Indian Institute of Information Technology, Allahabad

Predictive Analysis in E-Learning System

Submitted By:

Madhav Gupta (IIT2017031)

Sangeeta Meena (IIT2017032)

Ajay Raikar (IIT2017040)

Gurmeet Singh (IIT2017073)

Submitted To: **Signature:**

Dr. Ranjana Vyas **Date:20-11-2019**

ABSTRACT

Computer and Communication technologies results in a gradually scale up in many aspects of education, but making test papers manually seems to be a weak link. We introduce a paper where there will be automatically generated questions for practice of IIT-JEE which may occur in the future exams.

Every year more than 10 Lakh students appear for the exam and most of them prepare from the coaching institute given test series paper which have very low probability of appearing in the exam which results in sort of a poor performance of students.[2] Our data-driven framework for NLP (*Natural Language Processing)* leverages solution data from large number of IIT-JEE previous year question papers to predict whether the particular questions can appear in the coming exam with a higher probability. The success of NLP (Natural Language Processing) for text data is used in our model. First, we convert each questions to a cleaned text removing non meaningful words(stopwords, tokenization, stemming, lemmatization etc.) and eventually converting it to a series of numerical *features.* Second, we *classify* the features from questions to which topic they belong to by following the NCERT syllabus. We develop two different classification approaches, one that leverages generic classification algorithms *(K - Nearest Neighbors)* and one based on Random Forest*(Ensemble Learning using deep neural network and Adaboost techniques to boost the learning rate)* which includes all the different machine learning models like Logistic Regression, Naive Bayes Classifier, Support Vector Machines(SVM), Decision Trees etc.[15] Third we *automatically generate* potentially large number of tests paper based on their assigned topic and year of appearing. As a bonus, we can also do a *Time Series* analysis for how many questions can come for a particular chapter.For testing and validation of our project we use real world data so that we can demonstrate how it reduces the effort of human in education field.

Keywords

Automatic test paper generator, Machine Learning, Classification, Bayesian, IIT-JEE Mains, Natural Language Processing.

INTRODUCTION

Various platforms include massive open online courses (like nptel) , personalized learning systems and other tutoring systems.[1] While Computer and other technology provided an effective means to automate various processes and scale up the number of learners , still such platforms need people physically to do tasks like generation of question paper etc. This paper proposes various methods to automate the process of question paper generation which will be having a higher probability of coming in upcoming entrance exam using various machine learning techniques.

Methodology presented in this paper is a generic model which can be implemented in any academics branch, currently for demo we have implemented two subjects physics and chemistry inspired from Joint Entrance Exam(JEE). JEE Aspirants face many challenges during preparation for the exam.Many students opt for coaching from some private institutes which take a hefty amount of fees and still most of the students don't get a satisfactory output. During preparation, they have to study hard to cope up with the coaching schedule and complete the syllabus as a result they feel huge pressure while preparation.

Coaching Institutes for IIT JEE have become "Topper" Manufacturers with Top rank holders in the engineering entrance exam. Aspirants of JEE Main and Advanced exams prefer training from top coaching Institutes for IIT JEE in India to crack the engineering entrance exams as well as gain competitive edge. However, many aspirants who join the coaching institutes are not able to follow the competitive preparation strategies aimed to crack IIT JEE. Students are stressed because they are not able to cope with the coaching institute schedule; syllabus and they get anxious for cracking engineering entrance exam and most of them are financial weak so they can't afford to get into the

top coaching for IIT JEE and DPP's and books of the top institute are unaffordable for these people. This is the reason that mostly people who prepared from top institutes are selected or score good marks in IIT-JEE. This paper proposes a methodology to provide a platform for such financial weak people so they can prepare for IIT-JEE online. This preparation would help people to understand the topic clearly and solve various questions related to that topic. Also they will get their hands on questions of those topics which will have a higher probability of coming in the next upcoming exam.

There is a pressing need to find new ways and means to automate two critical tasks that are typically handled by the instructor or course assistants in a small as well as large scale course: (i) manually deciding what questions to give in a test (ii) manually selecting questions from book and different-different difficulty level questions from book.

In this paper, we study the problem of automatically generating test paper for IIT-JEE like [1]Entrance Exams that figure prominently in STEM(science, technology, engineering and mathematics) education. To the best of our knowledge, there exist no tools to automatically generate question papers for Entrance Exams. As a result, large-scale education platforms have resorted or totally depended upon Data Entry Analyst staff which institute particularly hires for the job of question paper setting. They also sometimes make error in question paper setting which results in giving marks to all of the candidate and which also have other consequences like re-exam.

