**Notice utilisateur (version open source)**

Cet outils possède deux fonctionnalités principales : classifier des documents (marquer les documents un par un, les catégoriser en sous-dossiers) & labelliser des documents (charger les “boîtes” reconnues par l’OCR ou les créer soi-même, taguer ces boîtes à la main, afficher et modifier les tokens reconnus par l’OCR, effectuer une labellisation Beginning-Inside-Outside, ajouter des métadonnées...).

**1./ Prérequis**

Système d’exploitation Windows.  
Il est conseillé de régler le paramètre de ‘Mise à l’échelle’ à 100% dans les paramètres d’affichage, autrement des effets graphiques indésirables vont apparaître.

**2./ Prise en main rapide**

Lancez l’exécutable LabelTool.exe du répertoire GUI\dist\LabelTool\.  
Vous avez alors le choix entre classifier des documents, labelliser des documents, ou faire les deux en même temps.   
Une fois l’option choisie, cliquez sur ‘Load Dataset’ pour charger un projet : il vous sera alors demandé de sélectionner un répertoire “input” contenant les documents à traiter et un répertoire “output” où le travail effectué sera sauvegardé.  
Appuyez sur la touche “?” pour afficher l’Help.

**3./ Fonctionnement de l’application**

Au lancement de l’application, vous êtes amené à choisir une tâche pour la session courante : classifier, labelliser, ou faire les deux.

**chargement et sauvegarde d’un projet**

Après avoir sélectionné la tâche à effectuer pour la session courante, vous êtes invité à renseigner les chemins d’accès des répertoires ‘Input Dataset’ et ‘Output Directory’. Il peut s’agir du même répertoire. Les chemins sont remémorés lors de la prochaine session.

* INPUT DATASET : il s’agit du répertoire contenant l’ensemble des documents à traiter. Ces documents doivent être de format TIFF, JPEG ou PNG pour être affichés dans l’interface. Vous serez par la suite en mesure de naviguer entre ces documents au sein de l’application. Le répertoire peut être sous format ZIP. Accès uniquement en lecture.
  + Classification : les documents sont simplement lus puis affichés.
  + Labellisation : les documents sont lus et affichés. Un document est identifié par son nom (ce qui précède l’extension du fichier).
* OUPUT DIRECTORY : il s’agit du répertoire où sont sauvegardés les traitements effectués lors de la session courante mais c’est aussi à partir de ce répertoire que l’utilisateur peut fournir une base, comme par exemple ce qui a déjà été effectué la fois précédente. Les sauvegardes sont au format JSON. Accès en lecture et écriture.
  + Classification : l’application y sauvegarde le fichier \_classes.json ainsi que le dossier \_archive si l’utilisateur le demande.
  + Labellisation : l’application y sauvegarde le fichier \_labels.json ainsi qu’un fichier <identifiant\_du\_fichier>.json pour chaque document traité au cours de la session.

**navigation dans l’interface**

L’écran principal de l’interface se constitue d’une fenêtre d’affichage pour visualiser les documents et de panneaux de contrôle de part et d’autre de cette fenêtre.  
  
Dans la fenêtre vous pouvez avec la souris:

* Dessiner sur le document à l’aide du bouton gauche ;
* Déplacer le document ou annulez le dessin en cours à l’aide du bouton droit ;
* Effectuer un zoom à l’aide de la molette.

Les panneaux de contrôles comportent le nécessaire pour effectuer la tâche souhaitée. La plupart des actions peuvent être réalisées à l’aide d’un raccourci clavier représenté par l’image d’une touche. Cliquez sur cette image pour changer le raccourci. Les raccourcis sont remémorés lors de la prochaine session.  
  
Naviguez entre les documents à l’aide des touches “↑” et “↓” ou bien cliquez sur “Go to...” pour vous rendre sur un document en particulier. Le nom du document est affiché en haut à gauche. Le dernier document traité est remémoré lors de la prochaine session.  
Naviguez entre les pages d’un document à l’aide des touches “←” et “→”.  
Utilisez le bouton “Focus” pour recentrer/redimensionner le document.  
Sauvegarder le travail réalisé à l’aide du raccourci Ctrl+S. Le projet est sauvegardé à chaque fois que l’utilisateur passe d’un document à l’autre.

**classification**

Sélectionnez l’option “Classify” au lancement de l’application.   
Cliquez sur “Load Dataset” et renseigner les chemins des répertoires d’entrée et de sortie.  
  
