CSCE 5290: Natural Language Processing Project Proposal

Project Proposal

Github: https://github.com/AraibUNT/NLPProject_CommonLit-Readabillity

Project Proposal

1. Project Title and Team Members

Title : CommonLit Readability Analysis

Team Member: Syed Araib karim, Amanullah Shareef and Soumya Bhandari

• Objectives

We'll create algorithms to rate the difficulty of reading passages for use in grades 3 through 12. We will achieve it by multiple models and try to compare their accuracy. To do so, we'll use our machine learning skills in conjunction with a dataset that contains readers of various ages and a big library of texts from diverse domains. Text coherence and semantics will be included in winning models.

• Motivation

We will be able to help administrators, teachers, and students. Curriculum planners and teachers who select passages for their schools will be able to evaluate works swiftly. Furthermore, these formulas will become more widely available.

• Significance

Successful completion will help administrators, teachers, and students. Curriculum planners and teachers who select passages for their schools will be able to evaluate works swiftly and precisely. Perhaps most crucially, students will gain from comments on the complexity and readability of their work, which will make it much easier for them to improve their reading skills.

• Features

We're predicting the reading ease of excerpts from literature. We've provided excerpts from several time periods and a wide range of reading ease scores. Note that the test set includes a slightly larger proportion of modern texts (the type of texts we want to generalize to) than the training set.

Increment 1 Guidelines:

Related Work (Background)

In CommonLit Readability Analysis, machine learning can help us to determine the proper reading level for a passage of material. Reading is a necessary ability for academic achievement. Students improve reading abilities organically when they have access to captivating passages with the appropriate amount of challenge. Currently, traditional readability methods or commercially accessible algorithms are used to match most instructional materials to readers. Each has its own set of problems. For model training we are going to use Adam optimizer, mean square error as loss function and root mean square error metrics.

Dataset

We are using the dataset from the Kaggle which includes the 2 datasets:

- train
- test

	id	url_legal	license	excerpt	target	standard_error
0	c12129c31	NaN	NaN	When the young people returned to the ballroom...	-0.340259	0.464009
1	85aa80a4c	NaN	NaN	All through dinner time, Mrs. Fayre was somewh...	-0.315372	0.480805
2	b69ac6792	NaN	NaN	As Roger had predicted, the snow departed as q...	-0.580118	0.476676
3	dd1000b26	NaN	NaN	And outside before the palace a great garden w...	-1.054013	0.450007
4	37c1b32fb	NaN	NaN	Once upon a time there were Three Bears who li...	0.247197	0.510845

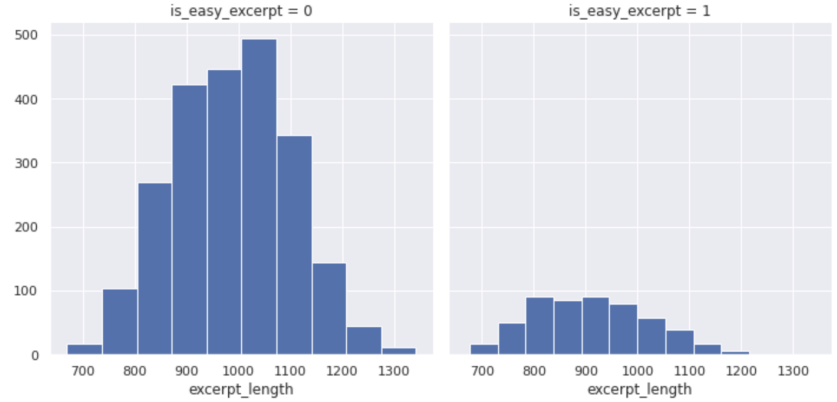
Features of dataset:

```
id              object
url_legal       object
license         object
excerpt         object
target          float64
standard_error  float64
dtype: object
```

Analysis:

- Checking Null values
- Removing & replacing null values.
- Finding new features:
    - Creating an excerpt length feature to track long passages.

