Department of Computer Science and Engineering

Major Project Report

on

Essay Grader and Recommendation using Machine Learning Techniques

Submitted

By

Praglva Jung Budhathoki (1SI17CS075)

Rabindra Basnet (1SI17CS080)

Rishav Karanjit (1SI17CS089)

Ritesh Prasad Yadav (1SI17CS090)

Batch ID: B26

Siddaganga Institute of Technology, Tumakuru – 572103

(An Autonomous Institution, affiliated to Visvesvaraya Technological University, Belagavi & Approved by AICTE, New Delhi, Program accredited by NBA, Accredited by NAAC & ISO 9001:2015 certified)

2020-2021

Table Of Content

<u>INTRODUCTION</u>

      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. Automated grading, if proven to match or exceed the reliability of human graders, will significantly reduce costs. The purpose of this project is to implement and train machine learning algorithms to automatically assess and grade essay responses. These grades from the automatic grading system should match the human grades consistently. Currently, automated grading is used instead of second graders in some high-stakes applications, and as the only grading scheme in low stakes evaluation.

OBJECTIVES

1. To score the essay based on:
   - Coherence,
   - Clarity,
   - Vocabulary,
   - Style,
   - Organization
2. To test for plagiarism in the essay.
3. To give following details of essay:
   - number of words,
   - unique words,
   - average word length per sentence,
   - number of grammatical mistake and number of spelling mistakes.
4. To recommend how an essay could be better.

LITERATURE SURVEY

Title of the work: Automatic Essay Scoring

Authors: Murray, and Orii

Publication: http://www.cs.cmu.edu/~norii/pub/aes.pdf

Description: In this paper, it is shown how linear regression can be used to automatically grade essays on standardized tests. We combine simple, shallow features of the essays, such as character length and word length, with part-of-speech patterns. Our combined model gives significant reduction in prediction error. We discuss which features were effective in predicting scores.

This paper explores the use of linear regression from text features to directly predict the score of given essays. Using $l1$ regularization, it takes a large feature space consisting of a variety of linguistic features and determine the most predictive ones. It is able to significantly reduce prediction error and obtain state-of-the-art results, comparable to human annotators.

Title of the work: Automatic Essay Grading with Machine Learning

Authors: Mahana, Johns and Apte

Publication: http://cs229.stanford.edu/proj2012/MahanaJohnsApte-AutomatedEssayGradingUsingMachineLearning.pdf

Description: The project aims to build an automated essay scoring system using a data set of ≈13000 essays from kaggle.com. These essays were divided into 8 different sets based on context. We extracted features such as total word count per essay, sentence count, number of long words, part of speech counts etc. from the training set essays. We used a linear regression model to learn from these features and generate parameters for testing and validation. We used 5-fold cross validation to train and test our model rigorously. Further, we used a forward feature selection algorithm to arrive at a combination of features that gives the best score prediction. Quadratic Weighted Kappa, which measures agreement between predicted scores and human scores, was used as an error metric. Our final model was able to achieve a kappa score of 0.73 across all 8 essay sets. We also got a good insight into what kind of features could improve our model, for example N-Grams and content testing features.

Title of the work: Develop an automated scoring algorithm for student-written essays
Authors: *Shihui Song, Jason Zhao*
Publication: https://nlp.stanford.edu/courses/cs224n/2013/reports/song.pdf
Description: In this project authors built an automated essay scoring system to score approximately 13,000 essays from an online Machine Learning competition Kaggle.com. There are 8 different essay topics and as such, the essays were divided into 8 sets which differed significantly in their responses to the features and evaluation. Our focus for this essay grading was the style of the essay, which is an extension on the studies conducted determining the quality of scientific articles by adding maturity A aspect of this project was to recognize the difference between the advanced nature of scientific articles to the coherency of middle to high school test essays. We evaluated Linear Regression, Regression Tree, Linear Discriminant Analysis, and Support Vector Machines on our features and discovered that Regression Trees achieved the best results with $\kappa = 0{:}52$

Title of the work: *Develop an automated scoring algorithm for student-written essays*
Authors: *Shihui Song, Jason Zhao*
Publication: https://nlp.stanford.edu/courses/cs224n/2013/reports/song.pdf
Description: Despite being investigated for over 50 years; the task of automated essay scoring continues to draw a lot of attention in the natural language processing community in part because of its commercial and educational values as well as the associated research challenges. This paper presents an overview of the major milestones made in automated essay scoring research since its inception.

