Max Fisher, Kyle Hultgren, Nick Allen
CSC 396-05

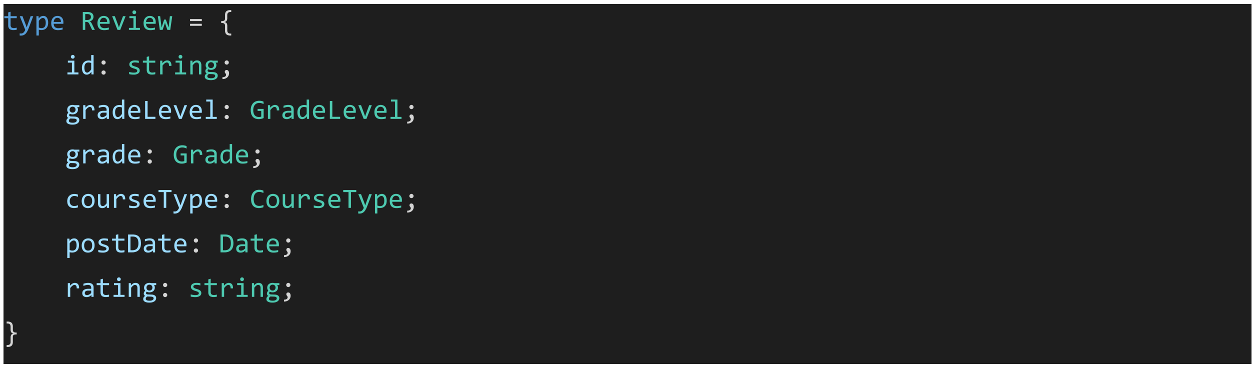
Project Report

Overview

The purpose of this project was to use Scala to search Polyratings for professors to match a prompt inputted by the user. Using the machine learning concepts of TF-IDF and cosine similarity we were able to search through the reviews. After entering a prompt the program will display the 10 most similar reviews.

Data

To obtain our data, we scraped information from the Polyratings website and converted it into Json format. The Json input file (ratings.json) contains information about student ratings, such as the grade they received, the supposed difficulty of the class, and comments about their rating.



```
type Review = {
    id: string;
    gradeLevel: GradeLevel;
    grade: Grade;
    courseType: CourseType;
    postDate: Date;
    rating: string;
}
```

The ratings.json file is just an array of these ratings objects. To see the datatypes in more detail see on the polyratings github here.

Code

Using Scala, we first used the Json input file to create a dataframe containing the rating ID, the professor's ID, and the student's review. From there, we were able to calculate term frequency of each word in each review and the inverse document frequency across the corpus. We then joined the dataframes and multiplied the term frequency by the inverse document frequency to obtain the TF-IDF for each wod in the ratings.

The user is then prompted to input a query to search the reviews. Ex: “One of the best professors ever. Easy grader and helps in office hours.” The term frequency of the input is calculated and then joined with the inverse term frequency of the corpus. We can then join the overall tf idf with the query terms to create vectors that can be compared with cosine similarity. The ratings with the top 10 cosine similarity scores are output as the result and logged to the console.