



PROJET ÉTUDIANT D'ENTREPRISE

Rapport d'étude

Outil de justice prédictive en droit social

Groupe P2E

CHAKROUN Karim
DELAGÉ Alexis
HAFFOUDHI Samy
MICHELON François
LONGATTI BAPTISTA Giovana
RADU Théodore

Correspondant Entreprise

SAMSON Franck

Tuteur Enseignant

ROUX Olivier

Sommaire

1. Présentation du projet dans son contexte	3
1.1 L'Entreprise	3
1.2 Le Projet	3
1.3 Le contexte et les contraintes	4
2. Démarche	5
2.1 Définition du cahier des charges	6
2.2 Maquette de l'interface	7
2.3 Choix des outils de développement	9
2.3.1 Site Web	9
2.3.2 Django	10
2.3.3 Replit	10
2.3.4 Bootstrap	10
2.4 Développement de l'interface	11
2.4.1 Présentation de l'interface	11
2.4.2 Base de données	12
2.4.3 Réutilisation des scripts précédents	13
2.4.4 Ajouts des fichiers et phase d'analyse	13
2.4.5 Recherche de fichiers	14
2.4.5.1 Recherche par dates	14
2.4.5.2 Recherche par Juridictions	14
2.4.5.3 Recherche par mots-clés	15
2.4.6 Partie Prédiction	15
2.4.7 Affichage des résultats	16
2.4.8 Gestion des utilisateurs	16
2.5 Déploiement du site	17
2.5.1 Stratégie de déploiement	17
2.5.2 Problèmes rencontrés	17
3. Résultats obtenus	18
3.1 Page de connexion	18
3.2 Recherche et prédiction	19
3.3 Ajout de fichiers	22
3.4 Informations	23
Conclusion	24
Remerciements	25
Table des illustrations	26
Bibliographie	27
Annexes	28
Table des annexes jointes au présent document	28
Annexe 1 : Rapport de projet du P2E précédent	28
Annexe 2 : Planning organisationnel	28
Annexe 3 : Schéma du fonctionnement de l'interface	28
Annexe 4 : Maquette du site web	28
Annexe 5 : Modèle Entités-Associations de la Base de Données	28

1. Présentation du projet dans son contexte

1.1. L'Entreprise

La Société nationale des chemins de fer français (SNCF) est l'entreprise ferroviaire publique française, officiellement créée par convention entre l'État et les compagnies de chemin de fer préexistantes, en application du décret-loi du 31 août 1937.

La SNCF est l'un des premiers groupes mondiaux de mobilité et de logistique, avec une présence dans 120 pays, 30 milliards d'euros de chiffre d'affaires en 2020 et 275 000 salariés en 2019. Depuis le 1^{er} janvier 2020, le groupe est répartie en 5 filiales indépendantes :

- **SNCF Voyageurs** : transport de voyageurs ;
- **SNCF Réseau** : gestion et entretien des infrastructures ferroviaires ;
- **Rail Logistics Europe** : activité de fret et logistique ferroviaire ;
- **Geodis** : logistique et transport routier de marchandises ;
- **Keolis** : transports en commun urbains.

1.2. Le Projet

En 2018 la SNCF a confié un projet P2E qui consistait à développer un outil permettant d'extraire les données pertinentes des fichiers PDF et d'analyser les décisions de justice. Cette année, le pôle juridique souhaite poursuivre ce projet par le développement de l'interface graphique de cet outil. Ainsi le travail à réaliser consistait à :

- Prendre en main des travaux précédents ;
- Définir le cahier des charges de l'interface ;
- Etudier les différentes solutions de mise en œuvre ;
- Développer un démonstrateur.

1.3. Le contexte et les contraintes

Le Pôle Juridique - Agence Juridique Voyageurs Atlantique (PJU-AJVA) de la SNCF traite des affaires de Droit Social devant les juridictions prud'homales notamment, mais aussi devant toute juridiction compétente en la matière. Le PJU-AJVA dispose d'une collection de décisions de justice en format PDF, difficilement exploitables. En effet, afin d'estimer les pertes ou gains potentiels liés à une affaire en cours, il est nécessaire d'examiner les décisions de justice similaires une par une ce qui prend un temps considérable. C'est dans ce contexte qu'il nous a été demandé de travailler sur un outil de justice prédictive qui s'appuierait sur l'ensemble des décisions de justice existantes. Il permettrait ainsi aux juristes du PJU de mieux apprécier l'issue d'une affaire basée sur les données renvoyées par notre programme.

Le projet précédent s'est donc divisé en plusieurs travaux : d'abord la traduction des fichiers PDF en fichiers TXT pour leur analyse, la recherche des données pertinentes (mots-clés, sommes perdues ou gagnées) dans les décisions de justice, la constitution d'un document résumé qui rassemble toutes les informations obtenues, le développement d'un réseau de neurones donnant les risques de perte pour une affaire, et enfin la mise en place d'une interface graphique pour faciliter l'utilisation de l'outil. Les résultats obtenus lors de ce premier projet sont présentés dans le rapport de ce groupe, en annexe 1.

Notre projet, prenant la suite de ce projet précédent, était donc de réaliser une interface graphique afin de pouvoir utiliser facilement les codes informatiques déjà développés. Pour mener à bien ce projet, il nous a été demandé de respecter certains critères :

- Utiliser des logiciels libres de droit ;
- Réaliser une interface compatible avec les travaux du précédent ;
- Mettre en place un système de mots de passe et d'identifiant ;
- Développer une interface simple d'utilisation et intuitive.

2. Démarche

Suite à nos premières réunions de groupe et rencontres avec l'entreprise, nous avons défini un planning pour les mois de février à juin. Ce planning peut être consulté sous la forme d'un Diagramme de Gantt (cf. Annexe 2) Notre démarche a été subdivisée en cinq grandes phases :

- Phase 1 : “Cadrage et Planning”, qui a pour objectif consolider les missions du projet, travailler sur l'organisation du groupe, définir des jalons et définir la démarche à suivre.
- Phase 2 : “Prise en main du P2E précédent”. Avant de commencer notre travail, il fallait d'abord prendre connaissance de ce qui a été développé dans le projet précédent, avec deux objectifs principaux : découvrir exactement notre base de départ et nous assurer de la compatibilité entre le travail des deux groupes.
- Phase 3 : Définition de l'interface
- Phase 4 : Réalisation de l'interface
- Phase 5 : Restitution

Afin de clarifier notre démarche pour le lecteur, celle-ci est présentée par étapes successives accomplies afin de comprendre le raisonnement et le déroulé du projet.

2.1. Définition du cahier des charges

Afin de définir le cahier des charges complet du projet, nous avons fait plusieurs réunions avec l'entreprise pour mettre au point les différentes fonctions de l'application. Nous avons ainsi abouti aux points suivants :

- pour un fichier donné, les points intéressants sont :
 - la condamnation : favorable, défavorable ou mixte pour la SNCF
 - le gain ou la perte financière : compensation ou amende attribuée à la SNCF, en euros
 - la date du jugement
 - le type de cour qui a jugé le dossier, c'est-à-dire 3 possibles ici :
 - le Conseil des Prud'Hommes
 - la Cour d'Appel
 - la Cour de Cassation
 - le lieu du jugement (c'est-à-dire la ville)
 - les mots-clés : une liste de mots-clés préalablement choisis et considérés comme importants dans l'analyse des fichiers
 - la lisibilité : critère permettant de décider si le fichier est lisible ou non, suivant les résultats de la reconnaissance optique des caractères sur le fichier scanné
- l'interface doit être capable de rechercher des fichiers suivants les critères suivants : date, type de cour, lieu, et mots-clés
- l'interface doit pouvoir prédire le résultat d'un procès (en terme de condamnation et de gain) à partir d'une liste de mots-clés donnée
- l'interface doit permettre d'ajouter des fichiers à la base de donnée, mais ne doit pas permettre de les modifier ou de les supprimer par la suite (sauf dans le cas où ils sont considérés "illisibles")
- l'accès à l'interface doit être protégé par un système d'authentification.

2.2. Maquette de l'interface

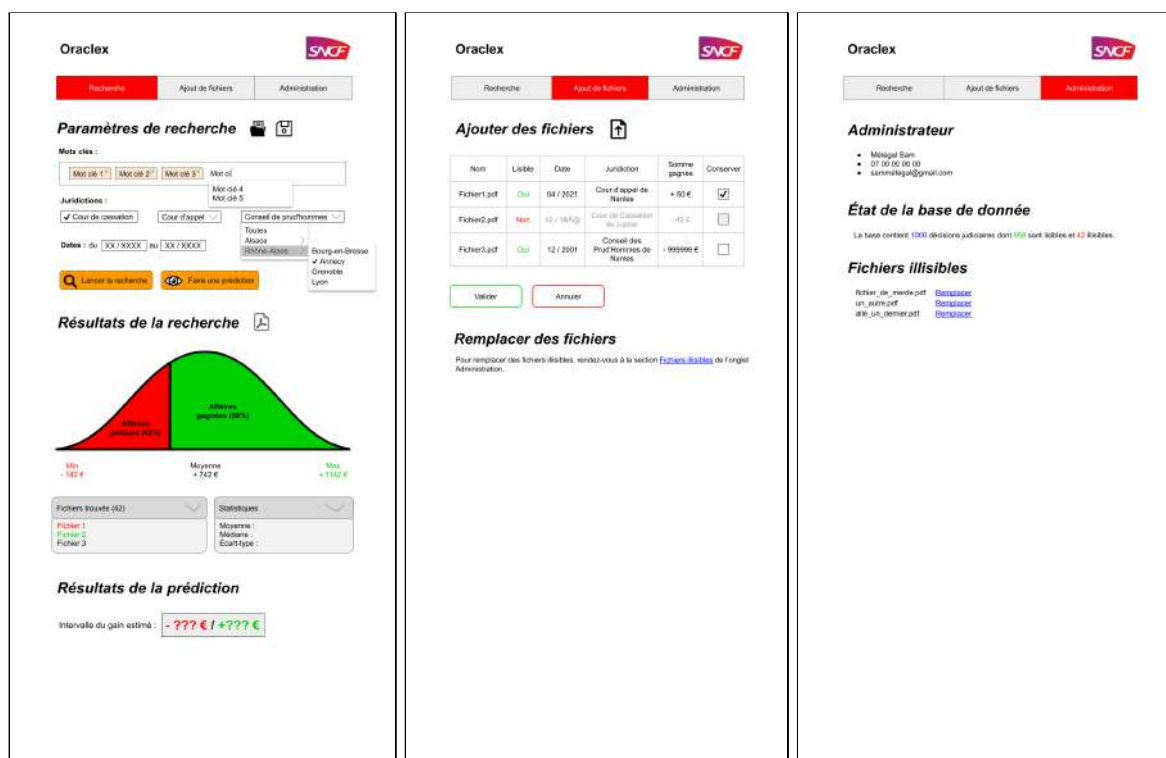


Figure 2.2 – Maquette de l'interface

Suite à la définition du cahier des charges, nous avons alors proposé une première idée d'interface. Celle-ci était non-fonctionnelle, uniquement visuelle, afin de donner une idée de ce à quoi pouvait ressembler l'interface finale. Elle nous a alors permis d'échanger avec l'entreprise sur les détails de l'interface jusqu'à l'élaboration de la maquette finale (cf Figure 1).

Afin de pouvoir identifier facilement le projet, nous avons choisi de lui donner le nom de code "Oraclex", contraction de *Oracle*, reflétant la partie prédiction du logiciel, et *Lex*, la loi, pour rappeler l'utilisation juridique à laquelle se destine cet outil. Ce nom n'est bien entendu pas définitif et pourra être changé si besoin lors de futures améliorations.

L'interface se découpe en 3 pages :

- Une première page présentant les paramètres de recherche, à savoir :
 - les mots-clés,
 - la date du jugement,
 - le type de cour,
 - et le lieu du jugement ;

avec ensuite les résultats de la recherche constitués de :

- la liste des fichiers correspondants,
- les statistiques de gain et perte (moyenne, médiane, écart-type, maximum et minimum),
- un graphique présentant la répartition des gains et des pertes,
- un graphique présentant la répartition des jugements favorables, défavorables, ou mixtes (c'est-à-dire à la fois favorable et défavorable) ;

et enfin les résultats de la prédiction par intelligence artificielle (cf. l'annexe 1).

- Une seconde page ensuite permettant d'ajouter des fichiers : il est possible d'importer plusieurs fichiers à la fois, le programme les analyse alors (par reconnaissance optique de caractères), et affiche les informations trouvées. L'utilisateur est libre ensuite d'ajouter ou non les fichiers dans la base de données du site, en fonction des résultats affichés.
- Une troisième page enfin présentant quelques informations générales sur le site :
 - le responsable du site : désigné pour une période donnée, ce responsable servira de référent au sein de l'entreprise pour gérer les données stockées et aider les utilisateurs en cas de soucis ;
 - le nombre total de fichiers stockés sur le serveur, avec le nombre de fichiers classés comme "illisibles" - ce nombre est important à connaître car il permet de juger du degré de prédiction du logiciel : plus la base de données est importante, plus les résultats seront précis et utiles ;
 - la liste des fichiers PDF considérés comme illisibles, afin de pouvoir les remplacer manuellement avec de nouvelles versions scannées.

2.3. Choix des outils de développement

2.3.1. Site Web

Pour la réalisation de l'interface graphique, nous avons le choix entre deux types d'interfaces : une interface locale (exécutable) ou bien une interface centralisée (site web). Nous avons dressé un comparatif des deux solutions, résumées dans la Table 1 ci-dessous : ce que nous avons considéré comme un avantage est indiqué en vert, alors que les inconvénients sont indiqués en rouge. Certains points n'ayant pas d'importance dans le choix de l'interface sont indiqués sans couleur, ils sont uniquement présents dans le but d'aider le lecteur à comprendre la différence entre les deux.

	Interface locale (logiciel)	Interface centralisée (site web)
Stockage des fichiers de jugements	Sur chaque ordinateur (les bases de données diffèrent alors d'un ordinateur à l'autre)	Sur un serveur centralisé (tous les ordinateurs ont alors accès à la même base de données)
Compatibilité	Uniquement un type d'appareil (Windows)	Tous les appareils (ordinateurs et téléphone)
Sécurité	Pas de connexion à internet, donc sécurité maximale	N'importe qui peut accéder au site, nécessite donc un mot de passe
Prix	Aucun frais	Location du serveur à payer (mais possibilité en théorie d'utiliser le serveur déjà existant de la SNCF, donc pas de frais supplémentaires)

Table 1 - Comparatif des choix d'interfaces

La solution retenue a donc été l'interface centralisée sous forme de site web, plus adaptée au projet, notamment pour pouvoir utiliser une base de données commune.

2.3.2. Django

Pour le langage de développement, nous avons choisi d'utiliser le framework Django. Celui-ci présente plusieurs avantages :

- gestion facilitée de bases de données importantes, donc plutôt adapté pour notre projet ;
- développement des scripts d'exécution en Python, qui était aussi le langage utilisé par le précédent P2E (cf. Annexe 1), ce qui permet de réutiliser directement les codes déjà écrits ;
- développement de l'interface graphique en HTML et CSS, les langages standards du web très souples et performants ;
- Django est gratuit et opensource.

De plus, un membre de l'équipe connaissait déjà Django, ce qui nous a permis d'avancer un peu plus vite sur le projet. Pour apprendre l'utilisation de ce framework, nous nous sommes basés sur un tutoriel proposé par Mozilla [\[1\]](#) et la documentation officielle de Django [\[2\]](#).

2.3.3. Replit

Pour développer notre site web à plusieurs de manière collaborative, nous avons fait le choix d'utiliser la plateforme replit.com. Celle-ci permet de travailler à plusieurs en temps réel, et offre un support assez complet de Django, ce qui permet notamment de tester très facilement le site.

2.3.4. Bootstrap

Bootstrap est une bibliothèque CSS permettant de réaliser des interfaces graphiques cohérentes et agréables facilement. Au cours du projet, nous avons beaucoup utilisé sa documentation [\[3\]](#) afin de mettre en page le site.

La SNCF a aussi développé une version adaptée de Bootstrap, reprenant la charte graphique de l'entreprise [\[4\]](#). Cette version adaptée nous a permis de réaliser rapidement un site reprenant les codes des sites de la SNCF, afin d'ancrer notre projet dans les produits de l'entreprise.

2.4. Développement de l'interface

2.4.1. Présentation de l'interface

Notre interface prend donc la forme d'un site web, facilement accessible pour n'importe qui en renseignant l'url adéquate et ses identifiants. Le site n'ayant pas pu être déployé "pour de vrai" suite à des difficultés avec le service informatique de l'entreprise, nous avons donc travaillé sur Replit, qui permet de développer et tester des sites web. Pour accéder à cette version test de notre site, merci de vous référer à la section 3.1 où vous trouverez l'adresse du site ainsi que les identifiants de connexion du compte test *Salomon*.

Attention, ce site est uniquement une version de développement ! Il n'est absolument pas sécurisé, n'importe qui peut accéder aux données même sans connaître les codes de connexion. Il doit donc uniquement être utilisé à des fins de tests et non à des fins de production.

ORACLEX / Recherche / Ajout de fichier / Informations / Salomon

RECHERCHE

PARAMÈTRES DE LA RECHERCHE

Date Minimale :

Date Maximale :

Type de juridiction (non-fonctionnel) :

Juridiction :

Mots-Clés :

Mots-Clés (non-fonctionnel) :

Figure 2.4.1 - Page d'accueil du site

2.4.2. Base de données

Afin d'organiser au mieux les données du site, nous avons fait le choix d'utiliser une base de données pour organiser et gérer les différentes informations stockées. En voici le modèle entité-association :

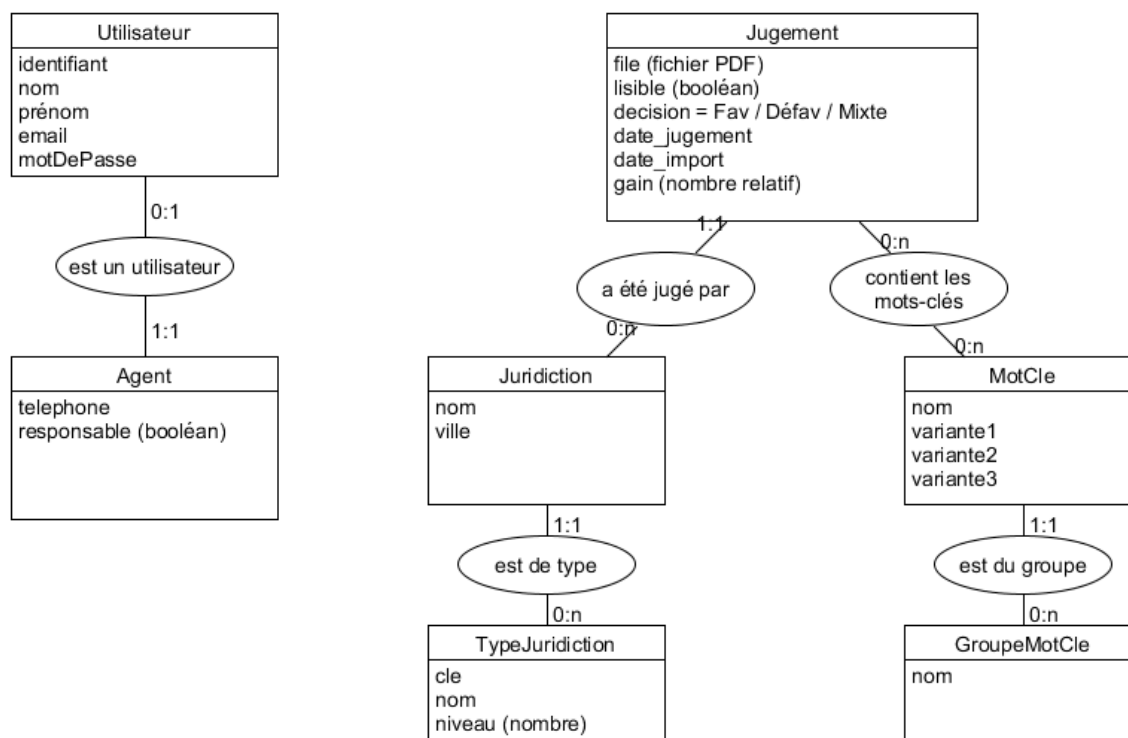


Figure 2.4.2 - Modèle entité-association de la base de données

Quelques explications sur les différentes tables :

- la table *Utilisateur* est une table déjà incluse dans *Django*, servant uniquement à l'identification des utilisateurs du site ;
- la table *Agent* correspond à une extension de la table *Utilisateur*, et permet notamment de stocker des informations supplémentaires ainsi que le rôle de l'Agent : il est soit responsable du site (l'attribut *responsable* est alors *vrai*), ou simple utilisateur (l'attribut est alors *faux*) ;
- la table *Jugement* représente les fichiers scannés importés sur le site : on y stocke également les informations lues sur ce fichier afin d'éviter d'avoir à le ré-analyser chaque fois ;
- la table *Juridiction* représente le conseil ou la cour qui a délivré le *Jugement* : une juridiction possède un nom, une ville, et un type (Cour d'Appel, Cour de Cassation, ou Conseil des Prud'hommes) - afin de permettre l'ajout futur d'autres types de cour pour d'autre services juridiques par exemple, les

différents types sont stockés dans une autre table *TypeJuridiction* ; afin d'être exhaustif, les différentes juridiction ont été importées depuis le site petitpois.justice.comarquage.fr, site servant de base de donnée à l'annuaire officiel sur www.annuaires.justice.gouv.fr ;

- enfin un *Jugement* est relié à plusieurs mots-clés, stockés dans la table *MotCle*, qui peuvent avoir plusieurs variantes (par exemple le mot-clé CDD aura comme variante C.D.D. et Contrat à Durée Déterminée) et appartenir à un groupe de mot-clé, stocké dans la table *GroupeMotCle*.

2.4.3. Réutilisation des scripts précédents

L'utilisation d'une telle base de données nous a néanmoins posé problème : en effet, les scripts développés par le groupe P2E précédent étaient conçu pour utiliser un simple fichier texte comme base de données (dénommé le "fichier résumé", cf. annexe 1). Nous avons donc dû réécrire ces scripts afin de les adapter à notre base de données, notamment les scripts pour la recherche de fichier ou encore la lecture du fichier PDF original. Nous n'avons malheureusement pas pu reprendre les autres scripts, comme cela sera détaillé ci-dessous.

2.4.4. Ajouts des fichiers et phase d'analyse

La première étape d'utilisation du site est d'ajouter des fichiers. Cet ajout fonctionne, l'utilisateur peut ajouter des fichiers PDF à la base de données qui sont ensuite stockés dans la base de données puis, en théorie, analysés. Cette analyse ne fonctionne pas actuellement, à cause d'un problème que nous n'avons pas réussi à résoudre à temps. Pour l'expliquer, voyons un peu en détails les différentes étapes de l'analyse :

1. D'abord, l'utilisateur importe le fichier sur le site.
2. Une fois validé, la première étape consiste à faire une OCR (*Optical Character Recognition*, ou Reconnaissance Optique de Caractères) qui permet au logiciel de transformer l'image importée en texte informatique exploitable. Les scripts précédents utilisaient la bibliothèque Python *pytesseract*, qui permet justement de faire une OCR. Malheureusement, cette bibliothèque ne fonctionnait pas lors de nos tests, probablement à cause d'une incompatibilité avec Replit, notre plateforme de développement. Pour résoudre ce problème, il faudra soit utiliser une autre bibliothèque, soit changer d'environnement de développement, ce qui dans les 2 cas était trop long pour nous dans le cadre d'un P2E.
3. Les étapes suivantes de l'analyse, lorsque l'OCR fonctionne sont alors la recherche des mots-clés, du gain, de la décision du jugement et du type de juridiction.

2.4.5. Recherche de fichiers

2.4.5.1. Recherche par dates




The form contains two side-by-side input fields. The left field is labeled 'Date Minimale :' and the right field is labeled 'Date Maximale :'. Both fields have a placeholder text 'AAAA MM' and a small blue icon with a calendar symbol on the left.

Figure 2.4.5.1 - Champs de recherche par date

Passons maintenant à la partie recherche. Concernant les champs de date, nos tests ont montré que ceux-ci fonctionnent sans problème : l'utilisateur entre les dates minimales et maximales souhaitées et l'algorithme ressort les fichiers correspondants.

2.4.5.2. Recherche par Juridictions



The form contains two side-by-side elements. The left element is labeled 'Type de juridiction (non-fonctionnel) :' and features a dropdown menu with a blue arrow icon. The right element is labeled 'Juridiction :' and features a search input field with a magnifying glass icon and the placeholder text 'Tapez le nom de la ville'.

Figure 2.4.5.2 - Champs de recherche par juridiction

Pour le type de juridiction, la recherche est censée être fonctionnelle, mais nous rencontrons un bug lors de son utilisation que nous n'avons pas réussi à identifier avant la fin du projet.

Pour la recherche de juridiction, la recherche est fonctionnelle mais ne renvoie aucun résultat puisque l'algorithme d'analyse actuel ne relève pas les juridictions dans l'analyse du document.

2.4.5.3. Recherche par mots-clés



Mots-Clés :

🔍 Tapez les mots-clés séparés par une virgule

Mots-Clés (non-fonctionnel) :

- ☐ Tous les mots-clés
- ☐ ABSENCE IRREGULIERE
- ☒ ACCROISSEMENT
- ☒ ACCROISSEMENT TEMPORAIRE
- ☐ ACTION EN JUSTICE
- ☐ ACTIVITE NORMALE ET PERMANENTE

Figure 2.4.5.3 - Champs de recherche par mots-clés

La recherche par mots-clés a été particulièrement difficile à implémenter : nous avons d'abord testé une recherche où l'utilisateur frappe ses mots-clés dans une barre de recherche (ligne du dessus). Cette méthode a l'avantage d'être rapide, mais peu robuste car une simple faute de frappe peut renvoyer une recherche fausse ou biaisée. Notons tout de même que ce champ de recherche fonctionne et peut être utilisé, à défaut d'être facile d'accès (on ne voit pas quels mots-clés sont disponibles).

La seconde méthode envisagée est une liste déroulante où l'utilisateur peut simplement cocher les mots-clés qui l'intéressent. Afin de diminuer la taille de la liste, il a été envisagé de regrouper les mots-clés par *groupe*, et aussi de fusionner les mots-clés ayant le même sens sous une même étiquette (cf. la section 2.4.2). Nous avons réalisé cette méthode ce champ via le système Bootstrap de la SNCF (cf. la section 2.3.4), mais il semble qu'il y ait une interférence entre ce système et Django, d'où le fait que ce champ ne fonctionne pas encore. Nous avons donc fait le choix de laisser le premier champ affiché, à des fins de tests.

2.4.6. Partie Prédiction

Concernant la prédiction, nous n'avons malheureusement pas pu utiliser les scripts précédents. En effet, ceux-ci incluent uniquement la structure du réseau de neurones, mais nous n'avons aucun exemple ni aucune indication sur comment l'utiliser. Cette partie ayant été réalisée dans les dernières étapes du projet (vu qu'elle nécessite que les autres parties fonctionnent), nous nous en sommes aussi rendus compte trop tard pour pouvoir contacter les étudiant·e·s du P2E précédents. Cette partie n'est donc pas fonctionnelle non plus.

2.4.7. Affichage des résultats

L'affichage des graphiques de résultats est réalisé à l'aide de la bibliothèque javascript chart.js. Pour plus de détails sur leur lecture, voir la partie 3.

2.4.8. Gestion des utilisateurs

Le site Oraclex permet la gestion de plusieurs utilisateurs. Chaque utilisateur peut alors modifier ses propres informations, son mot de passe, *et cætera*... Néanmoins, cette partie n'est pas fonctionnelle pour l'instant car la création et la validation de nouveaux comptes nécessite un serveur de mails, que nous n'avons pas à notre disposition pour l'instant. Un seul compte a donc été créé pour l'instant pour des phases de tests, et les pages permettant la modification de son compte n'ont pas été implémentées non plus.

2.5. Déploiement du site

2.5.1. Stratégie de déploiement

Le site présenté dans la partie précédente est une version de développement, c'est-à-dire qu'il n'est pas conçu pour être utilisé en production (pas de sécurités, pas de comptes utilisateurs multiples possibles, etc...). Afin d'être fonctionnel, le site doit être installé sur un serveur qui centralisera toutes les données.

Les caractéristiques de ce serveur doivent répondre aux critères de configuration minimale de django, et avoir un espace de stockage suffisant pour stocker toute la base de données (documents PDF et leur métadonnées principalement). Ce serveur est néanmoins payant (de l'ordre d'une dizaine d'euros par mois environ). En revanche, il sera peut-être possible d'intégrer ce site aux serveurs déjà existants de la SNCF afin d'éviter des frais supplémentaires.

2.5.2. Problèmes rencontrés

Afin de déployer le site sur des serveurs de la SNCF, nous avons fait une demande via notre correspondant entreprise auprès des services informatiques de la SNCF pour réaliser ce projet. Néanmoins, bien que la demande ait été faite dès les débuts du projet, les processus d'admission sont longs et les services informatiques ne sont pas forcément enclins à déployer ce site. On notera que cette difficulté étant assez contraignante, de nouvelles discussions seront nécessaires au sein de l'entreprise afin de trouver une solution à ce problème, qui va au-delà de la mission qui nous a été confiée (à savoir l'interface graphique).

3. Résultats obtenus

Dans l'objectif de simuler le fonctionnement envisagé de l'interface Oraclex, nous avons réalisé un démonstrateur hébergé sur la plateforme Replit. Attention : s'agissant uniquement d'une version de développement, celui-ci ne dispose pas d'un système adéquat de sécurisation des données et doit donc être uniquement utilisé à des fins de tests, et non de production ! Cela étant, le démonstrateur est consultable publiquement à l'adresse suivante : p2e46-oracle.delage.repl.co.

3.1. Page de connexion

L'accueil du visiteur sur le site s'effectue avec une page de connexion, l'invitant à s'identifier. Pour l'instant, la connexion est assurée par les identifiants temporaires suivants :

- Identifiant : salomon
- Mot de passe : mlkjhgfdsq

Figure 3.1.1 - Page de connexion du site

Qui plus est, l'interface du site propose aussi un système de déconnexion. Une interface de gestion de profil (cf. figure 3.1.2) avait été amorcée mais n'a pas abouti faute de temps.

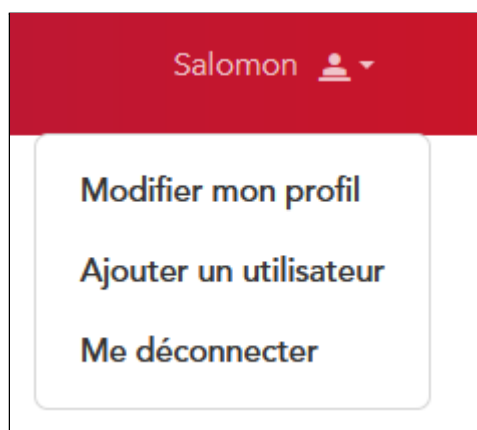


Figure 3.1.2 - Menu de l'utilisateur

3.2. Recherche et prédiction

La recherche d'informations dans la base de données ainsi que la prédiction de l'issue des jugements sont assurées par l'onglet "Recherche" du site.

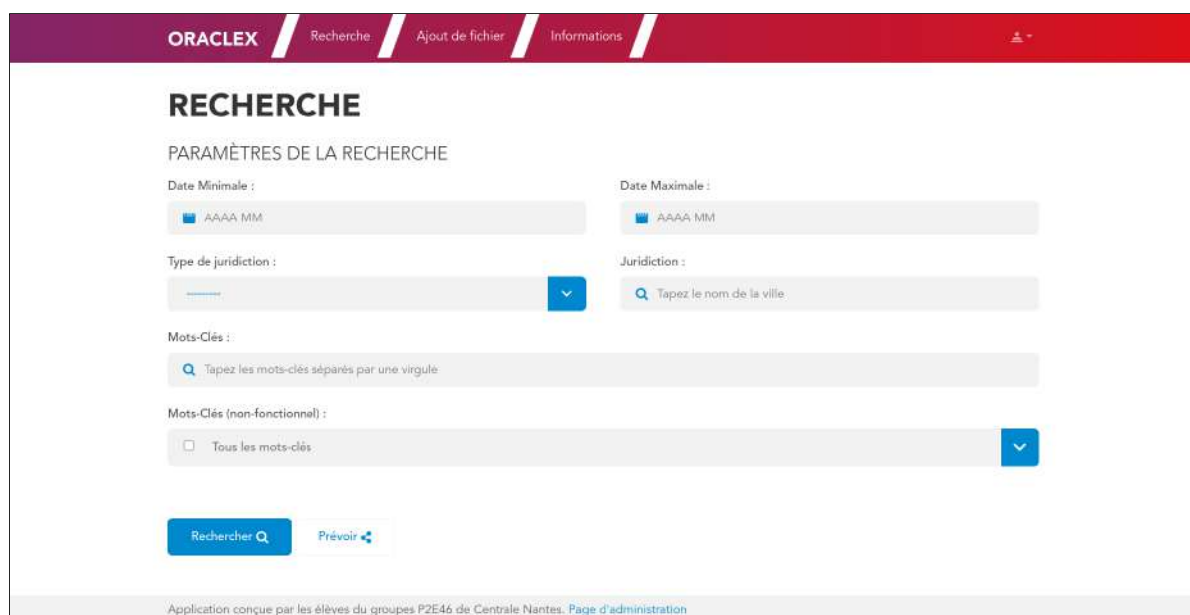
The screenshot shows the 'RECHERCHE' page of the ORACLEX application. The top navigation bar is red with the ORACLEX logo and links for 'Recherche', 'Ajout de fichier', and 'Informations'. The main content area has a title 'RECHERCHE' and a section 'PARAMÈTRES DE LA RECHERCHE'. It includes input fields for 'Date Minimale' and 'Date Maximale' (both with 'AAAA MM' placeholders), a 'Type de juridiction' dropdown, a 'Juridiction' search field with the placeholder 'Tapez le nom de la ville', a 'Mots-Clés' search field with the placeholder 'Tapez les mots-clés séparés par une virgule', and a 'Mots-Clés (non-fonctionnel)' section with a checkbox for 'Tous les mots-clés'. At the bottom are two buttons: 'Rechercher' and 'Prévoir'. A footer note states: 'Application conçue par les élèves du groupes P2E46 de Centrale Nantes. Page d'administration'.

Figure 3.2.1 - Page de lancement de recherche du site

Pour effectuer une recherche, plusieurs paramètres de filtrage sont mis à disposition. Nous pouvons ainsi sélectionner :

- les dates minimale et maximale des jugements à l'aide d'une entrée "date"
- le type de juridiction à l'aide d'un menu déroulant de sélection
- la juridiction à l'aide d'une barre de recherche
- la liste des mots-clés à l'aide d'un système de case à cocher et/ou d'une zone de texte

Plus précisément, en ce qui concerne la saisie des mots-clés, nous avons opté pour l'utilisation de la barre de saisie afin d'accélérer la sélection lorsque l'utilisateur a déjà connaissance de la liste. Cependant cette barre ne propose pas des propositions de complétion de la saisie (contrairement à la barre de saisie de la juridiction).

La liste de cases à cocher permet quant à elle le regroupement des mots-clés par catégories afin d'accompagner l'utilisateur dans la recherche, mais n'est pas encore fonctionnelle (cf. la section 2.5.4.3). Par ailleurs, nous avons ajouté l'option de sélection d'une juridiction même si elle n'est pas utilisée par le programme, afin de faciliter l'ajout de cette fonctionnalité plus tard.

Après validation des paramètres, la page des résultats s'affiche (cf. figure 3.2.2), avec :

- un résumé des paramètres de recherche ;
- un diagramme circulaire représentant la part de jugements favorables et défavorables ;
- un diagramme en barres représentant la répartition des gains ;
- quelques valeurs statistiques (moyenne, écart-type, médiane, minimum, maximum) ;
- la liste des fichiers trouvés avec un lien vers le fichier du jugement et le gain associé.

Concernant la partie prédiction, celle-ci n'a pas pu être ajoutée au site faute des codes nécessaires (cf. la section 2.4.6).

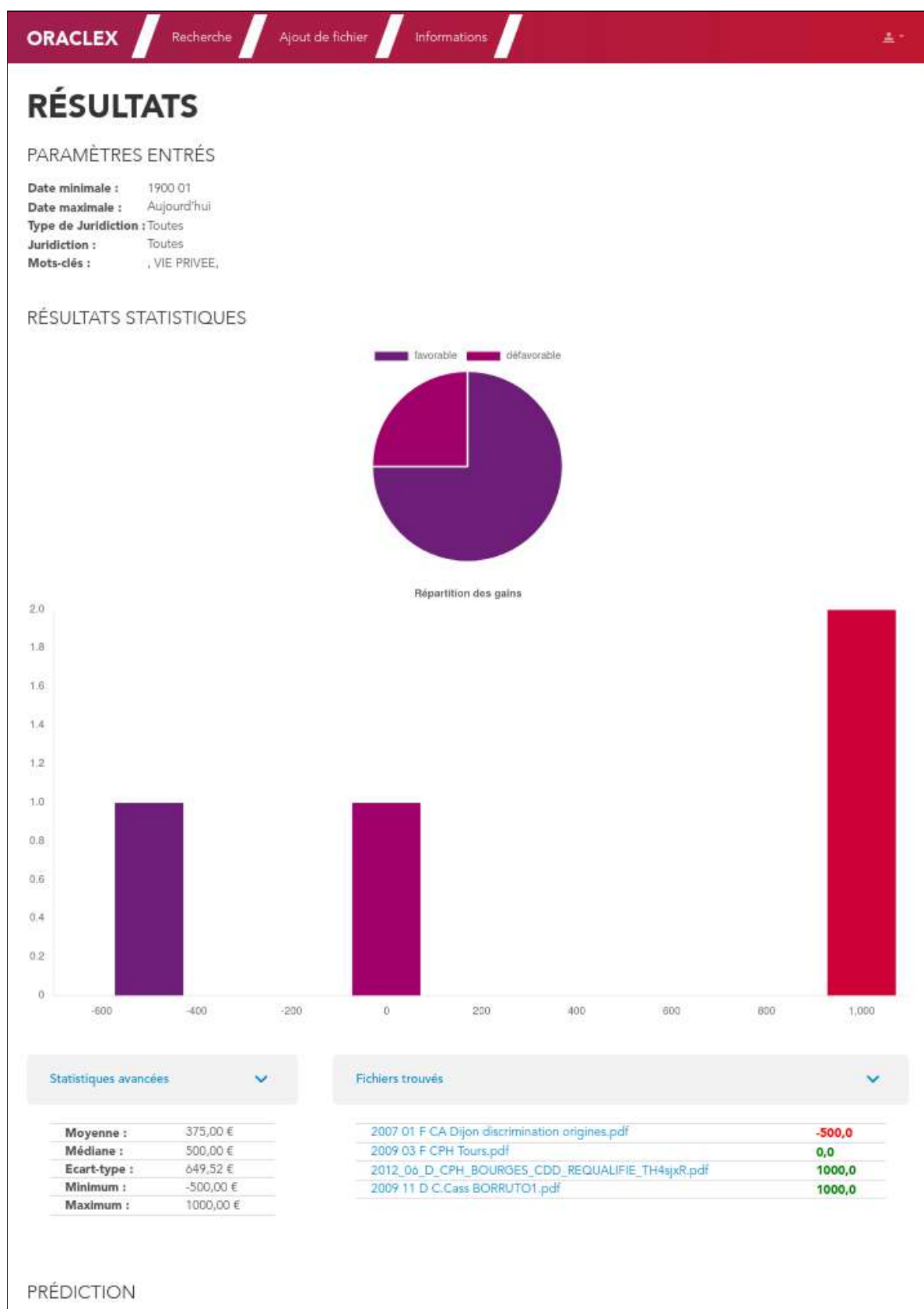


Figure 3.2.2 - Page des résultats de recherche du site

3.3. Ajout de fichiers

Oraclex est pourvu d'un système d'ajout de nouveaux jugements dans la base de données. Le bouton "Parcourir" ouvre l'explorateur de fichiers du système et permet de sélectionner un ou plusieurs fichiers (pour sélectionner plusieurs fichiers, il suffit de rester appuyer sur la touche *Ctrl* lors de la sélection).

Lors de l'importation des fichiers, ceux-ci sont analysés et un tableau récapitulatif affiche les informations clé extraites des jugements sélectionnés. La dernière colonne permet de sélectionner les fichiers que l'utilisateur souhaite conserver avant de valider définitivement l'importation.

Toutefois, comme décrit dans la section 2.4.4, l'analyse des fichiers n'est pas encore opérationnelle, et donc les valeurs clés ne s'afficheront pas pour l'instant.

ORACLEX Recherche Ajout de fichier Informations

AJOUT DE FICHIERS

ÉTAPE 1 : SÉLECTION DES FICHIERS

Fichiers : 1620828338.png

ÉTAPE 2 : VALIDATION DES FICHIERS

Nom	Lisible	Date	Juridiction	Somme gagnée	Conserver
					<input type="checkbox"/>

Application conçue par les élèves du groupes P2E46 de Centrale Nantes. [Page d'administration](#)

Figure 3.3 - Page d'ajout de fichiers du site

3.4. Informations

Enfin, Oraclex dispose d'une page d'informations affichant plusieurs données utiles :

- les informations de contact du responsable actuel du site ;
- l'état de la base de données :
 - nombre de jugements,
 - nombre de jugements lisibles, et
 - nombre de jugements illisibles ;
- et enfin la liste des fichiers illisibles, avec un lien permettant de les remplacer par une version améliorée via une redirection vers la page d'ajout de fichiers.

ORACLEX Recherche Ajout de fichier Informations

INFORMATIONS

REPOSABLE DU SITE

- **Franck Samson**
- Téléphone : 0123456789
- Email : franck.samson@sncf.fr

ÉTAT DE LA BASE DE DONNÉES

- Nombre total de Jugements enregistrés : **4**
- Nombre de Jugements lisibles : **0**
- Nombre de Jugements illisibles : **4**

FICHIERS ILLISIBLES À REMPLACER

Fichier	Date d'import	Modifier
2012_06_D_CPH_BOURGES_CDD_REQUALIFIE_TH4sXR.pdf	8 mai 2021 22:54	Remplacer
2007_01_F CA Dijon discrimination origines.pdf	14 mai 2021 11:44	Remplacer
2009_03_F CPH Tours.pdf	14 mai 2021 11:44	Remplacer
2009_11_D C.Cass BORRUTO1.pdf	14 mai 2021 11:53	Remplacer

Application conçue par les élèves du groupes P2E46 de Centrale Nantes. [Page d'administration](#)

Figure 3.4 – Page d'informations sur le site et sur la base de données

Conclusion

Pour conclure, l'interface développée répond plutôt bien au cahier des charges imposé : son format de site web permet de la rendre compatible avec tous les appareils, un système d'authentification sécurise son accès, et enfin l'interface elle-même permet de manipuler les jugements stockés dans la base de données.

Néanmoins, quelques fonctionnalités importantes, telles que l'ajout de fichier ou la prédiction, n'ont malheureusement pas pu être implémentées faute de temps. En effet, la constitution du cahier des charges et le choix des outils les mieux adaptés pour construire l'interface ont été particulièrement longs, nous laissant ainsi peu de temps pour le développement lui-même. Nous ne regrettons cependant pas d'avoir pris ce temps, car cela nous a permis de construire une interface robuste et claire, ce qui permettra au projet d'être facilement repris et amélioré dans le futur.

D'autres difficultés sont aussi apparues au cours du projet : parfois externes et indépendantes de notre volonté, comme pour la récupération des codes sources auprès du précédent P2E par exemple, mais aussi internes avec les difficultés pour apprendre et manipuler un nouveau langage en très peu de temps.

De nombreuses améliorations pourront encore être ajoutées à l'interface, à commencer par le support de la recherche par mots-clés, de l'ajout de fichier, et de la prédiction de résultats sur un jugement. Une réflexion sur la gestion des utilisateurs pourra aussi être menée, et tout le déploiement du site sur un serveur de production est encore à réaliser. De nouvelles fonctionnalités pourront aussi être développées, telles que le support de la recherche par ville, un ajout plus simple et intuitif des mots-clés, ou encore une meilleure reconnaissance de caractères pour le décryptage des jugements en s'appuyant sur les données déjà récoltées.

Finalement, le projet qui nous a été confié a beaucoup avancé et a répondu à la plupart des exigences, même si plusieurs fonctionnalités importantes manquent à l'appel. Nous aurions certes aimé avoir un peu plus de temps pour aller plus loin sur le développement du site, mais la limite temporelle fait partie des lois du P2E ; après tout, comme l'a dit un certain Ulpien : *Dura lex, sed lex !*

Remerciements

Nous tenions aussi à remercier les personnes suivantes pour leur aide dans ce projet étudiants-entreprise :

- Franck SAMSON, notre correspondant entreprise, pour avoir consacré beaucoup de temps à répondre à nos questions à toutes les réunions, par mail et même par téléphone ;
- Sarah BONNAMY, du service juridique de la SNCF aussi, pour nous avoir éclairé sur le fonctionnement de la justice en France afin de construire une interface la plus adaptée possible à leurs besoins ;
- Olivier ROUX, notre tuteur école, pour nous avoir suivi tout au long de ce projet et guidé sur la gestion d'un tel projet ;
- Henrique TOMAZ AMORIM, Aurèle HAINAUT, Orson JAY, Paul MARTEL, Emeric CLAUDEL, et Stephen JAUD, étudiant·e·s du P2E53 de 2019, pour avoir réalisé la première partie de ce projet et accepté de nous transmettre tout leur travail afin que nous puissions réaliser le nôtre.

Table des illustrations

Figure 2.2 - Maquette de l'interface	7
Figure 2.4.1 - Page d'accueil du site	11
Figure 2.4.2 - Modèle entité-association de la base de données	12
Figure 2.4.5.1 - Champs de recherche par date	14
Figure 2.4.5.2 - Champs de recherche par juridiction	14
Figure 2.4.5.3 - Champs de recherche par mots-clés	15
Figure 3.1.1 - Page de connexion du site	18
Figure 3.1.2 - Menu de l'utilisateur	19
Figure 3.2.1 - Page de lancement de recherche du site	19
Figure 3.2.2 - Page des résultats de recherche du site	21
Figure 3.3 - Page d'ajout de fichiers du site	22
Figure 3.4 - Page d'informations sur le site et sur la base de données	23

Bibliographie

[1] MDN Web Docs - Apprendre le développement web > Programmation de Sites Web côté serveur > Django Web Framework (Python) : <https://developer.mozilla.org/fr/docs/Learn/Server-side/Django>

[2] Django - Documentation de Django : <https://docs.djangoproject.com/fr/3.2>

[3] Bootstrap : <https://getbootstrap.com>

[4] SNCF Bootstrap - Design métier : <https://designmetier-bootstrap.sncf.fr/fr>

Annexes

Table des annexes jointes au présent document

- **Annexe 1 : Rapport de projet du P2E précédent**
- **Annexe 2 : Planning organisationnel**
- **Annexe 3 : Schéma du fonctionnement de l'interface**
- **Annexe 4 : Maquette du site web**
- **Annexe 5 : Modèle Entités-Associations de la Base de Données**

Annexes

Annexe 1 : Rapport de projet du P2E précédent

Annexe 2 : Planning organisationnel

Annexe 3 : Schéma du fonctionnement de l'interface

Annexe 4 : Maquette du site web

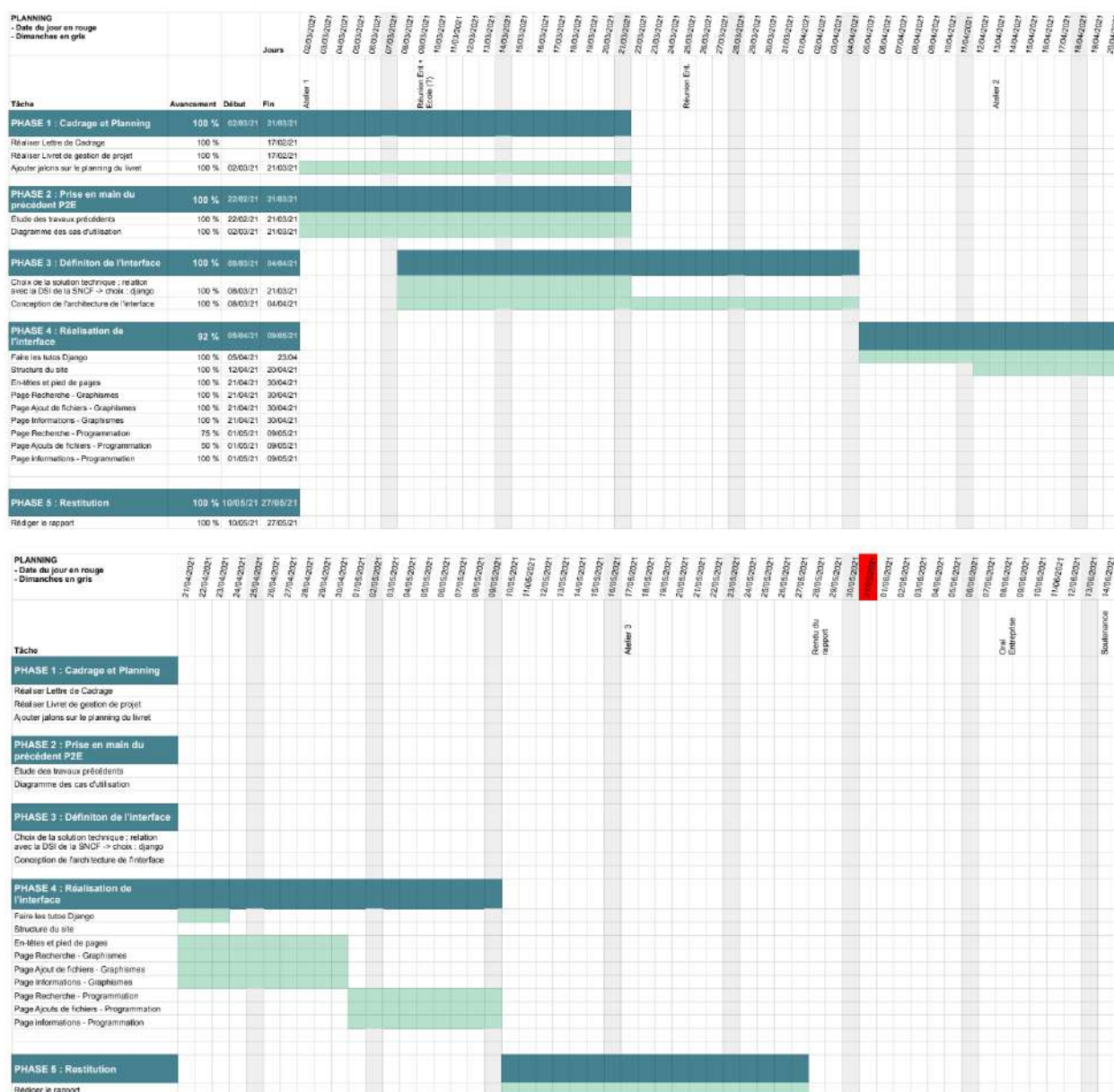
Annexe 5 : Modèle Entités-Associations de la Base de Données

Remarque : les documents annexes sont aussi disponibles dans le dossier accompagnant ce fichier, afin de les avoir en meilleure qualité.

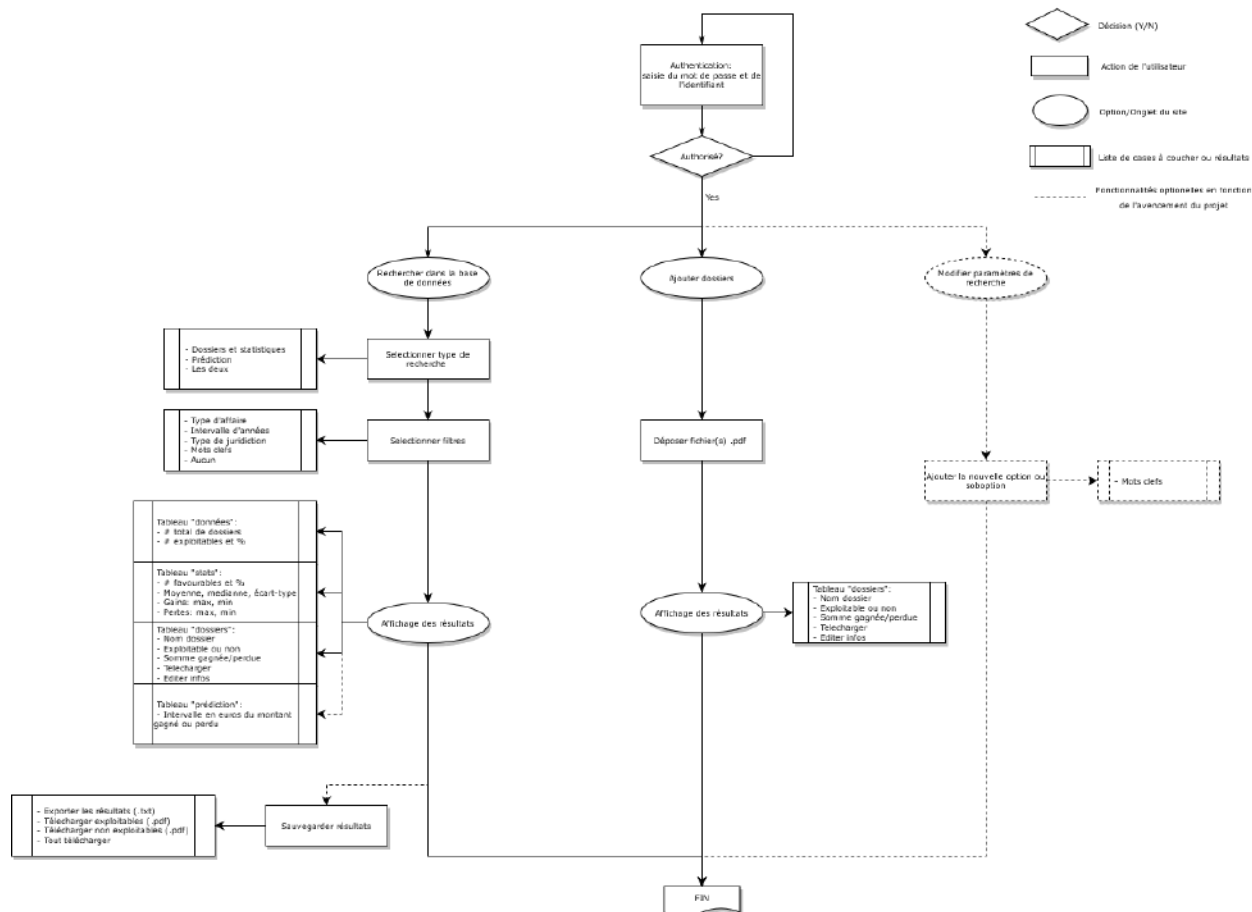
Annexe 1 : Rapport de projet du P2E précédent

Cette annexe n'est pas retranscrite ici, s'agissant d'un dossier complet.

Annexe 2 : Planning organisationnel




Annexe 3 : Schéma du fonctionnement de l'interface





Annexe 4 : Maquette du site web

Page 1 - Recherche

Oraclex

Recherche Ajout de fichiers Administration

Paramètres de recherche  

Mots clés :

Mot clé 1^x Mot clé 2^x Mot clé 3^x Mot clé 4 Mot clé 5

Mot clé 4 Mot clé 5



Juridictions :


☒ Cour de cassation ☐ Cour d'appel ☐ Conseil de prud'hommes


Toutes
Alsace
Rhône-Alpes

Bourg-en-Bresse
☒ Annecy
Grenoble
Lyon

Dates : du au

 Lancer la recherche  Faire une prédiction

Résultats de la recherche 



Min - 142 € Moyenne + 742 € Max + 1142 €


Fichiers trouvés (42)
Fichier 1
Fichier 2
Fichier 3

Statistiques
Moyenne :
Médiane :
Écart-type :

Résultats de la prédiction

Intervalle du gain estimé :


Page 2 - Ajout de fichiers

Oraclex 

Recherche

Ajout de fichiers

Administration

Ajouter des fichiers 

Nom	Lisible	Date	Juridiction	Somme gagnée	Conserver
Fichier1.pdf	Oui	04 / 2021	Cour d'appel de Nantes	+ 50 €	<input checked="" type="checkbox"/>
Fichier2.pdf	Non	42 / 1&A@	Cour de Cassation de Jupiter	- 42 E	<input type="checkbox"/>
Fichier3.pdf	Oui	12 / 2001	Conseil des Prud'Hommes de Nantes	- 999999 €	<input type="checkbox"/>

Valider

Annuler

Remplacer des fichiers

Pour remplacer des fichiers illisibles, rendez-vous à la section [Fichiers illisibles](#) de l'onglet Administration.

Page 3 - Informations/Administration

Oraclex 

Recherche

Ajout de fichiers

Administration

Administrateur

- Métégat Sam
- 07 00 00 00 00
- sammétégat@gmail.com

État de la base de donnée

La base contient 1000 décisions judiciaires dont 958 sont lisibles et 42 illisibles.

Fichiers illisibles

fichier_de_merde.pdf

[Remplacer](#)

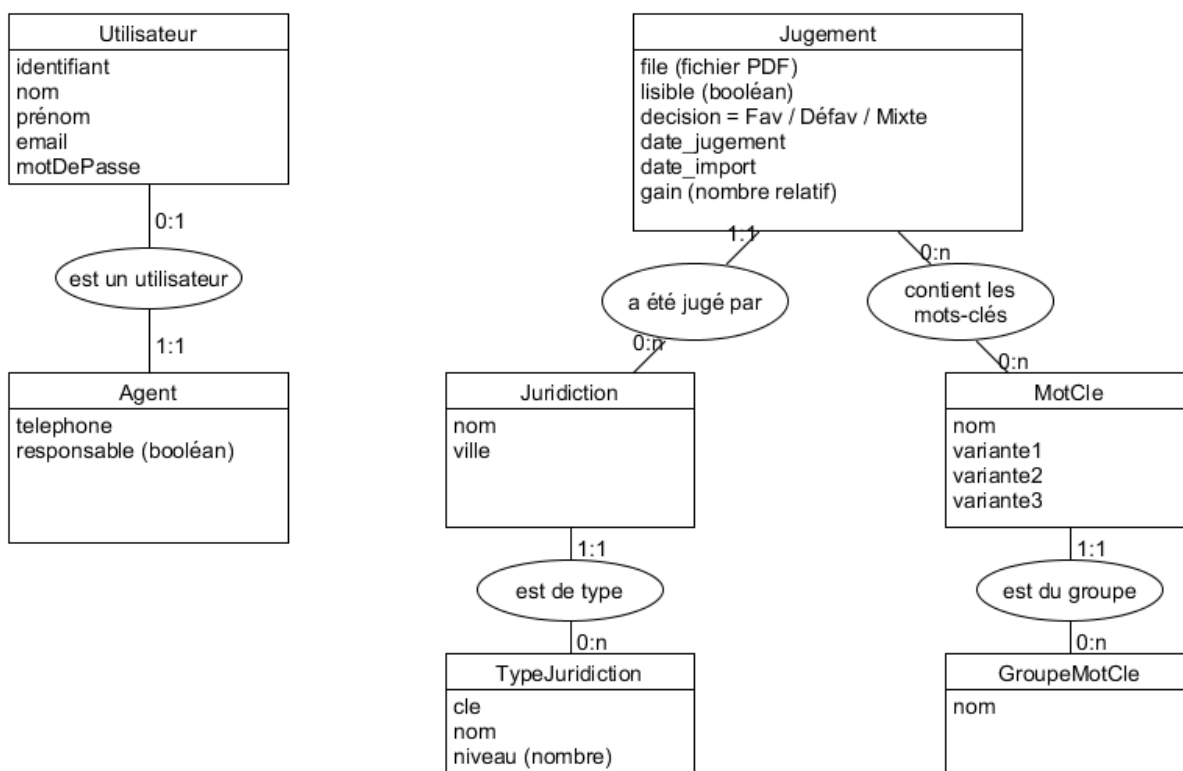
un_autre.pdf

[Remplacer](#)

allé_un_dernier.pdf

[Remplacer](#)

Annexe 5 : Modèle Entités-Associations de la Base de Données





PROJET ETUDIANT D'ENTREPRISE
Rapport d'étude

Outil de justice prédictive en droit social

Groupe P2E

TOMAZ AMORIM Henrique
HAINAUT Aurèle
JAY Orson
MARTEL Paul
CLAUDEL Emeric
JAUD Stephen

Correspondant Entreprise

SAMSON Franck

Tuteur Enseignant

LE BRIZAUT Jean-Sébastien

SOMMAIRE

1.	PRESENTATION DU PROJET DANS SON CONTEXTE.....	3
1.1.	<i>L'Entreprise.....</i>	3
1.2.	<i>Le Projet.....</i>	3
1.3.	<i>Le contexte et les contraintes.....</i>	3
2.	DEMARCHE.....	5
2.1.	<i>Conversion de fichiers.....</i>	5
2.2.	<i>Fichier résumé.....</i>	6
2.2.1.	Liste des informations à collecter.....	6
2.2.2.	Détails sur l'extraction des données.....	7
2.2.2.1.	<i>Extraction du nom du fichier.....</i>	7
2.2.2.2.	<i>Extraction du jugement.....</i>	7
2.2.2.3.	<i>Extraction de la date.....</i>	7
2.2.2.4.	<i>Extraction de la juridiction.....</i>	7
2.2.2.5.	<i>Extraction des mots-clés.....</i>	8
2.2.2.6.	<i>Extraction de la somme perdue ou gagnée.....</i>	8
2.2.3.	Détails concernant l'écriture des informations dans le fichier résumé.....	10
2.2.4.	Synthèse.....	10
2.3.	<i>Réseau Neurone.....</i>	10
2.3.1.	Le neurone.....	11
2.3.2.	L'association des neurones.....	11
2.3.3.	L'entraînement du réseau.....	13
2.4.	<i>Extraction d'informations à partir du fichier résumé.....</i>	16
2.4.1.	Les options de recherche.....	16
2.4.1.1.	<i>La liste de mots clés.....</i>	16
2.4.1.2.	<i>La date limite.....</i>	16
2.4.1.3.	<i>Une juridiction.....</i>	16
2.4.2.	Les résultats de la recherche.....	16
2.4.2.1.	<i>Les sommes perdues.....</i>	17
2.4.2.2.	<i>La qualité de la base données.....</i>	17
2.4.3.	Exemple d'utilisation.....	17
2.5.	<i>Critère de lisibilité.....</i>	19
2.5.1.	Première approche.....	19
2.5.2.	Deuxième approche.....	20
2.6.	<i>Interface graphique.....</i>	21
2.6.1.	Réflexion sur l'élaboration et le choix d'une interface.....	21
2.6.2.	Réflexion sur la sécurité du site.....	21
2.6.3.	Réflexions sur la forme de l'interface.....	21
2.6.4.	Implémentation du site web.....	22
3.	RESULTATS OBTENUS.....	24
3.1.	<i>Résultat de la réseau.....</i>	24
3.2.	<i>Interface graphique.....</i>	25
4.	CONCLUSION.....	26
4.1.	<i>Considérations pour l'amélioration.....</i>	26
5.	TABLE DES ILLUSTRATIONS.....	27
6.	BIBLIOGRAPHIE.....	28

1. Présentation du projet dans son contexte

1.1. L'Entreprise

La Société nationale des chemins de fer français (SNCF) est l'entreprise ferroviaire publique française, officiellement créée par convention entre l'État et les compagnies de chemin de fer préexistantes, en application du décret-loi du 31 août 1937.

La SNCF est l'un des premiers groupes mondiaux de mobilité et de logistique, avec une présence dans 120 pays, 30,5 milliards d'euros de chiffre d'affaires et 241 000 collaborateurs. Le groupe comprend 23 directions régionales et s'appuie sur cinq branches d'activité :

- SNCF Infra : gestion, exploitation, maintenance et ingénierie d'infrastructure ferroviaire;
- SNCF Proximités : transport public de voyageurs urbain, départemental et régional;
- SNCF Voyages : transport ferroviaire de voyageurs longue distance et à grande vitesse;
- SNCF Geodis : transport et logistique de marchandises;
- Gares & Connexions : gestion et développement des gares.

1.2. Le Projet

L'objectif de notre projet a été de concevoir un outil de justice prédictive en droit social permettant d'estimer le taux de perte d'une affaire en cours, et de quantifier la somme perdue potentielle. De plus, cet outil doit permettre de faciliter l'obtention et l'exploitation des données présentes dans la base des décisions de justices relatives à la SNCF, concernant le secteur du droit social. Cet outil représente un gain de temps considérable pour l'entreprise, sans avoir à changer la structure actuelle de leur base de données.

Le produit fini devait avoir la forme d'une interface relativement facile d'utilisation dans laquelle il faudrait rentrer certains mots-clés et détails de l'affaire permettant une évaluation du risque avec éventuellement un taux de rejet. L'outil doit également être dynamique, lorsqu'un fichier est rentré dans la base de données, il doit être traité immédiatement.

Notre projet s'est donc divisé en plusieurs travaux : d'abord la traduction des fichiers PDF en fichiers TXT pour leur analyse, la recherche des données pertinentes (mots-clés, sommes perdues ou gagnées) dans les décisions de justice, la constitution d'un document résumé qui rassemble toutes les informations obtenues, le développement d'un réseau de neurones donnant les risques de perte pour une affaire, et enfin la mise en place d'une interface graphique pour faciliter l'utilisation de l'outil.

Ces différentes étapes vont maintenant être détaillées dans la suite de ce rapport, où figurent les différentes avancées effectuées, les problèmes rencontrés et les solutions adoptées.

