EASE: Easy Automated Scoring of Essays

Background and Motivation

Today, state departments of education are developing new forms of testing and grading methods, to assess the new common core standards. In this environment the need for more sophisticated and affordable options is vital. For example, we know that essays are an important expression of academic achievement, but they are expensive and time consuming for states to grade them by hand. So, we are frequently limited to multiple-choice standardized tests.

We believe that automated scoring systems can yield fast, effective and affordable solutions that would allow states to introduce essays and other sophisticated testing tools. We believe that you can help us pave the way towards a breakthrough.

Through this project we try to see how well a computer program can perform a subjective human task.

This project covers the processes involved in modeling an automated essay scoring system, from data cleaning, feature generation to various models results comparisons.

Dataset

The data, acquired from Kaggle looks something like this:

	A	B	C	D	E	F	G
1	essay_id	essay_set	essay	rater1_dom	rater2_dom	rater3_dom	domain1_sco
2	1	1	Dear local newspaper, I think effects computers have on people are great lear	4	4		8
3	2	1	Dear @CAPS1 @CAPS2, I believe that using computers will benefit us in many	5	4		9
4	3	1	Dear, @CAPS1 @CAPS2 @CAPS3 More and more people use computers, but	4	3		7
5	4	1	Dear Local Newspaper, @CAPS1 I have found that many experts say that com	5	5		10
6	5	1	Dear @LOCATION1, I know having computers has a positive effect on people.	4	4		8
7	6	1	Dear @LOCATION1, I think that computers have a negative affect on us! How	4	4		8
8	7	1	Did you know that more and more people these days are depending on comp	5	5		10
9	8	1	@PERCENT1 of people agree that computers make life less complicated. I also	5	5		10
10	9	1	Dear reader, @ORGANIZATION1 has had a dramatic effect on human life. It h	4	5		9
11	10	1	In the @LOCATION1 we have the technology of a computer. Some say that th	5	4		9
12	11	1	Dear @LOCATION1, @CAPS1 people acknowledge the great advances that co	4	4		8
13	12	1	Dear @CAPS1 @CAPS2 I feel that computers do take away from peoples life a	4	4		8
14	13	1	Dear local newspaper I raed ur argument on the computers and I think they ar	4	3		7
15	14	1	My three detaileds for this news paper article is one state you opinion about t	3	3		6
16	15	1	Dear, In this world today we should have everyone useing computers. Compu	3	3		6
17	16	1	Dear @ORGANIZATION1, The computer blinked to life and an image of a blon	6	6		12
18	17	1	Dear Local Newspaper, I belive that computers have a negative effect on peop	4	4		8
19	18	1	Dear Local Newspaper, I must admit that the experts are centainly right cause	4	4		8
20	19	1	I aegre waf the evansmant ov tnachnolage. The evansmant ov tnachnolige is t	2	2		4
21	20	1	Well computers can be a good or a bad thing. I don'@CAPS1 realy see @CAPS	3	3		6

- essay_id: Unique reference number for the essay
- essay_set: [1,8] Describes the type of essay.
- essay: ASCII text of the essay
- rater1_domain1: Rater 1 score
- rater2_domain1: Rater 2 score
- domain1_score: Resolved final score.

essay_set	
1	1783
2	1800
3	1726
4	1772
5	1805
6	1800
7	1569
8	723

There are total 8 sets of essays, each with varying traits. For this project we selected the data sets 1,3,5 and 6 because they have the same attributes (2 rater scores and 1 domain score).

The table on the left shows the count of essays in each set.

Fortunately, the dataset was clean enough with merely <10 rows with missing values. Such rows were removed.