Ajouter/supprimer une classe :  
Sur le panneau de contrôle de gauche, utilisez “Add Class” pour ajouter une nouvelle classe. Vous êtes alors invité à définir un nom et une touche de raccourci pour cette nouvelle classe.  
Supprimez une classe en cliquant sur l’icône “poubelle” à côté de son nom. Si des documents ont été marqués par cette classe, une popup apparaît vous demandant confirmation pour déclasser ces documents, sinon, la classe est directement supprimée.  
  
Classer un document :  
Appuyez sur une touche raccourci de classe pour assigner le document à la classe correspondante, cette action entourera le nom de la classe choisie (en magenta) et passera au document suivant si l’option “Switch document once classified” est cochée.  
  
Visualiser le travail réalisé :  
Affichez l’état d’avancement de la classification sous forme d’un graphique circulaire à l’aide de la touche “µ”.  
  
Classer les documents en sous-dossier :  
Si vous souhaitez classer concrètement les documents dans des répertoires distincts selon les classes, cliquer sur “Archive documents”. Vous devrez alors choisir soit de déplacer les documents (et les fichiers qui les décrivent) dans le dossier \_archive, soit de les y dupliquer. Le dossier \_archive est alors créé, les fichiers \_classes.json et \_labels.json y sont déplacés.

**labellisation**

Sélectionnez l’option “Label” au lancement de l’application.   
Cliquez sur “Load Dataset” et renseigner les chemins des répertoires d’entrée et de sortie.  
  
Ajouter/supprimer un label :  
Sur le panneau de contrôle de droite, utilisez “Add local label” ou “Add general label” pour ajouter un nouveau label. Vous êtes alors invité à définir un nom et une couleur pour ce nouveau label. Ajouter un label général permet de le retrouver dans le panneau lors de la labellisation des documents suivants; à l’inverse un label local ne sera présent dans le panneau que pour la labellisation du document actuellement affiché.  
Le label “None” est toujours présent et ne peut être supprimé, c’est le label par défaut.  
Supprimez un label en cliquant sur l’icône “poubelle” à côté de son nom. Si des boîtes ont été taguées avec ce label, une popup apparaît vous demandant confirmation pour taguer ces boîtes avec le label “None”, sinon le label est directement supprimé.  
  
Afficher les boîtes de l’OCR :  
Cochez l’option “Show boxes”. Cela affichera (en bleu) sur le document les boîtes chargées depuis le fichier OCR correspondant.  
  
Labelliser :  
Si vous avez coché l’option “Show boxes”, et sélectionné la rubrique “Tool”, vous êtes en mesure de taguer les boîtes présentes sur le document à l’aide de l’outil “Tag”. Sélectionnez le label de votre choix puis cliquez sur une boîte pour la taguer. La boîte se colorise alors en la couleur du label. Il est possible de taguer plusieurs boîtes en même temps en dessinant un rectangle avec le bouton gauche de la souris. Taguer à nouveau une boîte à pour effet de la taguer avec le label “None”.  
  
*Note : le label sélectionné, s’il est général à l’ensemble des documents parcourus, reste sélectionné à chaque passage au document suivant. Il est donc possible de labelliser les documents par champ (par exemple parcourir l’ensemble du dataset et labelliser les “Total”, puis recommencer en labellisant les “Quantity”, etc...). Dans ce cas, décocher l’option “Auto refocus” peut être pertinent si les champs d’un même type se trouvent régulièrement au même endroit sur les documents.*  
  
Créer/supprimer une boîte :  
Via l’outil “Create”, vous êtes en mesure de créer une boîte en dessinant un rectangle. Décochez l’option “Box creation inside only” pour créer des boîtes en dehors du document.  
L’outil “Remove” permet de supprimer une boîte en cliquant dessus. Il est possible de supprimer plusieurs boîtes en même temps en dessinant un rectangle avec le bouton gauche de la souris.  
  
*Note : Si un label autre que “None” est sélectionné, la boîte créée est directement taguée par ce label.*  
  
Modifier une boîte :  
Cliquez sur l’outil “Move/Resize”. Les coins des boîtes deviennent sélectionnables et il est possible de les étirer.   
Maintenez le bouton gauche enfoncé sur une boîte pour la déplacer.  
  
Autre outils :  
“Encapsulate” : Sélectionnez une ou plusieurs boîtes en dessinant un rectangle avec le bouton gauche de la souris afin de créer la boîte la plus petite possible englobant l’ensemble boîtes sélectionnées.  
“Merge” : Sélectionnez plusieurs boîtes pour les regrouper en une seule la plus petite possible.  
  
