Notice sur Oraclex

Présentation des fonctionnalités développées et état des lieux de l'avancée du logiciel réalisé lors du stage de juillet 2021 par Alexis Delage, Théodore Radu et Anass Al Ammiri, sous la supervision de Franck Samson.

1. Fonctionnalités disponibles

La liste ci-dessous présente les différentes fonctionnalités développées pour Oraclex. Toutes ces fonctionnalités sont directement utilisables depuis l'interface, qui s'adapte aussi bien aux ordinateurs qu'aux mobiles (dans la plupart des cas).

1.1. Ajout de fichiers

L'outil *Ajout de fichiers* réalise plusieurs tâches successives que nous allons détailler ici :

- **Importation de fichiers :** l'utilisateur peut envoyer des fichiers au serveur, avec deux boutons :
    - le premier pour envoyer un ou plusieurs fichiers PDF (pour en sélectionner plusieurs, enfoncer la touche CTRL et cliquer sur chaque fichier)
    - le second pour envoyer un dossier complet : le logiciel cherche alors tous les fichiers PDF dans ce dossier et les importe

- **Conversion des fichiers PDF scannés en texte :**
    - Cette conversion ne s'opère que si le fichier PDF est effectivement scanné. s'il contient déjà du texte, la conversion ne sera pas effectuée et le texte sera directement utilisé.
    - La conversion, lorsqu'elle est faite, est faite avec *Tesseract-OCR* : ce module opensource utilise de l'intelligence artificielle pour décoder le fichier, en s'appuyant sur un panel de textes en français pour améliorer la reconnaissance.
    - Cette conversion renvoie un indice du taux de réussite de la conversion : 0% signifie que l'IA n'a rien réussi à analyser, et 100% signifie que le fichier a été parfaitement lu. Ce critère est visible sous l'appellation *Qualité* dans les détails d'un fichier.

- **Extraction des informations utiles :**
    - le logiciel essaie ensuite d'extraire dans le texte du fichier les informations suivantes : la date du jugement, le type de juridiction (Conseil des Prud'Hommes, Cour d'Appel, Cour de Cassation ou Tribunal Judiciaire), la ville où a eu lieu le jugement, la décision du jugement (Favorable, Défavorable ou Mixte), la somme gagnée (négative si perdue), et les mots-clés présents dans le texte
    - pour chaque information, on essaie d'abord de la trouver dans le nom du fichier, et, s'il elle n'y est pas, on la cherche dans le document lui même
    - pour optimiser l'extraction, le fichier doit être nommé de la manière suivante :

AAAA_MM_JJ_F/M/D_CA/CPH/CASS_VILLE_INFOS.pdf

exemple : 2021_07_23_D_CPH_ST MICHEL-CHEF-CHEF_blablabla.pdf

    - Quelques précisions sur certains champs :
        - il est important de bien **séparer** les différents mots et champs, par un tiret ou un espace : par exemple, pour un jugement de Conseil des Prud'Hommes Favorable, FCPH ne sera pas compris alors que F_CPH le sera.
        - Pour les noms de ville :
            - toujours écrire ST et jamais SAINT
            - séparer tous les mots par des tirets ou des espaces : ST-ETIENNE est compris mais pas STETIENNE
            - toujours écrire le nom de la ville en entier et avec les déterminants, comme sur une adresse postale : exemple : LE MANS sera compris mais pas MANS
        - Le champ INFOS ne sera pas analysé : il permet juste d'ajouter des informations à destination des juristes pour mieux identifier le fichier

- **Détection des doublons :** après l'extraction des données, le logiciel vérifie si le fichier est similaire ou non à un fichier déjà existant dans la base de données. Si c'est le cas, le fichier est mis en attente (cf. partie 1.5.1).

- **Enregistrement du fichier dans la base de données :** le fichier et ses informations sont sauvegardées, c'est la dernière étape du processus d'ajout automatique.

