Automatic Short Essay Grading

Dataset Description:

Training Dataset:

There are 10 datasets. Each dataset represents an Essay set. All responses were written by students primarily in Grade 10. All responses were hand graded and were double-scored.

Training data is provided in tsv form (Tab Separated Value)

    I.   File Name: train.tsv

           Size: 17207 X 5

    II.   File Name: train_rel_2.tsv (Some of the records are removed because of transcription error)

           Size: 17043 X 5

Features:

1. Id: A unique identifier for each individual student essay.
2. EssaySet: 1-10, an id for each set of essays.
3. Score1: The human rater's score for the answer. This is the final score for the answer and the score that you are trying to predict.
4. Score2: A second human rater's score for the answer. This is provided as a measure of reliability, but had no bearing on the score the essay received.
5. EssayText: The ascii text of a student's response.

Test Dataset:

    I.   File Name: test1.tsv

           Size: 5224 X 3

           Features:

1. Id:
2. EssaySet:
3. EssayText:

    II.   File Name: test1_soln.csv (Dataset having SCORE for the test data test1.csv)

           Size: 5224 X 4

           Features:

1. Id:
2. EssaySet:
3. EssayWeight:
4. EssayScore:

    III.   File Name: test2.tsv (Dataset without having SCORE)

           Size: 5100 X 3

           Features:

1. Id:
2. EssaySet:
3. EssayText: