Participation d'EDF R&D à DEFT 2022

P. Suignard, X. Huang, M. Bothua

Atelier du 27/06/2022

Plan de la présentation

- EDF R&D
- Pourquoi participer à ce concours ?
- Les tâches et les méthodes utilisées
- Résultats obtenus
- Conclusion

EDF R&D

- Structure

✓EDF R&D au service de toutes les entités du groupe EDF

✓Basée à Saclay, le nouveau centre de recherche

✓Environ 2000 personnes

EDF R&D

- Travaux sur le texte
  - ✓En appui aux différents métiers : EDF Commerce, Hydraulique, Eolien, Nucléaire, Enedis, RH, IT…
  - ✓En permanence : ~8 personnes, 1 doctorant, 2 stagiaires
  - ✓Thèmes : classification, clustering, orthographe, annotations, web sémantique, résumé, anonymisation de données, détection de nouveauté, plongements mots/documents…
  - ✓Sujets : mails, réclamations, comptes-rendus d'interventions, documents techniques, manuscrits anciens, conversations téléphoniques, réseaux sociaux, chatbots…
  - ✓Type de prestations : veille, développement, conseil, méthode, étude.

Pourquoi participer à ce concours ?

- Participation à la tâche de **base** et à la tâche **continue**
- La problématique abordée : calcul de **similarité entre paires de phrases**
- Permet de se comparer/partager/discuter avec les autres équipes
- Emulation interne
- Reconnaissance interne
- Les résultats contribuent directement à EDF Commerce et à d'autres entités du groupe EDF

Tâche de base : Evaluation automatique de copies d'après une référence existante

- Objectif : noter des copies
- Entrainement (données de 2021)
  - ✓On dispose de :
    - Question – q
    - Réponse attendue par l'enseignant – ra
    - Réponse de l'élève – r
    - Note obtenue
- Test
  - ✓On dispose de q, ra et r
  - ✓Il faut trouver la note

Qu'est-ce que le World Wide Web ?

- Système hypertexte fonctionnant sur internet
- Une des applications d'internet, comme courrier électronique, messagerie instantanée...