1.3. Le contexte et les contraintes

Le Pôle Juridique de la Délégation Juridique Territoriale Ouest (PJU-DJTO) de la SNCF traite des affaires de Droit Social devant les juridictions prudhommales notamment, mais aussi devant toute juridiction compétente en la matière. Le PJU dispose d'une collection de décisions de justice en format PDF, difficilement exploitable. En effet, afin d'estimer les pertes ou gains potentiels liés à une affaire en cours, il est nécessaire d'examiner les décisions de justice similaires une par une ce qui prend un temps considérable. C'est dans ce contexte qu'il nous a été demandé de travailler sur un outil de justice prédictive qui s'appuierait sur l'ensemble des décisions de justice

existantes. Il permettrait ainsi aux juristes du PJU de mieux apprécier l'issue d'une affaire basée sur les données renvoyées par notre programme.

L'une des contraintes qui a été évoquée est de n'utiliser que des logiciels libres de droit.

Les documents que nous avons à notre disposition ne sont par ailleurs pas soumis à une clause de confidentialité.

2. Démarche

2.1. Conversion de fichiers

La base de donnée d'affaires juridiques était constituée des fichiers du type PDF, dont les pages étaient les pages des documents scannés. On avait donc un vrai problème pour exploiter les données, parce qu'ils n'étaient pas lisibles pour l'ordinateur.

Il fallait convertir les documents PDF en texte (fichier du type txt) pour que l'ordinateur puisse le lire. On a trouvé une bibliothèque en langage Python qui faisait la conversion des images en texte et qui s'appelle Pytesseract.

Basé sur cette nouvelle bibliothèque, on a créé un programme qui prend tous les fichiers PDF localisés dans un dossier et convertit chaque une des pages en images et, ensuite, chacune de ces images en texte. Un exemple d'une conversion d'un morceau de fichier se trouve ci-dessous :

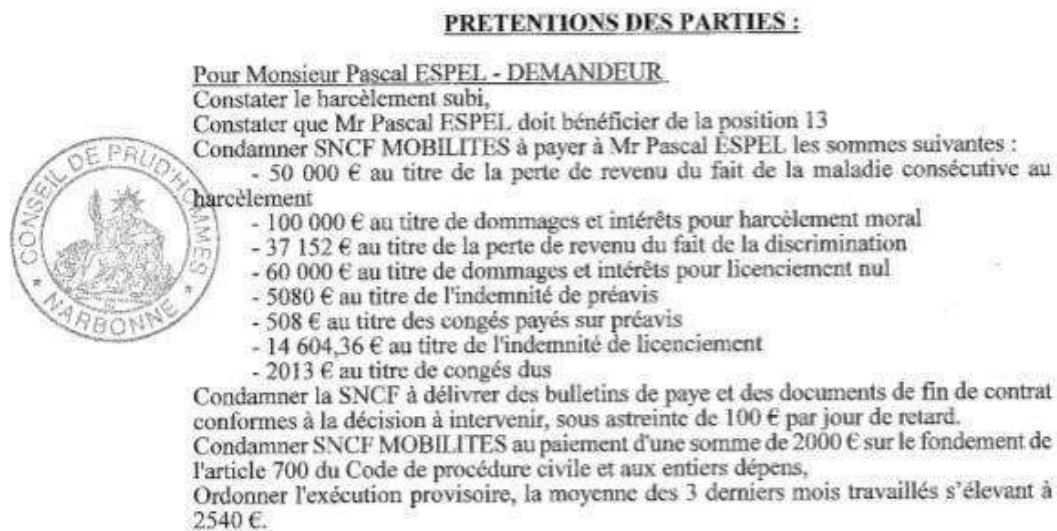


Figure 1 – Morceau d'un fichier PDF

PRETENTIONS DES PARTIES :

Pour Monsieur Pascal ESPEL - DEMANDEUR
 Constaté le harcèlement subi,
 W Constaté que Mr Pascal ESPEL doit bénéficier de la position 13
 WWW Condamner SNCF MOBILITES à payer à Mr Pascal ESPEL les sommes suivantes :
 «* ÎË”;\ – 50 000 € au titre de la perte de revenu du fait de la maladie consécutive au
 » "Ëgrcèlement
 ” – 100 000 € au titre de dommages et intérêts pour harcèlement moral
 – 37 152 € au titre de la perte de revenu du fait de la discrimination
 – 60 000 € au titre de dommages et intérêts pour licenciement nul
 – 5080 € au titre de l'indemnité de préavis
 – 508 € au titre des congés payés sur préavis
 – 14 604,36 € au titre de l'indemnité de licenciement
 – 2013 € au titre de congés dus
 Condamner la SNCF à délivrer des bulletins de paye et des documents de n de contrat
 conformes à la décision à intervenir, sous astreinte de 100 € par jour de retard.
 Condamner SNCF MOBILITES au paiement d'une somme de 2000 € sur le fondement de
 l'article 700 du Code de procédure civile et aux entiers dépens,
 Ordonner l'exécution provisoire, la moyenne des 3 derniers mois travaillés s'élevant à
 2540 €.

Figure 2 – Morceau d'un fichier PDF converti en texte

On peut bien observer que la conversion n'est pas parfaite, car l'algorithme de conversion reconnaît les caractères seulement si le fichier n'a pas d'autres éléments qui le confondent. C'est pour cette raison qu'on a développé un critère pour évaluer la qualité de la conversion et si le fichier devient vraiment exploitable à la fin de tout ce processus.

2.2. Fichier résumé

Après avoir rendu possible l'exploitation du contenu des fichiers de la base de données grâce à la conversion de PDF en fichiers textes, nous avons fait le bilan des informations « utiles » présentes dans chaque fichiers, celles-ci servant à la fois à l'outil de prédiction et à l'outil de recherche de données que nous devons implémenter. En conséquence, il a été décidé de créer un « fichier résumé » qui comporterait chacune des informations utiles de chaque fichiers de sorte que l'exploitation des fichiers .txt ne soit effectuée qu'une fois par le programme. Ce fichier a pour objectif de ne pas avoir à réutiliser des fichiers textes dont les données sont parfois difficiles à collecter.

2.2.1. Liste des informations à collecter

Les informations nécessaires à l'apprentissage de l'outil de prédiction et intéressantes pour les juristes de la SNCF sont les suivantes :

- Nom du fichier dans la base de donnée
- Jugement (Favorable/Défavorable/Mixte)
- Date du jugement final
- Juridiction (Conseil des Prud'hommes, Cours de Cassation, etc)
- Mots-clés permettant de caractériser la nature d'un fichier
- Somme perdue ou gagnée par la SNCF

2.2.2. Détails sur l'extraction des données

Durant tout le paragraphe, la démarche de recherche des informations est la même: la correspondance des motifs [1][2].

En effet, la mauvaise qualité de certains fichiers PDF peut conduire à une mauvaise traduction dans les fichiers textes (voir paragraphe sur la conversion des fichiers).

Par conséquent, la recherche d'informations n'est pas chose aisée et une analyse préalable des fichiers textes a été nécessaire dans le but d'évaluer les types d'erreurs les plus fréquents pouvant être trouvés au sein des textes.

Par chance certaines informations sont contenues dans le nom des fichiers ce qui rend leur extraction facile. Pour le reste des données, nous avons tenté de recenser le plus d'erreurs possible pour proposer une forme générale de motifs à rechercher tout en sachant qu'un motif présentant trop de variations pourrait afficher des données erronées.

La proposition des différents motifs à rechercher sera donc orientée vers la recherche de données valides, tout en sachant que certaines données seront écartées.

2.2.2.1. Extraction du nom du fichier

Le nom du fichier (avec l'extension .pdf) est récupéré grâce au nom des fichiers textes dont on change l'extension.

2.2.2.2. Extraction du jugement

Le jugement (Favorable/Défavorable/Mixte) est présent dans le nom du fichier sous la forme F/D/M. Par conséquent, une simple recherche sur l'une de ces trois lettres suffit à obtenir le jugement.

2.2.2.3. Extraction de la date

Là encore, la date est présente dans le nom du fichier mais sous deux différents formats :

- jj.mm.aaaa
- mm aaaa

Nous recherchons donc un motif correspondant à l'une de ces deux syntaxes dans le nom du fichier.

En ce qui concerne l'écriture dans le fichier résumé, le format de la date est « mm aaaa » pour ensuite l'exploitation des données. Il ne nous pas sembler pertinent de conserver le jour exact du jugement, cette information n'étant pas présente dans le nom de tous les fichiers et étant parfois inexploitable dans le contenu des fichiers.

2.2.2.4. Extraction de la juridiction

La juridiction est normalement indiquée au début du fichier et est souvent indiquée dans le nom du fichier. Pour l'instant, l'extraction de cette information s'effectue uniquement par analyse du nom du fichier. Toutefois, si cette information se révélait être utile pour la SNCF ou pour l'apprentissage de l'outil de prédiction, il serait possible de rechercher cette information dans le fichier dans une version ultérieure du programme, sous réserve que cette information puisse être extraite (contenu du fichier texte exploitable).

2.2.2.5. Extraction des mots-clés

a) Présentation de la liste de mots-clés

Les mots-clés sont des mots présents dans une liste qui nous a été fournie par les juristes de la SNCF. Ils ont deux objectifs :

- Caractériser chaque fichiers permettant aux utilisateurs de l'outil de recherche de rechercher des fichiers contenant certains mots-clés
- Savoir, grâce à l'outil de prédiction, si certains mots-clés permettent de prédire l'issue d'une affaire pour ainsi déterminer, dans le cadre d'une affaire en cours, si cette affaire a des chances d'être « gagnée » ou « perdue » selon les mots-clés qui apparaissent (voir paragraphe sur l'outil de prédiction pour plus de détails sur le fonctionnement).

b) Fonctionnement de la fonction d'extraction des mots-clés

Pour extraire ces données à l'intérieur des fichiers, il a fallu dans un premier temps dresser une liste, la plus complète possible, des différentes erreurs de syntaxe pouvant apparaître dans les fichiers textes.

Une fois cette liste d'erreur dressée, nous avons établi un motif permettant de rechercher les mots-clés dans les différentes pages d'un fichier pour ensuite les écrire dans le fichier résumé (voir fonction « extraction_mc » dans le programme portant sur l'écriture du fichier résumé).

Après avoir collecté tous les mots-clés ayant été repéré dans un fichier, une fonction permet de repérer les doublons pour ne dresser au final qu'une liste dans laquelle les mots-clés n'apparaissent qu'une fois.

c) Possibilités d'amélioration de la recherche de données concernant les mots-clés

Après réflexion avec le groupe de projet, il pourrait être intéressant, dans une version ultérieure du programme, de calculer le nombre d'occurrences de certains mots-clés apparaissant dans un fichier, pour leur donner un impact plus significatif lors de l'apprentissage de l'outil de prédiction. Les mots-clés pour lesquels cette donnée serait pertinente devront au préalable être indiqué par les juristes de la SNCF.

De plus, il n'est pour l'instant pas possible d'ajouter des mots-clés à la liste pré-établie compte-tenu notamment de la structure actuelle du réseau de neurone (voir paragraphe sur l'outil de prédiction).

2.2.2.6. Extraction de la somme perdue ou gagnée

a) Analyse préalable des fichiers txt/pdf

Cette analyse préalable a été nécessaire, toujours dans l'optique de définir un motif qui permettent de couvrir toutes les possibilités de syntaxe pouvant être rencontrées lors du parcours d'un fichier.

Dans un premier temps, nous avons analysé un échantillon représentatif de fichier pdf permettant de dresser une liste exhaustive de syntaxe concernant les sommes d'argent perdues ou gagnées par la SNCF dans le cadre d'une affaire. Cette liste est donnée ci-contre:

XXX.XXX € (« . » entre les milliers)

XXX,XX € (« , » entre euros et centimes)

XXXXXX€ (pas de points entre les milliers)

XXX XXX€ (espace entre les milliers)

XXX euros XX (« euros » avant centimes)

XXX.XX euros (« . » entre euros et centimes)

XXX.XXX,XX (« . » entre les milliers « , » entre euros et centimes)

le nom « euros » peut également se substituer au symbole « € ».

Dans un second temps, une analyse des fichiers txt a été nécessaire dans le but de repérer les possibles erreurs de syntaxe concernant notamment le symbole « € ». En outre les lettres « Ê » et « Ä » se sont régulièrement substituées au symbole « € » ce qui justifie leur présence dans le motif correspondant à une somme d'argent.

b) Recherche des sommes d'argent perdues ou gagnées à la fin d'un jugement

Suivant la recommandation des juristes de la SNCF, il a été mis en évidence que les sommes d'argent qui devaient être versées à l'issue d'une affaire étaient mentionnées dans une section suivant la mention « Par ces motifs ».

Par Conséquent, la recherche des sommes d'argent s'effectuent nécessairement après lecture de la mention « Par ces motifs » dans les fichiers textes. Le motif correspondant à cette mention prend en compte les éventuelles erreurs de syntaxe repérées dans les fichiers textes.

Note : Certains fichiers ne contiennent pas la mention « Par ces motifs » (voir fichier « 2010 05 F C.Cass NIERENGARTEN.pdf » dans le dossier « Discrimination » de la base de donnée de la SNCF) ou cette mention est désignée sous une autre appellation (par exemple « Décision »). Ces cas ne sont pour l'instant pas pris en charge par le programme. Cependant, la majorité des fichiers présentent la mention « Par ces motifs », ce qui permet d'avoir à priori un bon échantillon de fichier dont l'extraction de la somme est possible.

Si cette mention n'est pas trouvée ou si le critère de lisibilité (voir paragraphe sur le critère de lisibilité) la somme d'argent est absente du fichier résumée (fichier indiqué comme non lisible et nombre -1 à la place de la somme d'argent)

Ensuite, il s'agit de repérer les sous-sections dans lesquelles apparaissent les sommes d'argent.

Pour cela, on se base sur la convention adoptée dans les fichiers de décisions de justice concernant les condamnations : Chaque condamnation commence par le mot « condamne » (Note: certaines exceptions subsistent toutefois, mais en minorité).

De ce fait, on recherche chaque sous-sections commençant par le mot condamne (le motif tient compte des erreurs de syntaxe trouvées dans les fichiers txt).

En procédant ainsi, on peut extraire chaque somme d'argent de chaque sous-sections et savoir quel parti est condamné dans chaque sous-sections (voir le paragraphe ci-dessous concernant la qualification d'une somme perdue ou gagnée).

Puis après avoir extrait une expression, on s'assure grâce à une fonction externe que l'expression relevée est bien une somme d'argent (en effet un sous-motif du motif du recherche de la somme d'argent est un nombre quelconque, ce problème étant due à la forme très générale du motif).

Enfin, on convertit grâce à une fonction externe la chaîne de caractère obtenue en un nombre correspondant à la somme perdue ou gagnée par la SNCF.

c) Détermination du parti gagnant ou perdant

Pour savoir si une somme est perdue ou gagnée par la SNCF, on recherche dans chaque sous-section faisant intervenir le mot « Condamne », qui de la SNCF ou du civil apparaît en premier dans la structure de la phrase (exemple : « Condamne la SNCF à payer à Mr.XX la somme de XXX € »). On peut ainsi déterminer si la somme est perdue ou gagnée par la SNCF. Une liste de différents synonymes de « SNCF » nous a été fournie par l'entreprise dans le but de couvrir le plus de situations possible.

Par défaut, la somme est comptée négativement (somme perdue par la SNCF) car dans la plupart des cas, la SNCF ne gagne pas d'argent à l'issue d'une affaire.

2.2.3. Détails concernant l'écriture des informations dans le fichier résumé

Pour écrire les informations collectées dans chaque fichiers, on utilise une matrice donc chaque ligne représente un fichier et chaque colonne représente une des informations citée ci-dessus.

On écrit ensuite les informations contenues dans la matrice dans un fichier texte.

Ce mode de fonctionnement permet d'éviter le problème de l'ouverture « aléatoire » des fichiers textes contenus dans le répertoire.

2.2.4. Synthèse

En conclusion, la méthode adoptée permet d'extraire une grande partie des données utiles pour l'élaboration des outils de recherche et de prédiction bien que certains cas détaillés ci-dessus ne puissent pour l'instant être pris en charge par le programme.

Cependant, le problème principal reste la qualité des fichiers textes, rendant l'extraction de la somme d'argent difficile, cette donnée étant cruciale pour l'outil de prédiction ou pour les employés de la SNCF qui utiliseraient l'outil. Par conséquent, il convient d'insister sur la nécessité de rentrer des fichiers PDF de bonne qualité dans la base de données et de respecter au mieux la structures des documents détaillées ci-dessus, sur laquelle repose une majorité de la méthode d'extraction. D'autre part, une amélioration de l'outil de traduction des fichiers PDF en fichiers txt est également recommandée, toujours dans le but d'obtenir un maximum de données exploitables.