Feature Generation

A sample essay looks like:

```
"Dear @CAPS1 @CAPS2, I believe that using computers will benefit us in many ways like talking and becoming friends wi
ll others through websites like facebook and mysace. Using computers can help us find coordibates, locations, and abl
e ourselfs to millions of information. Also computers will benefit us by helping with jobs as in planning a house pla
n and typing a @NUM1 page report for one of our jobs in less than writing it. Now lets go into the wonder world of te
chnology. Using a computer will help us in life by talking or making friends on line. Many people have myspace, faceb
ooks, aim, these all benefit us by having conversations with one another. Many people believe computers are bad but h
ow can you make friends if you can never talk to them? I am very fortunate for having a computer that can help with n
ot only school work but my social life and how I make friends. Computers help us with finding our locations, coordiba
tes and millions of information online. If we didn't go on the internet a lot we wouldn't know how to go onto website
s that @MONTH1 help us with locations and coordinates like @LOCATION1. Would you rather use a computer or be in @LOCA
TION3. When your supposed to be vacationing in @LOCATION2. Million of information is found on the internet. You can a
s almost every question and a computer will have it. Would you rather easily draw up a house plan on the computers or
take @NUM1 hours doing one by hand with ugly erazer marks all over it, you are garrenteed that to find a job with a d
rawing like that. Also when appling for a job many workers must write very long papers like a @NUM3 word essay on why
this job fits you the most, and many people I know don't like writing @NUM3 words non-stopp for hours when it could t
ake them I hav an a computer. That is why computers we needed a lot now adays. I hope this essay has impacted your de
scion on computers because they are great machines to work with. The other day I showed my mom how to use a computer
and she said it was the greatest invention sense sliced bread! Now go out and buy a computer to help you chat online
with friends, find locations and millions of information on one click of the button and help your self with getting a
job with neat, prepared, printed work that your boss will love."
```

Since a computer algorithm cannot understand text as it is, we need to generate numerical features out of this. To do this we create the following features from the essays:

1. Lexical Features

These features include the parts of speech tags for the words in a given essay. The following were selected:

1. Number of Nouns.
2. Number of Proper nouns.
3. Number of Verbs.
4. Number of Adjectives.

These features were generated using the space library in Python.

2. Numerical meta features

These can directly be derived from the essay. They include:

5. Count of words.
6. Count of sentences.
7. Count of distinct words.
8. Count of syllables.

---

3. Derived Readability Indices

Readability Index

Readability index is designed to gauge the understandability of a written article. They check for features such as the number of words, sentences or syllables. Some also use dictionaries of difficult words. Text complexity is often compared to how well readers comprehend the text.

Every index is a little bit different and focusses on a particular aspects of text complexity. There are specialized indices for the domains of Military or Medical etc.
The simple readability index produces an approximate representation of the US grade level needed to comprehend a given text. Likewise, all readability indices have their own interpretations. There are multiple reading indices available. For this project the following common indices were selected

To know more, links to readability indices are given in the references section at the end of the report.

9. Simple Readability Index

It depends on the number of characters, count of words and count of sentences. The formula looks like :

```
(4.71 * (charCount/wordCount) + 0.5 * (wordCount/senCount) - 21.43)
```

10. Flesch Reading Ease

It depends on number of words, number of sentences and number of syllables.

```
FRE = 206.835 - float(1.015 * avg_sentence_length(text))
        - float(84.6 * avg_syllables_per_word(text))
```

11. Gunning Fog Grade

It depends on number of difficult words and number of words.

```
per_diff_words = (difficult_words(text) / addWordCount(text) * 100) + 5
grade = 0.4 * (avg_sentence_length(text) + per_diff_words)
```

12. Smog Index

Depends on syllable and sentence count.

```
SMOG = (1.043 * (30*(poly_syllab / addSentenceLength(text)))**0.5) + 3.1291
```

Here poly syllable count is the number of words of more than two syllables in a sample of 30 sentences.

13. Dale Chall Readability Index

Depends on number of difficult words, words and sentences.

```
raw_score = (0.1579 * diff_words) + (0.0496 * avg_sentence_length(text))

# If Percentage of Difficult Words is greater than 5 %, then;
# Adjusted Score = Raw Score + 3.6365,
# otherwise Adjusted Score = Raw Score
```

After generating these features for an essay we are able to create numerical features which look like this:

The feature generation process was the most critical component of this project. It took significantly more time as well to complete. Mentioned below is the time taken for generating these features for essay set 1.