Motivation

The primary motivation behind this project is to enable coaching institutes and jee aspirants to have an easy prediction of crucial questions for the coming examination so that the process of choosing questions manually can be automated.The manual selection of questions is a tedious task and it may result in some unnecessary selection of questions.Since Jee main is one of the sound examination in the country and many students opt for it that's why current project have been developed to help students in this competitive era.

Literature Review

Previously When machine learning algorithms where not known then it was difficult for teachers and instructors to give their students an idea about the probable questions for the exam.Analysing previous years questions rigorously and making predictions about the questions for the exam was lengthy and time killing task.But with the knowledge of Natural language processing ,text analysis and machine learning algorithms this become a do able task.so previously students used to struggle to have a planned routine for the preparation of the jee exam but now with the help of this E-learning model aspirants become able to do important set of problems ahead of the exam.Now the present model will help aspirants to have an insight about the important questions.In future ,our aim is to convert this model into a fully functional more interactive website,where students and instructors can communicate in a more vivid manner for jee exam.

Problem Definition

Developing an E-learning model to predict important set of questions for JEE aspirants so that manual selection of questions by instructors can be automated.This will enable students to practice crucial questions ahead of the exams to beat the competition.

Proposed Methodology

NLP FEATURE EXTRACTION

First step in our project to convert or transform the sets of questions dataset into a set of numerical features , after that these numerical features are used to cluster and classify the questions.

[1]Workhouse of NLP is the bag-of-words model - it has found a tremendous success in the text semantic analysis.In this model the text are seen as a collection of words and then the frequencies of these words are used as a numerical features to perform tasks like topic classification and document clustering in our project framework is to transform a collection of questions dataset into a set of numerical features. In later sections, we show

how the numerical features can be used to cluster and classify the questions. workhorse of NLP is the *bag-of-words* model; it has found tremendous success in text semantic analysis. This model treats a text document as a collection of words and uses the frequencies of the words as numerical features to perform tasks like topic classification and document clustering.

The pipeline of our project framework is composed of three main components:

.1 **Dataset Preparation :** We didn't find any dataset on the internet so we decided to make our own custom dataset from previous year *IIT-JEE* Mains paper we have considered last 15-20 years of papers in our dataset. Our dataset consists of questions, in which year they have appeared and corresponding topic to which it belongs i.e. *hydrocarbons*, *chemical_bonding* etc. heres a snapshot of our dataset.

	questions	year	topic
2	The ratio of mass percent of C and H of anorganic compound (CXHYOZ) is 6 : 1. Ifone molecul	2018	alkane
3	Which type of 'defect' has the presence ofcations in the interstitial sites ?	2018	solid state
4	According to molecular orbital theory,which of the following will not be a viable molecule ?	2018	chemical bonding
5	Which of the following lines correctly showthe temperature dependence ofequilibrium constant,	2018	chemical equilibrium
6	The combustion of benzene (l) gives CO2(g)and H2O(l). Given that heat of combustionof benze	2018	thermodynamics
7	For 1 molal aqueous solution of thefollowing compounds, which one willshow the highest freezi	2018	solutions
8	An aqueous solution contains 0.10 M H2Sand 0.20 M HCl. If the equilibriumconstants for the fo	2018	chemical equilibrium
9	An aqueous solution contains an unknownconcentration of Ba2+. When 50 mL of a1 M solutior	2018	chemical equilibrium
10	At 5188 C, the rate of decomposition of asample of gaseous acetaldehyde, initiallyat a pressure	2018	chemical kinetics
11	How long (approximate) should water beelectrolysed by passing through100 amperes current s	2018	electrolysis
12	The recommended concentration offluoride ion in drinking water is up to1 ppm as fluoride ion is	2018	cordination compounds
13	Which of the following compoundscontain(s) no covalent bond(s) ?KCl, PH3, O2, B2H6, H2SO4	2018	chemical bonding

Fig 1. chemistry dataset

2. **Feature Engineering :** Before building machine learning model for our project framework we need to first prepare our dataset in a way that our machine learning algorithm can make sense from it, in order to perform data preprocessing for feature engineering, we first removed all the stopwords (i.e. a, an, the, then, that, is, etc) after that we have performed word tokenization which separates the sentence into list of words. After that we have created the bag of words model using tf-idf model, here is a snapshot of the work showing feature engineering performed on dataset.

