Participation de l'équipe TGV à DEFT 2022 : Prédiction automatique de notes d'étudiants à des questionnaires en fonction du type de question

Vanessa Gaudray Bouju, Margot Guettier, Gwennola Lerus, Gaël Guibon, Matthieu Labeau, Luce Lefeuvre

SOMMAIRE

1 ÉTAT DE L'ART

ÉTUDE DES APPROCHES DE DEFT 2021 OÙ LA MÊME TÂCHE ÉTAIT PROPOSÉE.

Tâche envisagée : classification.

Principales approches mises en place :

- Calcul de scores de similarité (meilleurs résultats en comparant la réponse de l'élève avec celle du professeur sans prise en compte de la question) ;
- Plongements lexicaux (meilleurs résultats avec des modèles multilingues) ;
- Identification de caractéristiques ;
- Entraînement de classifieurs (meilleur résultat : 0,682 avec Random Forest).

2 REPRÉSENTATION DES DONNÉES

2.1 PRÉTRAITEMENTS RÉALISÉS

Nettoyage du corpus

- Suppression des balises HTML de structuration du corpus, des retours à la ligne et des espaces en trop ;
- Ajout des espaces manquantes entre les balises <p> pour récupérer les différents paragraphes de correction ;
- Remplacement des réponses ne contenant que des caractères non alphanumériques (« ??? », « :'(») par « NO_ANS ».

Séparation des réponses proposées et du barème.

Nature différente :

- Propositions de réponses sémantiquement et/ou formellement proches des réponses des élèves ;
- Barème peut donner des indications sans lien direct (« Compté 1 même si faute de frappe ou par ex vald »).

→ Création d'une nouvelle colonne dans le fichier des questions ne contenant que la partie réponse proposée, sans barème, afin de pouvoir plus facilement la prendre en compte seule.

2.2 VECTORISATION

TF-IDF

- Sur les caractères
  - Normalisés en minuscules
- Sans restriction de dimension
- De unigrams à 3-grams

Fonction de hachage (Weinberger et al., 2009)

- Sur les tokens
- Matrice sparse de taille fixe à 50 dim
  - Dimension petite motivée par la taille du corpus et le nombre de caractéristiques
- Dimensions testées : 50, 100, 300
  - Les collisions n'impactent pas la performance
- Données inconnues mappées sur 0

2.3 CARACTÉRISTIQUES

IDENTIFICATION DE DIFFÉRENTES CARACTÉRISTIQUES BASÉES SUR DIVERSES MÉTRIQUES

- Des marqueurs permettant d'identifier la tournure de la question (« qu'affiche », « quel code », « comment »…)
- Le nombre de mots et le nombre de caractères dans : la question, la réponse du professeur et celle de l'élève ;
- Le nombre de minuscules et de majuscules dans : la question ;
- Le nombre et le pourcentage de caractères de code dans : la question, la réponse du professeur et celle de l'élève ;
- Les ENT (entités nommées) pour le nom de langages informatiques (modèle camembert-ner) ;
- Les POS (catégories grammaticales) (modèle flair/upos-multi).

3 APPROCHES

3.1 APPROCHE PAR RÉGRESSION

Étapes

1. Vectorisation
2. SGD regressor (|0-1|)
3. Mapping des valeurs

Deux stratégies de classification

- SGD + arrondi vers valeurs réelles
- SGD + L2 distance

RMSE	MSE	MAE	Précision
0.5153	0.2650	0.4068	0.369

607 évaluations correctes
1 037 évaluations incorrectes

Avantages

- Réelle prédiction de la note
- Adaptable à n'importe quel jeu de test
- Convient à la tâche

Limites

- Catégories insuffisantes
- Dépendant de la stratégie de mapping
- Ne convient pas au défi

→ Changement de stratégie : approche par classification.

Objectif : Classifier les réponses par rapport aux notes (une note = une classe)
/!\ Disparités entre les fichiers.

3.2 APPROCHES PAR CLASSIFICATION

3.2.1 PROBABILITÉ DE NOTES D'ÉTUDIANTS

UTILISATION DE L'IDENTITÉ DE L'ÉTUDIANT COMME CARACTÉRISTIQUE

Description de l'approche :

Postulat : tous les étudiants n'ont pas les mêmes chances d'avoir 1 ou 0 selon leur niveau.

Calcul de statistiques pour chaque étudiant : quelle probabilité a-t-il d'avoir chacune des notes possibles ?

Mise en place d'un modèle probabiliste sur la base des probabilités calculées pour chaque étudiant.

