Automated Essay Grading for ELs: Sprint 3

Sally Shin
EC601 Fall 2022

TABLE OF CONTENTS

2

PROJECT RECAP

AIMS

- Kaggle competition
- Grade essays from English Learners while taking into account their skill levels
- Explore transformer NLP architecture through the project

GRADING CRITERIA

- **Cohesion**: How sentences or paragraphs flow naturally
- **Syntax**: Correct order of words in phrases and sentences
- **Vocabulary**: Number of unique words
- **Phraseology**: How different phrases may present
- **Grammar**: Correct usage of punctuation, word order, word usage, among others
- **Conventions**: mechanics of writing, notably capitalizations and punctuations

USERS

- Teachers who may need help in grading essays from students of all backgrounds
- Writers who wants to test out automated grading essays for their short essays

PREVIOUS GOALS

Accomplished:

- Configuring remote environment to run trainings
- Better understand workflow of transformer training

Not Accomplished:

- Finding ways to further quantify each grading categories

OVERALL TRAINING WORKFLOW

1. **Pre-preprocessing:** Clean the dataset into separate components
   a. Parse the Kaggle dataset for the essay (text) and its associated scores
   b. Remove unwanted characters like \n from the text
   c. Check that all data is present - no values such as NaN
2. **Preprocessing:** Transform dataset into expected model size inputs
   a. Tokenize the text
      i. Convert text to tokens using a byte-pair encoding tokenizer
   b. Assemble tokens into batches of input tokens to be fed into the model
      i. Padding or truncating batches to fit the model input size
3. **Training:** Use pre-trained models and run training loop
   a. Grab wanted pre-trained model (like Albert, Electra, MobileBert)
   b. Prepare model with desired optimizer and scheduler
   c. Load token batches into the model
   d. Retrieve model outputs
   e. Calculate loss using a loss function
   f. Run backward propagation & step forward
   g. Input next batch

   Run for every batch from dataset
   Then for each preset epoch
4. **Testing:** Use test data on model and convert result into wanted competition format

RESOURCES TO USE


```
import torch
import torch.nn as nn
import transformers
from transformers import (
    AutoModel, AutoConfig,
    AutoTokenizer, logging
)
```
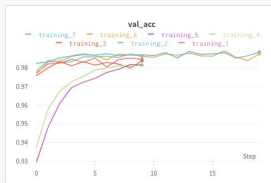
```
accelerator = Accelerator()

model, optimizer, training_dataloader, scheduler = accelerator.prepare(
    model, optimizer, training_dataloader, scheduler
)

for batch in training_dataloader:
    optimizer.zero_grad()
    inputs, targets = batch
    outputs = model(inputs)
    loss = loss_function(outputs, targets)
    accelerator.backward(loss)
    optimizer.step()
    scheduler.step()
```
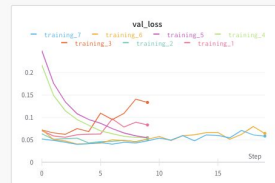

Preprocessing & Training:
- Using HuggingFace python packages
- Large set of pre-trained BERT models - including models of interest
- Also include packages for easy data tokenizing, data loading, model preparation & training

During Training:
- Use W&B
- Easily graph model outputs like loss and accuracy
- Also graph other metrics during training

GOALS FOR NEXT SPRINT

- Preprocess the Kaggle dataset for training

- Configure & Demo a working training loop

- (on side) Defining how to quantify skill levels of essays
  - For scaling input & output scores

RESOURCES

- https://huggingface.co/docs/transformers/index
- https://wandb.ai/site

Instructions for use

In order to use this template, you must credit **Slidesgo** by keeping the Credits slide.

You are allowed to:

- Modify this template.
- Use it for both personal and commercial projects.

You are not allowed to:

- Sublicense, sell or rent any of Slidesgo Content (or a modified version of Slidesgo Content).
- Distribute Slidesgo Content unless it has been expressly authorized by Slidesgo.
- Include Slidesgo Content in an online or offline database or file.
- Offer Slidesgo templates (or modified versions of Slidesgo templates) for download.
- Acquire the copyright of Slidesgo Content.

For more information about editing slides, please read our FAQs:
https://slidesgo.com/faqs

Fonts & colors used

This presentation has been made using the following fonts:

Denk One
(https://fonts.google.com/specimen/Denk+One)

