A Lab Report

for

Natural Language Processing (UCS615)

Submitted by:

Nikhil Jain	**101603206**
Prakhar Gupta	**101610066**

Submitted to:
Dr. Jasmeet Singh

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Computer Science and Engineering Department Thapar Institute of Engineering and Technology, Patiala

Jan-May, 2019

INDEX

Introduction

Analytics Industry is all about obtaining the "Information" from the data. With the growing amount of data in recent years, that too mostly unstructured, it's difficult to obtain the relevant and desired information. But, technology has developed some powerful methods which can be used to mine through the data and fetch the information that we are looking for.

One such technique in the field of text mining is Topic Modelling. Topic modeling is a type of statistical modeling for discovering the abstract "topics" that occur in a collection of documents. Topic modeling is a frequently used text-mining tool for discovery of hidden semantic structures in a text body. Topic Modelling is different from rule-based text mining approaches that use regular expressions or dictionary based keyword searching techniques. It is an unsupervised approach used for finding and observing the bunch of words (called "topics") in large clusters of texts.

Topics can be defined as "a repeating pattern of co-occurring terms in a corpus". A good topic model should result in – "health", "doctor", "patient", "hospital" for a topic – Healthcare, and "farm", "crops", "wheat" for a topic – "Farming".

Topic Models are very useful for the purpose for document clustering, organizing large blocks of textual data, information retrieval from unstructured text and feature selection. For Example – New York Times are using topic models to boost their user – article recommendation engines. Various professionals are using topic models for recruitment industries where they aim to extract latent features of job descriptions and map them to right candidates. They are being used to organize large datasets of emails, customer reviews, and user social media profiles.