The methods which will be used during the implementation of the project are:

NLP:-The purpose of semantic analysis is to draw exact meaning, or you can say dictionary meaning from the text. [11]The work of semantic analyzer is to check the text for meaningfulness.The first part of semantic analysis, studying the meaning of individual words is called lexical semantics. It includes words, sub-words, affixes (sub-units), compound words and phrases also. All the words, sub-words, etc. are collectively called lexical items. In other words, we can say that lexical semantics is the relationship between lexical items, meaning of sentences and syntax of sentences. This will determine the category of questions by knowing their meaning.This is used to know the chemistry and physics semantics but for maths we have to use MLP.

Data Preprocessing Steps:-

Import Libraries:- We are using Pandas for creating DataFrame from which we can work efficiently and manipulate data easily**.**

Import the Dataset:- After importing the libraries we are reading our dataset using read_csv method of ***pandas*** library.

Vectorization:-To convert the written text data into vector we will perform certain operation:-

1. Text Preprocessing :-
    - Remove Stopwords :- In order to perform well on the classification and prediction our machine learning algorithms need the data in the numerical data format called vector. In this step of text processing first we remove the ***stopwords*** from sentence here a particular question. Stopwords are the words which appear in a normal sentence too many times like a, an, the, where, there, then etc. We are performing this step using the stopwords library inbuilt in NLTK (Natural Language Toolkit) which includes all those words and more.

Example.

Original Sentence - Which type of ‘defect’ has the presence of cations in the interstitial sites

After Removing Stopwords - defect presence cations interstitial sites

- Tokenization:- Tokenization is the process of extracting features out of a sentence after cleaning the Text data from step 1 (stopwords removal) means taking only those words into account which will make sense in categorization of test data.Tokenization in python can be done by Python's NLTK library’s word_tokenize() function.

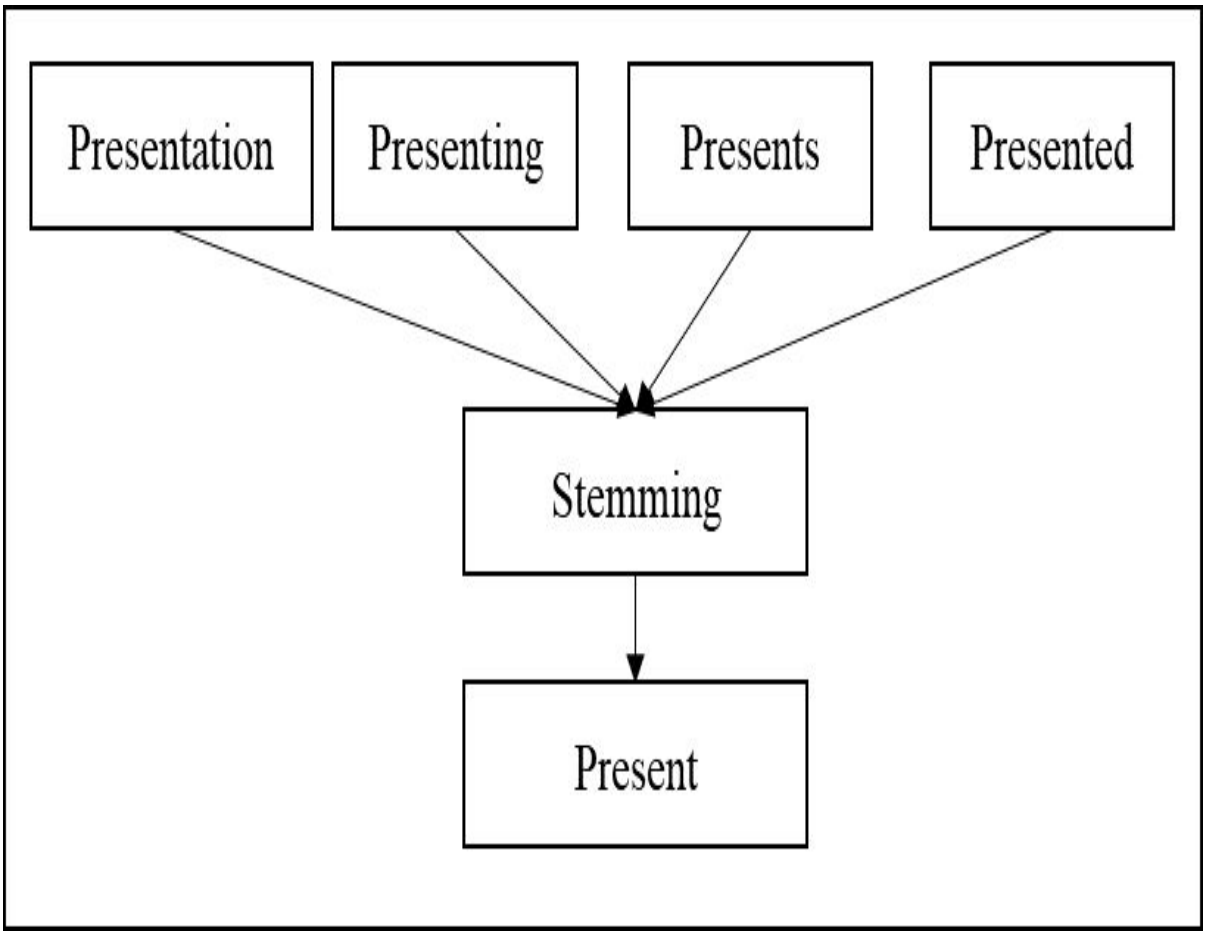
  After Tokenization the sentence would be.

