

JCS réalisera pour le compte du Client

I. Introduction

Les utilisateurs du site évoqués ci-après sont précisés ici :

- Les **utilisateurs** : tous les visiteurs externes du site qu'ils soient déjà client de SafeFlat ou non ;
- Les **clients du site** : tout utilisateur du site ayant déjà un abonnement et étant connecté sur le site ;
- Les **administrateurs** : ceux-ci ont un droit d'accès et de modification des données ainsi que la charge de la première validation des détections.

II. Qualités générales

Les qualités du site sont exprimées suivant les normes ISO/IEC 25010:2011 qui définissent les qualités d'un logiciel.

Critères essentiels :

- **Utilisabilité** : La plateforme doit être **intuitive**, facile à prendre en main et **accessible à tous les utilisateurs**. Le cheminement devra être clair et indiqué ;
- **Sécurité** : Le site doit permettre une navigation et un **traitement des données sécurisés**. L'hébergement des données doit être conforme au **RGPD** ;
- **Responsive** : la plateforme doit pouvoir fonctionner sur **différents types de navigateurs** (Chrome, Firefox, Safari, ...) ainsi que sur **PC** et **mobile** ;
- **Performance et efficacité** : la plateforme devra présenter un temps de réponse correct pour un utilisateur web, **inférieur à 2 secondes** ;
- **Maintenabilité** : le site étant voué à évoluer dans le temps et les technologies de récupérations de données à s'adapter aux sites concernés, le site doit être **facilement adaptable** et être compris par un développeur tiers. Pour ce faire, l'architecture des codes doit suivre les **bonnes pratiques** du développement.

Le site sera disponible en langue **française** uniquement.

III. Récupération des données

L'application web devra **consulter les données** d'une liste de sites leaders de la location de biens immobiliers. Les données seront consultées sur la **liste de sites suivante** et **uniquement** sur celle-ci:

- leboncoin.fr ;

- airbnb.fr ;
- abritel.fr ;
- seloger.com ;
- pap.fr ;
- gensdeconfiance.com .

Les données qui seront scrapées sur les sites en question seront les **descriptions textuelles** des biens.

Compte tenu des **sécurités** mises en place contre le scraping, JCS implémentera les techniques de scraping usuelles mais si le scraping de certains sites donne lieu à plus de **3 bannissements** d'adresses IP dû à des protections mises en place par les plateformes, JCS se réserve le droit de **ne pas réaliser** la récupération des données des sites en question.

Les sécurités élémentaires devront être **contournées**, potentiellement par l'utilisation de **proxys**.

De plus, JCS **décline toute responsabilité** quant à l'utilisation commerciale des scripts résultant de ces consultations de données.

L'outil de scraping sera développé en langage **Python** et traitera chaque jour les nouvelles annonces. Les bibliothèques Python **BeautifulSoup**, **Selenium** et **Request** pourront être utilisées ainsi que les potentielles **API** des sites scrapés.

À **chaque site** correspondra un script Python et chacun seront accompagnés d'une **documentation** sur le fonctionnement du script qui servira aussi de guide d'utilisation. Cette documentation n'excèdera pas **une page par script** et sera remise au format PDF.

IV. Interface web

Le front-end sera codé en **HTML/CSS**, **JavaScript** et **Python**.

JCS se chargera de déterminer, en collaboration avec SafeFlat, les **parcours utilisateurs** et la création des **wireframes** ainsi que la **charte graphique**. Le site devra être **épuré** et **moderne**. JCS autorisera un maximum de **3 itérations** pour le maquetage.

Le site comportera au maximum une **dizaine de pages** différentes, réparties entre **interface utilisateur** et **administrateur**.

a) Interface utilisateur

Du côté **utilisateur**, figureront :

- Une page d'**accueil** : cette page devra présenter le site avec des **textes fournis par le client** ;
- Une page d'**authentification** : cette page devra permettre aux clients de se connecter. Ceci à l'aide de leur **adresse mail** et **mot de passe** ou de leur compte **Google** ou **Facebook**. Les clients devront aussi pouvoir **récupérer leur compte** s'ils ont perdu leur mot de passe ;
- Une page d'**inscription** : cette page devra permettre aux utilisateurs non-client de le

devenir. Pour ce faire, ils devront renseigner leurs **informations personnelles** (adresse mail et mot de passe) qui seront stockées de manière **sécurisée**. Un **mail de confirmation** de l'adresse mail sera alors automatiquement envoyé à l'utilisateur. L'inscription pourra aussi se faire avec Google ou Facebook via le protocole **OAuth** ;

