

Cours Web, bases de données, localisation

CAHIER DES CHARGES

Projet : Création du site

Reconnaissance et recherche d'entités nommées










réalisé par Yuliya Masalskaya

Enseignante : Sarra El Ayari

INALCO, 2023

1) Description du contenu du dossier ren-site (site du projet)

Les fichiers créés lors du projet :

- 
index.html
1) Page d'accueil (page principale)
- 
registration.php
2) Création d'un compte utilisateur
- 
connexion.php
3) Connection au compte utilisateur
- 
logout.php
4) Déconnexion du compte utilisateur
- 
ren.php
5) Page 2 (Onglet Reconnaissance d'entités nommées)
- 
resultat.php
6) Page 3 (Affichage du résultat de REN)
- 
ren_3lang.py
7) Programme Python de REN (utilisant Spacy)
- 
script.js
8) Fonctions JQuery
- 
style.css
9) Fonctions CSS



ren.sql

10) Base de données

Le dossier contenant des logos et images utilisés sur le site: **images**.

2) Choix du projet

Le projet consiste en création d'un site qui permettra à un utilisateur de reconnaître des entités nommées (EN) dans un document qu'il devra insérer dans les formulaires affichés sur le site (zone d'insertion de texte ou sélection d'un fichier).

Le programme de reconnaissance d'entités nommées sera capable de détecter des EN en 3 langues : anglais, français et ukrainien.

La librairie utilisée pour reconnaître des entités nommées est Spacy (nous utilisons des modèles pré-entraînés spacy conçus pour cette tâche).

Pour profiter pleinement des fonctionnalités du site (en l'occurrence reconnaissance d'entités nommées (REN)), chaque utilisateur devra créer un compte utilisateur s'il n'en avez pas encore, ou de se connecter à son compte existant.

3) Structure de la base de données (BDD) utilisée

Les colonnes de la base de données relationnelle **ren** correspondent aux champs du formulaire de création d'un nouveau compte que chaque utilisateur devra remplir lors de la première visite du site s'il souhaite pouvoir profiter des fonctionnalités du site.

Table users : correspondant à l'identité de l'utilisateur

Serveur : MySQL 3306 » Base de données : ren » Table : users "table containing information about users"										
Parcourir Structure SQL Rechercher Insérer Exporter Importer Privileges Opérations Déclencheurs										
#	Nom	Type	Interclassement	Attributs	Null	Valeur par défaut	Commentaires	Extra	Action	
<input type="checkbox"/> 1	id	int			Non	Aucun(e)		AUTO_INCREMENT	Modifier	Supprimer Plus
<input type="checkbox"/> 2	firstname	varchar(50)	utf8mb4_0900_ai_ci		Non	Aucun(e)			Modifier	Supprimer Plus
<input type="checkbox"/> 3	lastname	varchar(50)	utf8mb4_0900_ai_ci		Non	Aucun(e)			Modifier	Supprimer Plus
<input type="checkbox"/> 4	email	varchar(80)	utf8mb4_0900_ai_ci		Non	Aucun(e)			Modifier	Supprimer Plus
<input type="checkbox"/> 5	profession	varchar(250)	utf8mb4_0900_ai_ci		Non	Aucun(e)			Modifier	Supprimer Plus
<input type="checkbox"/> 6	password	varchar(20)	utf8mb4_0900_ai_ci		Non	Aucun(e)			Modifier	Supprimer Plus

Colonnes de la table users de la bdd ren : Champs du formulaire affiché sur le site :

1. id
2. firstname

3. lastname
4. email
5. profession
6. password

[Connexion](#) / [Inscription](#)

Votre prénom
Votre nom
Votre adresse mail
Votre profession
Votre mot de passe
<input type="button" value="S'inscrire"/>

Table texts : correspondant au **texte** où des EN doivent être recherchées

#	Nom	Type	Interclassement	Attributs	Null	Valeur par défaut	Commentaires	Extra	Action
<input type="checkbox"/>	1	number	int		Non	Aucun(e)		AUTO_INCREMENT	Modifier Supprimer Plus
<input type="checkbox"/>	2	domaine	varchar(50)	utf8mb4_0900_ai_ci	Non	Aucun(e)			Modifier Supprimer Plus
<input type="checkbox"/>	3	theme	varchar(100)	utf8mb4_0900_ai_ci	Non	Aucun(e)			Modifier Supprimer Plus
<input type="checkbox"/>	4	language	text	utf8mb4_0900_ai_ci	Non	Aucun(e)			Modifier Supprimer Plus
<input type="checkbox"/>	5	iduser	int		Non	Aucun(e)			Modifier Supprimer Plus

