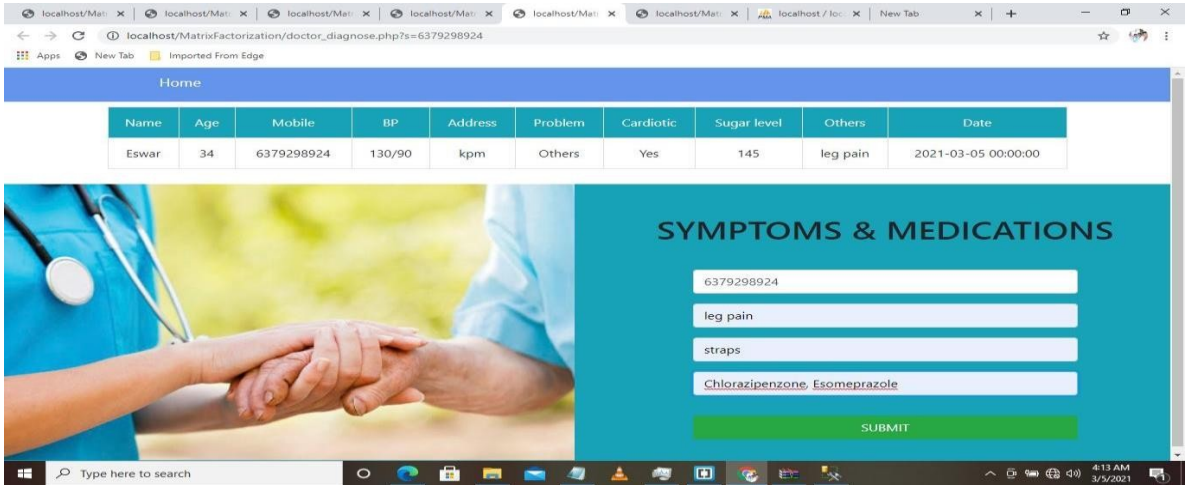



Symptom Based Clinical Document Clustering by Matrix Factorization
Home
Consult
Search
About Us
Patient Mobile No.
Search


Id
Mobile
Problem
Symptom
Medication
9972762345
abdomal pain
stomack pain, gastreties,
Mefinamic acid , pantoprazole
9972762342
abdomal pain
stomack pain, gastreties,
Mefinamic acid , pantoprazole
Enter Symtometic word
Search
"The physician treats with words; within the physician-patient social system, the patient is moved by fears and other sentiments, and these are modified by the physician's words and phrases. Physicians dispense not only medicines but words that influence medicines or, all by themselves, that affect the patient more than the medicine"


Symptom Based Clinical Document Clustering by Matrix Factorization
RECEPTIONIST
Login
LOGIN
9962
Login


Symptom
Medication
stomack pain, gastreties,
Mefinamic acid , pantoprazole
stomack pain, gastreties,
Mefinamic acid , pantoprazole



REFERENCES

[1] Fatih Altiparmak, Hakan Ferhatosmanoglu, Selnur Erdal, and Donald C. Trost, "Information Mining Over Heterogeneous And High-Dimensional Time-Series Data In Clinical Trials Databases", IEEE Transactions On Information Technology In Biomedicine, 2005, Vol. 10, no. 2.

[2] G. Hripcsak Et Al., "Mining Complex Clinical Data for Patient Safety Research: A Framework For Event Discovery", J. Biomed. Informat., 2003, Vol.36, no. 1, pp. 120–130.

[3] Jacquemin, C., "Spotting and discovering terms through natural language processing", MIT Press, 2001, http://books.google.com/books?id=W6AB06SBAGMC.

[4] Xu, H., Stenner, S.P., Doan, S., Johnson, K.B., Waitman, L.R., and Denny, J.C., "MedEx: a medication information extraction system for clinical narratives", Journal of the American Medical Informatics Association, 2010, pp. 19-24.

[5] W. Xu, X. Liu, and Y. Gong, "Document clustering based on non-negative matrix factorization", In Proceedings of the 26th Annual International ACM SI-GIR Conference on Research and Development in Informaion Retrieval (SIGIR), 2003, pp. 267-273.

[6] Hung Chim and Xiaotie Deng, IEEE "Efficient Phrase-Based Document Similarity For Clustering", IEEE Transactions On Knowledge And Data Engineering, 2008, Vol. 20, no. 9.

[7] F. H. Saad, B. D. L. Iglesia, and D. G. Bell, "A Comparison Of Two Document Clustering Approaches For Clustering Medical Documents", In Proc. Conf. Data Mining (DMIN), 2006.

[8] X. Wan and J. Yang, "Multi-document summarization using cluster-based link analysis", In Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in

Information Retrieval (SIGIR), 2007, pp. 299-306.

[9] X. Liu and W. B. Croft, "Cluster-based retrieval using language models", In Proceedings of the 27th annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), 2004, pp. 186-193.

[10] J. Silva, J. Mexia, A. Coelho, and G. Lopes, "Document clustering and cluster topic extraction in multilingual corpora", In Proceedings of the 1st IEEE International Conference on Data Mining (ICDM), 2001, pp. 513-520.

[11] C. Ding, T. Li, and M. I. Jordan, "Convex and semi-nonnegative matrix factorizations", IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, pp. 45–55.

[12] Deng Cai, Xiaofei He, Jiawei Han, and Thomas S. Huang, "Graph regularized non- negative matrix factorization for data representation", IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, pp. 1548–1560.

[13] R. Arora, M. Gupta, A. Kapila, and M. Fazel, "Clustering by left stochastic matrix factorization", In ICML, 2011.

[14] I. Dhillon, Y. Guan, and B. Kulis, "Kernel k-means, spectral clustering and normalized cuts", In KDD, 2004.

[15] Z. Akata, C. Thurau, and C. Bauckhage, "Non-negative matrix factorization in multimodality data for segmentation and label prediction," in Proc. 16th Comput. Vis. WinterWorkshop, 2011.

[16] Sonal Gupta, Diana L MacLean, Jeffrey Heer, and Christopher D Manning, "Induced lexico-syntactic patterns improve information extraction from online medical forums", J American Medical Informatics Association, 2014, pp.902-909.

[17] Yan Xu, Kai Hong, Junichi Tsujii, and Eric I-Chao Chang, "Feature engineering combined with machine learning and rule-based methods for structured information extraction from narrative clinical discharge summaries", American Medical Informatics Associatio, 2012, pp. 1-9.

[18] Aicha Ghoulam, Fatiha Barigou, Ghalem Belalem ,and Farid Meziane, "Using local grammar for entity extraction from clinical reports", International Journal of Artificial Intelligence and Interactive Multimedia, 2015, vol. 3, no. 3, pp.16-24.

[19] Louise Deleger, Katalin Molnar, Guergana Savova, Fei Xia, Todd Lingren, Qi Li, Keith Marsolo, Anil Jegga, Megan Kaiser, Laura Stoutenborough, and Imre Solti, "Large-scale evaluation of automated clinical note de-identification and its impact on information extraction", American Medical Informatics Association, 2012, pp. 1-11.

[20] Raymond J. Mooney and Razvan Bunescu, “Mining knowledge from text using information extraction”, SIGKDD Explorations, Volume 7, Issue, pp. 1-3. [21] Yefeng Wangand Jon Patrick, “Cascading classifiers for named entity recognition in clinical notes”,Workshop Biomedical Information Extraction, 2009, pp. 42-49.