Barlow Semi Condensed
(https://fonts.google.com/specimen/Barlow+Semi+Condensed)
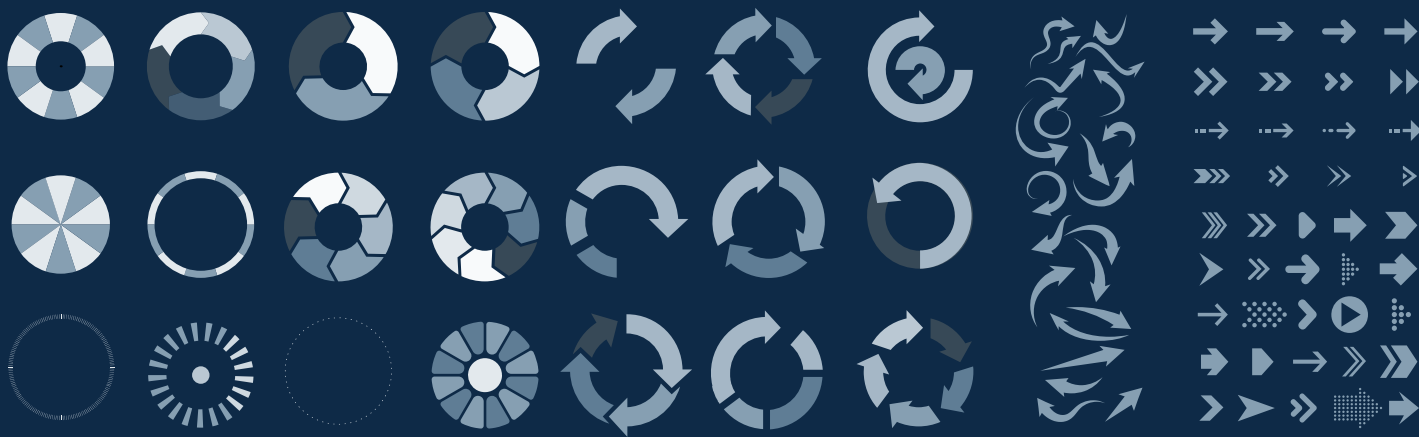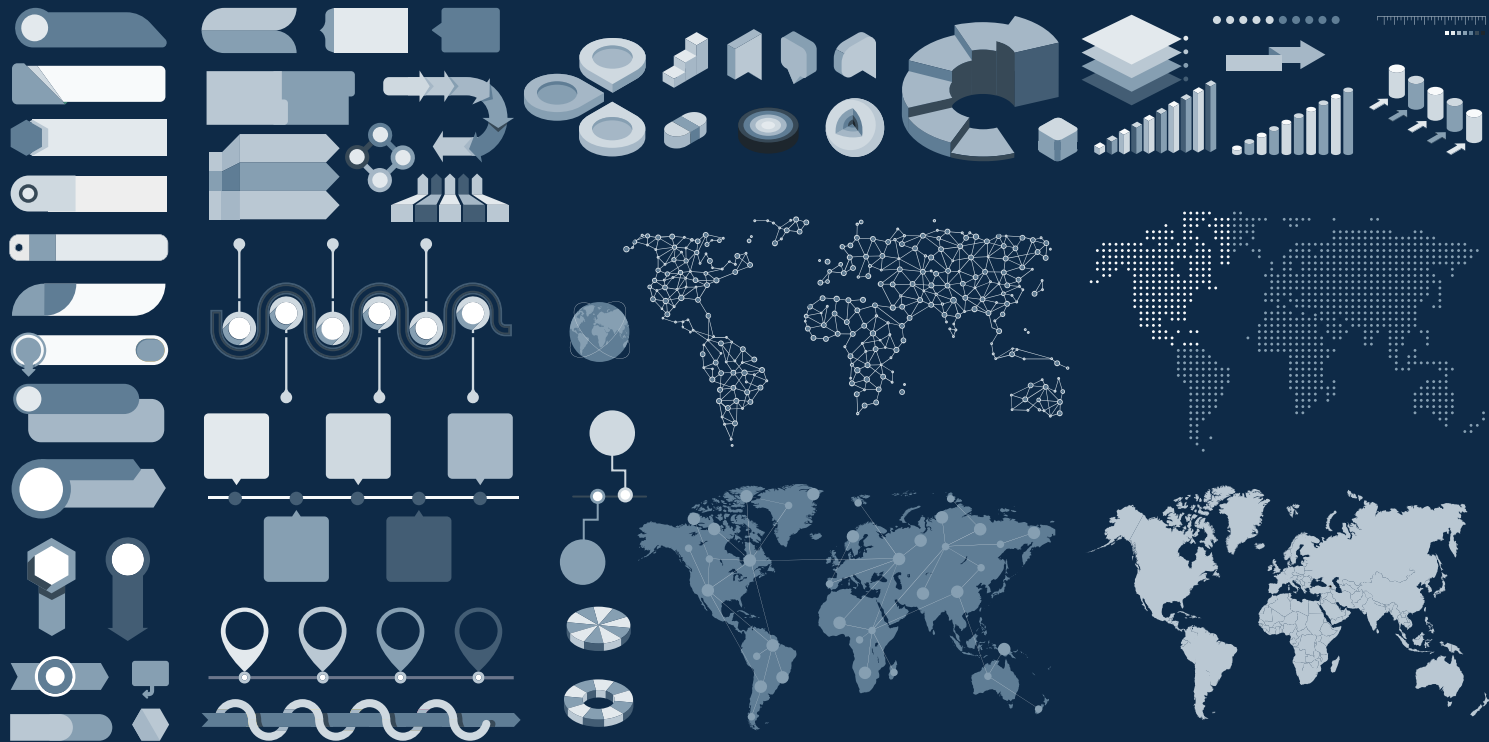
#00959c

#2ab0a7

#a5dddab7

#ff7c3e

#f4e9df

Use our editable graphic resources...

You can easily resize these resources keeping the quality. To change the color just ungroup the resource and click on the object you want to change. Then click on the paint bucket and select the color you want. Don't forget to group the resource again when you're done.

...and our set of editable icons

You can resize these icons keeping the quality.
You can change the stroke and fill color; just select the icon and click on the paint bucket/pen.

Business Icons

Avatar Icons

Creative Process Icons

Educational Process Icons

Help & Support Icons

Medical Icons

Nature Icons

Performing Arts Icons

SEO & Marketing Icons

Teamwork Icons