Ce sont les pages web accessible par tout navigateur => 0,5

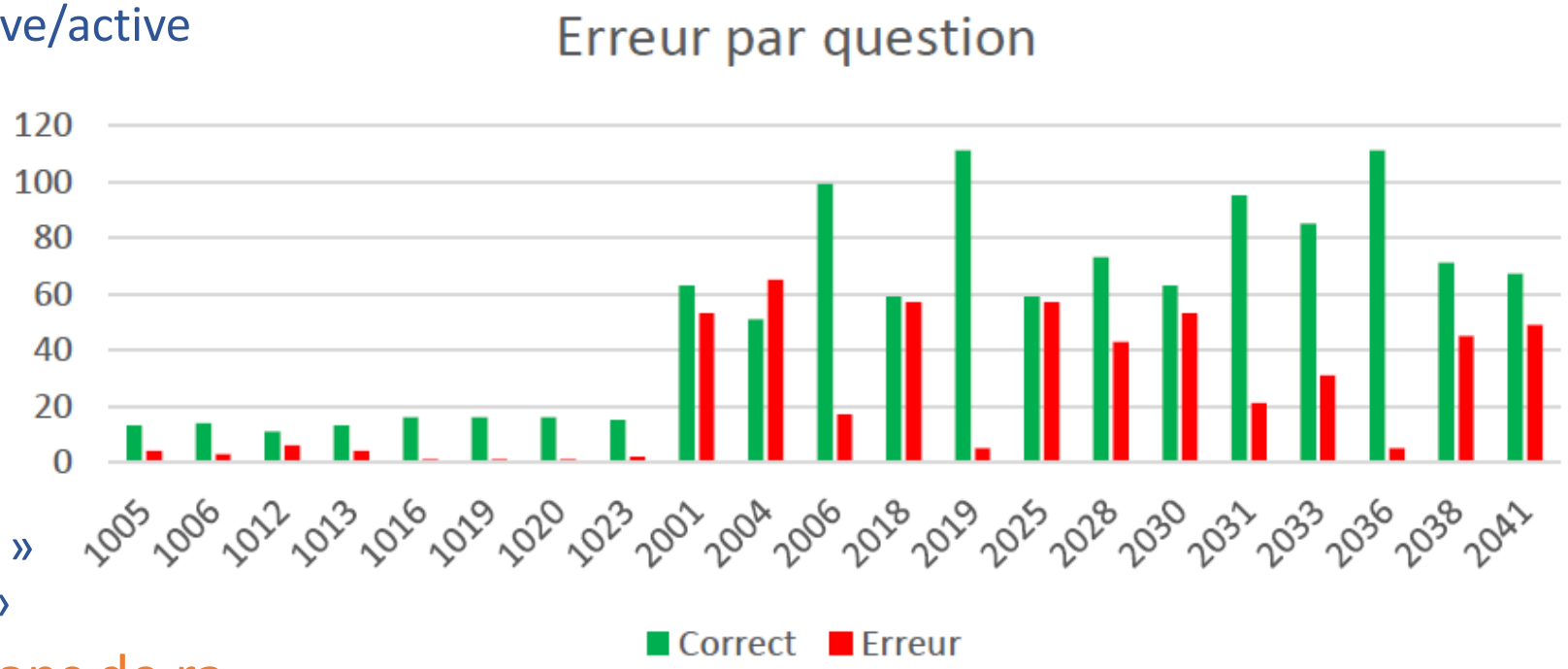
Le réseau mondial, Internet => 0

C'est le systeme hypertexte qui sert à consulter des documents et des pages hébergés sur le réseau internet=> 1

** Les fautes d'orthographes sont d'origine !*

Tâche de base : Evaluation automatique de copies d'après une référence existante

- Analyse des erreurs de notre système de 2021
  - 2004
    - « Mettre » voix passive/active
  - 2018
    - Même réponse 19 fois fausse et 40 fois exacte
  - 2041
    - Ra=« typage de contenu » alors que r=« type de données » « type des attributs »
  - => légères modifications de ra



Tâche de base : Evaluation automatique de copies d'après une référence existante

- Les étapes :
  - ✓Suppression des commentaires de l'enseignant (peuvent *perturber* le classifieur)
    - Si ceci 1, si cela 0,5, si autre 0 => Si ceci 1
    - « Réponse attendue », toutes les réponses sont attendues !
    - « 1 si petite faute sur le texte », comment définir une « petite faute » ?
    - « 1 point pour la définition, 1 point pour l'exemple », qu'est qu'une définition, qu'est-ce qu'un exemple ?
  - ✓Modification
    - changement de « i = mise en forme physique em= mise en forme logique » par « i = mise, met, mettre en forme physique em= mise, met, mettre en forme logique »
    - changement de « Les liens doivent décrire leur destination » par « Les liens doivent décrire leur destination, alt, alternatif »
- Méthode : Extraction de *features* (similarités croisées) + entraînement d'un classifieur

Tâche de base : Evaluation automatique de copies d'après une référence existante

- Les étapes :

✓Prétraitements

- Normalisation des balises "<", ">" en « < » et « > » ;
- Les balises , ,
 en début et fin de texte ont été supprimées ;
- Remplacement des caractères " " par un blanc ;
- Suppression des caractères \n et \t ;
- Insertion d'un caractère blanc avant « < » et après « > » pour faciliter la tokenisation en mots pour les calculs de similarité ;
- Utilisation du caractère blanc pour la séparation des phrases en tokens ;
- Passage en minuscule.

Tâche de base : Evaluation automatique de copies d'après une référence existante

- Les étapes (suites):
  - ✓Extraction de 42 *features* (basées sur Soft Cardinalité, Jaro Wikler, Damereau-levenshtein et Monge Elkan)
  - ✓Soft Cardinalité (utilisé pour SemEval 2013) ou similarité floue

$$|S|' = \sum_{i=1}^{n} \frac{1}{\sum_{j=1}^{n} sim(s_i, s_j)} \qquad sim(t_1, t_2) = \frac{2 * \left| t_1^{[2:3]} \cap t_2^{[2:3]} \right|}{\left| t_1^{[2:3]} \right| + \left| t_2^{[2:3]} \right|}$$

  - ✓Monge Elkan

$$sim_{MongeElkan}(S, T, sim_{mot}) = \frac{1}{|S|} * \sum_{i=1}^{|S|} \max_{j=1 \text{ à } |T|} sim_{mot}(s_i, s_j)$$

$$sim_{ME}(S, T, sim_{mot}) = \sqrt{sim_{MongeElkan}(S, T, sim_{mot}) * sim_{MongeElkan}(T, S, sim_{mot})}$$

Tâche de base : Evaluation automatique de copies d'après une référence existante

- Les étapes (suites):
  - ✓Les 42 *features*
    - q =question
    - a ou r = réponse
    - ra = réponse attendue
  - ✓Entrainement d'un classifieur (Random Forest)
  - ✓Prédiction de 3 classes :
    - 0 si note <0,25
    - 1 si note >0,75
    - 0,5 sinon
  - ✓+1 feature
    - Cos (a, q+ra)
    - sur bigrammes

