

OPENCLASSROOMS

SOUTENANCE: PROJET 8

COMMONLIT READABILITY PRIZE

Edward Levavasseur

Updated: 2021/07/09



Problématique

Can machine learning identify the appropriate reading level of a passage of text, and help inspire learning?

In this competition, you'll build algorithms to rate the complexity of reading passages for grade 3-12 classroom use. To accomplish this, you'll pair your machine learning skills with a dataset that includes readers from a wide variety of age groups and a large collection of texts taken from various domains.

1. Présentation des données
2. Elaboration des Word Embeddings
3. Réseau de Neurones
4. Résultats
5. Conclusion

PRÉSENTATION DES DONNÉES

Données Training:

- 2834 Textes
- 'target' : difficulté moyenne évaluée par des lecteurs

	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
5	f9bf357fe	NaN	NaN	Hal and Chester found ample time to take an in...	-0.861809	0.480936
6	eaf8e7355	NaN	NaN	Hal Paine and Chester Crawford were typical Am...	-1.759061	0.476507
7	0a43a07f1	NaN	NaN	On the twenty-second of February, 1916, an aut...	-0.952325	0.498116
8	f7eff7419	NaN	NaN	The boys left the capitol and made their way d...	-0.371641	0.463710
9	d96e6dbcd	NaN	NaN	One day he had gone beyond any point which he ...	-1.238432	0.465900

Données Test:

- 7 Textes
- Pas de données variable 'target'

	id	url_legal	license	excerpt
0	c0f722661	NaN	NaN	My hope lay in Jack's promise that he would ke...
1	f0953f0a5	NaN	NaN	Dotty continued to go to Mrs. Gray's every nig...
2	0df072751	NaN	NaN	It was a bright and cheerful scene that greete...
3	04caf4e0c	https://en.wikipedia.org/wiki/Cell_division	CC BY-SA 3.0	Cell division is the process by which a parent...
4	0e63f8bea	https://en.wikipedia.org/wiki/Debugging	CC BY-SA 3.0	Debugging is the process of finding and resolv...
5	12537fe78	NaN	NaN	To explain transitivity, let us look first at ...
6	965e592c0	https://www.africanstorybook.org/#	CC BY 4.0	Milka and John are playing in the garden. Her ...

Données Test:

- 7 Textes
- Pas de données variable 'target'

	id	url_legal	license	excerpt
0	c0f722661	NaN	NaN	My hope lay in Jack's promise that he would ke...
1	f0953f0a5	NaN	NaN	Dotty continued to go to Mrs. Gray's every nig...
2	0df072751	NaN	NaN	It was a bright and cheerful scene that gree...
3	04caf4e0c	https://en.wikipedia.org/wiki/Cell_division	CC BY-SA 3.0	Cell division is the process by which a parent...
4	0e63f8bea	https://en.wikipedia.org/wiki/Debugging	CC BY-SA 3.0	Debugging is the process of finding and resolv...
5	12537fe78	NaN	NaN	To explain transitivity, let us look first at ...
6	965e592c0	https://www.africanstorybook.org/#	CC BY 4.0	Milka and John are playing in the garden. Her ...

- Il faut:
 - Prédire la variable 'target'
 - soumettre les réponses à Kaggle pour évaluation

ELABORATION DES WORD EMBEDDINGS

Séparation des Données Training:

- Training Set: 99.3%
- Test Set 1: 7 observations
- Test Set 2: 7 observations
- Test Set 3: 7 observations

Séparation des Données Training:

- Training Set: 99.3%
- Test Set 1: 7 observations
- Test Set 2: 7 observations
- Test Set 3: 7 observations

10-fold Cross Validation:

- Division du Training Set en 10 (9.993 %):
 - Training : 9 sous-ensembles (89.93 %)
 - Validation : 1 sous-ensemble (9.993 %)

Word Embedding: TF-IDF

- Création d'un espace vectoriel :
 - Chaque mot du corpus de textes est une dimension
 - Chaque ligne est un texte
 - Chaque cellule est le:
 - la fréquence du mot dans le texte, fois le logarithme de la fréquence inverse des documents qui contiennent ce mot parmi tous les textes

