CSCI-B551- Elements of AI

Final Project Proposal

Title: Automated Essay Scoring (Referenced from Kaggle Competition)

Keywords: Machine Learning, Natural Language Processing, Convolutional Neural Networks, Recurrent Neural Networks, GloVe vectors, Word Embeddings.

Background Study:

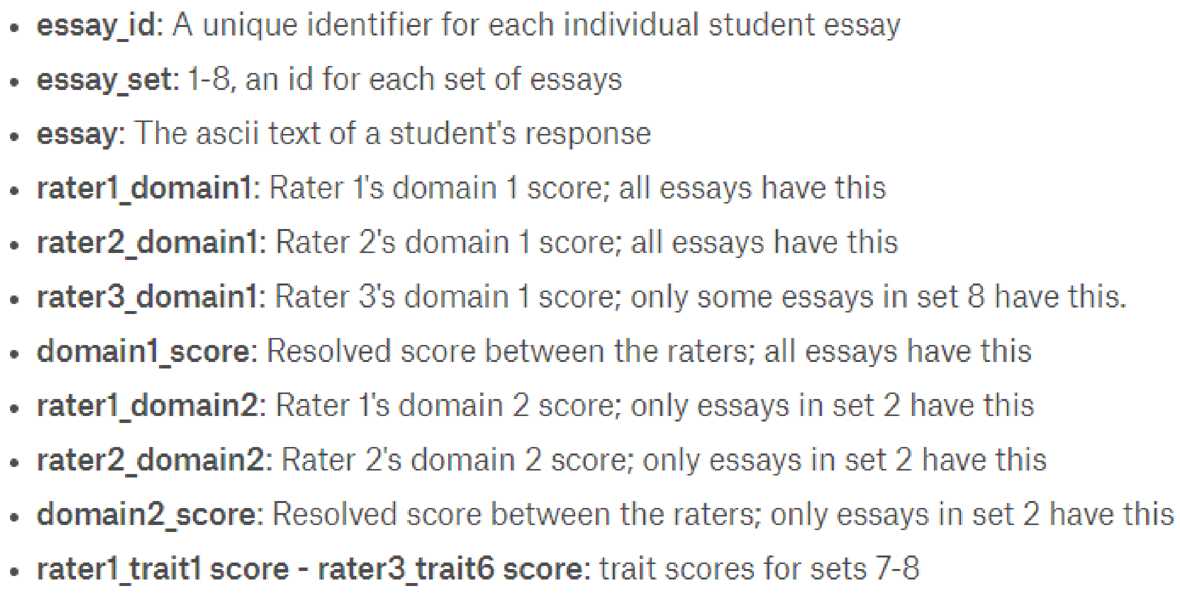
1. **Learning a New(s) Model:** An exploration of LSTM classification and Language Modelling for News Classification. [Link]
2. **GloVe:** Global Vectors for Word Representation. [Link]

Problem Statement:

In this problem, we are trying to build a model which will serve as fast, effective and affordable way to grade student-written essays. The competition has provided hand scored essays to build, validate and test model. Each essay belongs to one of the 8 essay set. Each essay is graded by two raters and the resolved score between them is considered as final score.

Data Description:

Training dataset contains 12979 records of graded essays categorized into 8 essay-sets.



- **essay_id**: A unique identifier for each individual student essay
- **essay_set**: 1-8, an id for each set of essays
- **essay**: The ascii text of a student's response
- **rater1_domain1**: Rater 1's domain 1 score; all essays have this
- **rater2_domain1**: Rater 2's domain 1 score; all essays have this
- **rater3_domain1**: Rater 3's domain 1 score; only some essays in set 8 have this.
- **domain1_score**: Resolved score between the raters; all essays have this
- **rater1_domain2**: Rater 1's domain 2 score; only essays in set 2 have this
- **rater2_domain2**: Rater 2's domain 2 score; only essays in set 2 have this
- **domain2_score**: Resolved score between the raters; only essays in set 2 have this
- **rater1_trait1 score - rater3_trait6 score**: trait scores for sets 7-8

https://www.kaggle.com/c/asap-aes/data

Current Progress:

1. **Domain Knowledge:**
   - Read and analyze **Learning a New(s) Model:** An exploration of LSTM classification and Language Modelling for News Classification.
   - Understood working of GloVe embeddings from Deep Learning with Keras[Link]

2. **Data Pre-processing and Analysis:**
   - Dataset has total 28 columns and 12979 rows. Every essay belongs to one of 8 essay-sets. Every essay has rater1_domain1, rater2_domain1, domain1_score columns.
   - Some essays's in essay-set 8 has rater3_domain1.
   - All the essays in essay-set 2 has rater1_domain2, rater2_domain2, domain2_score.
   - Essays in essay-set 7 and 8 has 5 rater-traits columns.
   - domain1_score has a fixed range from 0-60.
3. **Approach Discussion:**
   - Most of the columns in dataset are incomplete, so we have following columns which has complete information for all the essay-set.
     - **essay_id**
     - **essay_set**
     - **essay**
     - **rater1_domain1**
     - **rater2_domain1**
     - **domain1_score**
   - **Actual Problem Statement:** For the given essay from respective essay-set, we'll try to predict domain1_score for the respective essay. Since, the domain1_score has a fixed range from 0-60. We will frame this problem as a classification problem by making domian1_score as categorical variable.
   - **Dependent and Independent Variables:**
     We will use essay-set and essay as our independent variables(IV) and domain1_score as our dependent variable(DV).
   - **Approach:**
     First, we'll vectorize the textual data (essay) using GloVe and generate numeric vector representation of every essay. Next, we'll append the essay-set number to the essay-vector. We'll train the RNN network to classify each essay based on domain1_score(0-60).
4. **Base Model**
   - I designed basic LSTM model by vectorizing essays using Word2vec. Accuracy on validation set was 39.47% after 5 epochs.

Upcoming Planned Tasks:

- Using GloVe for word embedding.
- Using CNN to learn co-occurrence between words in essay to generate more meaningful vectors and RNN for classification.