$\lvert a\rvert'$	$\frac{\lvert a \cap ra\rvert'}{\min(\lvert a\rvert', \lvert ra\rvert')}$	$cosinus_{bi}(a, ra)$
$\lvert q\rvert'$	$\frac{\lvert a \cap ra\rvert'}{\max(\lvert a\rvert', \lvert ra\rvert')}$	$cosinus_{bi}(q, ra)$
$\lvert ra\rvert'$	$\frac{\lvert a \cap ra\rvert' * (\lvert a\rvert' + \lvert ra\rvert')}{2 * \lvert a\rvert' * \lvert ra\rvert'}$	$cosinus_{tri}(a, q)^1$
$\lvert a \cup q\rvert'$	$\lvert a \cup ra\rvert' - \lvert a \cap ra\rvert'$	$cosinus_{tri}(a, ra)$
$\lvert a \cup ra\rvert'$	$cosinus_{mot}(a, q)^2$	$cosinus_{tri}(q, ra)$
$\lvert q \cup ra\rvert'$	$cosinus_{mot}(a, ra)$	$cosinus(a, q \cup ra)$
$\lvert a \cap ra\rvert'$	$cosinus_{mot}(q, ra)$	$cosinus_{bi}(a, q \cup ra)$
$\lvert a \backslash ra\rvert'$	$similarité_{Damereau\_Levenshtein}(a, q)^3$	$cosinus_{tri}(a, q \cup ra)$
$\lvert ra \backslash a\rvert'$	$similarité_{Damereau\_Levenshtein}(a, ra)$	$sim_{ME}(a, q, sim_{DamereauLevenshtein})^4$
$\frac{\lvert a \cap ra\rvert'}{\lvert a\rvert'}$	$similarité_{Damereau\_Levenshtein}(q, ra)$	$sim_{ME}(a, ra, sim_{DamereauLevenshtein})$
$\frac{\lvert a \cap ra\rvert'}{\lvert ra\rvert'}$	$similarité_{JaroWinkler}(a, q)^3$	$sim_{ME}(q, ra, sim_{DamereauLevenshtein})$
$\frac{\lvert a \cap ra\rvert'}{\lvert a \cup ra\rvert'}$	$similarité_{JaroWinkler}(a, ra)$	$sim_{ME}(a, q, sim_{stricte})^5$
$\frac{2 * \lvert a \cap ra\rvert'}{\lvert a\rvert' + \lvert ra\rvert'}$	$similarité_{JaroWinkler}(q, ra)$	$sim_{ME}(a, ra, sim_{stricte})$
$\frac{\lvert a \cap ra\rvert'}{\sqrt{\lvert a\rvert' * \lvert ra\rvert'}}$	$cosinus_{bi}(a, q)^6$	$sim_{ME}(q, ra, sim_{stricte})$

Tâche de base : Evaluation automatique de copies d'après une référence existante

- Run 1 et 2 :
  - ✓3 classes à prédire :
    - 0 si la note était inférieure à 0,25
    - 0,5 si la note était comprise entre 0,25 et 0,75
    - 1 si la note était supérieure à 0,75.
  - ✓Random Forest avec 100 et 200 arbes
- Run 3 :
  - ✓classifieur entrainé à prédire la « vraie » note comprise entre 0 et 1
  - ✓la valeur prédite a ensuite été arrondie au dixième le plus proche
  - ✓100 arbres Random Forest