Word Embedding: TF-IDF

- Création d'un espace vectoriel :
 - Chaque mot du corpus de textes est une dimension
 - Chaque ligne est un texte
 - Chaque cellule est le:
 - la fréquence du mot dans le texte, fois le logarithme de la fréquence inverse des documents qui contiennent ce mot parmi tous les textes
- TF-IDF entraîné sur les données Train
- Transformation des données:
 - Train
 - Validation
 - Test 1
 - Test 2
 - Test 3

Word Embedding: CountVectorizer

- Création d'un espace vectoriel :
 - Chaque mot du corpus de textes est une dimension
 - Chaque ligne est un texte
 - Chaque cellule est le:
 - le nombre de fois que le mot a été utilisé

Word Embedding: CountVectorizer

- Création d'un espace vectoriel :
 - Chaque mot du corpus de textes est une dimension
 - Chaque ligne est un texte
 - Chaque cellule est le:
 - le nombre de fois que le mot a été utilisé
- CountVectorizer entraîné sur les données Train
- Transformation des données:
 - Train
 - Validation
 - Test 1
 - Test 2
 - Test 3

Semantic Embedding: Doc2Vec

- Création d'un espace vectoriel :
 - Transformation de chaque texte en vecteur, où les dimensions sont un hyper-paramètre
- Application de Doc2Vec sur:
 - Train
 - Validation
 - Test 1
 - Test 2
 - Test 3

Semantic Embedding: Doc2Vec

- Création d'un espace vectoriel :
 - Transformation de chaque texte en vecteur, où les dimensions sont un hyper-paramètre
- Application de Doc2Vec sur:
 - Train
 - Validation
 - Test 1
 - Test 2
 - Test 3
- Problème:
 - les valeurs du vecteurs sont légèrement différentes à chaque fois (seed ne peut pas être fixée).

Réseau de NEURONES

- Couches

- Dense(2000)
- Dense(1000)
- Dense(500)
- Dense(250)
- Dense(125)
- Dense(60)
- Dense(30)
- Dense(15)
- Dense(1)

- Hyper-paramètres

- Dropout Rate: 0.1
- Optimizeur: 'sgd'
- Loss: MSE

- 10-fold cross-validation
 - 1 prédiction par itération
 - Sur données Train, Validation, Test 1, Test 2 et Test 3

```
*      excerpt Predicted_target_0 Predicted_target_1 ... Predicted_target_7 Predicted_target_8 Predicted_target_9
0 Her husband gasped at the audacity of the idea... -0.861321 -0.715166 ... -0.629515 -0.885164 -0.415509
1 At last they managed to leap from the logs, no... -0.445580 -0.186719 ... -0.214262 -0.307410 -0.230487
2 Those who have not learned to read the ancient... -1.397059 -1.571614 ... -1.715401 -1.789596 -1.440662
3 Colors are everywhere in nature, and they comm... -0.635872 -0.418479 ... -0.687495 -0.485997 -0.598047
4 The first well was put down about eight years ... -1.739423 -1.437097 ... -1.607612 -2.002952 -1.770393
5 Depression is the most common mental illness. ... 0.232145 0.184169 ... 0.025086 -0.726862 -0.005836
6 King Edward, be it remembered, was a man of ma... -1.536290 -1.388514 ... -1.905120 -1.749020 -1.708939
7 Meanwhile, Hercules travelled constantly onwar... -1.259620 -1.537378 ... -1.512524 -2.032029 -1.749643
```

- Calcul de la médiane des prédictions
- Calcul de la moyenne des prédictions

```
3
0 Her husband gasped at the audacity of the idea... Predicted_target_0 -0.861321 Predicted_target_1 -0.715166 ... Predicted_target_9 -0.415509 MedianPredictions -0.713076 MeanPredictions -0.684579
1 At last they managed to leap from the logs, no... -0.445580 -0.186719 ... -0.230487 -0.251685 -0.287543
2 Those who have not learned to read the ancient... -1.397059 -1.571614 ... -1.440662 -1.617835 -1.630733
3 Colors are everywhere in nature, and they comm... -0.635872 -0.418479 ... -0.598047 -0.611902 -0.593320
4 The first well was put down about eight years ... -1.739423 -1.437097 ... -1.770393 -1.794803 -1.801557
5 Depression is the most common mental illness. ... 0.232145 0.184169 ... -0.005836 0.024072 -0.146582
6 King Edward, be it remembered, was a man of ma... -1.536290 -1.388514 ... -1.708939 -1.755907 -1.742901
7 Meanwhile, Hercules travelled constantly onwar... -1.259620 -1.537378 ... -1.749643 -1.643511 -1.650096
```

