Automated Essay Grading System

By : Team Vectors
Team members:Mohit Sharma
Vivek Singh

1. Abstract :

Essays are paramount for assessing academic excellence along with linking the different ideas with the ability to recall but are notably time consuming when they are assessed manually. Manual grading takes a significant amount of evaluator's time and hence it is an expensive process. Automated grading if proven effective will not only reduce the time for assessment but comparing it with human scores will also make the score realistic. The purpose of this project is to implement and train machine learning models to automatically assess and grade essay responses. The dataset we used is made available on the internet by Hewlett Foundation for one competition on Kaggle.com. We used a linear regression model for training our model. The features extracted from essays include number of sentences, words, comma, exclamation mark, brackets, nouns, verbs, adjectives and misspelled words etc. We tested our model on a testing dataset and calculated the RMSE(Root Mean Square Error) for each essay set. We have found that RMSE values for set 1-6 are lower than set 7 and 8, this is because essays in these sets are much more content specific. This is our future plan to employ more advanced NLP features so that our model could grade the content specific essays also efficiently.

2. Introduction :

Essays are crucial testing tools for assessing academic achievement, integration of ideas and ability to recall, but are expensive and time consuming to grade manually. Manual grading of essays takes up a significant amount of instructors' valuable time, and hence is an expensive process.

In this project we tried to make a system which automatically grades the essay, by considering various features, thus reducing the assessment time as well makes the evaluation process much more efficient and transparent. Thus if implemented correctly this project proved to be very much beneficial for both students and professors.

The project aims to develop an automated essay assessment system by use of machine learning techniques by classifying a corpus of textual entities into a small number of discrete categories, corresponding to possible grades.
The dataset available to us consists of scores given by 2 raters. On analysis of the dataset, we conclude that each rater had given the marks from 0 to 30(both inclusive) to each essay, which again bifurcates the essay into 8 sets depending upon the scale of marking.

3. Objective

In the project we first clean the dataset, particularly by removing those columns which contain almost 70-80% of NULL values. We then extracted useful features from the dataset. Total 15 features were extracted which includes number of sentences, words, comma, exclamation mark, brackets, nouns, verbs, adjectives, foreign words and misspelled words etc.
We have used the linear regression technique to train our model. We have used polynomial regression (degree =2,3 and 4) also, but this leads to overfitting of data, thus the output we get with degree 2,3 and 4 is not much accurate.

4. Technology Used:

We used Linear Regression as our learning model to learn parameters based on the features. Scores were predicted for a distinct set of test essays.

We have also tested polynomial regression of degree 2, 3, 4, 5 but due to increased number of features in higher degree regression model overfitting(model was performing very good on training data but on test set performance was very poor) occurred.

So currently we are using linear regression as our learning model.

5. Problems Faced :

Data collection was not a problem. A minor problem occurred during cleaning the data like for our model two steps cleaning is required one for converting essay into list of words and second for converting essay into list of sentences. In first case punctuations(like ., ?, !) were not needed in second case punctuations(like ., ?, !) were required to break into sentences.

Feature extraction is complicated and lengthy but no problem occurred. Total 15 features were extracted from each essay in the training data set.

Model prediction has not caused any big problem, only a lot of time was consumed when we tested a polynomial regression model of degree=4, 5. It took almost 20 minutes for degree=5 to calculate model parameters based on the training data set.

6. Datasets Used:

We used a data set from a competition on kaggle.com, by the William and Flora Hewlett Foundation. The data consists of 8 sets of essays in ASCII text, written by students from Grade 7 to Grade 10. Essays in each of the 8 sets have unique characteristics that are used for grading. This ensures that the automated grader is trained effectively across different types of essays. Each essay has one or more human scores and a final resolved score. Each essay is approximately 150 to 550 words in length. Some essays are more dependent upon source materials than others.

Essay set	Total number of Essays
1	1783
2	1800
3	1726
4	1772
5	1805
6	1800
7	1569
8	723
TOTAL	**12978**

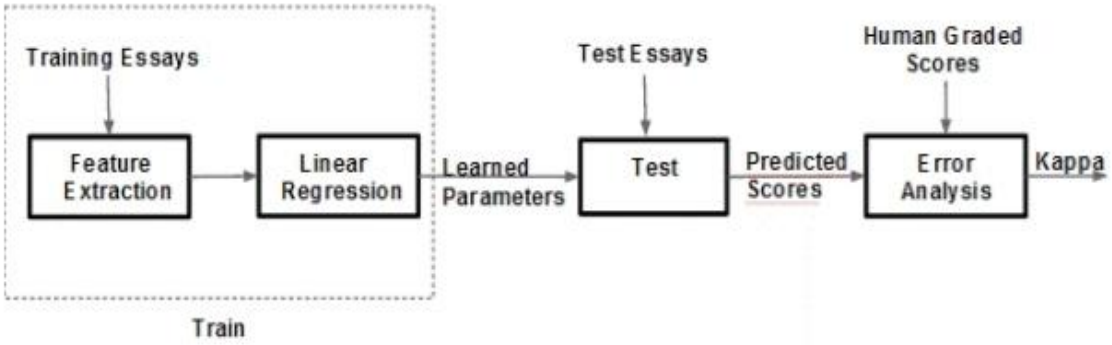
Each essay set is very different in terms of marking scale used and essay content and essay topic is also different for different sets.

Essay Set	Scale of Marking
1	2-12
2	2-10
3	0-3
4	0-3
5	0-4
6	0-4
7	0-30
8	0-60

7. Models Used and Implementation:

An overview of related prior work (see references) indicates that linear regression works well for essay grading applications, We chose linear regression as our learning model. We have to split the training data set into a 9:1 ratio. 90% is used to train our model and 10% is used for testing purposes. This was done to guard against overtting.

Splitting is done for each essay set(1-8) separately. We have used domain wise linear model(model parameters are calculated for each essay set separately).



First step is to clean the training data set. In cleaning process all alphanumeric text, square brackets ([]), special characters (@,#) were removed. After that first some features like comma, exclamation, quotation mark, punctuations were counted.

In second step, list of sentences is created for each essay by removing all punctuations except (! , ?, .). Then after sentences list has been created all punctuations are finally removed and then all stopping words like is, are, the, a are removed and a fresh list of words is created for each essay.

Finally remaining features like verb, adverb, adjective, noun, pronoun, word count, sentence count is extracted from word list and sentence list.

After computing all 15 features of each essay in training data these features along with the resolved score of each essay(90%) is provided to the linear model to compute model parameters for each essay set separately.
So overall we have an 8x15 matrix of model parameters. Here each row represents each essay set and columns represent learned model parameters.

The score of each test essay set is calculated using learned parameters.
Predicted scores are then used to compute error against resolved scores.

Why we used linear model?

Numerical features : Features such as total word count per essay, average word length per essay, sentence count indicate language fluency and dexterity.

Parts of speech count: Various parts-of-speech such as nouns, adjectives, adverbs and verbs are good proxies to test vocabulary.

Correct word spelling indicates command over language and facility of use.

Structure and Organization : Punctuation is a good indicator of a well structured and organized essay.

All these features are related to the score of the essay. Many previous works have shown that a linear model is a very good score predictor when these features are used to compute the score of an essay.

9. Result And Performance:

1.

	Essay_set	Essay_number	Predicted	Actual
0	7	8	15.783	17
1	7	12	19.4422	17
2	7	23	15.0723	17
3	7	33	17.5959	17
4	7	45	16.0556	17
5	7	54	17.9126	17
6	7	65	16.982	17
7	7	69	18.8437	17
8	7	85	16.489	17
9	7	104	18.0507	17
10	7	109	16.087	17
11	7	121	14.0104	17
12	7	131	17.1512	17
13	7	138	14.6214	17
14	7	149	17.1654	17

Result of our linear regression model

Above picture shows the predicted score by our Linear Regression model for essay set 7 and actual score 17 from the test set. About 12 essays have their predicted score in between 15 to 19 which is a good prediction score and 3 values are away from actual score 17.

2.

	Essay_set	RMSE
0	1	0.894278
1	2	0.49595
2	3	0.568122
3	4	0.606688
4	5	0.551004
5	6	1.24174
6	7	3.27394
7	8	4.41838

RMSE of our Model(Set-wise)

Above picture shows root mean square error values of the test set for each essay set. Set 1-6 have a small value of RMSE. Set 7-8 are having a considerable amount of error in predicted values and this is because essays in set 7, 8 are content specific and richer essays and have complex sentence structure.

10. Conclusion And Further Work:

Our model works relatively better on non-context specific essays. Performance on content specific and richer essays can be improved by incorporating content and advanced NLP features like Bag of Words(BoW), bi-gram, tri-gram etc. Features that are grammar and usage specific can further enhance the prediction model.

Our future work is on incorporating advanced NLP features in our model to improve prediction score mainly in set 7 and 8 which are content specific and richer essays and have complex sentence structure.

References :

1. https://www.kaggle.com/sakshisaku3000/automated-essay-grading-using-nlp-part1
2. http://cs229.stanford.edu/proj2012/MahanaJohnsApte-AutomatedEssayGradingUsingMachineLearning.pdf
3. https://www.kaggle.com/irfanmansuri/the-havelett-foundation-automated-scoring
4. https://iopscience.iop.org/article/10.1088/1742-6596/1000/1/012030/pdf