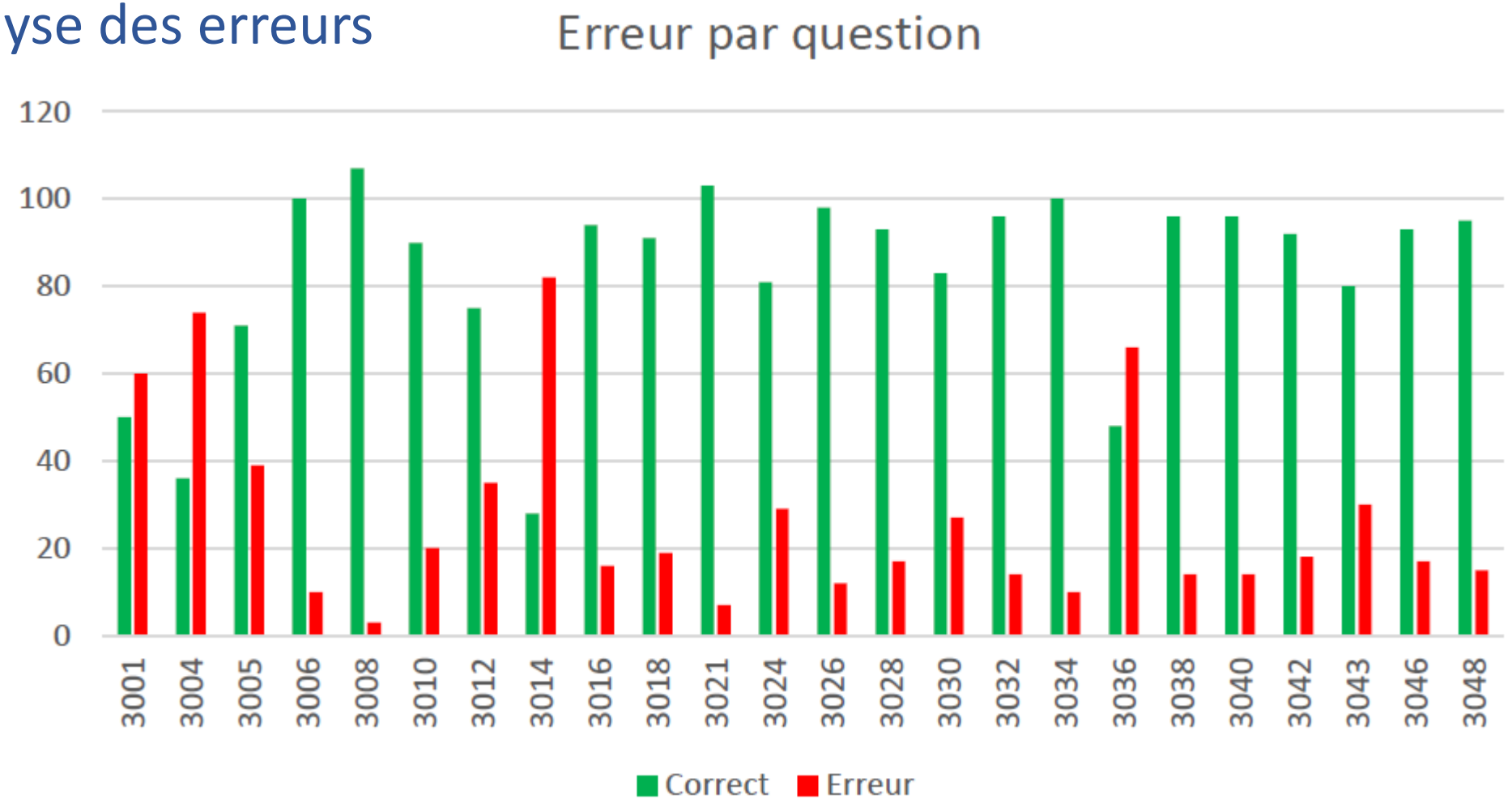
Tâche de base : Evaluation automatique de copies d'après une référence existante

- Résultats

Run	Evaluation
Run 1 :	0,752
Run 2 :	0,756
Run 3 :	0,323
Maximum	0,756
Médiane	0,524
Moyenne	0,542
Minimum	0,323

Tâche de base : Evaluation automatique de copies d'après une référence existante

- Analyse des erreurs



Tâche continue

I. Présentation

- **Objectif**

Concevoir un système de prédiction en intégrant des interactions entre système et serveur (annotateur et oracle)

- **Données à disposition**

Les fichiers « trainT2-Q.tab » et « trainT2-R.tab » à disposition**.**

- **Restructuration des données:**

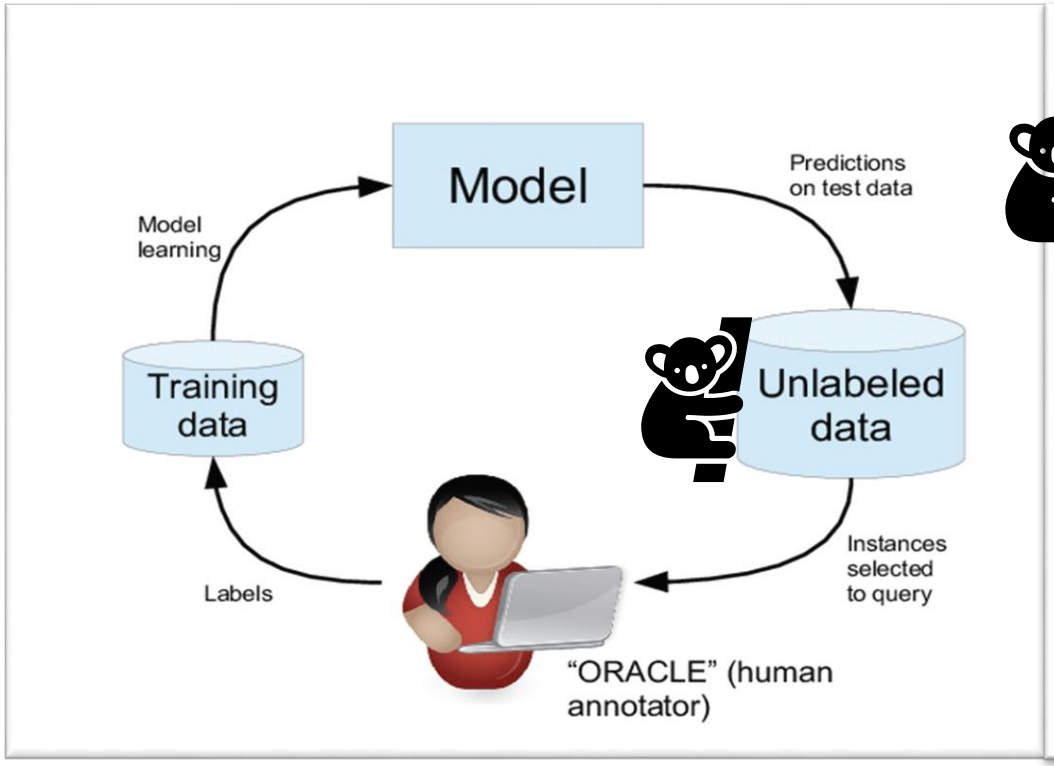
{'1001': [['student101','Ce sont les pages web accessible par tout navigateur.'], ['student108', 'Un réseau mondial'],…}