  sentence = [“defect”, “presence”, “cations”, “interstitial”, “sites”]

- Lemmatization :- *Lemmatization* means replacing the words having similar meanings into one common words used for those words. Cause every single word contribute to a considerable increase in features (word columns) which can slow down our *ML Algorithms* and even decrease precision, So what we do is we make sure that all those similar words are getting replaced by one common word like good, better, best all these three words are somehow saying the words so we are going to replace it with good. This is called *lemmatization*. *NLTK* library have built in class for performing this lemmatizing task.

- Stemming :- *Stemming* is the process of modifying the words which are derivative of some other words and which may represent the same meaning as the previous one, so why not modify it to the previous word itself that what *Stemming* do. For example *force*, *forces*, *forced* all are indicating the same word *force* so we modify it to the appropriate word *force*. There are two

stemmer in NLTK library one is PorterStemmer and the second one is Snowball Stemmer.



2. Bag of words:- Basically in this we contain all the tokens and their frequency.So using a word count we convert word or sentence to the vector , for example where we count occurrence of word in a document w.r.t list.[4] For example- vector conversion of sentence “There used to be iron Age” can be represented as :

   Example:-

   “There” = 1 , ”was”= 0 , ”to”= 1 ,”be” =1 , ”used” = 1 , ”Stone”= 0 , ”Bronze” =0 , “Iron” =1 ”Revolution”= 0 , ”Digital”= 0 ,”Age”=1 , ”of”=0 ,”Now”=0 , ”it”=0 , ”is”=0.

   “There used to be iron age” = [1,0,1,1,1,0,0,1,0,0,1,0,0,0,0].

token	porter_stemmed	snowball_stemmed
very	veri	veri
orderly	orderli	order
methodical	method	method
looked	look	look
ticking	tick	tick
sonorous	sonor	sonor
his	hi	his
flapped	flap	flap
newly	newli	newli
pitted	pit	pit

Fig 3. PorterStemmer and SnowballStemmer

token	stemmed	lemmatized
Very	veri	Very
orderly	orderli	orderly
methodical	method	methodical
looked	look	looked
ticking	tick	ticking
sonorous	sonor	sonorous
his	hi	his
flapped	flap	flapped
newly	newli	newly

Fig 4. Stemming and Lemmatization

Comparative study of different-different machine learning algorithms

1.Logistic regression :- Logistic regression is the right algorithm to start with classification algorithms.The output of the logistic regression will be a probability ($0 \leq x \leq 1$), and can be used to predict the binary 0 or 1 as the output (if $x<0.5$, output= 0, else output=1). Basically it is used for prediction or knowing the probability.We can use this technique in various fields to predict future circumstances.The logistic model is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick.

2.SVM (Support vector machine) :- In this if we are given a labelled training dataset of supervised learning , the algorithm will create an optimal hyperplane which categorizes new example. Basically the SVM helps in the separation of classes. Through SVM we can distinguish between the classes and can separate it with the help of the hyperplane.If the separation of two classes is difficult we may use the transformation can assume one more axis for simplification and separate the classes.This transformation is known as kernels.

There are some terms that need to be known while using SVM like Regularization , Gamma. Regularization tells the SVM optimization how much you want to avoid misclassifying each training example . [7]The gamma parameter defines how far the influence of a single training example reaches, with low values meaning ‘far’ and high values meaning ‘close’. In other words, with low gamma, points far away from plausible separation line are considered in the calculation for the separation line.