Progress in dimension-specific scoring research is hindered in part by the scarcity of annotated corpora needed for model training. An issue that is often under-emphasized is which corpora one should choose for data annotation. We envision that in the long run, substantial progress in AES research can only be made if different researchers create their annotations on the same corpus. For instance, having a corpus of essays that are scored along multiple dimensions of quality can facilitate the study of how these dimensions interact with each other to produce a holistic score, allowing us to train joint models that enable these challenging dimension-specific scoring tasks to help each other via multi-task learning.

Title of the work: Automated Essay Scoring based on Two-Stage Learning

Authors: Jiawei Liu, Yang Xu, and Yaguang Zhu

Publication: https://arxiv.org/pdf/1901.07744

Description: Current state-of-the-art feature-engineered and end-to-end Automated Essay Score (AES) methods are proven to be unable to detect adversarial samples, e.g. the essays composed of permuted sentences and the prompt-irrelevant essays. Focusing on the problem, we develop a Two-Stage Learning Framework (TSLF) which integrates the advantages of both feature-engineered and end-to-end AES methods. In experiments, we compare TSLF against a number of strong baselines, and the results demonstrate the effectiveness and robustness of our models. TSLF surpasses all the baselines on five-eighths of prompts and achieves new state-of-the-art average performance when without negative samples. After adding some adversarial essays to the original datasets, TSLF outperforms the features engineered and end-to-end baselines to a great extent, and shows great robustness.

Title of the work: Automated Essay Scoring

Authors: Semire DIKLI

Publication:

https://www.researchgate.net/publication/26415982_Automated_Essay_Scoring/fulltext/0e6053e6f0c46d4f0ab03488/Automated-Essay-Scoring.pdf

Description: The impacts of computers on writing have been widely studied for three decades. Even basic computers functions, i.e. word processing, have been of great assistance to writers in modifying their essays. The research on Automated Essay Scoring (AES) has revealed that computers have the capacity to function as a more effective cognitive tool. AES is defined as the computer technology that evaluates and scores the written prose.

The AES systems described in this article employ various techniques to provide immediate feedback and scoring. While E-rater and IntelliMetric use NLP techniques, IEA is based on LSA. Moreover, PEG utilizes proxy measures to assess the quality of essays. Unlike PEG or IEA, e-rater and IntelliMetric systems have instructional applications (Criterion and My Access) as well. AES is a developing technology. Many AES systems are used to overcome time, cost, and generalizability issues in writing assessment. The accuracy and reliability of these systems have been proven to be high. The search for excellence in machine scoring of essays is continuing and numerous studies are being conducted to improve the effectiveness of the AES systems.

Title of the work: Human versus Automated Essay Scoring: A Critical Review

Authors: Beata Lewis Sevcikova

Publication: https://osf.io/preprints/socarxiv/6h4fd/download

Description: The paper reflects these developments as it charts the paths recently taken in the field, evaluates automated and human essay scoring systems in academic environments and analyzes the implications that both systems offer. In recent years, ways and opportunities for giving feedback have changed as computer programs have been more widely used in assessing students writing. Numerous researchers have studied computerized feedback and its potential. Different problems, such as quality of this type of feedback, validity, and reliability have been analyzed. This critical review examines two major types of academic writing support. The objective of the study based on the literature review is to examine the potential support of human and automated proofreaders for teaching and learning purposes.

Title of the work: Automated Essay Scoring: Applications to Educational Technology

Authors: Darrell Laham, Peter Foltz

Publication:

https://www.researchgate.net/profile/Darrell_Laham/publication/239061100_Automated_Essay_Scoring_Applications_to_Educational_Technology/links/5c212399458515a4c7f6edd3/Automated-Essay-Scoring-Applications-to-Educational-Technology.pdf

Description: The Intelligent Essay Assessor (IEA) is a set of software tools for scoring the quality of essay content. The IEA uses Latent Semantic Analysis (LSA), which is both a computational model of human knowledge representation and a method for extracting semantic similarity of words and passages from text. Simulations of psycholinguistic phenomena show that LSA reflects similarities of human meaning effectively. To assess essay quality, LSA is first trained on domain-representative text. Then student essays are characterized by LSA representations of the meaning of their contained words and compared with essays of known quality on degree of conceptual relevance and amount of relevant content. Over many diverse topics, the IEA scores agreed with human experts as accurately as expert scores agreed with each other. Implications are discussed for incorporating automatic essay scoring in more general forms of educational technology.

