An Interactive Visual Analytics System for Incremental Classification Based on Semi-supervised Topic Modeling

Kaushal Mhalgi
kmhalgi@asu.edu

Kalpana Algotar
kalgotar@asu.edu

Anchit Bhattacharya
abhatt22@asu.edu

Varun Rao Veeramaneni
vveeram2@asu.edu

Aakash Rastogi
arastog9@asu.edu

Abstract

Text labeling for documents which are not annotated is a laborious task and consumes a great amount of time. Here we have introduced an interactive visual analytics system that makes use of modified Gibbs MedLDA which classifies the documents incrementally. A scatterplot is designed to show documents and topics to help the user pick the desired label. Once the user assigns the label, Gibbs MedLDA is applied to classify the documents which fall under the given label. A scatterplot with classifier boundary, word cloud to pick the most implied label and matrix view to visualize the weights of the classifier is also shown. The user assigns the labels iteratively. We extended the model by providing Dendrogram and TreeMap which further helps the user to label effectively.

Index Terms – Document Labelling, Dendrogram, Gibbs MedLDA, Visual Analytics, TreeMap, Visualization

1. Introduction

This paper tries to use interactive visualization to help users to label a collection of unannotated text datasets. Semi-supervised topic modeling methods such as Gibbs MedLDA is used for this purpose. Gibbs MedLDA is a combination of LDA and SVM based classifiers that return both the topic information and classification information of the documents. The topic information is provided to help the user get a high-level overview of the document topics to create the labels. The Gibbs MedLDA is extended to allow for multi-label models and active learning algorithm where the user can label documents which are closer to the classification boundary i.e. low label confidence from the classifier. The topic and the classification information is visualized in many ways. For visualizing the topic information, a scatter plot with document and topic distributions are shown, word cloud for each topic is displayed, and for classification visualization, a weighted topic view for classifiers is provided. Additionally, document visualizations are added to allow the user to explore the text collection better. In that regard, a text list is provided with recommended labeling which is sorted based on the labeling uncertainty of the text, to help the users to label these documents. The full plain text would be visible on clicking text list.

2. Visualization Design

The Implementation of visualizations are divided into three parts Topic visualization, Classification visualization and Document visualization based on their usage and the content.

2.1. Topic Visualization

Given the documents, our first aim is to help users quickly grasp the content of the text collections. We preprocess the document collection by performing lemmatization and removing stop words. We generate a document-topic scatterplot as shown in fig 1. Using one hot encoding Topics are encoded as a T * T matrix M. The topic proportions of the documents is a D* T matrix. We concatenate the above two matrices and employ t-SNE using PCA.

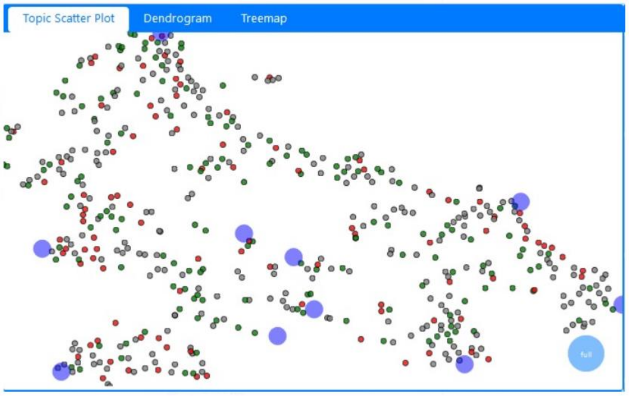


Fig 1: The topic scatterplot shows the document and topic distributions.

Each document is represented using a circle and each sector of the circle represents a label and is represented using different colors. The larger blue circles represent the topics and the size is directly proportional to the percentage of the topic in the text collection. The cluster of documents near the topic will let the user know that they belong to that topic. This scatterplot is updated after each training. Because the topic scatter plot shows only a small number of keywords, users may not be able to understand the topic meaning accurately. Thus, we provide a word cloud view(shown in fig.2) in our system. The size of a word represents the score value of each phrase. Users can browse the keywords for every topic by clicking the topic circle in the topic scatter plot.

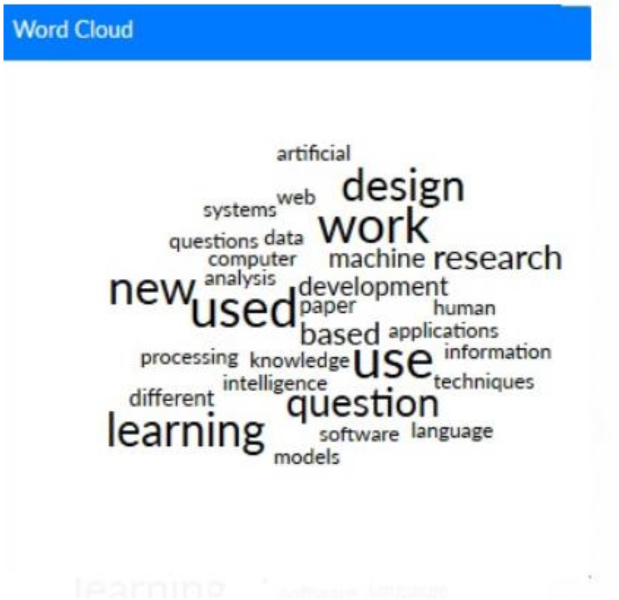


Fig 2: Word Cloud showing the keywords of a topic to convey users the meaning of the topics

To ensure the continuity of the t-SNE result and reduce the time cost, we use the previous t-SNE result as the initial value, and run t-SNE for ten iterations, as ten iterations are adequate to reach convergence.

2.2. Classification Visualization

Modified Gibbs MedLDA generates topics and suggests a few labels. The user can manually assign a label to the document and the algorithm again suggests documents that might come under the same category. The user then assigns a label to any of the suggested documents which he thinks appropriate. This process is iteratively done to label all the documents. Label list visualization guides the user while labeling the document. Two more visualizations are created to guide the user while labeling the documents and are described below.

Here is the Algorithm used (referred from Yuyu Yan et al. [1])

```
Input: text collection $C = (w_d, y_d)_{d=1}^{D}$
for t=1 to T do
    Assign labels to documents.
        Train the text collection $C$ via Gibbs MedLDA.
        for each label $l$ in $L$ do
            for each document $d$ in $D$ do
                Predict its label $y_{dl}$, the score is the absolute prediction value $score(y_{dl}) = |\eta_l^T . \bar{z}_d|$
            end for
            calculate the threshold value $bl$ according to the score values of annotated documents, $bl = max(w1\_, w2mean(score(y_{dl}), \forall d, y_{dl} \neq 0))$, where $w1$, $w2$ are the weights to balance values. In this paper, we set $w1$ as 2, and $w2$ as 0.8.
            Select a set of documents $Cs = \{dl|score(y_{dl}) > bl , y_{dl} =0\}$, and update the labels $y_{dl} = \mathrm{sign}(\eta_l^T . \bar{z}_d)$
        end for
end for
```

The weight of the topic in each label i.e. classifier-topic relationship is shown in the following topic of weight visualization. The size of the circle represents the weight that the topic has for a label as predicted by the classifier. This visualization updates each time when the user assigns a new label. This visualization guides the user to understand the topic weight and assign an appropriate label.

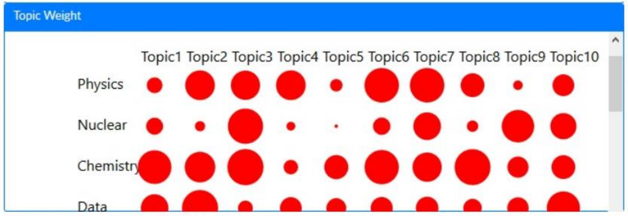


Fig 3: topic weight showing the topic weights for each classifier

The label list is created to show the basic information regarding the classification. The green part represents annotated positive, correctly annotated positive and the red part represents annotated negative, correctly annotated negative to guide the user regarding the accuracy of the classifier.