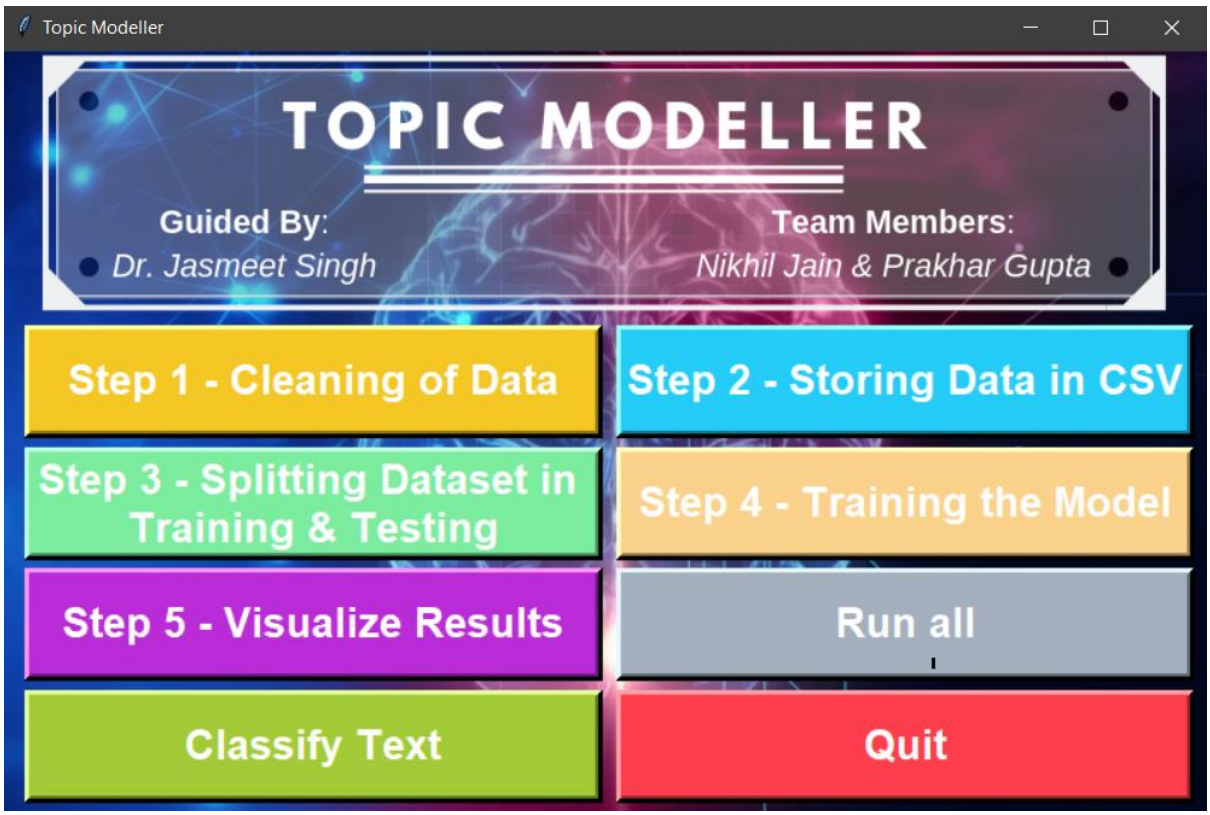

Fig1. GUI of our project

About Dataset

The 20 Newsgroups data set is a collection of approximately 20,000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups. It was originally collected by Ken Lang, for his "Newsweeder: Learning to filter netnews" paper.

The data is organized into 20 different newsgroups, each corresponding to a different topic. Some of the newsgroups are very closely related to each other (e.g. comp.sys.ibm.pc.hardware / comp.sys.mac.hardware), while others are highly unrelated (e.g misc.forsale / soc.religion.christian). A list of the 20 newsgroups, partitioned (more or less) according to subject matter is shown in fig1. The number of documents in each category is given in table 1.

comp.graphics comp.os.ms-windows.misc comp.sys.ibm.pc.hardware comp.sys.mac.hardware comp.windows.x	rec.autos rec.motorcycles rec.sport.baseball rec.sport.hockey	sci.crypt sci.electronics sci.med sci.space
misc.forsale	talk.politics.misc talk.politics.guns talk.politics.mideast	talk.religion.misc alt.atheism soc.religion.christian

Fig 2: 20 Newsgroup Categories

Table 1: Category wise training and testing set size

Category	**Total**
alt.atheism	799
comp.graphics	973
comp.os.ms-windows.misc	985
comp.sys.ibm.pc.hardware	982
comp.sys.mac.hardware	963
comp.windows.x	988
misc.forsale	975
rec.autos	990
rec.motorcycles	996
rec.sport.baseball	994
rec.sport.hockey	999
sci.crypt	991
sci.electronics	984
sci.med	990
sci.space	987
soc.religion.christian	997
talk.politics.guns	910
talk.politics.mideast	940
talk.politics.misc	775
talk.religion.misc	628
Total	**18846**

Steps of Working

1. Step 1 – Data cleansing

   In this step, data cleansing was performed. All 18846 documents were read and all stopwards, punctuations were removed. Documents headers were removed and all words were converted to lower case. All pre-processed documents were stored under same directory structure as original directory structure.

2. Step 2 – Storing data in csv file

   Each document in dataset was scanned and results were stored in 20 csv files, one for each category. Each csv file contain path to document, fileid and category of document.

3. Step 3 – Train Test Split

   Each of the 20 csv files generated in step 2 were read and each document was classified as testing set or training set. 60% of dataset was used as training set and 40% dataset was used as testing set. Two csv files were generated, one for testing and training set each. In training set csv, path to all training documents, fileids and categories were stored. Similarly, in testing set csv, path to all testing documents, fileids and categories were stored.