Table entities : correspondant aux entités retrouvées dans chaque texte

#	Nom	Type	Interclassement	Attributs	Null	Valeur par défaut	Commentaires	Extra	Action
<input type="checkbox"/>	1	number	int		Non	Aucun(e)		AUTO_INCREMENT	Modifier Supprimer Plus
<input type="checkbox"/>	2	numbertext	int		Non	Aucun(e)			Modifier Supprimer Plus
<input type="checkbox"/>	3	entity	varchar(250)	utf8mb4_0900_ai_ci	Non	Aucun(e)			Modifier Supprimer Plus
<input type="checkbox"/>	4	label	varchar(30)	utf8mb4_0900_ai_ci	Non	Aucun(e)			Modifier Supprimer Plus

4) Ergonomie

a) Contenu des pages et placement de ses éléments

Le site est composé de 3 pages :

- 1) **Page principale** (onglet **Accueil**) correspondant au fichier **index.html** est un page de description des fonctionnalités du site et de leur utilisation) et création d'un compte utilisateur / connexion au compte utilisateur existant

Sur la page principale

En haut à gauche, il y a le menu du site comportant 3 onglets : Accueil, Reconnaissance d'entités nommées et Recherche d'entités nommées (*partie du projet non-réalisée*) qui se répète sur les 3 pages du site.

En haut à droite, il y a les formulaires de connexion au compte d'utilisateur et de création de compte utilisateur. Au lancement du site, ces formulaires sont cachés (*grâce aux fonctions JQuery, librairie de JavaScript \$('#register-form').hide(); \$('#login-form').hide();*).

Le corps principal du site se divise en 1 section qui est déclinée en des sous-sections :

- I) Section comportant les formulaires à remplir ;
- II) Section « Description des fonctionnalités et leur utilisation »

2) **Page 2** (onglet **Reconnaissance d'entités nommées**) correspondant au fichier **ren.php**. Cette page est déclinée en 2 sous-sections :

- zone comportant le prénom de l'utilisateur qui s'est connecté (avec un bouton de déconnection)
- zone avec 2 champs correspondant à l'insertion du texte où l'utilisateur souhaite rechercher des entités nommées : un champ via fichier et l'autre via champ d'insertion (*textarea*).

Sous les champs d'insertion de texte, il y a un bouton qui lance le processus de reconnaissance d'entités nommées (*réalisé avec la librairie spacy dans un programme Python*).

3) **Page 3** correspondant au fichier **resultat.php** est une page d'affichage du résultat de REN

Dans le corps de la page, sous forme d'un tableau HTML, s'affiche le résultat du REN : les EN retrouvés dans le texte et leur label.

b) Couleur, police, taille des éléments du contenu

- 1) Les couleurs du site sont **verte, jaune et noire**. Les fond du site est en blanc.
- 2) La police du texte est principalement "**Roboto**", sans-serif en noire correspondant à un site académique.
- 3) Dans la zone du bandeau du site, nous avons utilisé une image de fond, empruntée sur le site [Named Entity Extraction with Python - NLP FOR HACKERS](#)
- 4) Le logo utilisé sur notre site a été conçu et créé à l'aide de Canva par nos efforts. Il est possible de l'agrandir lors du survol de la souris.

5) Les messages affichés lors de la connexion et la création du compte sont en rouge pour attirer l'attention de l'utilisateur.

5) Fonctionnalités du site relevant du côté serveur (assuré par PHP)

- 1) **Insertion de données d'utilisateur** dans la BDD se fait à l'aide du formulaire html sur la page d'accueil et le code **php** dans le fichier **registration.php**.
- 2) **Connexion au compte utilisateur** : pour réaliser cette fonctionnalité, nous utilisons un formulaire html sur la page d'accueil et le code **php** dans le fichier **connexion.php**.