2.3. Réseau Neuronne

D'après explications précédents, il faut développer un logiciel de justice prédictive. Nous allons un ressource informatique très efficace pour travailler avec les prédictions basées sur une base de données existant.

Les réseaux neuronne sont très connus actuellement car elles représentent le coeur de l'intelligence artificielle. Nous allons exploiter un peu de la structure que nous avons utilisé pour proposer une solution pour le problème de prédictions sur les cas de discrimination.

2.3.1. Le neurone

Le neurone est la cellule la plus basique d'un réseau neurone. En effet, il représente une abstraction. Un neurone peut être actif – quand il retourne une valeur proche de 1 – ou inactif – quand sa sortie est proche de zéro. Mathématiquement, le neurone a comme entrée une fonction linéaire et, comme résultat, il retourne un nombre entre zéro et un qui est obtenu à partir de son entrée.

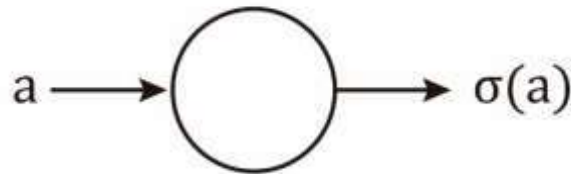


Figure 3 – Schéma d'un neurone

Sur la figure précédente la sortie est définie comme étant $\sigma(z)$, où z est l'entrée (fonction linéaire). La fonction σ est telle que $\sigma(z) \in [0,1]$ et est appelée fonction d'activation. Pour notre réseau nous avons utilisé une fonction très courante, la sigmoïde définie ci-dessous :

$$\sigma(z) = \frac{1}{1+e^{-z}} \quad (1)$$

Finalement, c'est possible de créer une association de multiples neurones, comme il sera exposé ensuite.

2.3.2. L'association des neurones

Le réseau neurone est constitué de plusieurs neurones interconnectés comme illustré ci-dessous :

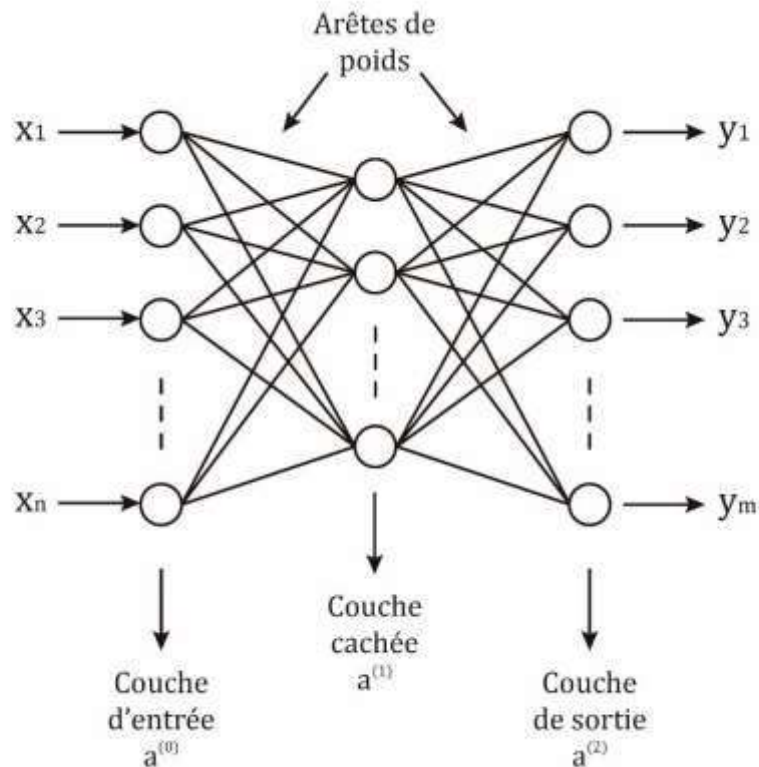


Figure 4 – Schéma d'une réseau neuronne

La première couche (couche d'entrée) recevra les données sur lesquels le réseau va opérer les calculs pour fournir une réponse à un problème posé, qui correspond au résultat obtenu à la dernière couche.

On considère que les couches sont nommées selon la notation $a^{(l)}$, avec $l \in \{0, 1, 2, \dots, L\}$. Comme spécifié avant, l'entrée de chaque neurone sera une fonction linéaire qui dépend des sorties de la couche précédent. En considérant $w_{jk}^{(l)}$ comme l'arête de poids entre les neurones $a_k^{(l-1)}$ et $a_j^{(l)}$ et $b_j^{(l)}$ un scalaire qui sert à fonction de *offset*, l'entrée $z_j^{(l)}$ du j-ème neurone de la l-ème couche s'écrit (pour $l \neq 0$) :

$$z_j^{(l)} = \sum_{k=1}^n w_{jk}^{(l)} a_k^{(l-1)} + b_j^{(l)} \quad (2)$$

Basé sur cette idée, on peut écrire la sortie d'une couche $a^{(L)}$ sur une notation vectoriel :

$$\vec{a}^{(l)} = \sigma(\vec{z}^{(l)}) = \sigma(W^{(l)}\vec{a}^{(l-1)} + \vec{b}^{(l)}) \quad (3)$$

Cette notion s'applique de manière récursive sur les plusieurs couches existantes.

Pour notre problème, on a adopté seulement trois couches, car la solution ne traite pas une grande quantité de données. La configuration adopté pour la couche initiale d'entré a été d'une centaines de neurones où chacun correspondait à la présence ou absence d'un certain mots clé pour le fichier analysé. En résumé, si un fichier possède un certain mot, le neurone pertinent sera actif, sinon il sera inactif.

La configuration de la couche intermédiaire est arbitraire et choisie selon le critère de précision que nous voudrions avoir dans les réponses.

Finalement, pour la couche finale de sortie, nous avons considéré une dizaine de neurones où chacun correspond à une gamme de perdre d'argent. C'est à dire que si un tel neurone est actif, il veut dire que l'entreprise perdra une somme d'argent d'une ordre de grandeur spécifique, comme entre 1000 et 5000 euros, par exemple. D'après cette logique, il faut que seulement un neurone soit actif à la fois.

Un soucis qui apparaît naturellement dans ce contexte est comment définir les plusieurs paramètres du réseau (c'est à dire $w_{jk}^{(L)}$ et $b_j^{(L)}$). C'est pour cette raison que nous avons besoin des données précédents. Il sont importants pour *entraîner* le réseau d'une manière tel que la sortie soit la plus précise possible pour les cas que nous connaissons déjà la réponse. Ensuite, nous parlerons de l'entraînement du réseau et comment l'effectuer.

2.3.3. L'entraînement du réseau

On considérera qu'on a à notre disposition un conjoint des données assez grande, organisé d'une manière telle que l'entrée \vec{x} , correspondant à liste de valeurs indiquant la présence ou absence de mots clés, donne la réponse \vec{y}^* , liste des valeurs (situées entre zéro et un) qui indique l'ordre de grandeur de la somme d'argent perdu.

Ensuite, il est possible de définir une fonction d'erreur calculée entre la sortie donnée actuellement pour le réseau ($\vec{a}^{(L)}$) et la réponse correcte :

$$C = \frac{1}{2} \|\vec{a}^{(L)} - \vec{y}^*\|^2 = \frac{1}{2} \sum_{i=1}^m (a_i^{(L)} - y_i^*)^2 \quad (4)$$

Comme on a que la sortie \vec{y} dépend des paramètres du réseau (w et b), la fonction d'erreur C est aussi fonctions de ces paramètres. On peut, donc, la minimiser pour qu'on puisse avoir la réponse la plus précise pour les donnés qu'on a. Puisqu'elle est une fonction de plusieurs variables, la méthode la plus approprié pour trouver un minimum local est la méthode du gradient descendant.

A partir des expressions ci-dessus et basé sur Michael Nielsen (2018) [3], on déduit les dérivés suivantes :

$$\delta_j^{(L)} = (a_j^{(N)} - y_j^*) \sigma'(z_j^{(L)}) \quad (5)$$

$$\delta_j^{(l)} = \left((W^{(l+1)}) \delta_j^{(l+1)} \right)_j \sigma'(z_j^{(l)}) \quad (6)$$

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = a_k^{(l-1)} \delta_j^{(l)} \quad (7)$$

$$\frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)} \quad (8)$$

Avec les expressions ci-dessus, on peut calculer le gradient de C par rapport aux paramètres du réseaux (∇C). La méthode du gradient descendant nous dit maintenant qu'il suffit de soustraire le gradient de paramètres initiales multiplié pour un numéro h (appelé *learning rate* et assez petit pour assurer qu'on ne dépasse pas le minimum) jusqu'à moment où on atteint un point proche du minimum local.

À la pratique, le critère qu'on a utilisé pour finir le calcule est l'échelle de C , c'est à dire qu'il faut continuer le processus jusqu'au moment où C est moins grand qu'un limite déterminé (ϵ dans

notre cas) et aussi le nombre d'itérations (s'il est plus grand qu'un limite fixé, il faut arrêter forcément).

Tous les concepts présentés ci-dessus sont définies basé sur une seule entrée, un seul donnée. Cependant, dans la réalité, pour vraiment entraîner la réseau pour qu'elle fasse de prédictions réalistes, il faut faire cet entraînement sur une base de données assez grande.

On utilisera encore la méthode du gradient descendant, mais on va calculer ce gradient pour un petit ensemble de données de notre base et on prendra le gradient moyen pour ce petit ensemble de données. On fera l'entraînement en prenant un séquence de petits ensembles jusqu'on ait a convergence du processus.

Pour écrire l'algorithme, on introduira l'opération suivante entre deux vecteurs de même taille :

$$\vec{a} \odot \vec{b} = (a_1 b_1 \ a_2 b_2 \ : \ a_n b_n) \quad (9)$$

En forme d'algorithme, on a :

```
// Définition aleatoire initiale des paramètres
pour k allant de 1 à L faire
    W(k) = W_initial_aleatoire
    b(k) = b_initial_aleatoire
    // W et b sont matrices
fin pour

fonction sigmoid (X) :
    renvoyer 1/(1+e^(-X))

fonction sigmoid_derive (X) :
    renvoyer (sigmoid(X))*(1- sigmoid(X))

fonction forwardPropagation ( a(0) ) :
//Calcul des couches
pour k allant de 1 à L faire
    Z(k) = W(k)*a(k-1)+b(k)
    a(k)= sigmoid(Z(k))
fin pour

fonction erreur (X, y*) :
// Calcul d'écart C en utilisant  $a_i^{(L)}$  qui dépend des paramètres de
// la réseau. Variable prediction garde la réponse de la réseau
// pour l'entrée X(i)

    C = 0

    pour i allant de 0 à n-1 faire
        prediction = forwardPropagation ( X(i) )
        C = C +  $\frac{1}{n} ||prediction - y^*(i)||^2$ 
```



```

    fin pour
renvoyer C

fonction backPropagation(X, y*, bs, W, b, Z, a, h) :

pour chaque X, y* faire
    a(0) = X
    delta(L) = (a(L)-y*)  $\odot$  sigmoid_derive( Z(L) )

    gradB(L) = deltaL
    gradW(L) = a(L-1)*deltaL

    pour k allant de L-1 à 1 faire
        delta(k) = (W(k-1)*delta(k+1))  $\odot$  sigmoid_derive( Z(k) )
        gradB(k) = delta(k)
        gradW(k) = a(k-1)* delta(k)
    fin pour

    pour k allant 1 à L faire
        W(k) = W(k) - (h/bs)*gradW(k)
        b(k) = b(k) - (h/bs)*gradB(k)
    fin pour
fin pour

renvoyer W, b

// Entraînement
// Soit X le vecteur de données d'entrée
// Soit y* le vecteur de données de sortie
// Soit bs la taille du petit ensemble de données sur lequel on //
va entraîner la machine
X = (X0, X1, ... , X(n-1))
Y* = (y*(0), y*(1), ... , y*(n-1))

tant que ( C >  $\epsilon$  et nb_iteration < LIMITE ) faire
    pour k allant de 0 à (n-n%bs)/bs faire
        // La ligne ci-dessous change W et b plusieurs fois
        // à chaque petit ensemble de données
        // La notation est : X[k:k+p] = (Xk, X(k+1), ..., X(k+p))

        W,      b      =      backPropagation(X[k*bs :      k*bs+bs],
        y*[k*bs : k*bs+bs], bs, W, b, Z, a, h)
    fin pour
    C = erreur(X, y*)
fin tant que

```

Fin algorithme

2.4. Extraction d'informations à partir du fichier résumé

Le fichier résumé condense les informations pertinentes des différentes décisions de justice. De plus, grâce à sa structure pensée dans cette optique, il est possible d'automatiser des tâches simples qu'un humain pourrait faire en lisant une par une les différentes décisions de justices. Ainsi, nous avons développé un outil de recherche qui donne des statistiques pertinentes sur des décisions de justices répondant à certains critères que doit définir l'utilisateur (un juriste de la SNCF par exemple).

2.4.1. Les options de recherche

Pour présenter cet outil il convient de préciser quels informations doit donner l'utilisateur à l'outil. Ces informations correspondent aux critères que devra respecter une décision de justice pour être pris en compte par l'outil :

- 1) Une liste de mots clés
- 2) Une date limite (facultatif)
- 3) Une juridiction (facultatif)

2.4.1.1. La liste de mots clés

Cette liste est l'information 'principale' que devra renseigner l'utilisateur. Elle correspond simplement à des mots spécifiques qui doivent apparaître dans les décisions de justice présent en compte.

2.4.1.2. La date limite

Au travers des échanges avec des juristes de la SNCF, nous avons mis en évidence le fait que les lois et la juridiction évoluent au cours du temps. Ainsi, il ne serait pas très pertinent de prendre en compte certaines décisions de justice datant d'avant une certaine date limite.

Précisons que nous avons décidé que ce critère soit facultatif.

2.4.1.3. Une juridiction

Ce critère permet de préciser dans quelle juridiction les décisions de justices doivent se placer pour être pris en compte.

Précisons encore une fois que ce critère est facultatif.

2.4.2. Les résultats de la recherche

C'est en discutant avec des juristes de la SNCF que nous avons établi les différentes informations que souhaiterait connaître un utilisateur de cet outil. Nous avons établi 2 axes de renseignement : les sommes perdues et la qualité de la base de données.

2.4.2.1. Les sommes perdues

L'outil va ainsi donner des statistiques simples (médiane, minimum, maximum, moyenne et écart-type) sur les sommes perdues des décisions de justice répondant aux critères de recherche. Également, le pourcentage de cas favorable parmi les dossiers sélectionnés est donné. Tous ces chiffres permettent de donner une première idée de la répartition des pertes :

- La **moyenne** donne une prédiction grossière des pertes d'une affaire qui se place dans le même contexte que celui de la recherche effectuée.
- l'**écart-type** correspond à la précision de la prédiction par la moyenne (plus il est petit par rapport à la moyenne plus sûre sera la prédiction)
- Le **minimum** et le **maximum** renseignent sur l'amplitude de la distribution des pertes
- La **médiane** joue un rôle similaire à la moyenne mais n'est pas beaucoup influencée par des valeurs exceptionnelles.

En complément de ces chiffres, l'outil retourne le nom de tous les fichiers répondants aux critères.

2.4.2.2. La qualité de la base données

Pendant la conduite du projet on a été confronté au problème de la mauvaise qualité de certains dossiers qui aboutit à l'incapacité complète ou partielle d'extraire les informations du dossier vers le fichier résumé. Ce phénomène étant un aspect non négligeable (~40% des dossiers) de notre étude, nous avons décidé de renseigner les dossiers non pris en compte pour le calcul des statistiques pour que l'utilisateur ne puisse pas passer à côté d'une décision de justice potentiellement très pertinente (énorme perte par exemple). Nous avons également ajouté le pourcentage de dossiers non utilisés.

Précisons que nous ne prenons pas en compte les dossiers jugés de mauvaise qualité par notre critère de lisibilité car même si certaines informations ont tout de même pu être extraites (comme certains mots-clés par exemple) il est fort probable que la somme perdue détectée soit mauvaise, ce qui fausserait les statistiques.

2.4.3. Exemple d'utilisation

L'interaction avec l'utilisateur est une question traitée dans la partie sur l'interface graphique. Ici ne sera présenté que le résultat brut d'une recherche effectuée avec le seul mot-clé "discrimination".

```

mediane : -2500.0 €
moyenne : -8625.105 €
ecart-type : 18498.15041886716 €
perte maximale : 72235.50000000001 €
dossier (min) : 2012 07 D CPH Perpignan discr syndicale.pdf
gain maximum : 500.0 €
dossier (max) : 2018 11 CA DIJON F.PD.pdf

% de cas favorables : 51.42857142857142

dossiers exploités 20 :
2009 11 D C.Cass BORRUT01.pdf -2500.0 €
2010 01 F CPH Paris ROLLAND.pdf 0.0 €
2010 09 F CPH Strasbourg.pdf 200.0 €
2011 06 D CA Orléans Ygnace.pdf -9300.0 €
2011 11 F CPH Paris destierdt clos versa.pdf 0.0 €
2012 01 D CA Metz.pdf -40000.0 €
2012 07 D CPH Perpignan discr syndicale.pdf -72235.50000000001 €
2013 05 F CPH THIONVILLE_STRAPPAZZON_discrimination et autres.pdf -640.14 €
2014 02 F CPH PAU BOURCQ Qualif E discrimination.pdf 0.0 €
2015 03 F CPH METZ - OBORA (déroulement de carrière - discriminations).pdf 0.0 €
2015 04 D CA ORLEANS ROLLAND.pdf -5000 €
2015 07 D CA METZ - Discrimination - résiliation judiciaire - STRAPPAZZON.pdf -36726.46 €
2016 10 F CPH Bordeaux.pdf -6800.0 €
2018 06 Jugement CPH Paris TAHER F.PD.pdf 0.0 €
2018 11 CA DIJON F.PD.pdf 500.0 €
2018 11 CPH CHAMBERY F.pdf 0.0 €
2018 11 F CA Limoges.pdf 0.0 €
2018 12 cph strasbourg F.pdf 0.0 €
2019 01 F CPH Narbonne.pdf 0.0 €
TGI Bayonne 16.11.17.pdf 0.0 €

% de dossiers non exploitables : 42.857142857142854

dossiers non exploités 15 :
2010 02 D CPH Paris.pdf
2010 05 F C.Cass NIERENGARTEN.pdf
2010 06 D CPH Tours ygnace.pdf
2010 09 F CPH CHARLEVILLE DISCRI SYND GIR.pdf
2011 10 F CPH Paris Houpin.pdf
2012 02 F CPH REIMS discrimination syndicale.pdf
2013 04 F CA REIMS GRATIOT_Discrimination syndicale.pdf
2013 11 D CPH Brest reconnaissance de diplome.pdf
2014 05 D CPH METZ - Déroulement de carrière - Discrimination par l_âge - VOGELGESANG.pdf
2015 03 D CPH Libourne.pdf
2015 11 D CPH Toulouse.pdf
2016 06 F CA PAU (nécessité d'invoquer un motif discriminatoire).pdf
2018 09 F CPH Limoges.pdf
COTE Jgt Départage 14.03.2013 .pdf
XURIACH - Jgt départage 30 05 2013.pdf

```

Figure 5 – Le résultat brut d’une recherche effectuée avec
le seul mot-clé “discrimination”

Sous forme graphique:

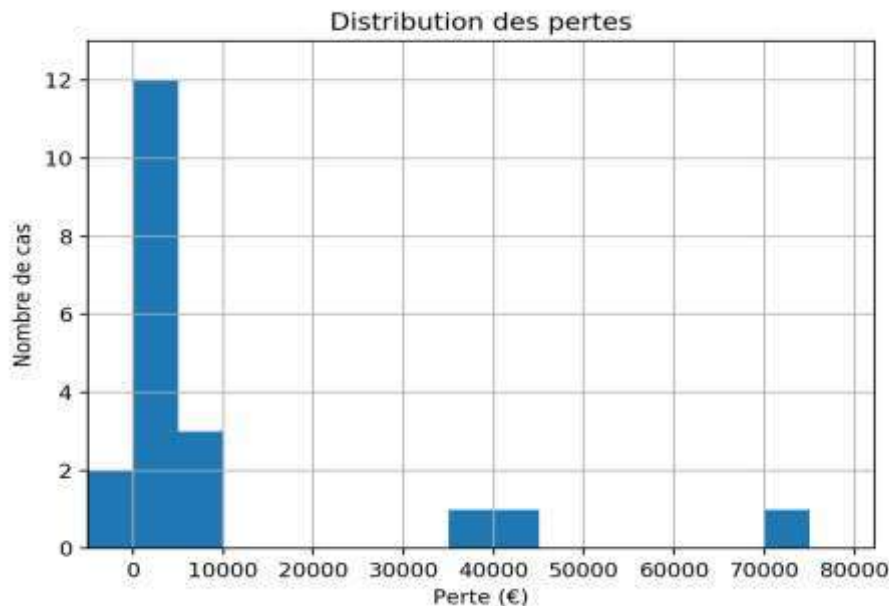


Figure 4 – Le résultat brut d’une recherche effectuée avec le seul mot-clé “discrimination” sous forme graphique

2.5. Critère de lisibilité

Assez tôt dans la conduite du projet, nous avons été confronté au problème que posait la mauvaise qualité des photocopies de certaines décisions de justice ce qui amenait à une traduction en texte partiellement voire entièrement illisible. Nous avons alors compris qu’il était impossible d’analyser l’entièreté de la base de données. Ainsi, Il a été décidé d’écarter les dossiers dits ‘illisibles’ des manipulations ultérieures pour que les résultats que l’on fournisse se basent sur des informations sûres.

Un algorithme a dû être développé pour automatiser la sélection des dossiers lisibles. un tel algorithme doit prendre en entrée un fichier texte (dans notre contexte, la traduction en texte d’une décision de justice) et doit retourner un booléen (True si le texte est ‘lisible’, False sinon).

2.5.1. Première approche

Lorsque l’on regarde une traduction de mauvaise qualité, on observe l’apparition de caractères particuliers (caractères spéciaux, lettres majuscules avec accents ...) :

— 9 0{}{}ÉÜÜ € à fitre &" indemnité pam: via!atian des dispaſitions âes articles L. 12Ü-4 Éi L. 1252-45 et ſuivants du Cade du Üavaih

- 1 SÜÜ,ÜÜ € Sur lE: fanäement df: i'arücie "7ÜB du NÜÜVÉAI CÜEİÉ de FI'ÜÜÉÖHIÉ ÜIVİ1É..

7ar'jugemenä du 3 juin 28%, le CÜ 3 :ü da pmd'hames & dii que l i Hcenciemænf repasait Sur una... causa réeHe ai séräeuse et: & èbÜ l é MÜIISİE CHÜÜE de i'ememble de ses damandes E=i ia 5Üiëiä A..Éİ).T. ÉRANCÉ de sa demanda reaanvantianneiia en paiement {? une: 1dæmnité sur İE: fÜ d l i de E'articia 708 du NÜHVÉ&H Cade fié pncédwa civile.

Appekan'z de cette décisima Mü i 'ül' Cî- AOÜİ damagda à la Cüüi' de :

— dénigrer ie licænciemant äëpüÜ & cle: causa réaâla ei: sériæuse :i de ÜÜ ôæ i SÜH mpiüÿæ fa saciétä AË.T. FRANCE à lui vâr3er 13 gamma da 15 ÜÜÜ,,ÜÜ € à titre d*indamné yüüi limenciem& sans cau5a réeli&: et 3ériause,

»- mndamær SÜI1 &mpıayaur ia sæiäiä A.D."IZ FRANCE à lui vârser la wma de 1 28038 € à titre de rap;3ei de 33i&iæ plus ia sammæ df: 12fLÜG € à titre ÉËÉ c9ngä5 payés a ërenta

— cnnämnar 3311 empiayeur ia sæiétë A.i).T_ FRANCE à Eui vârssr ia gamma de 3 933,7ä! € à titre d'inäæmnité de præavim üÜÜ' t la game de 3 3.,37 € au titæ des cangës payés añ'éranſü

— candamner 13 SÜCİÉİÉ A.Dİİ. FRANCE & lui varser la gamme de 4 58038 € à titre d'" indemnité pam viaiatimn df: l"a icie L. 1264 du CÜdë au Havaii,

Cette première approche consistait tout simplement à calculer la proportion de ces caractères particuliers. Il suffisait maintenant de déterminer une proportion seuil à partir duquelle on considère que le texte est illisible.

Cependant, aucune valeur n'était vraiment satisfaisante. En effet, en pratique il n'est pas raisonnable de répertorier tous les caractères particuliers qui peuvent traduire une non lisibilité. C'est pourquoi nous calculions en réalité le complémentaire de cette proportion, c'est à dire la proportion de caractères 'normaux' (lettres de l'alphabet, chiffres, ponctuation simple et les espaces). Nous avons alors remarqué que la différence de proportion entre les textes visuellement lisibles et les textes visuellement illisibles était trop faible pour pouvoir fixer un seuil satisfaisant.

Nous étions obligé de fixer un seuil qui englobait aussi des fichiers lisibles pour s'assurer que tous les dossiers visuellement illisibles soit jugés 'illisibles' par l'algorithme. Ce compromis nous a poussé à réfléchir à une autre méthode.

2.5.2. Deuxième approche

Cette fois-ci, au lieu d'essayer de comprendre pourquoi le fichier est illisible visuellement, nous avons réfléchi à pourquoi un texte un illisible à la lecture. Nous sommes venu à la conclusion que cela dépendait de quelle lettre suit quelle lettre. Ainsi, l'algorithme aurait pu consister à calculer la fréquence d'apparition des doublets de lettre et comparer ces fréquences à des valeurs de références. Cependant, cela aurait voulu dire qu'il fallait calculer au minimum $26 \times 26 = 676$ fréquences sur des textes d'environ 20 000 lettres. Autrement dit que chaque fréquence aurait

en moyenne été calculée à partir de $20\,000/676 = 30$ doublets, ce qui n'aurait pas donné un résultat très précis.

C'est pourquoi la deuxième approche consiste plus simplement à calculer la fréquence d'apparition des lettres (dont accentuées) qu'on compare à des valeurs de références (calculés sur l'ensemble des articles en français de Wikipédia).

Pour détailler l'implémentation de cette comparaison prenons X_{ref} un vecteur dont chaque composante correspond à la fréquence de référence d'une lettre. Notons maintenant X le vecteur dont chaque composante correspond à la fréquence d'une lettre dans le texte. Ainsi, le résultat de l'algorithme est la valeur booléenne de $\|X - X_{ref}\|^2$ où ϵ est une constante à déterminer. À la suite de plusieurs observations sur la valeur de $\|X - X_{ref}\|^2$ pour différents dossiers nous avons fixé $\epsilon = 35$.

2.6. Interface graphique

2.6.1. Réflexion sur l'élaboration et le choix d'une interface

Une fois que nous avons réussi à avoir des résultats satisfaisant pour l'outil de prédiction et de recherche de statistiques il fallait pouvoir donner accès à la SNCF à ces résultats. Nous avons donc dû créer une interface graphique car il n'était pas concevable de laisser les juristes de la SNCF manipuler python. Nous avons donc demandé conseil à M. Martin responsable de l'option informatique à l'Ecole Centrale de Nantes. M. Martin nous a proposé deux possibilités : une première consistant à développer une application qui s'installerait sur chaque ordinateur puis une deuxième consistant à créer un site web. Nous avons préféré la deuxième option, en effet elle répondait au mieux aux besoins de la SNCF car cet outil a pour but d'être utilisé par tous les juristes de la SNCF sur le territoire français. L'application ne pouvait pas répondre correctement à ce besoin car si le logiciel doit être modifié il faut le modifier sur chaque ordinateur alors que pour le site web il suffit de mettre à jour le serveur. L'élaboration du site web nécessite d'avoir accès à un réseau et de mettre en place une sécurité. En effet un site internet est accessible par tous le monde qui obtient l'url et les dossiers de décisions de justice ne sont pas voués à être diffusés en toute liberté sur le web.

2.6.2. Réflexion sur la sécurité du site

Un système simple d'identifiant/mot de passe devrait suffire d'autant plus que pour l'intranet de la SNCF les employés ont déjà des identifiants/mot de passe propre à la SNCF. Il suffirait donc de relier l'annuaire de la SNCF au site. De plus s'il on développe le site sur un serveur interne à la SNCF cela évite qu'une personne étrangère à la SNCF n'est accès à l'outil. Pour mettre en place une telle sécurité il faut travailler en collaboration avec la DSI de la SNCF.

2.6.3. Réflexions sur la forme de l'interface

Au fur et à mesure des réunions avec l'entreprise nous avons convenu que l'outil serait composé de deux parties principales : partie critères de recherche et une partie avec les résultats. Les critères de recherche sont les suivants :

- Type d'affaire (Discrimination, harcèlement, CDD, etc...)
- Juridiction (Cours d'appel, cours de Cassation, conseil de prud'hommes, etc...)

- L'année à partir de laquelle on souhaite considérer les dossiers
- Des mots-clefs (ils nous ont été donnés par les juristes de la SNCF)

De plus l'utilisateur peut choisir d'utiliser l'outil de prédiction, l'outil de recherche et de statistiques ou bien les deux.

Ensuite l'interface présente les résultats suivants correspondant à l'outil de recherche :

- Pourcentage des affaires favorables
- Moyenne des sommes perdues
- Médiane des sommes perdues
- Ecart type
- Plus grosse somme perdue avec le nom du dossier correspondant
- Plus grosse somme gagnée ou plus petite somme perdue avec le nom du dossier correspondant
- Pourcentage des dossiers exploités
- Liste des dossiers exploités
- Liste des dossiers non exploités qui correspondaient aux critères de recherche mais dont la somme perdue ou gagnée était illisible.

Si l'outil de prédiction est sélectionné le seul résultat est un intervalle ou l'on situe la somme estimée perdue ou gagnée.

De plus l'interface doit permettre de pouvoir ajouter un fichier à la base de données des décisions de justice. Il a été convenu qu'il fallait ajouter la possibilité d'écrire la somme perdue ou gagnée, cela permet de faciliter la lisibilité et de réduire le nombre de dossiers inexploitable.

L'interface doit aussi afficher la liste de tous les dossiers inexploitable. Ainsi l'utilisateur peut directement voir quels dossiers ne sont pas pris en compte et peut potentiellement scanner une nouvelle fois le dossier et rentrer la somme à la main. Ces deux dernières options sont importantes car elles permettent de réduire le nombre de dossiers déchets et donc gagner en précision que ce soit pour l'outil de statistiques ou l'outil de prédiction.

2.6.4. Implémentation du site web

Lors de notre entretien avec M. Martin, il nous a expliqué que nous n'avions pas les connaissances requises pour développer une telle interface, en effet ces compétences ne font pas partie de nos programmes de première année ou même de classe préparatoire. De plus comme nous l'avons vu précédemment, la mise en place de la sécurité nécessaire au fonctionnement du site il faut travailler en collaboration étroite avec la DSI de la SNCF ce qui n'était pas possible au vu du temps qu'il restait. M. Martin nous a donc conseillé de créer une page html qui servira de modèle et de base si ce projet est amené à être repris. Nous avons donc dû apprendre à maîtriser les bases du langage html et du langage CSS afin de pouvoir créer ce modèle. Voici le modèle graphique que nous avons amené à M. Martin pour qu'il nous conseil (cette simple interface a été faite sur PowerPoint et n'est donc pas fonctionnelle) :

Recherche/Prédiction

Dossiers non exploitables

☐ Recherche et statistiques

☐ Prédiction pour une affaire en cours

Type d'affaire : *Discrimination* ↓

Prendre en compte les dossiers depuis : *année* ↓

Type de juridiction : *Cours de cassation* ↓

Dossiers comprenant les mots clefs suivant : *Mot clef* ↓ +

Résultats de la recherche prenant en compte les dossiers exploitables :

Pourcentage d'affaire favorable : -- %

Moyenne des sommes perdues par affaire : -- €

Ecart type : --

Somme perdue maximale -- € (nom du dossier)

Somme perdue minimale -- € (nom du dossier)

Nom des dossiers exploités : ...

Nom des dossiers non exploités : ...

Pourcentage de dossiers non exploités : --%

Résultat de la prédiction :

Somme perdue estimée entre : -- € et -- €

Recherche / prédiction

Dossiers non exploitables

Les dossiers ci-dessous ne sont pas exploitables car les scans ne sont pas de assez bonne qualité :

Nom_dossier

Nom_dossier

Nom_dossier

Nom_dossier

Nom_dossier

Nom_dossier

Nom_dossier

Nom_dossier

Nom_dossier

Nom_dossier

Nom_dossier

Nom_dossier

Figure 6 – Proposition d'interface graphique

3. Résultats obtenus

3.1. Résultat de la réseau

Pour nos données, on a choisi la structure de réseau suivante :

- 169 neurones d'entrée (un associé à chaque mot)
- Deux couches internes de 100 neurones
- 41 neurones de sortie qui groupent les pertes d'argent des affaires juridiques dans intervalles de 5000 euros. Le premier neurone sera actif quand on gagne entre 1 et 5000 euros. Ensuite, tous les autres neurones sont relatifs aux pertes suivants les intervalles de 5000 (0 à 5000, 5000 à 10000, jusqu'à 200000 euros).

Le résultat de l'entraînement pour nos données pour réseau décrite ci-dessus peut être montré pour le graphique de l'erreur en fonctions des itérations d'apprentissage :

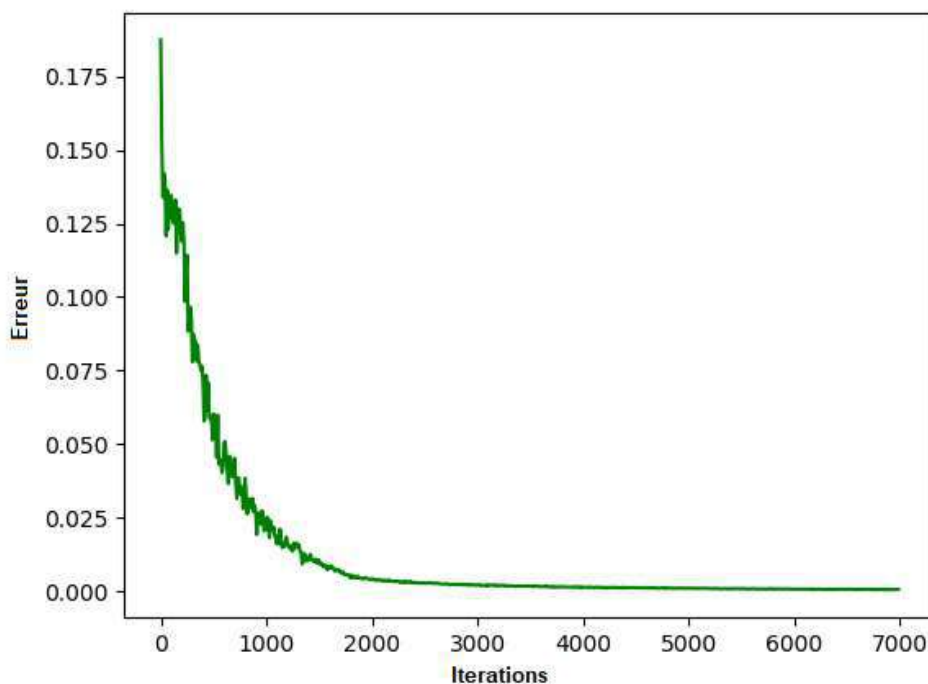


Figure 7 – Erreur du réseau en fonction de nombre d'itérations

Aussi, il est possible de regarder les prédictions qu'elle fait pour chacun des fichier qu'on avait. La plupart des intervalles dans la prédiction sont complètement compatibles avec les pertes réelles. Une exemple de sortie est montré ci-dessous :

0: 2007 01 F CA Dijon discrimination origines.pdf [0, 5000]
1: 2009 03 F CPH Tours.pdf [0, 5000]
2: 2009 11 D C.Cass BORRUTO1.pdf [-5000, 0]
3: 2010 01 F CA Paris dequeker.pdf [0, 5000]
4: 2010 01 F CPH Paris ROLLAND.pdf [0, 5000]
5: 2010 02 D CPH Paris.pdf [-5000, 0]
6: 2010 02 D CPH Toulouse RAOUL.pdf [-20000, -15000]
7: 2010 02 F CA Paris blanco.pdf [0, 5000]
8: 2010 05 F CPH Toulouse malet.pdf [0, 5000]
9: 2010 06 D CA Bordeaux FALA.pdf [-5000, 0]
10: 2010 06 D CPH Tours ygnace.pdf [-5000, 0]

On a bien le nom du fichier et la perte, exprimée sous forme d'intervalle.

3.2. Interface graphique

Comme prévu nous n'avons pas pu relier les codes pythons et l'interface html ni la publié sur le web, nous n'avons donc juste produit un modèle en langage html afin de donner une idée de la forme. L'esthétisme de ce modèle reste aussi extrêmement sobre, nous nous sommes concentrés sur le côté exploitable de l'interface plutôt que le visuel. Le modèle d'interface que nous avons conçu comporte 4 pages. La première est la page d'index où l'utilisateur rentre son identifiant et son mot de passe afin de s'identifier. La deuxième page est la page principale du site, elle comprend la section où l'on choisit les critères de recherches, la section des résultats de l'outil de recherche et de statistiques, la section du résultat de l'outil de justice prédictive et deux liens permettant d'accéder aux deux autres pages. L'une comprend la liste des dossiers inexploitable et sur l'autre page on peut ajouter des fichiers à la base de données.

L'alternative pour le moment est donc de laisser l'utilisateur manipulé Python à l'aide d'une notice, ainsi il aura accès aux résultats donnés par les codes qui sont fonctionnels.

4. Conclusion

4.1. Considérations pour l'amélioration

Comme exposé précédemment, on a utilisé l'outil du Réseau Neurone pour réaliser la prédiction des pertes des affaires juridiques. Cependant, un outil comme celui marche mieux pour une très grande quantité de données. Dans ce projet, on a fait attention seulement au cas de Discrimination et C.D.D, qui sont seulement une petite partie de l'ensemble des cas et données qu'on avait.

Le réseau neurone se base sur l'ensemble des mots clés qui ont été retirés des fichiers et donc, pour qu'elle puisse réaliser des prédictions plus fiables et aussi plus générales, il faut avoir une énorme quantité de fichiers et mots clés. Au cas contraire, on se place dans une situation de risque dans laquelle le réseau sera soumis au phénomène de *overfitting*. C'est à dire qu'elle obtiendra les valeurs de poids (W) et *bias*(b) qui sont les plus appropriés pour donner une réponse très exacte pour notre ensemble de données, mais qui ne sont pas le plus approprié pour une vraie prédiction.

On listera, ci-dessous, une liste de mesures et précautions à prendre pour améliorer et agrandir l'outil prédictive et, aussi, éviter les soucis décrits :

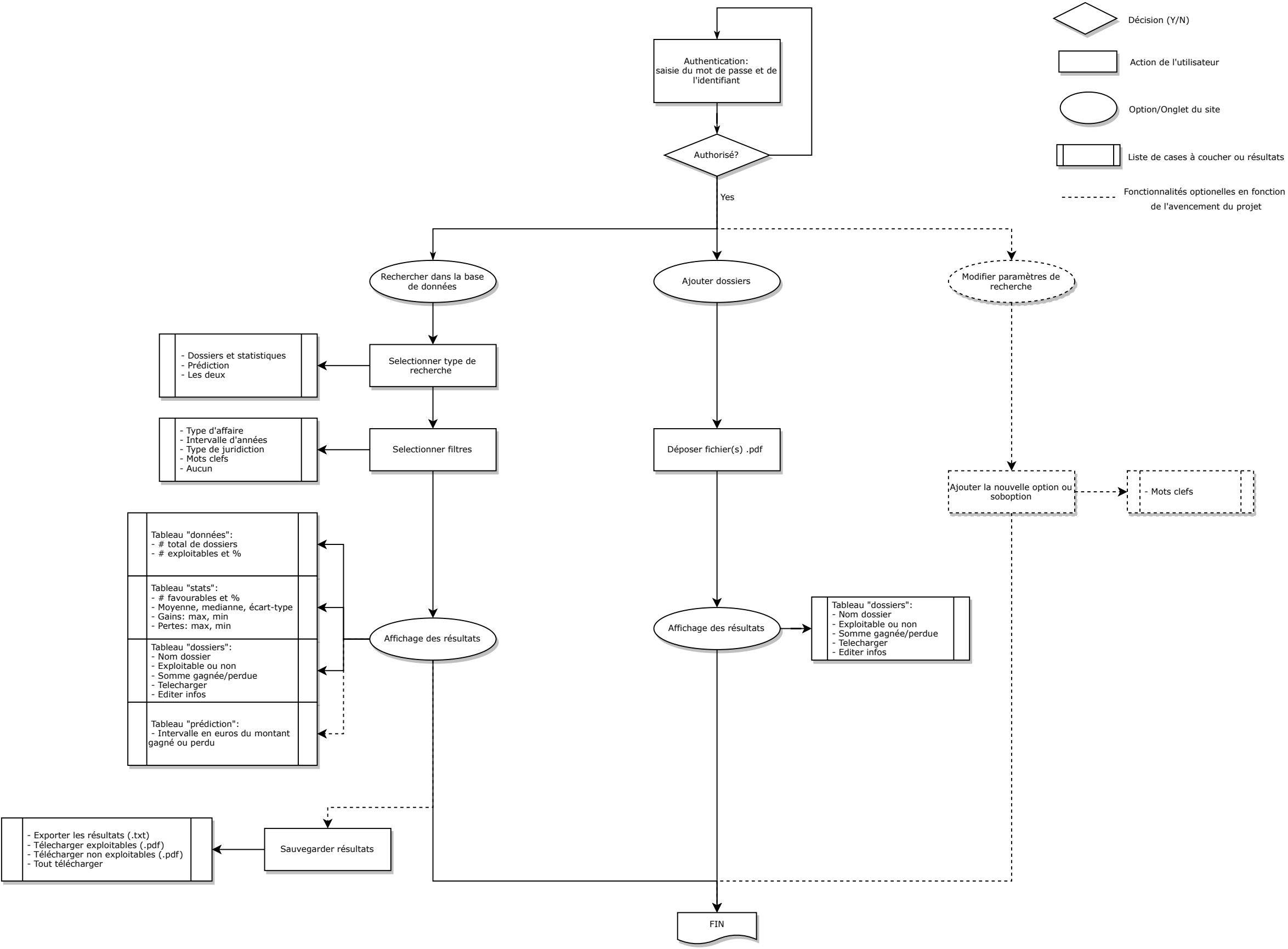
- Améliorer la qualité des scans des documents effectués. Il ne doit pas y avoir des taches, des ombres ou des caractères incompréhensibles. Le contraste doit être assez fort pour que le logiciel puisse reconnaître la plupart des caractères.
- Possiblement, au moment d'ajouter un nouveau fichier dans la base de données, l'utilisateur pourrait écrire le chiffre correspondant à la perte manuellement pour assurer l'exploitabilité du fichier.
- Agrandir la liste de mots clés pour que le réseau puisse traiter plus des cas et d'affaires.
- On a considéré un réseau avec deux couches intermédiaires. Au cas où il soit nécessaire de traiter beaucoup d'affaires différents, il faudrait ajouter des couches.

5. Table des illustrations

<i>Figure 1 – Morceau d'un fichier PDF</i>	<i>5</i>
<i>Figure 2 – Morceau d'un fichier PDF converti en texte</i>	<i>6</i>
<i>Figure 3 – Schéma d'un neurone</i>	<i>11</i>
<i>Figure 4 – Schéma d'une réseau neuronne</i>	<i>12</i>
<i>Figure 5 – Le résultat brut d'une recherche effectuée avec</i>	<i>18</i>
<i>Figure 6 – Proposition d'interface graphique</i>	<i>23</i>
<i>Figure 7 – Erreur du réseau en fonction de nombre d'itérations</i>	<i>24</i>

6. Bibliographie

- [1] <https://docs.python.org/fr/3/howto/regex.html#more-metacharacters>
- [2] <https://openclassrooms.com/fr/courses/235344-apprenez-a-programmer-en-python/233857-manipulez-les-expressions-regulieres>
- [3] <http://neuralnetworksanddeeplearning.com/chap2.html>
- [4] <https://towardsdatascience.com/how-do-we-train-neural-networks-edd985562b73>
- [5] <https://techburst.io/improving-the-way-we-work-with-learning-rate-5e99554f163b>
- [6] <https://adventuresinmachinelearning.com/stochastic-gradient-descent/>



Paramètres de recherche



Mots clés :

Mot clé 1

Mot clé 2

Mot clé 3

Mot clé 4

Mot clé 5

Mot clé 6

Juridictions :

☒ Cour de cassation

☐ Cour d'appel

☐ Conseil de prud'hommes

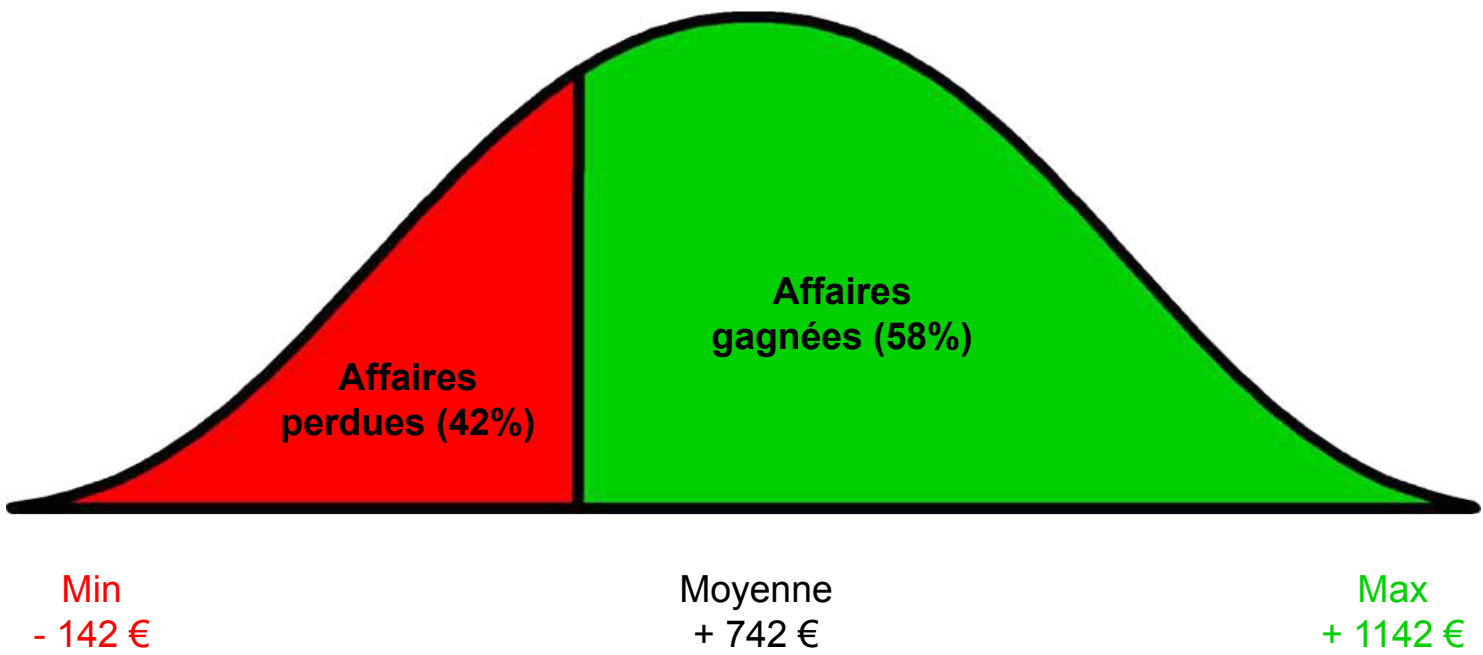
Dates : du au

 Lancer la recherche

 Faire une prédiction

- Toutes
- Alsace
- Rhône-Alpes
- Bourg-en-Bresse
- ☒ Annecy
- Grenoble
- Lyon

Résultats de la recherche



Fichiers trouvés (42)	Statistiques
Fichier 1 Fichier 2 Fichier 3	Moyenne : Médiane : Écart-type :

Résultats de la prédiction

Intervalle du gain estimé : - ??? € / + ??? €

Recherche

Ajout de fichiers

Administration

Ajouter des fichiers



Nom	Lisible	Date	Juridiction	Somme gagnée	Conserver
Fichier1.pdf	Oui	04 / 2021	Cour d’appel de Nantes	+ 50 €	<input checked="" type="checkbox"/>
Fichier2.pdf	Non	42 / 1&À@	Cour de Cassation de Jupiter	- 42 £	<input type="checkbox"/>
Fichier3.pdf	Oui	12 / 2001	Conseil des Prud'Hommes de Nantes	- 999999 €	<input type="checkbox"/>

Valider

Annuler

Remplacer des fichiers

Pour remplacer des fichiers illisibles, rendez-vous à la section [Fichiers illisibles](#) de l’onglet Administration.



Recherche	Ajout de fichiers	Administration
-----------	-------------------	----------------

Administrateur

- Métégal Sam
- 07 00 00 00 00
- sammétégal@gmail.com

État de la base de donnée

La base contient 1000 décisions judiciaires dont 958 sont lisibles et 42 illisibles.

Fichiers illisibles

fichier_de_merde.pdf	Remplacer
un_autre.pdf	Remplacer
allé_un_dernier.pdf	Remplacer