*Note : Si un label autre que “None” est sélectionné, les boîtes créées sont directement taguées par ce label.*  
  
Annuler un changement :  
Utilisez le raccourci Ctrl+Z.  
  
ATTENTION : L’ordre des boîtes OCR n’est pas forcément conservé dans le fichier JSON de sortie lors de la création d’une boîte (“Create”), d’un recouvrement (avec “Move/Resize”), de l’encapsulation (“Encapsulate”) et de la fusion (“Merge”). Taguer une boîte (“Tag”) ou la supprimer (“Remove”) ainsi que déplacer/redimensionner sans recouvrement (“Move/Resize”) n’affecte pas l’ordre.  
  
Visualiser et modifier les tokens OCR :  
Cochez l’option “Show tokens” pour afficher les tokens reconnus par l’OCR pour chaque mot du document. Vous pouvez modifier directement la graphie d’un token en cliquant dessus.  
  
Métadonnées :  
Appuyez sur la touche “TAB” pour ouvrir la fenêtre des métadonnées. A la manière de la création de label, vous pouvez définir une métadonnée locale ou générale.  
  
Labellisation B-I-O :  
Si vous avez coché l’option “Show boxes”, et sélectionné la rubrique “BIO labelling”, vous êtes en mesure d’affecter à chaque boîte une lettre (B pour Beginning, I pour Inside et O pour outside) manuellement ou en créant un groupe avec “Select chunk”. Utilisez l’id pour différencier différents groupes.  
  
*Note : Il est conseillé de décocher l’option “Show background” pour une meilleure lisibilité.*  
  
ATTENTION : la labellisation B-I-O peut être lente et affecter la fluidité des mouvements du visuel du document.  
  
Visualiser le travail réalisé :  
Cochez la vignette en haut à gauche pour approuver un document. Cela signifie qu’il a été convenablement labellisé.  
Affichez l’état d’avancement de la labellisation sous forme d’un graphique circulaire à l’aide de la touche “µ”.

**format des fichiers du projet**

Le fichier \_classes.json est accédé au cours de la session.   
Voici un exemple :

{  
 "documents\_by\_classes": [  
 {  
 "name": "invoice",  
 "key": 105,  
 "size": 3,  
 "batch": [  
 "D:\\datasets\\data\_set\\CD.433348341360438018",  
 "D:\\datasets\\data\_set\\CD.433418710104615695",  
 "D:\\datasets\\data\_set\\CD.369066493198464905"  
 ]  
 },  
 {  
 "name": "purchase\_order",  
 "key": 112,  
 "size": 1,  
 "batch": [  
 "D:\\datasets\\data\_set\\CD.433383525732526861"  
 ]  
 }  
 ],  
 "classes\_by\_documents": {  
 "D:\\datasets\\data\_set\\CD.369066493198464905": "invoice",  
 "D:\\datasets\\data\_set\\CD.433348341360438018": "invoice",  
 "D:\\datasets\\data\_set\\CD.433383525732526861": "purchase\_order",  
 "D:\\datasets\\data\_set\\CD.433418710104615695": "invoice"  
 },  
 "last-document": "CD.433348341360438018"  
}

* "documents\_by\_classes" : liste des classes, une classe étant représentée par son nom "name", l’identifiant ASCII d’une touche raccourci "key", le nombre de documents qui la composent "size" et le lots de documents "batch" sous forme de liste contenant les documents de la classe représentés par leurs identifiants ;
* "classes\_by\_documents": dictionnaire permettant d’accéder à la classe d’un document à partir de son identifiant ;
* “last-document” : dernier document traité.

Le fichier \_labels.json est accédé au cours de la session.   
Voici un exemple :

{  
 "labels": [  
 {  
 "name": "None",  
 "color": [  
 0,  
 127,  
 255  
 ]  
 },  
 {  
 "name": "Total",  
 "color": [  
 255,  
 85,  
 85  
 ]  
 },  
 {  
 "name": "Quantity",  
 "color": [  
 85,  
 85,  
 255  
 ]  
 },  
 {  
 "name": "Address",  
 "color": [  
 255,  
 170,  
 85  
 ]  
 },  
 {  
 "name": "Logo",  
 "color": [  
 85,  
 255,  
 85  
 ]  
 }  
 ],  
 "metadata": [  
 "Bad\_quality",  
 "Total",  
 "Type (train/ validation/ test)"  
 ],  
 "last-document": "CD.433383525732526861",  
 "nb-approved": 3  
}

* “labels” : liste des labels, un label étant représenté par son nom “name” et sa couleur “color” ;
* “metadata” : liste des noms des métadonnées ;
* “last-document” : dernier document traité.