Dans les cas de ces fonctionnalités, nous nous connectons à la base de données **ren** grâce à la classe PDO :

```
// Connexion à la base de données
$conn = new PDO('mysql:host=' . $server . ';dbname=' . $db, $login, $password);
$conn->setAttribute(PDO::ATTR_ERRMODE, PDO::ERRMODE_EXCEPTION);
```

Nous utilisons également la méthode *bindValue()* pour lier des valeurs récupérées dans les champs du formulaires (en l'occurrence email et mot de pass) aux colonnes de la base de données ce qui permet de nous assurer si un utilisateur avec les données insérées existent déjà. C'est très utile pour contrôler l'insertion de données dans la bdd et pour la sécurité de connexion (contrôle d'erreur dans les données insérées) :

```
// Requête SQL pour vérifier si l'email et le mot de passe existent dans la base de données
$sql = "SELECT * FROM users WHERE email = ? AND password = ?";
$stmt = $conn->prepare($sql);
$stmt->bindValue(1, $email);
$stmt->bindValue(2, $passworduser);
$stmt->execute();
```

- 3) **Lancement du programme Python** de reconnaissance d'entités nommées : cette fonctionnalité réalisée grâce à la fonction php *shell_exec()* permet d'exécuter une commande dans le shell du système d'exploitation et de récupérer la sortie résultante.

```
// Appel au script Python pour l'analyse des entités nommées
$command = "python ren_3lang.py $langue \"$texte\"";
$resultat = shell_exec($command);

// Convertissement du résultat JSON (sortie du code Python) en tableau PHP
$resultat_array = json_decode($resultat, true);
```

Elle prend en paramètre une chaîne de caractères contenant la commande à exécuter et retourne la sortie de cette commande en tant que chaîne de caractères pour la stocker dans une variable et traiter ultérieurement.

Source : [Exécuter Python en HTML | Delft Stack](#)

- 4) **Interrogation de la BDD** après le traitement du texte inséré pour récupérer les EN retrouvés précédemment (si aucune, affichage 0) pour l'affichage à l'utilisateur sous forme d'un tableau HTML (*non-réalisée*). Cette fonctionnalité correspondrait à l'onglet du site Recherche d'entités nommées mentionné ci-haut.

6) JQuery

La partie prévue initialement : « Sur la 2^{ème} page, il y a une possibilité de lancer une requête JS : si l'utilisateur ne souhaite pas lancer le processus de REN sur un nouveau texte (pas de nouveau formulaire envoyé), il peut interroger la BDD pour afficher les EN et leur ~~nombre~~ label dans un texte traité auparavant stockées dans la BDD. Le résultat de la requête est à afficher sur la 3^{ème} page du site sous forme d'un tableau HTML ». n'a pas été réalisée.

En revanche, nous utilisons des fonctions JQuery, telles que hide(), show() et prevetDefault() pour afficher ou cacher les formulaires affichés sur la page d'accueil :

```
// Permet de gérer le clic sur le lien "Inscription"
$('.link-register').click(function(e) {
    e.preventDefault();
    $('#login-form').hide();
    $('#register-form').show();
});
```

Cela permet de cacher les formulaires Connexion et Inscription initialement lors du lancement du site (pour raison esthétique). Puis si l'utilisateur click sur le lien Connexion le formulaire respectif s'affiche et le formulaire Inscription est caché. Le même se passe lors du click sur le lien Inscription.

[Connexion](#) / [Inscription](#)

Votre email
Votre mot de passe
<input type="button" value="Se connecter"/>

[Connexion](#) / [Inscription](#)

Votre prénom
Votre nom
Votre adresse mail
Votre profession
Votre mot de passe
<input type="button" value="S'inscrire"/>

7) HTML

Structure des pages HTML :

Header

Metadonnés

Librairies et fichiers de style

Body

Bandeau avec le logo REN (créé par nous-même avec Canva) et le titre du programme Reconnaissance et Recherche d'entités nommées.



Accueil

Reconnaissance d'entités
nommées

Recherche d'entités
nommées

Menu

Partie principale (main) **différente pour chaque page** comprenant le contenu des pages du site :

[Déconnexion](#)

Bienvenue, Luca !