Feature	Time in Seconds
Noun count	250
Proper noun count	478
Verb count	220
Adjective Count	220
Numerical features count	2.8
Simple Readability Index	0.15
Flesch readability index	1.78
Gunning Fog Grade	23
Smog readability index	22
Dale chall readability index	23
Total Time	**992**

It took around 16 minutes for the 1800 essays in set 1 to get the features. This is much more than the time taken to actually run train the models.

After generating the features, an essay which looks like this:

```
"Dear @CAPS1 @CAPS2, I believe that using computers will benefit us in many ways like talking and becoming friends wi
ll others through websites like facebook and mysace. Using computers can help us find coordibates, locations, and abl
e ourselfs to millions of information. Also computers will benefit us by helping with jobs as in planning a house pla
n and typing a @NUM1 page report for one of our jobs in less than writing it. Now lets go into the wonder world of te
chnology. Using a computer will help us in life by talking or making friends on line. Many people have myspace, faceb
ooks, aim, these all benefit us by having conversations with one another. Many people believe computers are bad but h
ow can you make friends if you can never talk to them? I am very fortunate for having a computer that can help with n
ot only school work but my social life and how I make friends. Computers help us with finding our locations, coordiba
tes and millions of information online. If we didn't go on the internet a lot we wouldn't know how to go onto website
s that @MONTH1 help us with locations and coordinates like @LOCATION1. Would you rather use a computer or be in @LOCA
TION3. When your supposed to be vacationing in @LOCATION2. Million of information is found on the internet. You can a
s almost every question and a computer will have it. Would you rather easily draw up a house plan on the computers or
take @NUM1 hours doing one by hand with ugly erazer marks all over it, you are garrenteed that to find a job with a d
rawing like that. Also when appling for a job many workers must write very long papers like a @NUM3 word essay on why
this job fits you the most, and many people I know don't like writing @NUM3 words non-stopp for hours when it could t
ake them I hav an a computer. That is why computers we needed a lot now adays. I hope this essay has impacted your de
scion on computers because they are great machines to work with. The other day I showed my mom how to use a computer
and she said it was the greatest invention sense sliced bread! Now go out and buy a computer to help you chat online
with friends, find locations and millions of information on one click of the button and help your self with getting a
job with neat, prepared, printed work that your boss will love."
```

Was now represented like this:

nounCount	propnCount	verbCount	adjCount	senCount	wordCount	distinctCount	syllableCount	avgSPerWord	readabilityIndex	riFRE	riGF	riSI	riDC
78	5	73	19	17	338	128	452	1.32	9.943070	74.98	10.071284	3.129	1.033

Now that we derived numerical attributes out of the essay we can prepare and run our models.

Model Preparation

To prepare our new data set for modeling we split it into training and testing. We split the dataset to retain 80% for training and 20% for testing.

The data sets were then scaled using a Z score scaler from sklearn library in python.

Models

Logistic Regression

Logistic regression is a simple model. All features were given as an input and the output was one of the domain score classes, treating this as a classification problem.

The R squared values obtained were as below:

Essay Set	Train	Test
1	0.54	0.46
3	0.379	0.367
5	**0.622**	**0.601**
6	0.390	0.358

K Nearest Neighbors

KNN is one of the most common classification algorithms. Again all features were given as input and the domain score class was the output. These results are for n set to 7.

The R squared values obtained are as below:

Essay Set	Train	Test
1	**0.715**	**0.658**
3	0.48	0.35
5	0.660	0.544
6	0.50	0.32

The model actually performed good for essay set 1! Although there may be slight overfitting as indicated from the gap between Test and Training scores.

Neural Network

The final model used was a neural networks. Neural networks have proven to be robust good models which simply work better exploiting the computation power we have easily available these days.