MOTIVATION

Essay writing is the requirement for many placement exams, aptitude tests like SAT, GRE and GMAT and language test like IELTS, TOEFL and PTE. Although many of the tests requires an essay writing, no reliable and free essay grader is available.

Some of the essay grader available are eRater – by educational testing service(ETS), Intellimetric – by Vantage Learning and Project Essay Grade by Measurement, Inc. According to our study, these essay rater are very reliable but the problem with this essay rater is they are not free and not only this they are not even cheap because these essay rater takes atleast $20 per essay which is very high because students will be practicing tons of essay.

Some online essay grading service like Kaplan inc, Princetonreview and some other websites provide free essay grading. However, there free service have very poor rating citing this essay service to be not viable at all. Moreover, this essay grading service may check spelling and number of worlds correctly and grade the essay but some factors like coherence, writing style, vocabulary and organization are not tested although they are very important for all essay writings.

Finally, no essay grading service gives recommendation of how to improve the quality of the essay – not even the paid ones give. We felt that student should be told what and where to improve. Only after giving recommendation student will realise there mistake and correct it next time.

Hence, due to aforementioned reasons we, as student, felt great need of essay grader and recommendation. This will help us to improve our performance in placement exams, aptitude exams and language test exams.

METHODOLOGY

FLOWCHART:

```
                              ┌──────────────┐
                              │    Start     │
                              └──────┬───────┘
                                     │
                                     ▼
                              ╱ Input essay ╲
          ┌──────────────────▶  document    │