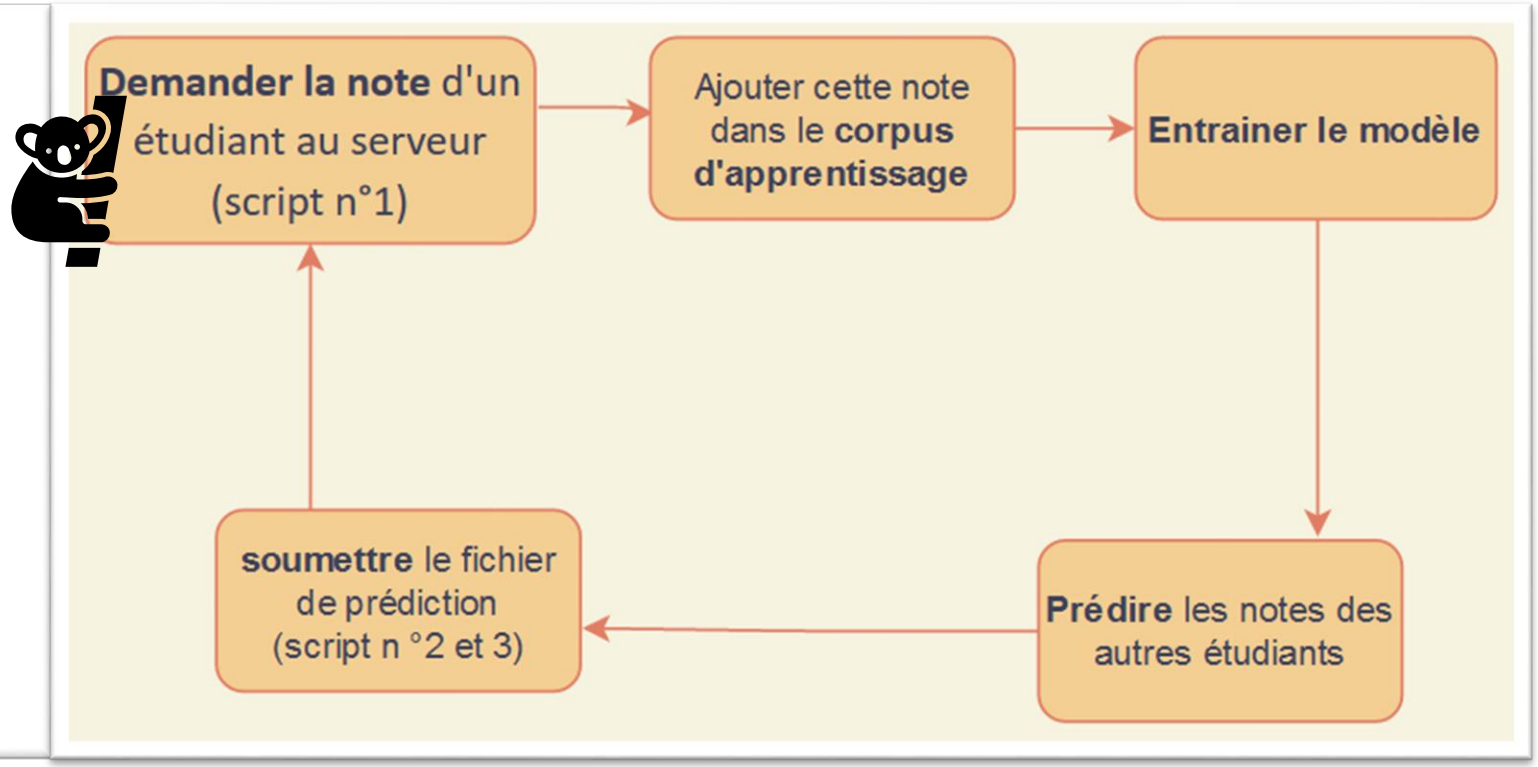
- **Etapes de traîtement**

On fait référence aux principes de l'*active learning*.

