Examen blanc 2

/20

Introduction aux humanités numériques – L1HN001

Nom et prénom d'étudiant·e :

Numéro d'étudiant·e :

Date d'examen : 7 décembre 2023, 08h-09h

Question 1 **(4 points)**

Dans la phrase suivante, soulignez toutes les entités nommées :

Emmanuel Macron, né le 21 décembre 1977 à Amiens, est président de la France.

Question 2 **(1 point)**

Parmi les options ci-dessous, choisissez celle(s) qui correspond(ent) à (aux) expression(s) régulière(s) valide(s) :

- ❑ `^jour`
- ❑ `[[^jour`
- ❑ `[a-zA-Z0-9_])`
- ❑ `[a-zA-Z0-9_]`

Question 3 **(1 point)**

Comment s'appelle l'étape initiale d'une chaîne de traitement d'OCR ?

- ❑ exportation
- ❑ prédiction
- ❑ transcription
- ❑ segmentation

Question 4 **(1 point)**

Remplissez le blanc avec le mot approprié :

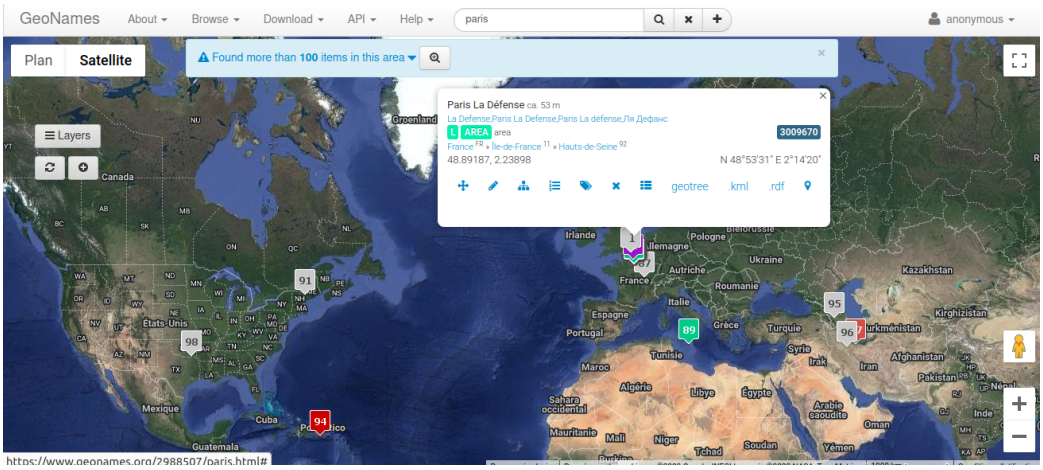
Transkribus est une plateforme de la _____ automatique des textes.

Question 5 **(2 points)**

L'illustration suivante représente un exemple de :

(plusieurs réponses possibles)

- ❏ l'annotation des entités nommées suivant le format BIO
- ❏ la cartographie des entités nommées
- ❏ la campagne d'évaluation de la reconnaissance d'entités nommées
- ❏ la désambiguïsation des entités nommées



Question 6 **(1 point)**

Quel est l'intérêt d'utilisation des expressions régulières ?

Question 7 **(2 points)**

Répondez par VRAI (V) ou FAUX (F) aux phrases suivantes :

		V	F
(a)	L'approche à base de règles pour effectuer certaines tâches de TAL nécessite un entraînement d'un modèle.	☐	☐
(b)	Généralement, les logiciels d'OCR / HTR permettent d'exporter des transcriptions dans différents formats.	☐	☐

Question 8 **(2 points)**

Étant donné la phrase *Je suis en train de réviser pour mon $2^{ème}$ partiel*,

formulez les regex qui permettent de capturer les mots suivants :

1. *2ème* _____
2. *Je* _____

Pour tester vos regex, vous pouvez utiliser un site dédié comme, p. ex. `https://regex101.com/`.

Attention : les réponses qui utilisent uniquement les caractères littéraux *2ème* et *Je* ne sont pas acceptées.

Question 9 **(3 points)**

1. Chargez le corpus *Le tour du monde en quatre-vingts jours* (utilisé lors de la séance du 23 novembre 2023, « Reconnaissance d'entités nommées ») dans Voyant Tools depuis votre ordinateur ;
2. Combien de *tokens* et de *types* ce corpus contient-il ?

   Tokens : _____

   Types : _____
3. Quelle est la fréquence absolue du mot `fogg` ? _____

Question 10 **(3 points)**

1. Dirigez-vous vers la page de l'outil Tanagra.

   Sélectionnez `French_lg` comme modèle de reconnaissance d'entités nommées et chargez le même corpus que dans la question 9.
2. Parmi les entités nommées récupérées par Tanagra, en trouvez-vous une localisée en Espagne ? Si oui, laquelle ?