          │                   ╲             ╱
          │                        │
          │                        ▼
          │                   ╱ Plagiarism Check ╲       ┌────────────────┐
          │                  ◁   needed?        ▷──Yes──▶│ Output Essay   │
          │                   ╲                ╱          │   Details      │
          │                        │                     └────────┬───────┘
          │                        No                             │
          │                        ▼                              │
          │                 ┌────────────────┐◀───────────────────┘
          │                 │ Extract Essay  │
          │                 │    details     │
          │                 └────────┬───────┘
          │                          │
          │                          ▼
          │                 ┌────────────────┐
        Yes                 │ Output Essay   │
          │                 │    Score       │
          │                 └────────┬───────┘
          │                          │
          │                          ▼
          │                  ╱ Need Essay ╲             ╱ Output       ╲
          │                 ◁ Recommendation ▷──Yes──▶  │ improvement  │
          │                  ╲            ╱              ╲  details     ╱
          │                        │                          │
          │                        No                         │
          │                        ▼                          │
          └─────────────────▶╱ More essay? ╲◀────────────────┘
                             ╲             ╱
                                   │
                                   No
                                   ▼
                              ┌──────────────┐
                              │     End      │
                              └──────────────┘
```
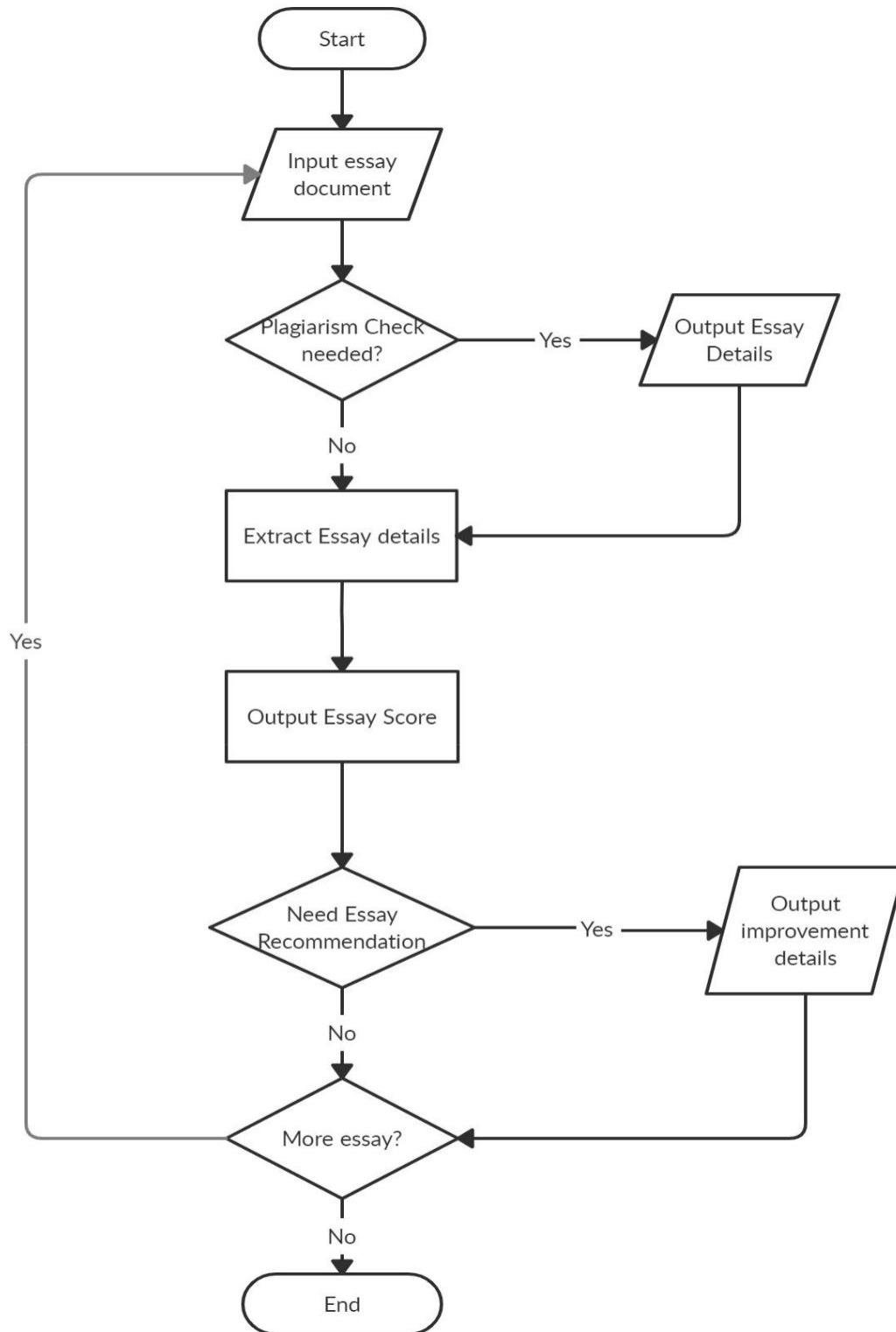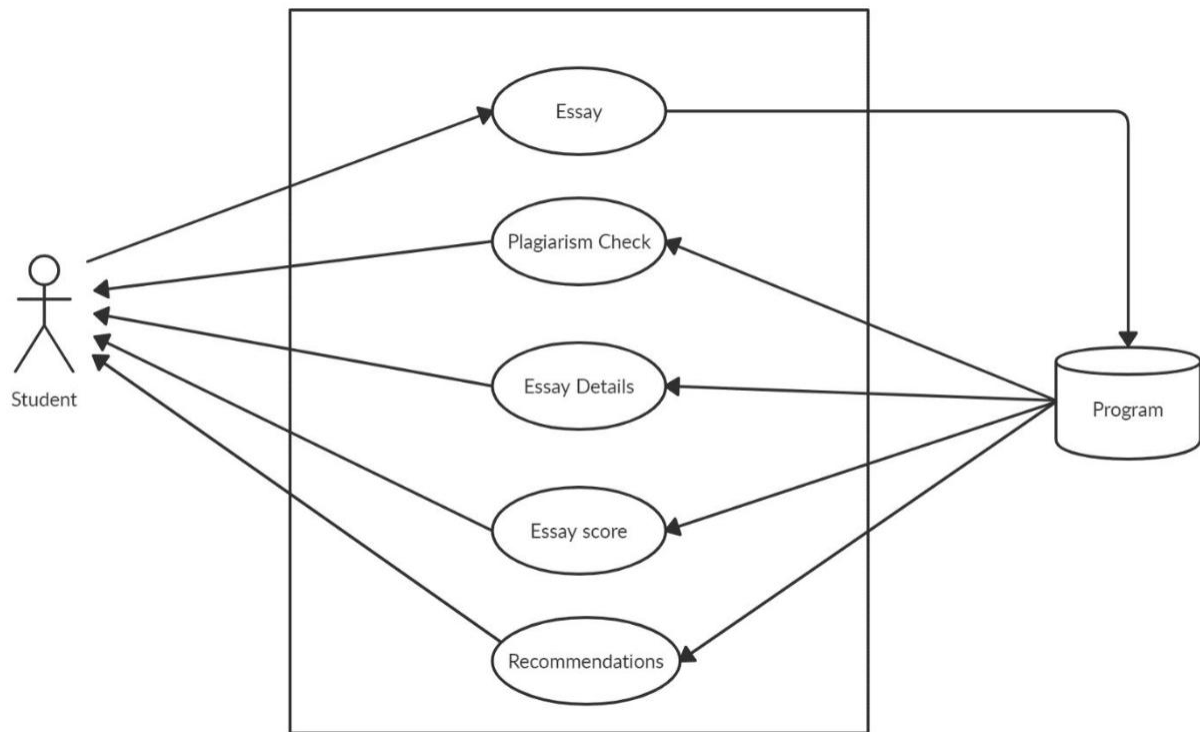
USE CASE DIAGRAM:

TOOLS AND TECHNOLOGIES

Languages and Frameworks used:

Frontend:
- ➢ pyqt5(python framework),css

Backend:
- ➢ Python

Database:
- ➢ Mysql

◆ We chose to implement our model in Python 3, since there exist a diverse set of libraries for working with natural language processing.
◆ We will use the Natural Language Toolkit (NLTK) and textmining for most NLP tasks.
◆ We also used scikit-learn to implement regularized linear regression (in addition to our own implementation). Other libraries (numpy, scipy, xlrd, xlwt, re) have been used for various tasks.

Learning Model:

➢ Linear regression works well for essay grading applications, We chose linear regression as our learning model. We did not have a separate validation test essay set, so we used 5-fold cross validation to train and test our learning model across the whole range of essays at our disposal. This was done to guard against overfitting.

EXPECTED OUTCOMES

➢ Our project is expected to score essay with very high accuracy.
➢ Our plagiarism test will be very helpful if someone has copied essay from internet.
➢ The word detail, unique word detail and other other details will help student in determining which part they should improve.
➢ Our project will be capable to tell user what to change and where to change in the essay so that students can improve future essay writing

Phase 2 Outcomes

Model of the prototype

We have used iterative model in our project. In the Iterative model, iterative process starts with a simple implementation of a small set of the software requirements and iteratively enhances the evolving versions until the complete system is implemented and ready to be deployed.

Iterative process starts with a simple implementation of a subset of the software requirements and iteratively enhances the evolving versions until the full system is implemented. At each iteration, design modifications are made and new functional capabilities are added. The basic idea behind this method is to develop a system through repeated cycles (iterative) and in smaller portions at a time (incremental).

The following illustration is a representation of the Iterative and Incremental model
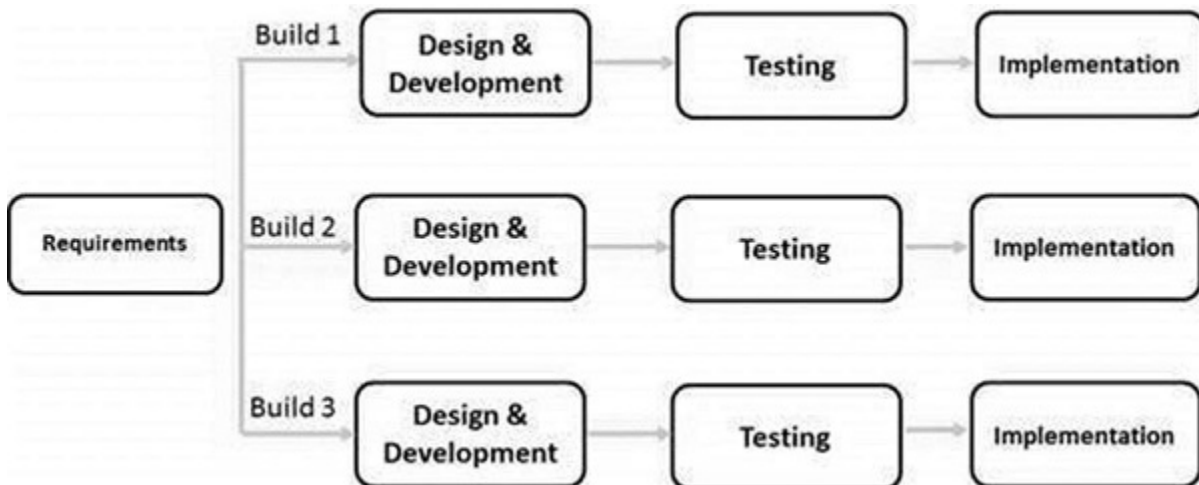
Iterative and Incremental development is a combination of both iterative design or iterative method and incremental build model for development. "During software development, more than one iteration of the software development cycle may be in progress at the same time." This process may be described as an "evolutionary acquisition" or "incremental build" approach."