- Une page d'**espace client** : cette page permettra au client la **gestion de toutes ses propriétés sécurisées** par le site et de ses abonnements. La gestion se divisera en plusieurs actions : validation ou infirmation des **alertes reçues, modifications** du type d'abonnement et modifications des informations sur la propriété. Le client du site pourra également **ajouter des propriétés** à surveiller en renseignant les données nécessaires (photo du bien, description détaillée, etc.) à travers cette page. L'ajout d'une propriété fera l'objet d'un **nouvel abonnement** et donc redirigera sur la page de paiement ;
- Une page de **paiement** : cette page devra permettre aux clients de **finaliser leur abonnement** au site. Pour ce faire, le service **Stripe** sera utilisé aux **frais de SafeFlat** ;
- Une page **CGU** : cette page contiendra les **conditions générales** d'utilisation du site, dont le texte sera **fourni par SafeFlat**.

b) Interface administrateur

Du côté **administrateur**, figureront :

- Une page de **connexion** des administrateurs : cette page devra garantir un **accès sécurisé** aux pages administrateur ;
- Une page de **validation des biens** : cette page devra permettre aux administrateurs de **visualiser les données** des biens inscrits pour **valider** le respect des CGU, notamment la validation de la possession effective du bien renseigné ;
- Une page de **données clients** : cette page devra permettre aux administrateurs de **visualiser les données** des clients : nom, prénom, adresse mail, les biens enregistrées et leurs caractéristiques ainsi que le type d'abonnement pour chaque bien ;
- Une page de **traitement des alertes** : cette page permettra aux administrateurs de **valider ou infirmer** les alertes remontées par l'algorithme avant d'envoyer une **notification aux propriétaires** concernés. Cette notification sera envoyée **automatiquement** par **mail** lors de la confirmation de l'administrateur. Le template du mail sera **rédigé par JCS** et confirmé au travers d'au plus **3 itérations** avec le client ;
- Une page de **gestion des scrapers** : cette page devra permettre aux administrateurs de gérer **en temps réel** l'utilisation des **scrapers** de données sur **chaque site** et d'accéder à un **dashboard** simple de données liées aux scrapers.

V. Algorithme de détection

Le site devra pouvoir **comparer** les données relevées sur les sites web cités plus tôt avec les données des biens protégés pour **signaler** d'éventuelles sous-locations.

Cette détection sera réalisée en plusieurs temps:

- Dans un premier temps, **un premier modèle** d'algorithme n'utilisant **pas de Machine Learning** (ML) sera développé pour donner un ordre d'idée des performances atteignables ;
- Un **deuxième modèle** utilisera de l'**Intelligence Artificielle** (IA) et notamment du **Natural Language Processing** (NLP) sera développé. Cette version fera alors l'objet d'un **livrable intermédiaire** ;
- Un **dernier modèle** sera développé en **affinant le modèle précédent**. Ce modèle devra être **connecté aux bases de données** et analyser **automatiquement** toutes les propriétés protégées à un intervalle **à définir** lors d'échanges avec SafeFlat.

Dans un processus d'**amélioration continue**, une **documentation détaillée** du modèle utilisé et des approches essayées sera fournie au **format PDF** et fera entre 2 et 5 pages.

VI. Avenant à la convention d'étude

Si lors du développement du site, JCS rencontre un **problème non lié à son développement** mais issu d'un problème provenant d'une **source extérieure**, un **avenant** à la convention d'étude en délai, budget et méthodologie pourra être fait.

VII. Déploiement du site

Le back-end sera codé en **Django**. Le site sera hébergé sur **EC2**, service de Amazon aux frais du client.

L'API et le front devront être **sécurisés** en passant par le protocole **https** et les bonnes pratiques sur les compartiments et les droits d'accès seront mis en place.

MÉTHODOLOGIE

I. Maquettage

Cette phase consistera en l'élaboration de la charte graphique et de l'identité visuelle du site ainsi que des parcours utilisateurs.

Sous-phases:

a. Définition des parcours utilisateurs (Offert, semaine 1)

L'intervenant définira l'ensemble des situations d'utilisations et des parcours de l'utilisateur sur le site.

b. Création des wireframes (1 JEH à 450 €, semaine 1 à 2)

L'intervenant créera un schéma fonctionnel de l'application au moyen de croquis/dessins et en croisant les différents parcours utilisateurs, précédemment identifiés.

c. Création des designs (Offert, semaine 2 à 3)

L'intervenant créera les wireframes et la charte graphique.

d. Retours maquettes (Offert, semaine 3)

L'intervenant se concertera quotidiennement avec le client, dans un but de validation des différents éléments fonctionnels et de design.