Résultats et limites :

- Trop dépendant des données cibles
- Pour un élève « moyen », revient au hasard
- Besoin de considérer le contenu de la réponse

Meilleur résultat
0.49 en précision

3.2.2 CLASSIFICATION DES QUESTIONS

IDENTIFICATION DE LA PRÉSENCE DE 3 TYPES DE QUESTION DANS LE CORPUS.

Réponse langue formelle

Question : Quel code permet d'afficher la note de Jacques Blanc ?<code> <?php $notes=array("Jean Bleu"=>14, "Jacques Blanc"=>15); ?> </code>

Réponse : echo$notes["Jacques Blanc"]

Réponse attendue

Question : Qu'affiche le code suivant:<code> <?php function add_presence ($note){ $note++; }; $note=14; add_presence($note); echo $note; ?> </code>

Réponse : 14

Réponse langue naturelle

Question : Qu'est-ce que le World Wide Web ?

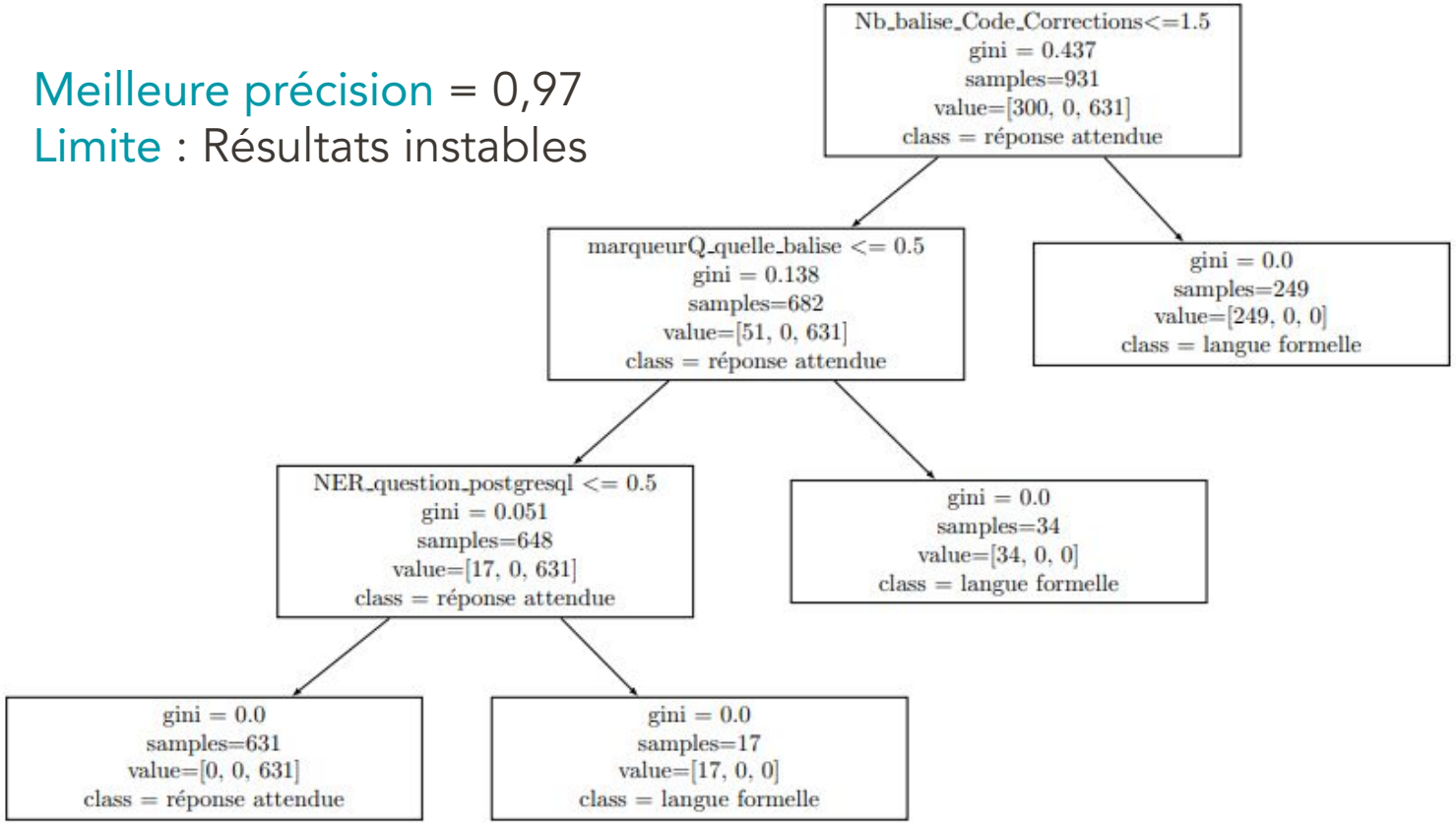
Réponse : Ce sont les pages web accessible par tout navigateur