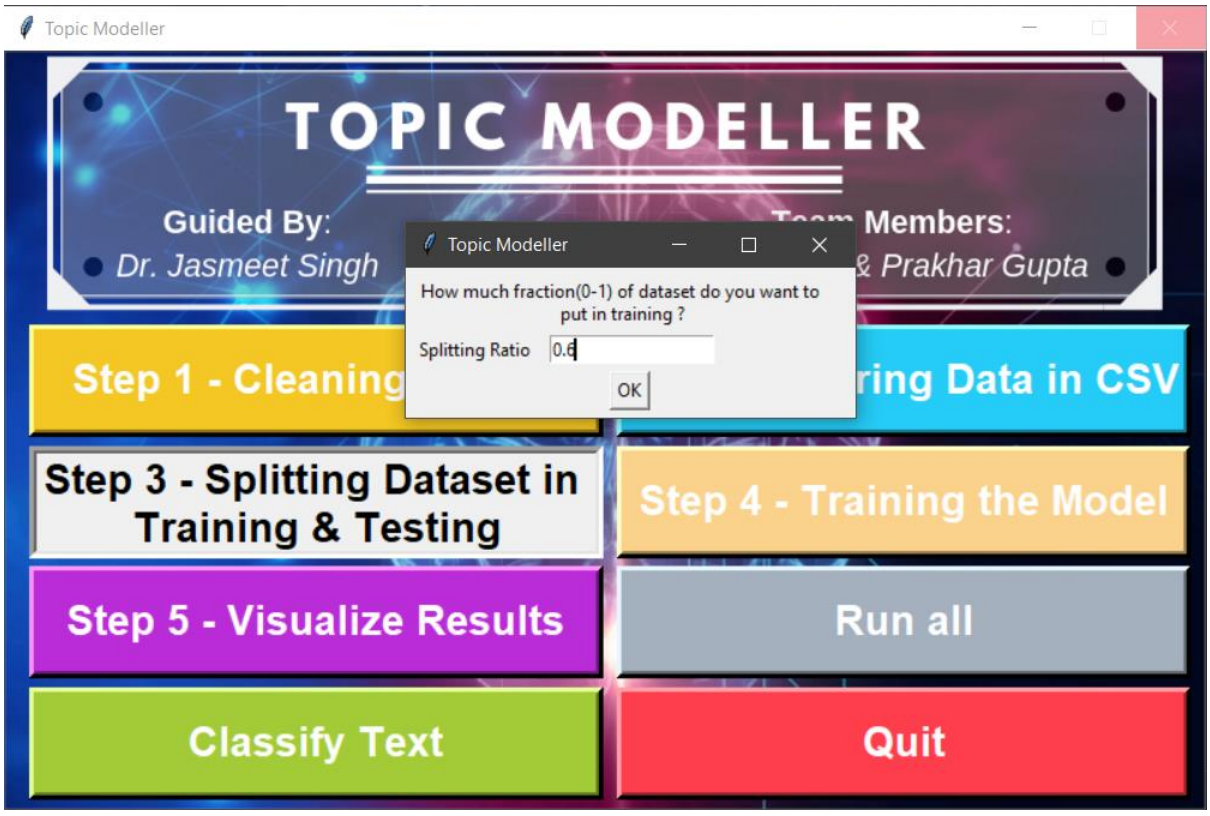

Fig3. Asking training size from user

4. Step 4 – Model Building

   a. Loading training set and test set.
   b. Find total words in entire dataset and find its frequency distribution.
   c. Convert each document as vector of words and make Term Document Matrix.
   d. Build the model using Naïve Bayes Classifier.