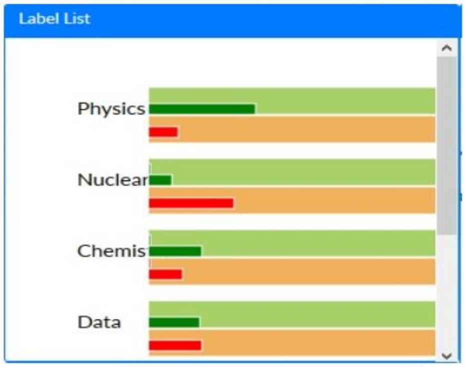
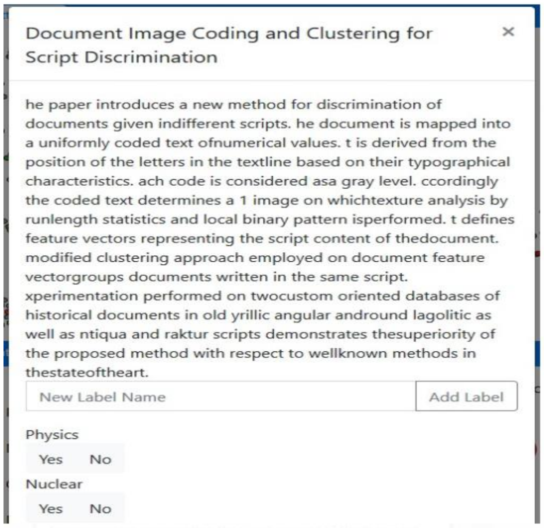


Fig 4: label list showing the classification results of the label

2.3. Document Visualization

To help users to explore the text collection, we provide a text list and a plain text view. The text list displays(fig 6) a list of documents, including titles and other meta information.

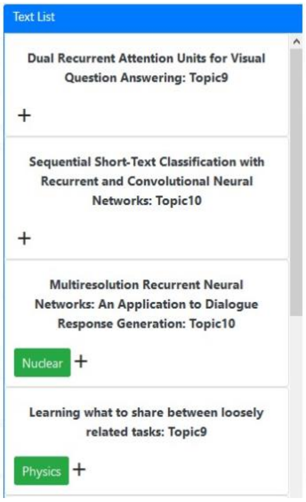


Document Image Coding and Clustering for Script Discrimination ×

he paper introduces a new method for discrimination of documents given indifferent scripts. he document is mapped into a uniformly coded text ofnumerical values. t is derived from the position of the letters in the textline based on their typographical characteristics. ach code is considered asa gray level. ccordingly the coded text determines a 1 image on whichtexture analysis by runlength statistics and local binary pattern isperformed. t defines feature vectors representing the script content of thedocument. modified clustering approach employed on document feature vectorgroups documents written in the same script. xperimentation performed on twocustom oriented databases of historical documents in old yrillic angular andround lagolitic as well as ntiqua and raktur scripts demonstrates thesuperiority of the proposed method with respect to wellknown methods in thestateoftheart.

Fig 5: Plaint text view showing the content of the document

Users can go through the text list to see the full text in the plain text view of the document for confirmation as shown in fig 5. when the document is found, they may add label to the document from one of the previously existing labels or they may create a new label and assign to the document.



Text List

Dual Recurrent Attention Units for Visual Question Answering: Topic9

+

Sequential Short-Text Classification with Recurrent and Convolutional Neural Networks: Topic10

+

Multiresolution Recurrent Neural Networks: An Application to Dialogue Response Generation: Topic10

Nuclear +

Learning what to share between loosely related tasks: Topic9

Physics +

Fig 6: New label view containing a new label with an initial training set.

The proposed system suggests few documents which can be labeled with the new label added and the user could mark the other suggested documents to fall in under it. Users can check if the suggested documents are related by clicking on the document to see the plain text view.

3. Extension

For a better understanding of the relationship between the topics and documents, we have used multiple data visualization techniques in extension to the paper like dendrogram and Treemap. In the scatter plot there is a possibility that the documents are present at the boundary of the multiple topics or the distance of the documents is equivalent from multiple topics. In this case, it's difficult for the user to know the topics assigned by the classifiers to those documents. For making such conditions simple, we have introduced Dendrogram and Treemap.

TreeMap represents the hierarchical data in the form of nested rectangles. All the topics are shown as the rectangles, whose size depends upon the proportion of that topic in the text collections. When the user will click on the topic, then the zoom-in view of that topic will be displayed. This view will show the user, all those documents which are classified under that topic by the classifier.