Tâche continue

1.3. Etapes de traîtement

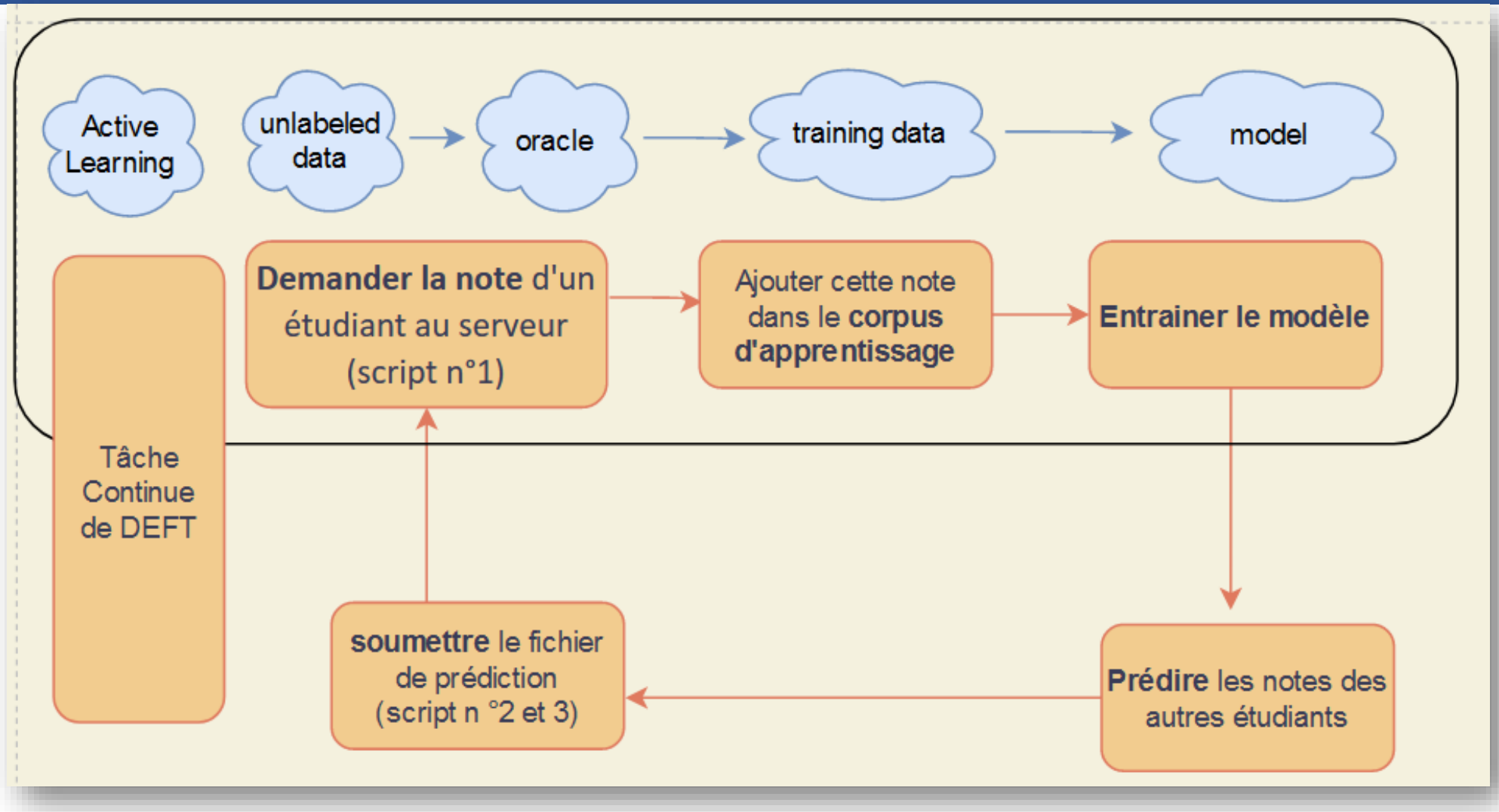




Active Learning

Tâche Continue de DEFT

Tâche continue



Etapes de traîtement de la Tâche Continue en comparaison avec celle de l'Active Learning en général