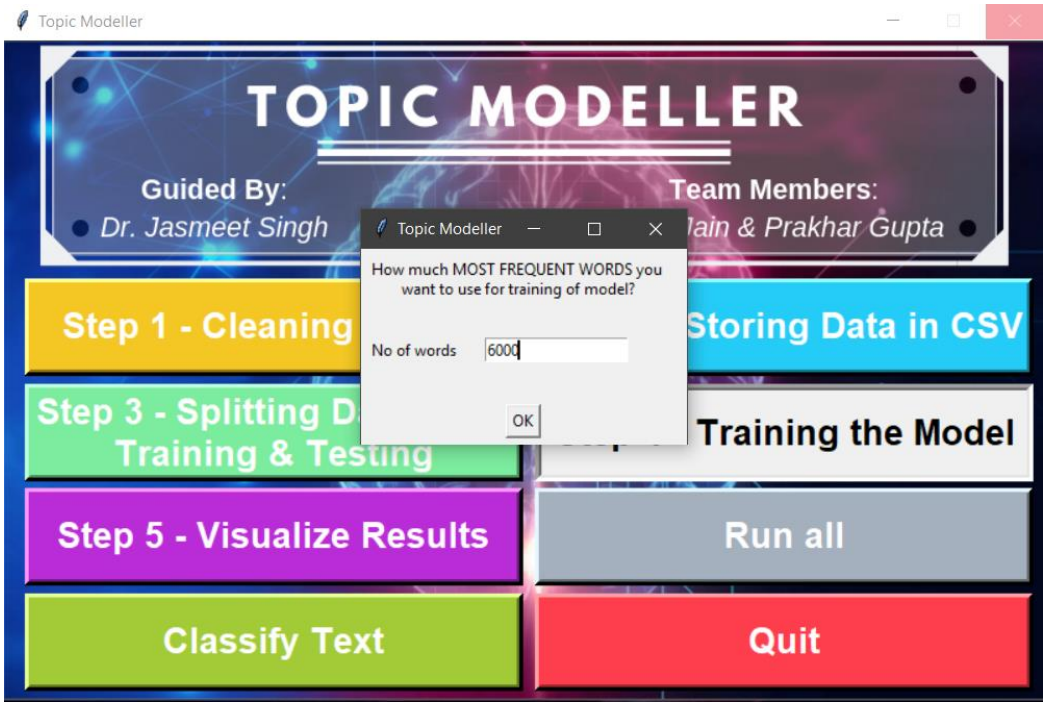

Fig4. Asking user on no. of words to train model.