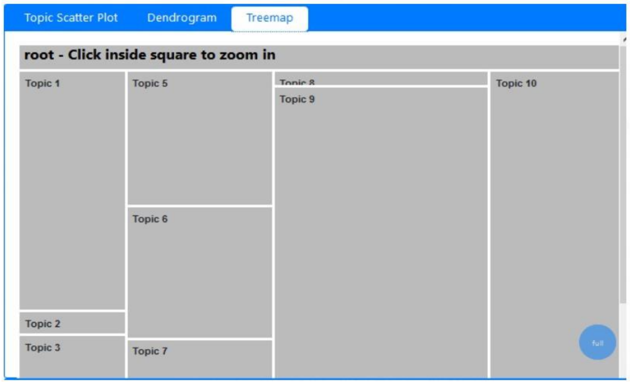
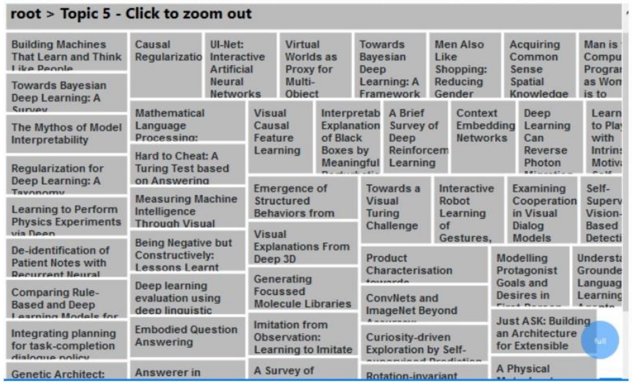


Fig 7: Zoom out view of the TreeMap showing all the topics in the topic view.

After clicking on the zoom-in strip, which is present at the top of the Treemap view, it will navigate us to the topic view. In the topic view (fig 7), we can explore other topics by clicking on a topic that will take the user to document view (as shown in fig. 8)of that topic.



root > Topic 5 - Click to zoom out
Building Machines That Learn and Think Like People
Towards Bayesian Deep Learning: A Survey
The Mythos of Model Interpretability
Regularization for Deep Learning: A Taxonomy
Learning to Perform Physics Experiments via Deep
De-identification of Patient Notes with Recurrent Neural
Comparing Rule-Based and Deep Learning Models for
Integrating planning for task-completion dialogue policy
Genetic Architect:
Causal Regularizatio
UI-Net: Interactive Artificial Neural Networks
Virtual Worlds as Proxy for Multi-Object
Towards Bayesian Deep Learning: A Framework
Men Also Like Shopping: Reducing Gender
Acquiring Common Sense Spatial Knowledge
Man is Compu Progra as Won is to
Mathematical Language Processing:
Hard to Cheat: A Turing Test based on Answering
Measuring Machine Intelligence Through Visual
Being Negative but Constructively: Lessons Learnt
Deep learning evaluation using deep linguistic
Embodied Question Answering
Answerer in
Visual Causal Feature Learning
Interpretab Explanation of Black Boxes by Meaningful
A Brief Survey of Deep Reinforcem Learning
Context Embedding Networks
Deep Learning Can Reverse Photon
Learn to Pla with Intrin Motiv
Emergence of Structured Behaviors from
Visual Explanations From Deep 3D
Generating Focussed Molecule Libraries
Imitation from Observation: Learning to Imitate
A Survey of
Towards a Visual Turing Challenge
Interactive Robot Learning of Gestures,
Examining Cooperation in Visual Dialog Models
Self-Superv Vision-Based Detecti
Product Characterisation towards
ConvNets and ImageNet Beyond Accuracy
Curiosity-driven Exploration by Self-supervised Prediction
Rotation-invariant
Modelling Protagonist Goals and Desires in First-Person
Understa Grounde Languag Learning
Just ASK: Building an Architecture for Extensible
A Physical
full

Fig 8: Zoom in view of the TreeMap showing a topic and its associated documents in the document view.

A Dendrogram is a data visualization technique that represents the relationships in the form of a hierarchical organization. In dendrogram visualization(fig 9), at the first level, we have the root node, which is present at the center of the circle. At the next level, we have topics created by the Gibbs MedLDA algorithms, which are attached from the root node by edges or branches. At the last level, we have the documents, which are connected to their associated topics. As we click on the document circle, its name will be displayed. When we click on the topic circle, we will be able to see the word cloud associated with that topic. When we click on the full tab, which is present at the bottom of the dendrogram, we will be able to see the complete view of the dendrogram.