Tâche continue

II. Méthode

- **Méthode d'apprentissage** : 50% données pour apprentissage, 50% pour test
- **Choix de classifieur** : RamdomForest
- **Feature** : Concaténation d'une réponse avec la question correspondante. Vectorisation avec SentenceTransformer.
- **Stratégie d'interrogation du serveur** : Echantillonnage d'Incertitude

(Demander l'étiquette des données pour lesquelles le modèle actuel a le moins de certitude)

Tâche continue

Question 1001 → Pour chaque étudiant n : → n = 0 générer des résultats avec la même note et les même scores de confiance (1classe) → n = 1, 2.. jusqu'à obtenir 3 classes différentes 😄

```
Entrée [307]:  1  train_pool
```

Out[307]:

	num_ques	num_stud	grade		sim_cos
0	1001	student101	0.5	Ce sont les pag	[tensor(0.8095)]
1	1001	student108	0	Un réseau mondia	[tensor(0.8723)]
2	1001	student116	0	C est le réseau	[tensor(0.8937)]
3	1001	student121	0	Le réseau mo	[tensor(0.8364)]
4	1001	student122	0	C est un reseau p	[tensor(0.7278)]
5	1001	student13	0	Le world Wide	[tensor(0.8725)]
6	1001	student3	1	C est le syste	[tensor(0.7721)]

2002_student68.csv

	QNUM	e_dem	etudiant	note	Confiance
2	2002	student68	student10	0	0.51
3	2002	student68	student100	1	0.57
4	2002	student68	student101	1	0.53
5	2002	student68	student102	0	0.6
6	2002	student68	student103	1	0.49
7	2002	student68	student104	1	0.53
8	2002	student68	student105	1	1
9	2002	student68	student106	0	0.57
10	2002	student68	student107	1	0.59

1001	student101	student58	0.5	1
1001	student101	student62	0.5	1
1001	student101	student7	0.5	1
1001	student101	student74	0.5	1
1001	student101	student95	0.5	1
1001	student101	student96	0.5	1
1001	student101	student98	0.5	1

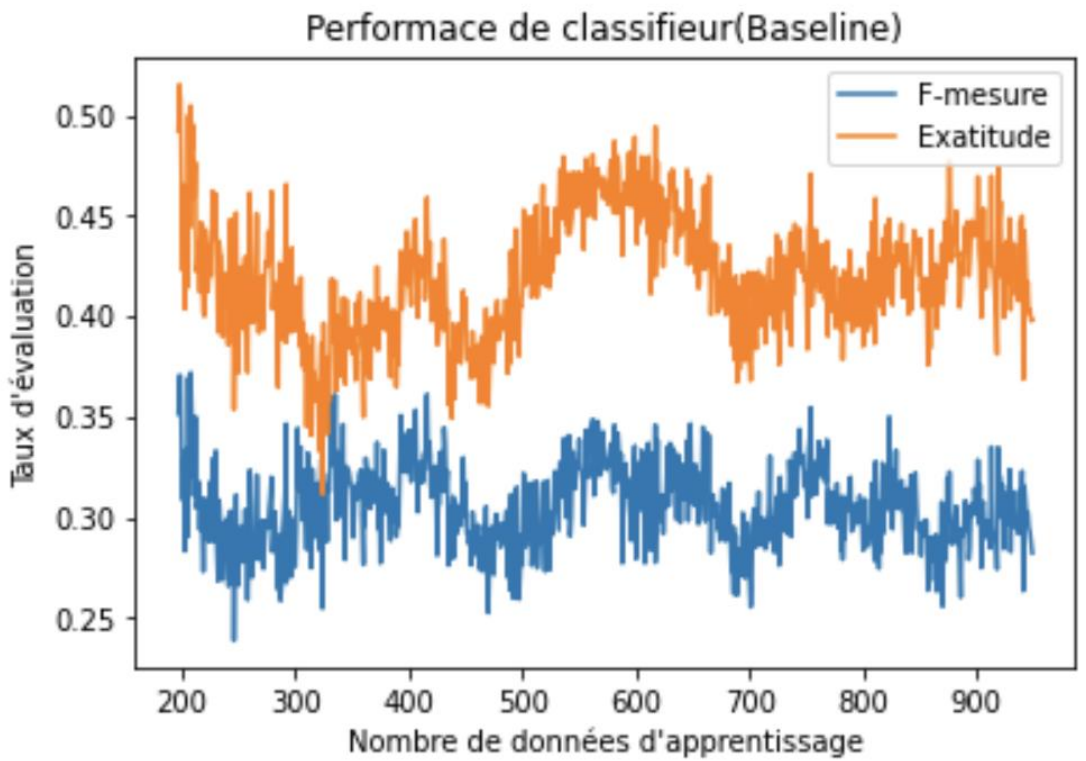
Demander au serveur la vraie note et l'ajouter dans le corpus Train