- Defining a flag variable ‘is_easy_excerpt’; either 0 or 1 depending if the difficulty is less than 0 or greater than 1.

- Creating a distribution graph based on the flag ‘is_easy_excerpt’
  - 

Implementation:

1. Find new features from the dataset.
   - num_words
   - num_unique_words: Number of unique words in the text
   - num_chars: Number of characters in the text
   - num_stopwords: Number of stopwords in the text
   - num_punctuations: Number of punctuations in the text
   - num_words_upper: Number of title case words in the text
   - num_words_title: Number of title case words in the text
   - mean_word_len: Average length of the words in the text

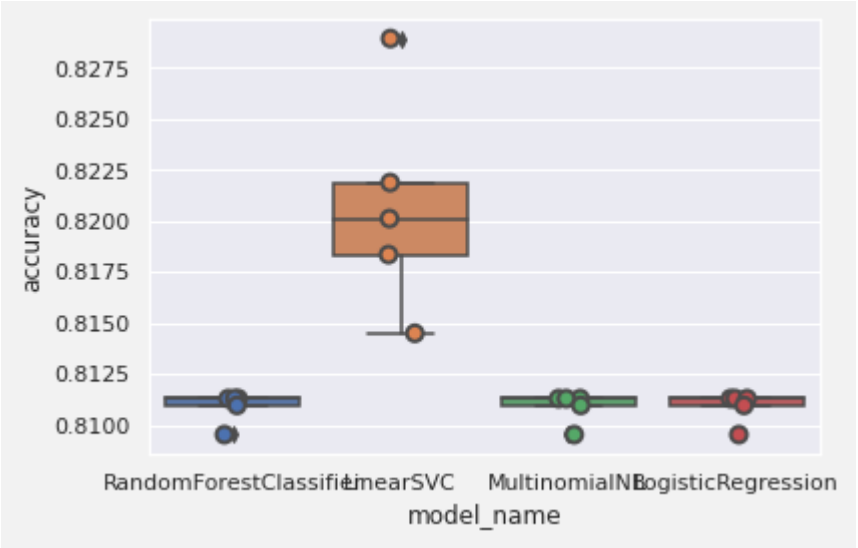
num_words	num_unique_words	num_chars	num_stopwords	num_punctuations	num_words_upper	num_words_title	mean_word_len
179	114	992	88	27	0	16	4.547486
169	127	937	73	56	6	28	4.550296
166	128	908	75	47	2	29	4.475904
164	118	909	70	33	0	8	4.548780

2. Finding unigram and bigrams from excerpt:

```
# '0':
  . Most correlated unigrams:
       . said
```

```
        . mother
    . Most correlated bigrams:
        . little boy
        . little girl
# '1':
    . Most correlated unigrams:
        . said
        . mother
    . Most correlated bigrams:
        . little boy
        . little girl
```

3. Trying baseline models such as
   - Linear regression,
   - SVM,
   - multinomial,
   - LogisticRegression



4. Using SVM and finding RMSE:

   The RMSE (1.0345) value proves that the model is not overfitting, however it needs a lot of improvement.

Project Management:

Work completed:

- EDA
- Feature Engineering

- Baseline model

Responsibility (Task, Person):

- EDA - Soumya Bhandari
- Feature Engineering - Syed Araib Karim & Amanullah
- Baseline models - Syed Araib Karim & Amanullah.
- Report - Amanullah

Contributions (members/percentage):

Syed Araib karim(50%), Amanullah Shareef(20%) and Soumya Bhandari (30%)

Work to be completed & Responsibility (Task, Person):

- More EDA required - Soumya Bhandari
- Finding features which we can relate to the easiness of the passage. - Syed Araib Karim & Amanullah
- Transformer models - Syed Araib Karim & Amanullah

References/Bibliography:

https://www.kaggle.com/c/commonlitreadabilityprize