5. Step 5 – Saving trained model.
   Model was saved using pickle library of python.

6. Step 6 – Calculating accuracy
   Using training set, accuracy of trained model was calculated.

7. Step 7 – Classify use's text
   The text was classified among 20 categories given in table 1

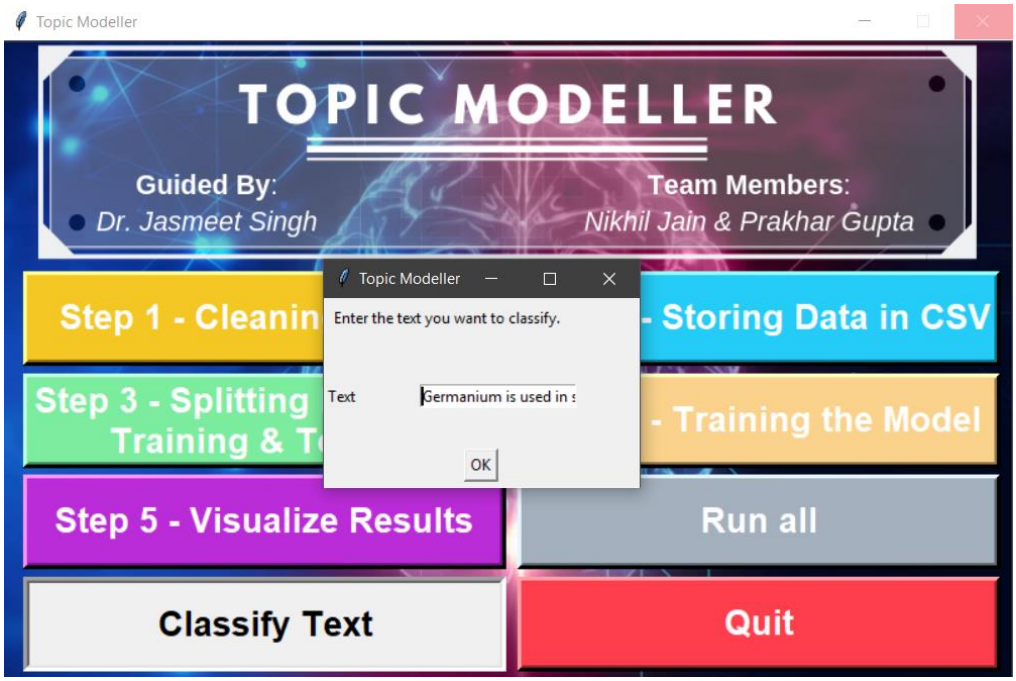

Fig5. User text input

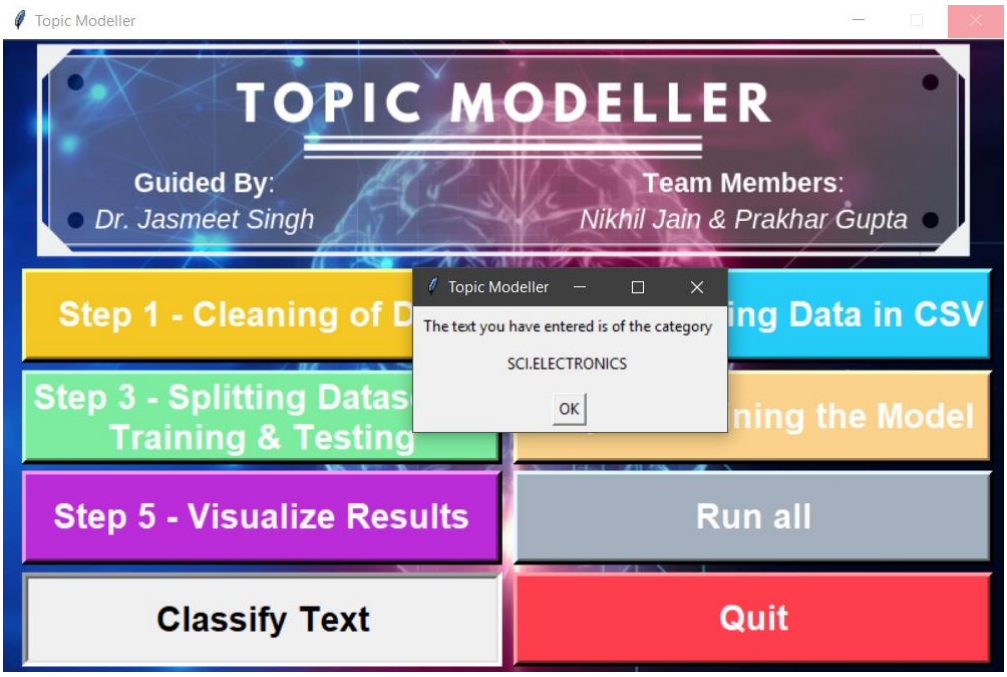

Fig6 . User's text classified

Results

We successfully trained model using 6000, 8000, 10000 most frequent words. Testing was done using test set. Results of accuracy, time taken to build the model, and time taken to extract the features are plotted in figure 7-9.