Tâche continue

IV. Résultats (1)

La performance du modèle pour la phase d'apprentissage

Classifieur : Randamforest
Feature : Embedding: Réponse + Question



	Average	Max	Min
Accuracy	**0.42**	0.51	0.31
F-mesure	0.31	0.37	0.24

Tâche continue

IV. Résultats (1)

La méthode de l'apprentissage active permet de **réduire** le nombre de données d'entrainement et de **sélectionner** des données les plus representatives (utiles).

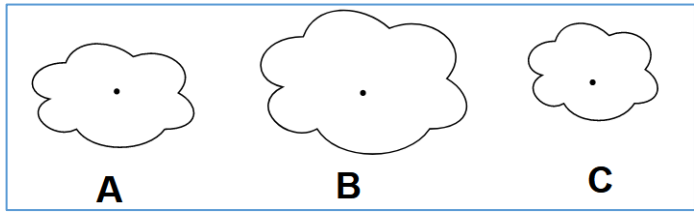
	Nb de données	Remarques
Corpus d'apprentissage	949	24.8% du corpus complet d'apprentissage
Corpus complet d'entraînement (ensemble des fichiers train fournis)	3 820	• Question 1001 – 1028 (20 questions, 17 étudiants): 340 • Question 2002 – 2046 (30 questions, 116 étudiants): 3480

Tâche continue

V. Perspectives

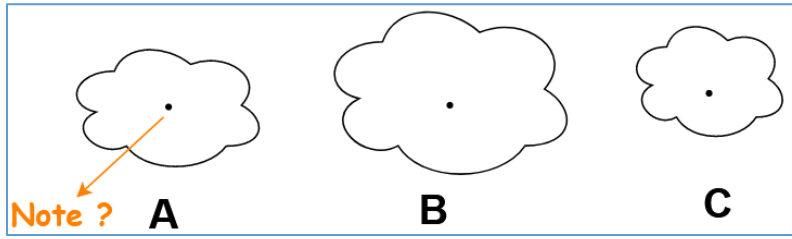
- **Réduire le nombre de fois de requête au serveur**.
- **Améliorer le niveau d'exactitude** :

1) On peut d'abord obtenir **trois clusters**.



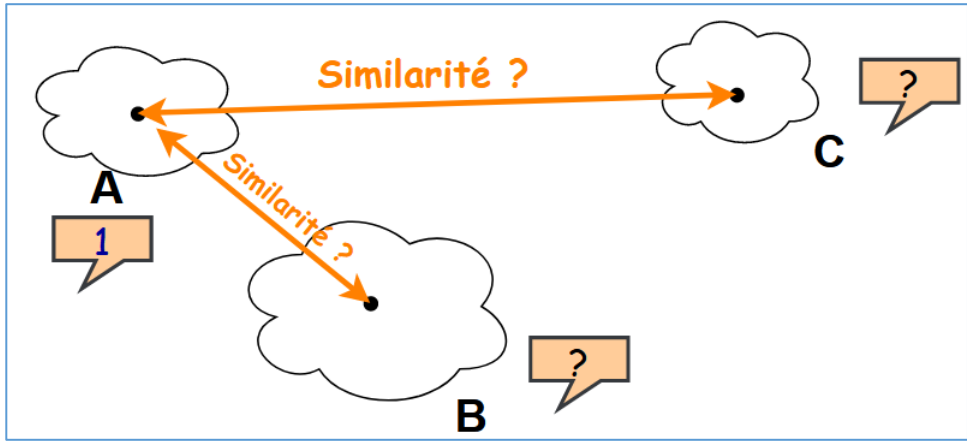
2) **Trouver** la centroïde d'un cluster et **demander** la note de ce centroïde.

3) **Calculer** la similarité entre les centroïdes :



Les deux centroïdes les plus lointains auront les étiquettes 0 et 1.

4) Une fois que **les étiquettes des trois clusters** sont déterminées, toutes les réponses dans les clusters seront notées.



Conclusion

- Participer à la campagne DEFT 2022, nous a permis de tester plusieurs méthodes de **calcul de similarité** et de **l'Active Learning**.
- Les résultats obtenus sont très satisfaisants, classés **premiers** sur les 2 tâches
- Les méthodes que nous avons mises en œuvre sont facilement **transposables** à d'autres tâches et peuvent intéresser plusieurs entités du groupe EDF