II. Setup

Cette phase consistera en la création de l'architecture du back-end et de toutes les ressources nécessaires au fonctionnement du site.

Sous-phases:

a. Définition de l'architecture du back-end (1 JEH à 450 €, semaine 4 à 5)

L'intervenant créera la structure du back-end en Django.

b. Définition de la structure de la base de données (Offert, semaine 5 à 6)

L'intervenant définira tous les champs et relations de la base de données.

c. Setup GitHub (Offert, semaine 4)

L'intervenant créera les comptes propriétaires, s'occupera du set-up des équipes et des différents répertoires de code.

III. Interface utilisateur

Cette phase consistera en l'implémentation des différentes pages et interfaces visible par les utilisateurs du site.

Sous-phases:

a. Page d'accueil (1 JEH à 450 €, semaine 5 à 6)

L'intervenant développera la page d'accueil, à la fois le front-end et le back-end.

b. Page d'authentification (2 JEH à 450 €, semaine 6 à 7)

L'intervenant développera la page d'authentification en mettant en place l'architecture back-end pour l'authentications, les routes d'authentications avec mots de passe, les tokens et les refresh tokens ainsi que la récupération de mot de passe.

c. Page d'inscription (1 JEH à 450 €, semaine 7 à 8)

L'intervenant développera la page d'inscription notamment avec le formulaire d'inscription et de connexion avec une option de connexion OAuth via Facebook ou Google.

d. Page de paiement (1 JEH à 450 €, semaine 8)

L'intervenant développera la page de paiement notamment avec l'API de Stripe et gèrera les confirmations de paiement.

e. Espace client (1 JEH à 450 €, semaine 8 à 9)

L'intervenant développera l'espace client notamment la visualisation de toutes les propriétés du client ainsi que le formulaire de saisie de nouvelles propriétés.

f. Page CGU (Offert, semaine 9)

L'intervenant développera la page des conditions générales d'utilisations.

IV. Interface administrateur

Cette phase consistera en l'implémentation des différentes pages et interfaces visible par les administrateurs du site.

Sous-phases:

a. Page de connexion des administrateurs (Offert, semaine 10 à 11)

L'intervenant développera la page de connexion sécurisée aux serveurs en mode administrateurs qui permettra l'accès aux pages décrites ci-dessous.

b. Page de validation des biens (1 JEH à 450 €, semaine 11 à 12)

L'intervenant développera la page de validation des biens permettant aux administrateurs de valider la concordance des données fournies par les clients entre elles afin de sécuriser la bonne utilisation du site.

c. Page de données clients (Offert, semaine 11 à 12)

L'intervenant développera la page de consultations de données en récupérant les données de la base de données et en les affichant pour l'administrateur.

d. Page de traitement des alertes (1 JEH à 450 €, semaine 12 à 13)

L'intervenant développera la page de traitement des alertes permettant de trier à la main les alertes remontées par l'IA et d'envoyer un mail au client en cas de suspicion.

e. Dashboard scraping (1 JEH à 450 €, semaine 12 à 13)

L'intervenant développera la page qui affichera le dashboard de l'utilisation des scrapers en agissant sur les scripts Python.

f. Notification mail (Offert, semaine 13)

L'intervenant développera la procédure d'envoi de mail automatique en Python et rédigera le template du mail envoyé.

V. Extraction de données

Cette phase consistera en la récupération de premières données et la compréhension de celle-ci.

Sous-phases:

a. Extraction NLP (1 JEH à 450 €, semaine 1 à 4)

L'intervenant développera un modèle de NLP pour extraire toutes les informations des annonces scrapées.

b. Nettoyage des données (1 JEH à 450 €, semaine 4 à 6)

L'intervenant transformera les données un format standard tout en étant capable de détecter les collisions.

c. Compréhension des données (1 JEH à 450 €, semaine 5 à 7)

L'intervenant tentera de mieux comprendre les données, les corrélations, distributions etc. pour la mise au point du premier modèle.

d. Harmonisation des données (1 JEH à 450 €, semaine 7 à 8)

L'intervenant homogénéisera les données récoltées des différents sites web et leur format afin de pouvoir appliquer le même modèle à différents sites.