The architecture of the neural network looks like this:

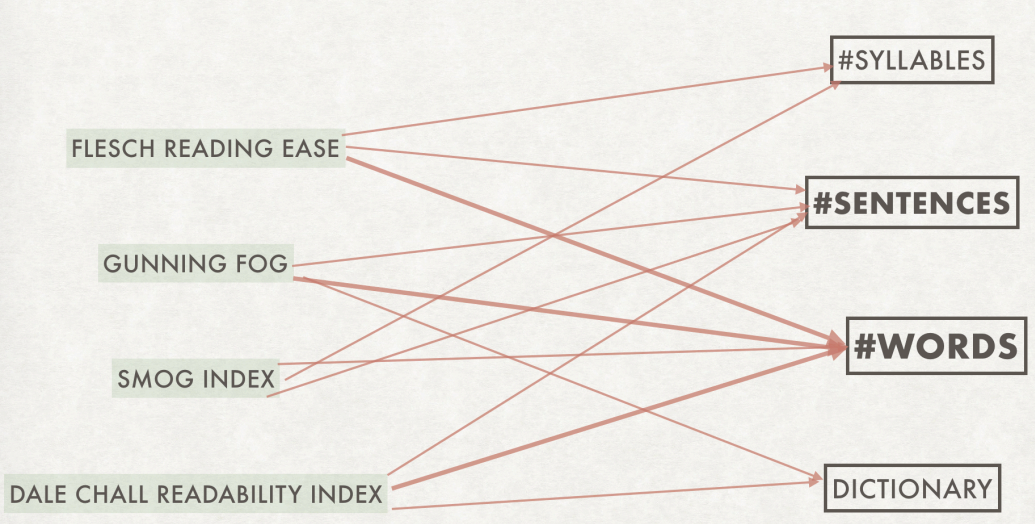
Input Layer	14 neurons
Hidden Layer 1	72 neurons
Hidden Layer 2	64 neurons
Hidden Layer 3	48 neurons
Output Layer	1 neuron
Learning Rate	0.001
Optimizer	Adam
Loss Function	Mean Squared Error
Activation Function	Relu
Validation Set	0.20
Epochs	20

The R squared values obtained are as below:

Essay Set	Train	Test
1	**0.78**	**0.72**
3	0.59	0.50
5	0.74	0.67
6	0.64	0.50

The performs performs quite well for essay set 1. And also essay set 5. But essay set 3 and 6 still don't have that good performance.

Conclusions:



- All readability indices depend on basic features. The above diagram represents the prominence of those features. It is easy to see that number of words is the most important or most used feature in readability indices used. Next was number of sentences and the lesser used ones were Dictionary to find difficult words and the number of syllables.
- Essay set 1 showed the most computer predictable behavior.
- Essay sets had varying predictability.
- Care has to be taken to not overtrain the neural network. The optimal learning rate and number of epochs have to be chosen.

Whats ahead:

- Factor analysis to find out which features or combination of were the most important indicators of the domain score.
- If we read more about the essay sets, maybe we can generate essay set specific features to bridge the gap between the varying results for essay sets.
- Hyper parameter tuning is always a challenge for models.

References:

1. Simple readability index : https://en.wikipedia.org/wiki/Automated_readability_index
2. Flesch-Kincaid readability index: https://en.wikipedia.org/wiki/Flesch%E2%80%93Kincaid_readability_tests
3. Gunning fog index: https://en.wikipedia.org/wiki/Gunning_fog_index
4. Smog index: https://en.wikipedia.org/wiki/SMOG
5. Dale chill: https://en.wikipedia.org/wiki/Dale%E2%80%93Chall_readability_formula
6. Dataset link: https://www.kaggle.com/c/asap-aes
7. TensorFlow: https://www.tensorflow.org/
8. Spacy: https://spacy.io/
9. Textstat: https://pypi.org/project/textstat/0.1.4/