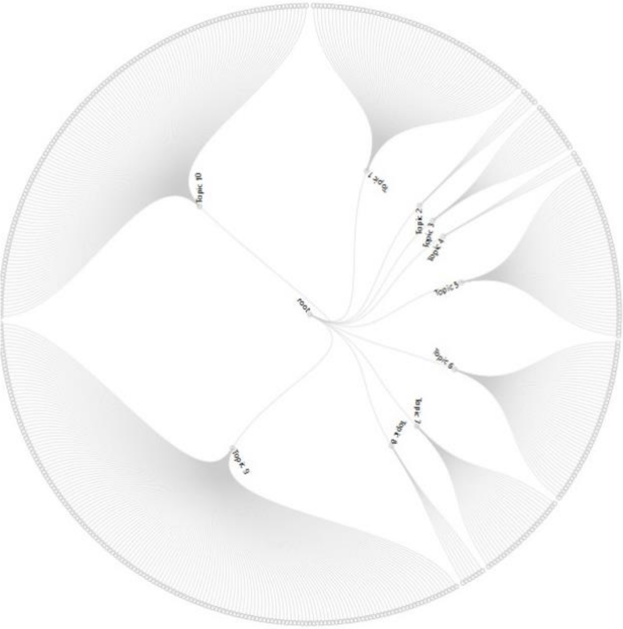


Fig 9: Dendrogram showing the relationship between the topics and the documents

4. Case Study

Here, we are presenting case studies using the two different datasets from Kaggle. One is food reviews and second is article – summary. The goal of this case study is to walk over the whole process and to know how the topic-document, topic-keywords, document distribution over different topics, document-label is generated through backend using Python and different visualization like scatter plot, dendogram, TreeMap, word cloud, Topic weight view, Label List view using d3.js shows the relationship between document, topic, keyword, and visualization allows the user to label the document and how to retrain the classifier. This case study shows the power of machine learning with visualization will allow the user without domain knowledge to label the large amount of text documents.

The process starts with collecting the data. Second is text processing. Because text data is not suitable to process with machine learning algorithms. We need to convert text data into vectorize format to process with machine learning algorithms. In text processing, we used the nltk's stopword list to remove the unnecessary words from the documents, removed special characters from the documents, convert all texts into lower case. Third, we applied TextBlob to text collections and collected noun-phrases from the collections and converted into vectorize format. So, data is ready for processing with machine learning. Fourth, we built GridSearchcv LDA to dynamically select the number of topics for the text collections. After finding the best number of topics, we built LDA to generate the topic-document, topic-keywords, topic-word distribution from vectorized input. At the end we built SVM classifier to train the text collections with label.

The interactive visualization is built using the d3.js, ajax, javascript. The data generated in backend is passed as an input in d3.js to generate the different types of visualization like scatter plot, dendogram, TreeMap. These plots show the relationship between document and topics. We created word cloud to visualize the number of keywords is associated with selected topic. The interactive visualization allows the user to label the key documents. After labeling the document, d3.js calls the backend' s SVM classifier to retain the model based on label given by the user. With the updated topic and classification information, the user continues to label documents to refine the classifiers until a satisfactory result is obtained. We design several views to help users explore the text collection and refine the classifiers. We evaluated our system via two case studies and a user study.

5. Discussion

The number of topics into which the documents are to be arranged is currently given by the user. In the future using state of the art machine learning algorithms likes of grid search algorithm the number of topics to be divided can be automated. The Scatterplot used in the project for Topic visualization does not work efficiently with a large number of topics. We have solved this issue using Treemap and dendrogram. Dendrogram used int the extension is radian but a general dendrogram can show any number of topics. The modified Gibbs MedLDA used in this project works only for static data. This can be further extended to work with real-time data

References

[1] Yuyu Yan, Yubo Tao, Sichen Jin, Jin Xu, Hai Lin, " An Interactive Visual Analytics System for Incremental Classification Based on Semi-supervised Topic Modeling", Proceedings of IEEE Pacific Visualization Symposium(PacificVis 2019), pages 148-157, 2019