Data Mining Project Proposal

Team Members: Rishabh Dev, Mattew Faubert, Akash Jassal, Iman Kiani Nezhad

Title: “A Machine-Assisted Approach to Short Answer Question Grading using Data Mining Techniques.”

Introduction: The traditional method of grading short answer questions can be time-consuming and repetitive for teachers, especially when dealing with a large number of responses. This project proposes a new approach to machine-assisted grading, called "powergrading," which aims to amplify the grader's effort by grouping similar responses and to provide rich feedback to groups of answers. This project will be implemented using Data Mining techniques and will be completed within a period of 9 weeks.

Objective:
The objective of this project is to develop a machine-assisted approach for grading short answer questions that amplifies the grader's effort by grouping similar responses and providing rich feedback to groups of answers using data mining and clustering techniques. The project aims to reduce the time and effort required for grading short answer questions and improve the efficiency and effectiveness of the process for teachers and educators.

Methods:
- Familiarize with analogous dataset to understand how to clean the data.
- Preprocessing the student answers using techniques such as Text normalization, Lemmatization, and stop-word removal.
- Train a similarity metric between student responses using techniques such as Cosine similarity, Jaccard similarity, or Euclidean distance.
- Group responses into clusters and subclusters using techniques such as K-means, Hierarchical clustering, or DBSCAN.
- Use an answer key to perform automatic actions for grading.

Proposed technologies and languages:
- Python for implementing data preprocessing, similarity metric calculation, and clustering algorithms.
- Natural Language Processing libraries such as NLTK, spaCy or Gensim for text preprocessing.
- Scikit-learn for clustering algorithm.
- Any database management system for storing student answers and answer key.

Evaluation:
- Compare the results of “powergrading” with a traditional LDA-based approach on a test corpus of 10 questions answered by 698 respondents.
- Measure the progress of grading with a fixed amount of human actions.

Timeline:

- **Week 1:** Familiarize with the existing literature and methods on automated short answer question grading and data mining techniques, and familiarize with analogous datasets to understand how to clean the data.
- **Week 2-3:** Develop the preprocessing pipeline, similarity metric, and clustering algorithm.
- **Week 4-5:** Implement the clustering algorithm and test it on the test corpus.
- **Week 6-7:** Integrate the automatic grading feature and test it on the test corpus.
- **Week 8-9:** Write a report on the project and prepare for the presentation.

Expected Outcomes:
- Streamline the grading process for teachers.
- Provide rich feedback to groups of similar answers.
- Discover modalities of misunderstanding among students.
- Reduce the time and effort required for grading short answer questions.

Impact:
This project has the potential to significantly improve the efficiency and effectiveness of short-answer question grading for teachers and educators. Using data mining techniques, the ability to group similar responses and provide rich feedback to groups of answers can save time and effort while also helping teachers better understand their students' understanding of the material. Additionally, the ability to automatically perform actions when an answer key is available can further reduce the workload of teachers.

Report:
The final report for this project proposal will present a comprehensive overview of the development and implementation of the efficient short answer grading approach using Data Mining techniques. It will include an introduction that provides background information on the problem of short answer question grading and the motivation behind the proposed solution. The report will then detail the objective, methods, technologies, and languages used in the project, including the preprocessing of student answers, the calculation of similarity metrics, and the use of clustering algorithms. It will also include the evaluation of the project, comparing the results of the proposed approach with a traditional LDA-based approach on a test corpus of 10 questions answered by 698 respondents, and measuring the progress of grading with a fixed amount of human actions. The report will also provide results and conclusion of the project, discussing the impact of the proposed approach on the efficiency and effectiveness of short answer question grading for teachers and educators.