Les fichiers <identifiant\_du\_document>.json sont accédés au cours de la session.   
Voici un exemple :

"info": {  
 "filesize": 178339,  
 "engine": "ScanSoft",  
 "recognition-info": "",  
 "page-conf-level": "20;",  
 "pdf": "yes",  
 "nbPages": 1,  
 "res-X": 300,  
 "res-Y": 300,  
 "pg-size-X": 2479,  
 "pg-size-Y": 3499,  
 "BIO\_labelling": true,  
 "approved": true,  
 "class": null,  
 "pages": [  
 {  
 "numero": 1,  
 "res-X": 300,  
 "res-Y": 300,  
 "pg-size-X": 2479,  
 "pg-size-Y": 3499  
 }  
 ]  
 },  
 "labels": [  
 {  
 "name": "None",  
 "color": [  
 0,  
 127,  
 255  
 ],  
 "occurrence": 430  
 },  
 {  
 "name": "Total",  
 "color": [  
 255,  
 85,  
 85  
 ],  
 "occurrence": 1  
 },  
 {  
 "name": "Quantity",  
 "color": [  
 85,  
 85,  
 255  
 ],  
 "occurrence": 11  
 },  
 {  
 "name": "Address",  
 "color": [  
 255,  
 170,  
 85  
 ],  
 "occurrence": 12  
 },  
 {  
 "name": "Logo",  
 "color": [  
 85,  
 255,  
 85  
 ],  
 "occurrence": 1  
 }  
 ],  
 "metadata": {  
 "Bad\_quality": "",  
 "Total": "",  
 "Type (train/ validation/ test)": ""  
 },  
 "boxes-by-page": [  
 {  
 "page-number": 1,  
 "boxes": [  
 {  
 "left": 206,  
 "top": 37,  
 "right": 779,  
 "bottom": 173,  
 "confidence": 20,  
 "token": "bartholus",  
 "label": "Logo",  
 "BIO-prefix": "O"  
 },  
 {  
 "left": 1073,  
 "top": 139,  
 "right": 1295,  
 "bottom": 187,  
 "confidence": 30,  
 "token": "RECU",  
 "label": "None",  
 "BIO-prefix": "O"  
 },  
 {  
 "left": 1328,  
 "top": 147,  
 "right": 1384,  
 "bottom": 193,  
 "confidence": 50,  
 "token": "08",  
 "label": "None",  
 "BIO-prefix": "O"  
 },  
...

* "info" : contient différentes information sur le document :
  + "filesize", "engine", "recognition-info", "page-conf-level", "pdf", "nbPages", "res-X", "res-Y" sont inutiles dans cette version open source
  + "pg-size-X" largeur de l’image ;
  + "pg-size-Y": hauteur de l’image ;
  + "BIO\_labelling" indique si il y a eu labellisation B-I-O ;
  + "approved": indique si le document a été approuvé par l’utilisateur (si la labellisation est convenable) ;
  + "class": indique la classe du document s’il y a eu classification ;
  + "pages" liste des pages, représentées par leur numéro "numero" et leur dimension "res-X", "res-Y", "pg-size-X" et "pg-size-Y" ;
* “labels” : liste des labels, un label étant représenté par son nom “name”, sa couleur “color” et le nombre de boîtes pour lesquelles il a été utilisé “occurrence” ;
* “metadata” : liste des noms des métadonnées avec leurs valeurs ;
* “boxes-by-page” : contient les boîtes regroupées par page :
  + “page-number” : le numéro de la page ;
  + “boxes” : les boîtes représentées par :
    - “left”, “top”, “right”, “bottom” : les coordonnées ;
    - “confidence” : la confiance OCR ;
    - “token” : le token OCR ;
    - “label” : le label de la boîte ;
    - “bio-prefix” : la lettre de la boîte en labellisation B-I-O (optionnel).

**Quitter l’application**

Quitter l’application avec la croix en haut à droite ou la touche “Echap”. Cela ne sauvegarde pas ce qui était en train d’être fait sur le document en cours. Il faut utiliser le raccourci Ctrl+S avant de fermer l’application si l’on souhaite sauvegarder les modifications.