VI. Premiers modèles

Cette phase consistera en la création des premiers modèles de comparaison des données récupérées et des propriétés protégées.

Livrables :

- Script Python de la première version de l'IA ;
- Notice d'utilisation du script

Sous-phases:

a. Etat de l'art (Offert, semaine 1 à 2)

L'intervenant fera une revue de la littérature sur ce qui a déjà été fait dans le domaine et identifiera des pistes de modèles à essayer, les performances atteignables, des datasets utiles pour entraîner les modèles.

b. Premier algorithme (Offert, semaine 8 à 9)

L'intervenant implémentera à la main un premier algorithme simple qui n'utilise pas de Machine Learning pour établir un point de référence des performances possibles.

c. Modèle d'IA (1 JEH à 450 €, semaine 9 à 10)

L'intervenant implémentera un premier modèle de ML (en apprentissage semi-supervisé) simple et mesurera ses performances.

d. Collecte de données (2 JEH à 450 €, semaine 10 à 13)

L'intervenant collectera des données cohérentes pour entraîner le modèle.

VII. Modèle avancé

Cette phase consistera en l'amélioration du modèle déjà développé.

Sous-phases:

a. Amélioration du modèle (2 JEH à 450 €, semaine 13 à 15)

L'intervenant entraînera, améliorera, ajustera finement le premier modèle jusqu'à atteindre les objectifs de performance voulus puis établira les variables importantes.

b. Différentiation (1 JEH à 450 €, semaine 15 à 17)

L'intervenant créera des variantes des modèles qui s'adaptent à chaque site/groupe de sites et intégrer tous les sites scrapés avec succès.

c. Connection au site (Offert, semaine 17 à 18)

L'intervenant connectera l'algorithme de décision à la BDD back-end pour qu'il puisse avoir accès au portfolio des propriétés à protéger ainsi qu'à toutes les annonces scrapées.

d. Documentation du modèle (Offert, semaine 18 à 19)

L'intervenant rédigera une documentation qui explique le modèle utilisé, les approches essayées etc. utilisable pour de futurs procédés d'amélioration continue.

VIII. Scraping

Cette phase consistera en la récupération des données de tous les sites concernés. Chaque sous-phase aura lieu pour chaque site.

Sous-phases:

a. Etude et analyse (1 JEH à 450 €, semaine 1 à 2)

L'intervenant étudiera l'architecture du site par la définition des données à récupérer, la définition des pages à scraper et l'étude des conditions de récupération des données autorisées par le site.

b. Etude de la littérature existante (1 JEH à 450 €, semaine 2 à 6)

L'intervenant étudiera l'existence d'API développées par les sites web ou de bibliothèques développées par d'autres utilisateurs.

c. Script de scraping (4 JEH à 450 €, semaine 6 à 14)

L'intervenant développera le script qui récupérera les données par le code source des pages en évitant les sécurités et détections. Ce script devra aussi parser avec BeautifulSoup les données récupérées pour stocker dans un dataframe, converti plus tard en CSV, uniquement les données qui nous intéressent.

d. Scraping des nouveautés (1 JEH à 450 €, semaine 14 à 17)

L'intervenant scrapera uniquement les nouveautés, possiblement plusieurs pages en parallèle.

e. Automatisation (1 JEH à 450 €, semaine 16 à 18)

L'intervenant développera un script qui scrapera le site à intervalle régulier et qui exportera les données vers la BDD.

f. Documentation (Offert, semaine 19)

L'intervenant rédigera une documentation sur le fonctionnement du script pour pérenniser l'utilisation.

IX. Déploiement

Cette phase consistera en le déploiement du site pour ouverture au public.

Livrables :

- Site déployé sur le serveur et droits d'accès ;
- Notices d'utilisation des scripts de scraping pour chaque site scrapé
- Rapport d'utilisation de la version finale de l'IA au format PDF.

Sous-phases:

a. Installation de la solution (1 JEH à 450 €, semaine 18 à 19)

L'intervenant déploiera l'application sur EC2.

b. Mise en place de la CI/CD (1 JEH à 450 €, semaine 18 à 19)

L'intervenant mettra en place sur git des pipelines d'intégration et de déploiement continus du back-end et du front-end.

c. Sécurisation (Offert, semaine 19)

L'intervenant sécurisera l'API et le front en passant à du https et en appliquant les bonnes pratiques sur les compartiments et les droits d'accès.

Paul Robin Matthieu Lucas