- **Ajout en parallèle :** afin d'optimiser la vitesse de traitement des fichiers, une fonctionnalité a été ajoutée afin d'analyser plusieurs fichiers en même temps. Ces fichiers sont analysés sur le serveur, l'utilisateur peut donc fermer la page une fois que la liste des fichiers importés apparaît, même si ceux-ci sont encore en cours d'analyse.

- **Historique :** un historique, personnel à chaque utilisateur, permet de voir l'état d'avancement de l'analyse des fichiers importés. Cet historique est présenté sous la forme d'un tableau, avec des fonctionnalités de tri selon chaque colonne. Les fichiers peuvent être ouverts pour voir les informations extraites par le logiciel.

1.2. Recherche

1.2.1. Formulaire de recherche

Le formulaire de recherche permet de chercher parmi toutes les décisions de justice ajoutées selon plusieurs critères :

- **les mots-clés :** pour entrer un mot-clé, commencez à le taper, puis cliquez sur la proposition appropriée qui apparaît. Il faut ensuite taper sur ENTREE pour valider le mot-clé. Vous pouvez aussi valider des mots-clés qui n'existent pas (ils ne sont alors pas proposés), mais ils ne seront pas pris en compte dans la recherche. les résultats affichés seront les décisions contenant TOUS les mots-clés (recherche en ET)
- **le type de juridiction :** CPH / TJ / CA / CASS
- **la juridiction :** tapez le nom de la ville et choisissez la bonne juridiction
- **la date du jugement :** entrez les mois et années entre lesquelles vous souhaitez rechercher des décisions. Les mois de début et de fin sont tous les deux inclus.
- **les fichiers illisibles :** il est possible d'inclure les fichiers illisibles dans la recherche (définis par le critère de lisibilité, cf ci-dessous), néanmoins cela peut fausser les résultats, c'est donc déconseillé

1.2.2. Résultats de recherche

Les résultats se présentent sous la forme de deux onglets :

- un premier onglet avec des résultats statistiques sur le gain et les résultats des jugements, sous forme de graphiques
- un second onglet avec la liste des décisions trouvées : vous pouvez cliquer sur une décision pour voir toutes les informations extraites de la décision (le responsable peut aussi les modifier si besoin), à savoir :
    - le lien pour télécharger le fichier PDF original
    - la date du jugement
    - la juridiction (ou tribunal) qui a prononcé le jugement
    - la décision finale : Favorable, Défavorable, ou Mixte
    - la somme gagnée, ou perdue si négative
    - les mots-clés détectés dans le fichier
    - la qualité du fichier lors de la conversion du scan en texte (cf partie *Conversion* dans le 1.1)
    - le critère de lisibilité : un fichier est considéré comme lisible si sa qualité est supérieure à 60 %, et que la date du jugement et la juridiction ont bien été trouvées
    - la date d'import du fichier sur le logiciel
    - l'utilisateur ayant importé le fichier sur le site

Les critères de recherche entrés apparaissent dans les formulaires où ils ont été entrés, afin de les vérifier à posteriori. À noter que pour les mots-clés, ceux qui ont validés appraissent au-dessus du champ de recherche et non dedans.

1.3. Prédiction

La prédiction utilise l'intelligence artificielle pour construire un réseau de neurones, basé sur toutes les décisions importées dans la base de données (y compris les illisibles : il suffit juste de détecter des mots-clés et une décision pour pouvoir entraîner le modèle).

En tapant des mots-clés dans le champ de recherche, le réseau de neurone va émettre une prédiction quant au résultat de l'affaire, sur une échelle de 0 % à 100 % :

- 0 % : l'IA est certaine que le jugement sera défavorable
- 100 % : l'IA est certaine que le jugement sera favorable
- 50 % : l'IA hésite et ne dispose pas de suffisamment de données pour donner une bonne prédiction

À noter que plus le nombre de mots-clés entrés est important, plus la prédiction sera précise : une prédiction optimale est obtenue pour une dizaine de mots-clés environ.

Concernant l'ajout de fichiers, lorsqu'une décision est ajoutée à la base de données, cela ne se répercute pas directement sur le réseau de neurones. Afin de prendre en compte cette modification, le serveur reconstruit tout le réseau de neurones le samedi à minuit, avec toutes les nouvelles décisions ajoutées dans la semaine.

1.4. Gestion des utilisateurs

Chaque utilisateur doit avoir un compte pour utiliser le logiciel. Ce compte pourra être uniquement créé par le responsable du site (cf. 1.5.4). Pour utiliser le site, l'utilisateur doit ensuite se connecter au site, puis il peut se déconnecter lorsqu'il a terminé. il peut aussi modifier ses informations personnelles, à savoir :

- identifiant et mot de passe : attention à ne pas trop les modifier, c'est ce qui vous permet de vous connecter au site
- nom et prénom : c'est le nom qui s'affiche en haut à droite du site, et aussi le nom qui apparaîtra au responsable dans la liste des membres lorsqu'il élire un autre responsable (cf 1.5.5)
- email et mot de passe : ces champs seront affichés à tous les utilisateurs lorsque vous êtes élu responsable, pour qu'ils puissent vous contacter facilement

1.5. Super-pouvoirs du responsable

Le responsable du site dispose de quelques fonctionnalités supplémentaires, destinées à assurer la bonne maintenance du site. Si ces fonctionnalités ne sont pas suffisantes, vous pouvez toujours contacter l'un des créateurs du site pour qu'il vous dépanne.

1.5.1. Gestion des doublons

La page de gestion des doublons permet de gérer les conflits de fichier lorsqu'un doublon est détecté : vous avez le choix entre conserver l'ancien fichier, le nouveau, ou bien les deux s'ils s'avère que ce ne sont pas en fait des doublons.

1.5.2. Ajout de mots-clés

Le responsable peut aussi ajouter des mots-clés sur le site : à chaque mot-clé ajouté, toutes les décisions sont automatiquement ré-analysées afin de trouver le nouveau mot-clé dans chacune d'elle, cette opération peut donc prendre un peu de temps.

De plus, ajouter un mot-clé "casse" le réseau de neurones pour la prédiction : la première fois que vous essayez de faire une prédiction après avoir ajouté un mot-clé, le réseau de neurones va s'auto-détruire pour se reconstruire entièrement avec la nouvelle liste des mots-clés. Cette opération étant aussi plutôt longue, il est conseillé de lancer une prédiction manuellement (peut importe laquelle) juste après avoir ajouté un mot-clé, pour le reconstruire.

1.5.3. Modification des informations d'un fichier

Lorsque la page de détails d'un fichier est ouverte (par exemple depuis les résultats de recherche, l'historique d'ajout, la liste des fichiers illisibles ou encore depuis la page des doublons), vous pouvez modifier les données du fichier, et aussi si besoin supprimer le fichier. Cela n'est pas recommandé mais reste disponible en cas de problème ou de bug informatique sur l'analyse d'un fichier.

1.5.4. Ajout d'utilisateur

Le responsable peut ajouter un utilisateur depuis le menu principal.

1.5.5. Changement de responsable

Le responsable peut aussi attribuer le rôle de responsable à une autre personne. Attention, une fois le formulaire validé, il perd immédiatement les droits de responsable et ne peut donc plus modifier à nouveau le responsable, assurez-vous donc que le nouveau responsable a bien accès à son compte (notamment qu'il n'a pas perdu son identifiant ou son mot de passe) avant de le changer.

2. Améliorations possibles

2.1. Ajout de fichiers

La détection de la somme n'est pas encore parfaite et pourrait être améliorée, notamment pour détecter qui paie à qui en prenant en compte plus de formulations classiques.

2.2. Recherche

La recherche pourrait être améliorée, notamment pour le formulaire permettant d'entrer les mots-clés : il y aurait possibilité de les afficher au même endroit lorsque la recherche a été lancée, plutôt qu'ils disparaissent comme c'est le cas actuellement. La gestion de la validation de chaque mot-clé pourrait aussi être améliorée.

Les graphiques de résultats pourraient aussi être plus clairs, notamment celui avec la répartition des gains (les couleurs et l'axe des abscisses ne sont pas très lisibles pour l'instant).