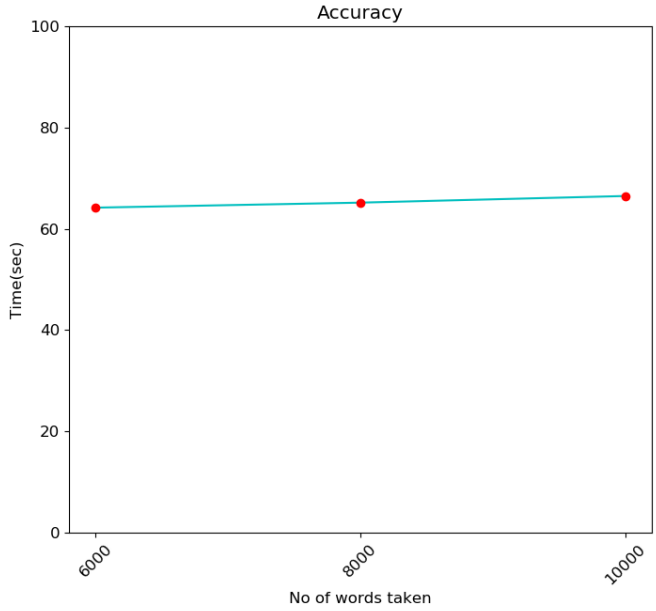

Fig7. Graph of accuracy

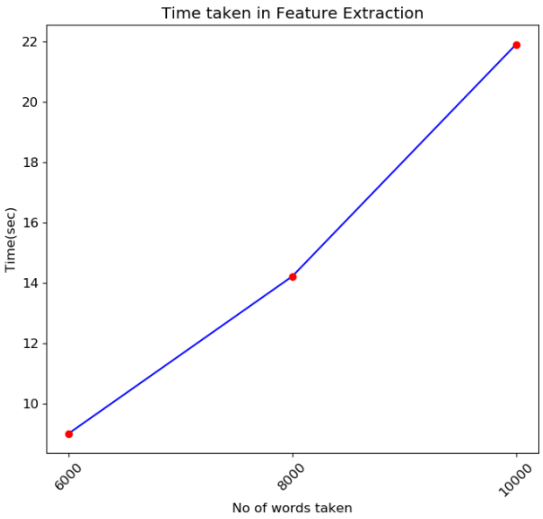

Fig8. Feature extraction

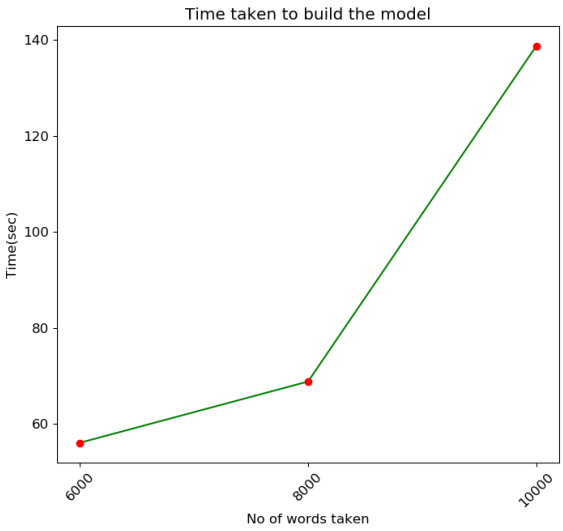

Fig9. Model building

Applications

1. Email classifier
2. Sentiment analyser
3. Question answering
4. Information retrieval system

Future Scope

We have build our model with an accuracy of 65%. We could improve accuracy by modifying the data preprocessing steps. If provided with powerful GPUs, the training time could be reduced by a significant amount. At present we have used only one classifier i.e Naive Bayes classifier. We can use more than one classifier and then make predictions from each of these models. We can then use machine learning approaches like ensembling to further improve the accuracy of our model. At present we used boolean term-document matrix to train our model. Other options available are tf-idf based term-document matrix.

References

1. Home Page for 20 Newsgroups Data Set [Online] Available: http://qwone.com/~jason/20Newsgroups/ [Accessed May 3, 2019].
2. Bakalov A, McCallum A, Wallach H, Mimno D (2012) Topic models for taxonomies. In: Proceedings of the 12th ACM/IEEE-CS joint conference on digital libraries, pp 237–240
3. Bell, M.: 1983, 'Materials available worldwide for teaching applications of mathematics at the school level', in Zweng, M. et al. (eds.), Proceedings of the Fourth International Congress on Mathematical Education, Birkhäuser, Boston, pp. 252–267
4. Berry, J. et al. (eds.): 1984, Teaching and Applying Mathematical Modelling, Ellis Horwood, Chichester.
5. Berry, J. et al. (eds.): 1986, Mathematical Modelling Methodology, Models and Micros, Ellis Horwood, Chichester.
6. Berry, J. et al. (eds.): 1987, Mathematical Modelling Courses, Ellis Horwood, Chichester.