Choisissez une des 3 langues proposées : Anglais ▼

Insérez le texte à analyser :

Choisissez un fichier à analyser :

Choose File No file chosen

Analyser

Footer avec crédits, nom du site en version sigle et lien vers la page d'accueil

8) CSS / Bootstrap

Des fonctions CSS et Bootstrap permettant de réaliser l'ergonomie du site sont assez nombreuses. Nous avons plus privilégié CSS à Bootstrap (question d'habitude uniquement mais en espérant la changer bientôt).

La partie majeure du style est gérée par des fonctions CSS dont celles que nous avons utilisées pour la première fois : "letter-spacing", "box-shadow", "grid-area", "overflow", "list-style-type". En fait, lors de la création du site, nous avons utilisé le livre « Premiers pas en CSS3 et HTML5 » de Francis Draillard d'où ces fonctions ont été empruntées.

Parmi les fonctions Bootstrap utilisées "text-center", "mb-(valeurs différentes)" et "container". Nous avons également utilisé la classe "btn btn-success" pour gérer la couleur du bouton de lancement de REN.

9) Programme Python :

Pour pouvoir effectuer le processus de reconnaissance d'entités nommées dans le texte inséré par l'utilisateur, nous avons créé un programme Python (fichier **ren_3lang.py**) se basant sur des modèles pré-entraînés Spacy en 3 langues (langues possibles pour REN).

```
# Chargement des modèles spacy pour les langues FR; EN; UK (ukrainien)
nlp_fr = spacy.load("fr_core_news_sm")
nlp_en = spacy.load("en_core_web_sm")
nlp_uk = spacy.load("uk_core_news_sm")
```

Ce programme est lancé grâce à la méthode shell_exec() de php (cf point 5 ci-haut).

La sortie du programme est un tableau json

```
{"entites": [{"entite": "Paris", "label": "LOC"}]}
```

que nous parcourons dans une boucle grâce au code php :

```
// Parcours des entités et leur insertion dans la base de données
foreach ($resultat_array['entites'] as $entite) {
    $entity = $entite['entite'];
    $label = $entite['label'];
```

pour récupérer chaque entité et son label et la sauvegarder dans la bdd (table entites) et également pour l'afficher dans un tableau html sur la page de résultat de REN :

Bienvenue, Luca !

Résultat de l'analyse :

Entité	Label
Paris	LOC
la France	LOC
XIIe	MISC
internationales134	LOC
Ville Lumière	LOC
Seine5	MISC
Parisiens5	MISC

10) Conclusion

Expérience

Le projet que j'ai réalisé a été une occasion précieuse pour mettre en pratique et renforcer mes compétences en développement web, notamment en HTML et CSS. J'ai eu l'opportunité de travailler sur la partie frontend du site, en améliorant son interface utilisateur, ainsi que sur la partie backend, en gérant le serveur et la base de données. J'ai été confronté à plusieurs défis auxquels j'ai su trouver des solutions de manière autonome, ce qui m'a permis d'acquérir une plus grande confiance en mes capacités.

De plus, cette expérience m'a permis d'approfondir considérablement mes connaissances en PHP : même si son implication dans le projet se limitait à une petite partie, cela a été une avancée significative pour moi, car j'ai pu explorer ce nouveau langage et mieux comprendre ses fondamentaux.

Ouverture

- 1) Initialement prévue comme partie du projet, la page permettant d'interroger la base de données pour d'y récupérer des entités nommées



rech_en_bdd.html

déjà existantes n'a pas été terminée (fichier [rech_en_bdd.html](#)). Sa finalisation pourrait donc être la suite de ce projet.

En fait, l'existence de telle page permettra aux utilisateurs de récupérer des entités nommées par domaine ou par thème (dans notre bdd, ce sont des attributs de la table **texts**) ce qui est utile lors de construction de bases connaissance ou de glossaires thématiques d'entités nommées.

- 2) Nous n'avons pas non plus proposé aux utilisateurs de possibilité de sauvegarder le résultat de reconnaissance et de recherche d'entités nommées sous un fichier ce qui serait fait très pratique.

11) Références :

- 1) Draillard Francis, Premiers pas en css3 et html5 (9e édition). Ed. Eyrolles, 2022. 704 p.