2.3. Prédiction

Pour la prédiction, il y a toute la partie de prédiction du gain potentiel qui pourrait être ajoutée à la prédiction de la décision.

2.4. Gestion des utilisateurs

Une amélioration possible serait de connecter et de créer les comptes en envoyant des mails de confirmation afin de sécuriser l'accès. Néanmoins, cela nécessite un vrai serveur de déploiement pour le faire.

3. Différences et limitations des versions déployées

Pour vous permettre de tester le logiciel, Oraclex est actuellement disponible à deux emplacements différents :

- une version en ligne, accessible par tous à l'adresse oraclex.herokuapp.com
- une version locale, installée sur l'ordinateur personnel de Franck Samson

3.1. Version en ligne sur Heroku

La version en ligne est hébergée sur Heroku, avec le compte personnel de Alexis Delage, à l'adresse oraclex.herokuapp.com. Cet hébergement a l'avantage d'être gratuit, car il est conçu à des fins de démonstrations, mais il dispose par conséquent de nombreuses limitations :

- lorsque le site est inutilisé, le serveur est automatiquement éteint, il peut donc mettre un certain temps à redémarrer lorsqu'on se connecte au site
- on ne peut pas non plus importer des fichiers sur ce serveur (limitation à cause du fait que ce soit gratuit) : l'ajout de fichiers ne fonctionne donc pas. Pour corriger ce problème, il faudrait intégrer un second serveur pour stocker les documents (AWS par exemple)
- le serveur étant éteint la nuit, la mise à jour du réseau de neurones ne se fait pas non plus du coup (mais vu qu'on ne peut pas ajouter de fichiers, cela n'a pas beaucoup d'importances…). Néanmoins, le réseau initial que nous avons implanté a été entraîné auparavant avec plus de 600 décisions, ce qui lui permet de fonctionner et de renvoyer des résultats pour les démos.

Attention, pour accéder au site, il vous faudra un compte : à l'heure où ces lignes sont écrites, seul Franck Samson possède un compte (dont l'identifiant et le mot de passe (temporaire, il faudrait le changer) sont stockés sur son ordinateur personnel), et il est en plus le responsable du site. Il lui reviendra donc d'ajouter des utilisateurs (cf 1.5.4) pour que d'autres personnes puissent y accéder.

3.2. Version locale

La version locale, installée sur l'ordinateur personnel de Franck Samson, est là pour pallier les problèmes de la version en ligne : l'import de fichiers et toutes les autres fonctionnalités sont opératrices, son seul inconvénient est justement de ne pas pouvoir être accessible en ligne depuis n'importe quel ordinateur.

Conclusion

Et bah… ça marche pas trop mal, nous on est content de ce qui a été fait.

Ce stage nous a permis d'une part de découvrir un peu plus le fonctionnement d'une grande entreprise, et aussi de nous perfectionner dans notre domaine en travaillant sur un vrai projet, alors on tenait à vous remercier pour nous avoir confié ce projet et pour votre suivi régulier tout au long du développement. Si ce logiciel existe, c'est aussi grâce à toute l'équipe du service, qui nous ont fait part de leurs retours et commentaires constructifs pour toujours améliorer le projet.

Bien sûr, quelques points pourraient encore être améliorés, mais le plus gros du travail est fait : le logiciel en lui-même est fonctionnel et utilisable au quotidien par les juristes. Bien sûr, il faudra aussi trouver un moyen de le rendre disponible en ligne en l'installant sur un vrai serveur (soit en voyant avec le service informatique de la SNCF, soit en louant un serveur à part).

En cas de besoin, le code source complet du logiciel ainsi que la procédure d'installation peuvent être retrouvés sur ce repo gihub : github.com/hydrielax/oraclex. Cela pourrait être utile pour de futurs travaux sur le projet.

Bonne vacances (pour ceux qui en prennent !) et bonne continuation à toute l'équipe que nous avons côtoyé avec joie pendant ces 4 semaines !

Oraclexement,

Anass, Théodore et Alexis