3.2.2 CLASSIFICATION DES QUESTIONS

ARBRE DE DÉCISION APPRIS

Meilleure précision = 0,97

Limite : Résultats instables



3.2.2 CLASSIFICATION DES QUESTIONS

ARBRE DE DÉCISION MANUEL

Algorithme 1 : Arbre de décision manuel

```
suppression des NO_ANS
if % moyen des % de caractères code dans les réponses > 3:
    classif = "langue formelle"
else:
    if longueur moyenne des réponses < 70 caractères:
        classif = "réponse attendue"
    else:
        "langue naturelle"
```

Résultats

- Données d'entraînement : exactitude de 0,98
- Données de validation : exactitude de 0,9

Avantages

- Maîtrise des paramètres
- Plus de stabilité

→ Nous choisissons de conserver cet arbre pour la suite.

3.2.2 CLASSIFICATION PAR QUESTIONS

RECHERCHE D'HYPER PARAMÈTRES POUR CHACUNE DES 3 APPROCHES ENVISAGÉES (3 471 RUNS)

	Langue formelle	Langue naturelle	Réponse attendue
1		*Featsonly sur tout le dataset* val 75,41 % (SGD)	
2	*Featsonly* val 75,41 % (SGD)	*Featsonly* val 62,27 % (RF)	*Featsonly* val 75,13 % (GOSS)
3	*Tfidf on characters* val 75,04 % (DART)	*Token2Hash* val 57,09 % (SGD)	*Token2hash* val 73,30 % (GBDT)

4 AUTRES APPROCHES ENVISAGÉES

4.1 PRÉDICTION STRUCTURÉE OU AGRÉGÉE

Stratégie adoptée

token2hash, tfidf char

Hyper paramètres

~ pseudo grid search

Algo principal

Structured Perceptron avec Gradient Boosting & Viterbi

Stratégies agrégées

token2hash, featsonly*, tfidf char, questions+answers (question, correction, note)

Algo

Gradient Boosting

Meilleur résultat
0.53 en précision

4.2 EMBEDDINGS ET SIMILARITÉ

OBJECTIF : AJOUTER LES SCORES DE SIMILARITÉ AUX CARACTÉRISTIQUES FOURNIES AUX MODÈLES

Test de différentes mesures

- Calculant la proximité globale entre deux phrases (similarité cosinus) ;
- Repérant les différences plus subtiles (distance de Levenshtein, « character error rate »).

Deux méthodes de calcul

- Similarité entre la réponse proposée (sans le barème) et la réponse de l'élève
- Idem mais en supprimant les déterminants + les mots communs entre la réponse de l'élève et la question s'ils ne sont pas dans la réponse proposée.

Réalisées sur les embeddings issus du modèle multilingue *stsb-xlm-r-multilingual.*

Tendances observées

- Réponses en langue naturelle sous-évaluée
- Réponses en langue formelle surévaluées

→ Ne suffit pas seule mais caractéristique intéressante pour une approche par type de question (seuil adapté à chacun).

4.3 STRATÉGIE POUR CHAQUE TYPE DE QUESTION

Réponses attendues

- Prise en compte du barème : récupération de chaque proposition et de la note associée puis recherche de matching direct (avec une marge d'erreur basée sur la distance de Levenshtein).

Réponses en langue formelle

- Intégration d'outils de vérification de code.

Réponses en langue naturelle

- Identification des mots-clés dans les propositions de réponse, les associer à un lexique de synonymes et rechercher les différentes formes dans la réponse de l'élève.

5 RÉSULTATS

Résultats

- Run 1 (featsonly sur tous) : 0,491
- Run 2 (featsonly par type) : 0,536
- Run 3 (hashingtrick par type) : 0,624

Strategie	Run	Val	Test	QFormelle	QAttendue	QNaturelle
all_featsonly	1	0, 754	0, 491			
qtype_featsonly	2	0, 709	0, 536	**0, 7541**	**0, 7513**	**0, 6227**
qtype_vect	3	0, 687	**0, 624**	0, 7504	0, 7329	0, 5792
all_vect		0, 644				
structured		0, 53				
student_history		0, 49				
regression_mapping		0, 36				

Conclusion

- L'hypothèse de l'influence du type de question est validée
- Les caractéristiques ne suffisent pas / sont à revoir
- Avec les données de test, la performance des méthodes est inversée par rapport aux données de val → pose la question de la généralisation de chacune