
FINAL CAPSTONE PROJECT PROPOSAL



Final Capstone Project Proposal

Prepared for: Thinkful

Prepared by: Daniel Rios

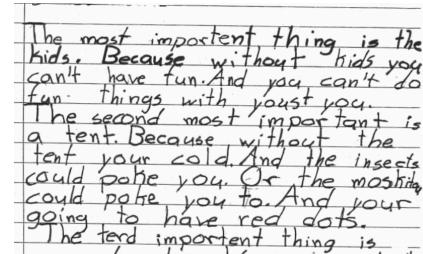
August 1, 2021

FINAL CAPSTONE PROJECT PROPOSAL

DEVELOPING AUTOMATED ASSESSMENT OF STUDENT WRITING EXCERPTS FOR TELPAS

Problem

Every year, the Texas Education Agency requires schools to administer the Texas English Language Proficiency Assessment System (TELPAS) to monitor English Learners' mastery of the English language. Part of the assessment includes a writing portion, where students respond to a prompt in English; an example is shown to the right. Currently, the evaluation of the writing excerpts is conducted by educators and other school staff. Problems with bias exist, and so some cross-evaluation methods are implemented. The state assessment process could benefit from an automated process that evaluates student writing excerpts for English proficiency.



The most important thing is the kids. Because without kids you can't have fun. And you can't do fun things with your kids. The second most important is a tent. Because without the tent your cold. And the insects could bite you. Or the mosquitos could bite you too. And you're going to have red dots. The third important thing is ..

Solution

Before a model that evaluates the student writing samples could be created, there would first have to be a model that transcribes the student writing into text data created. After this initial handwriting model is made, another model that parses the text for data such as the student's use of vocabulary, parts of speech, sentence structure, and other linguistic features would be trained to predict its TELPAS proficiency score. This model evaluation of student writing samples could be implemented to process the state's TELPAS writing samples, and could be expanded to the state's standardized assessment, STAAR, writing samples. For the purposes of this project, I will be focusing working on the second half of the aforementioned solution, and will be creating a model that parses English Learners' writing samples and evaluates them on an English proficiency scale.

Data Source

Unfortunately, the TEA only has one .pdf documents with 50 pages worth of student writing samples, and transcription does not usually happen. Thus, I do not have easily available text data from TEA. During the 2019-2020 school year, however, students completed the TELPAS exam online and had to type their responses. If TEA decides to transition to this method instead, then the need for a handwriting transcription model would not need to be created. Since I need to apply Natural Language Processing techniques to incorporate my specialization, I want to create the model that parses text data to predict a proficiency score. I could create a dataset from the TEA's released .pdf documents by transcribing them and scraping for the score, but there are less than 50 publicly available writing samples. I was, however, able to parse 1,144 .xml files with student responses to the formerly named Cambridge English: First (FCE) exam, as provided by Yannakoudakis (2011), for

FINAL CAPSTONE PROJECT PROPOSAL

text data and individual answer scores on the FCE proficiency scale . An image of one of the .xml files can be found at the end of this proposal.

Techniques and Possible Challenges

Generally speaking, this project is being approached as a supervised learning task with the goal of predicting a student's response score using linguistic features. A further breakdown of skills, methods, and techniques follows, as does a brief summary of some possible challenges.

- Dataset generation
 - Parsing of .xml files using Python
 - Data pipelining with numpy, pandas, and spacy
- Model preparation
 - Feature engineering
 - NLP methods for linguistic features
- Supervised learning models
 - Linear regression
 - Support vector machines
 - Gradient boosting
- Unsupervised learning models
 - Dimensionality reduction methods
 - K-means
 - DBSCAN

The dataset generation itself proved to be a challenge. I had to read through a majority of the lxml.etree package to learn about parsing .xml files, especially given the structure of the .xml files I'm working with. Other possible challenges include the feature generation. I now have a DataFrame with individual responses and their scores, but applying newly learning methods from the NLP module could be challenging. Finally, exploring various kinds of models and implementations of feature sets will surely be time-consuming, so ensuring a thorough review of the data and feature engineering will help alleviate the challenge of time.

Dataset Reference

Yannakoudakis, H., Briscoe, T., and Medlock, B. (2011). A New Dataset and Method for Automatically Grading ESOL Texts. In Proceedings of the 49th Annual Meeting of the Association for Computation Linguistics: Human Language Technologies, pages 180 - 189. ACL.

FINAL CAPSTONE PROJECT PROPOSAL

Image of .xml File

```
<?xml version="1.0" encoding="UTF-8"?>
<learner><head sortkey="TR1005*0100*2000*01">
  <candidate><personnel><language>Turkish</language><age>26-30</age></personnel><score>19.0</score></
  <text>
    <answer1>
      <question_number>1</question_number>
      <exam_score>2.3</exam_score>
      <coded_answer>
        <p>Dear Helen<NS type="RP"><i>. </i><c>, </c></NS></p>
        <p>I have received your letter<NS type="MP"><c>, </c></NS> which <NS type="UV"><i>is</i></NS
        c></NS> in July<NS type="RP"><i>, <NS type="MD"><c>the</c></NS></i><c>. The</c></NS> rest o
        <p>I would prefer to stay in <NS type="MD"><c>a</c></NS> tent. I think it is more enjoyable
        <p>I'm <NS type="S"><i>crayz</i><c>crazy</c></NS> about tennis and swimming. I used to take
        NS>.</p>
        <p>Do I have to bring some cash <NS type="RT"><i>on</i><c>with</c></NS> me? Do you think <N
        there?</p>
        <p><NS type="DA"><i>Your</i><c>Yours</c></NS> sincerely,</p>
      </coded_answer>
    </answer1>
    <answer2>
      <question_number>3</question_number>
      <exam_score>2.2</exam_score>
      <coded_answer>
        <p>Shopping is part of our life. I don't like shopping. Can you <NS type="RV"><NS type="S">
        i></NS> long <NS type="FN"><i>queue</i><c>queues</c></NS>. <NS type="AS"><i>All day you had
        Your time <NS type="AGV"><i>have</i><c>has</c></NS> been stolen. I can't cope with that. I
        <p>There are a lot of <NS type="AGN"><i>option</i><c>options</c></NS> and <NS type="AGN"><i
        shopping you have to carry <NS type="MD"><c>a</c></NS> <NS type="S"><i>havy</i><c>heavy</c>
      </coded_answer>
    </answer2>
  </text>
</head></learner>
```