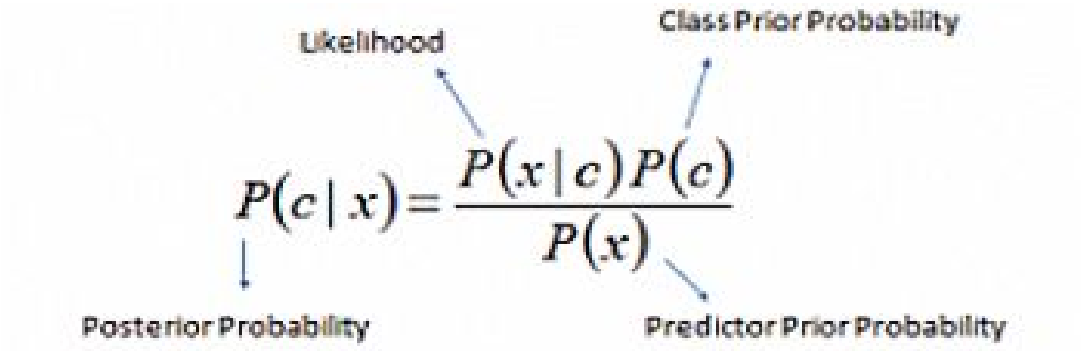
3.Naive Bayes :- Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature means no two features will be dependent on each other . In this first we convert training dataset into frequency table and then we create probability table (probability of outcome), by using Naive Bayesian

equation to calculate the posterior probability for each class , The class with the highest posterior probability is the outcome of prediction.

Naive Bayes uses a similar method to predict the probability of different class based on various attributes. This algorithm is mostly used in text classification and with problems having multiple classes. [6] Naive Bayes classifiers mostly used in text classification (due to better result in multi class problems and independence rule) have a higher success rate as compared to other algorithms. As a result, it is widely used in Spam filtering (identify spam email) and Sentiment Analysis (in social media analysis, to identify positive and negative customer sentiments)

Mathematics Behind Naive Bayes

Bayes Theorem provides a way of calculating posterior probability p(c|x) form p(c) , p(x) and p(x|c)



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

IDF Inverse Document Frequency

$$idf(t, D) = log \frac{|D|}{1 + |\{d \in D : t \in d\}|}$$

Challenges faced during project :-

1.The collection of dataset of past years from institute was a major problem faced by us.

2. After that we have to manually create the datasets for physics and chemistry from last year JEE MAINS papers, which took a lot of effort and then we processed the same in our source code accordingly.

3.Till Now we haven't been introduced the course for Machine learning and basic Natural Language Processing so first we have to go through the basics of Machine Learning and Natural Language Processing.

4. There haven't been any Word2Vec for Institutional (e.g STEM) like texts there are pre trained Word2Vec for sentiment analysis, review classification, news classification etc. That's why we have to make our own dataset from scratch and make a Tf-IDF vectorizer to make predictions. And we can include this into our future scope to train an Word2Vec from scratch.

Implementation Plan

1. **Week-1 :** Requirement Analysis
2. **Week-2 :** Research
3. **Week-3** : Modeling designing and pseudocode implementation
4. **Week-4** : Review
5. **Week-5** : Prototype
6. **Week-6** : User testing
7. **Week-7** : Project Deployment
8. **Week-8 :** Final Presentation

Conclusion

The current model helps instructors to generate questions that are most probable to come in the coming exams.Questions categorized by the model are quite accurate and crucial for the examination point of view . Students will also be provided with sample questions. It will help them to practice questions without joining any institutes ,following schedule of any famous Coaching, they don't have to be worried about the heavy fee of the classes.The Model is quite interactive that can be easily used by the students as well as instructors.

References

[1]. Lan, Andrew S., et al. "Mathematical language processing: Automatic grading and feedback for open response mathematical questions." *Proceedings of the Second (2015) ACM Conference on Learning@ Scale*. ACM, 2015.

[2]. web.ece.rice.edu

[3]. towardsdatascience.com

[4]. medium.com

[5]. www.analyticsvidhya.com

[6]. experfy.com

[7]. m.jagranjosh.com

[8]. doublesharpevideo.com

[9]. pythonprogramminglanguage.com

[10]. Lan, Shiting, et al. "Mathematical language processing: automatic grading and feedback for open response mathematical questions." U.S. Patent No. 10,373,512. 6 Aug. 2019.

[11]. Manning, Christopher, et al. "The Stanford CoreNLP natural language processing toolkit." *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*. 2014.

[12]. Bird, Steven, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[13]. Chowdhury, Gobinda G. "Natural language processing." *Annual review of information science and technology* 37.1 (2003): 51-89.

[14]. Sriram, Bharath, et al. "Short text classification in twitter to improve information filtering." *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2010.

[15]. McCallum, Andrew, and Kamal Nigam. "A comparison of event models for naive bayes text classification." *AAAI-98 workshop on learning for text categorization*. Vol. 752. No. 1. 1998.