RÉSULTATS

Modèle préliminaire - 2 Validation sets

R2 - TF-IDF :

Train : 0.981
Validation 1 : 0.862
Validation 2 : 0.896
Test 1 : 0.919
Test 2 : 0.824
Test 3 : 0.947

RMSE - TF-IDF :

Train : 0.308
Validation 1 : 0.826
Validation 2 : 0.715
Test 1 : 0.648
Test 2 : 0.987
Test 3 : 0.455

(a) TF-IDF

R2 - CountVectorizer :

Train : 0.974
Validation 1 : 0.859
Validation 2 : 0.894
Test 1 : 0.929
Test 2 : 0.826
Test 3 : 0.958

RMSE - CountVectorizer :

Train : 0.358
Validation 1 : 0.838
Validation 2 : 0.722
Test 1 : 0.603
Test 2 : 0.982
Test 3 : 0.405

(b) CountVectorizer

R2 - Doc2Vec :

Train : 0.957
Validation 1 : 0.82
Validation 2 : 0.819
Test 1 : 0.952
Test 2 : 0.744
Test 3 : 0.827

RMSE - Doc2Vec :

Train : 0.46
Validation 1 : 0.945
Validation 2 : 0.941
Test 1 : 0.499
Test 2 : 1.188
Test 3 : 0.819

(c) Doc2Vec

Median Predictions - R2 :

Test 1 : 0.9358209015773998

Test 2 : 0.8700344197855303

Test 3 : 0.9033267873899911

Mean Predictions - R2 :

Test 1 : 0.9354950071897643

Test 2 : 0.8706475724950339

Test 3 : 0.9070023926908595

- RMSE sur soumission Kaggle: **0.750**

Median Predictions - R2 :

Test 1 : 0.8889602945703001
Test 2 : 0.9279415129321984
Test 3 : 0.9676183605256818

Mean Predictions - R2 :

Test 1 : 0.890805813714494
Test 2 : 0.926357823488308
Test 3 : 0.9671768611977828

Final Predictions - R2 :

Test 1 : 0.8899309282394382
Test 2 : 0.9271605817264256
Test 3 : 0.9674360595870752

(a) R2

Median Predictions - RMSE :

Test 1 : 0.6501217693178905
Test 2 : 0.6286083598152495
Test 3 : 0.41139320098665966

Mean Predictions - RMSE :

Test 1 : 0.6446965056403581
Test 2 : 0.6354785420312691
Test 3 : 0.4141882248484389

Final Predictions - RMSE :

Test 1 : 0.64727407107668
Test 2 : 0.6320054406835749
Test 3 : 0.4125495987113205

(b) RMSE

- RoBERTa - Tokenizer:
 - Méthode pré-entraînée
 - Transformation de chaque texte en vecteur
- Application de CountVectorizer aux valeurs des vecteurs
- "Fitting" du réseau de neurones sur le train

- RoBERTa - Tokenizer:
 - Méthode pré-entraînée
 - Transformation de chaque texte en vecteur
- Application de CountVectorizer aux valeurs des vecteurs
- "Fitting" du réseau de neurones sur le train
- Amélioration des Résultats
- Problème:
 - Nécessite une connexion à internet
 - Kaggle refuse la connexion à internet dans cette compétition

CONCLUSION

Conclusion

- Participation à la compétition CommonLit
 - Prédiction de la difficulté de Textes
 - Supervised learning
- Word Embedding
- Réseau de Neurones
- Résultats - Soumission:
 - $RMSE : 0.755$
- Résultats - Variante Roberta:
 - $RMSE : 0.6$ environ