« Aide à la classification de documents hétérogènes »

Manuel d’utilisation

BDL - Mars 2004

Adrien Missemer
Céline Monthéard
Nicolas Boulanger
Rana BouJaoudé

Conventions typographiques, remarques générales

Les boutons sont désignés de la façon suivante : OK

Les touches du clavier sont notées : [ENTER]

Les fichiers sont notés : **`path\file`**

Le code SQL est noté : `select * from documents`

I. Manuel d'installation

MyDOX est un programme écrit en Java. Il a été testé sous Windows avec le JDK1.4.2_03.

1°) Création de la base des documents

a) Installation d'Oracle 9i

MyDOX utilise les fonctionnalités d'Oracle Text inclues dans Oracle 9i Database. Le logiciel a été testé avec la version 9.2.0.1 (donc la Release 2).

L'ordinateur sur lequel MyDOX sera utilisé doit donc disposer au minimum d'un client Oracle 9i, et d'un moyen de se connecter à un serveur Oracle 9i.

Oracle InterMedia Text doit être installé correctement sur le serveur.

b) Création de la base

La création de la base de documents est faite par le script **`install/install.sql`**. Pour pouvoir l'exécuter l'utilisateur doit avoir reçu le rôle `CTXAPP` dans Oracle. Le script appelle `french_stoplist.sql` et `english_stoplist.sql` qui contiennent les mots non indexés par Oracle Text.

```
SQL> connect ctxsys/ctxsys;
SQL> grant CTXAPP to bdl_projet;
```

Après exécution du script, la base doit contenir, en plus des tables d'indexation (pour les 3 index `BDOCSINDEX, DOXINDEX, SUMMARYINDEX`) les tables :

- `BYNARYDOCUMENTS`
- `DOCUMENTS`
- `GIST_RES`
- `THEMES_RES`

et les vues suivantes :

- `SUMMARIES`
- `THEMES`

c) Modification du PATH

Le répertoire **`$oracle$\bin`** (où **`$oracle$`** désigne le répertoire d'installation d'Oracle) doit se trouver dans le PATH, pour que la JDBC puisse accéder aux bibliothèques dynamiques d'Oracle.

d) Java DataBase Connectivity

Les pilotes JDBC pour Oracle9i sont installés par Oracle, dans **`oracle\jdbc\lib`**. Pour la JDK1.4, le pilote à utiliser est **`ojdbc14.jar`**.

2°) Compilation

Le logiciel est livré avec le code compilé (avec la JDK1.4.2_03), dans **`lib\MyDOX.jar`**.

Dans **`install`**, le script **`make.bat`** permet éventuellement de le recompiler, après avoir été modifié.

Le script **`jarify.bat`** comprime les .class obtenus en **`lib\MyDOX.jar`**.

Enfin le script **`clean.bat`** supprime les .class temporaires.

Au lieu de lancer successivement les 3 scripts, tapez simplement **`install`**.

3°) Modification de MyDOX.bat

Modifiez **`MyDOX.bat`** pour définir le nom d’utilisateur, le mot de passe et la chaîne hôte Oracle à utiliser.

4°) Lancement du programme

Lancez **`MyDOX.bat`**.

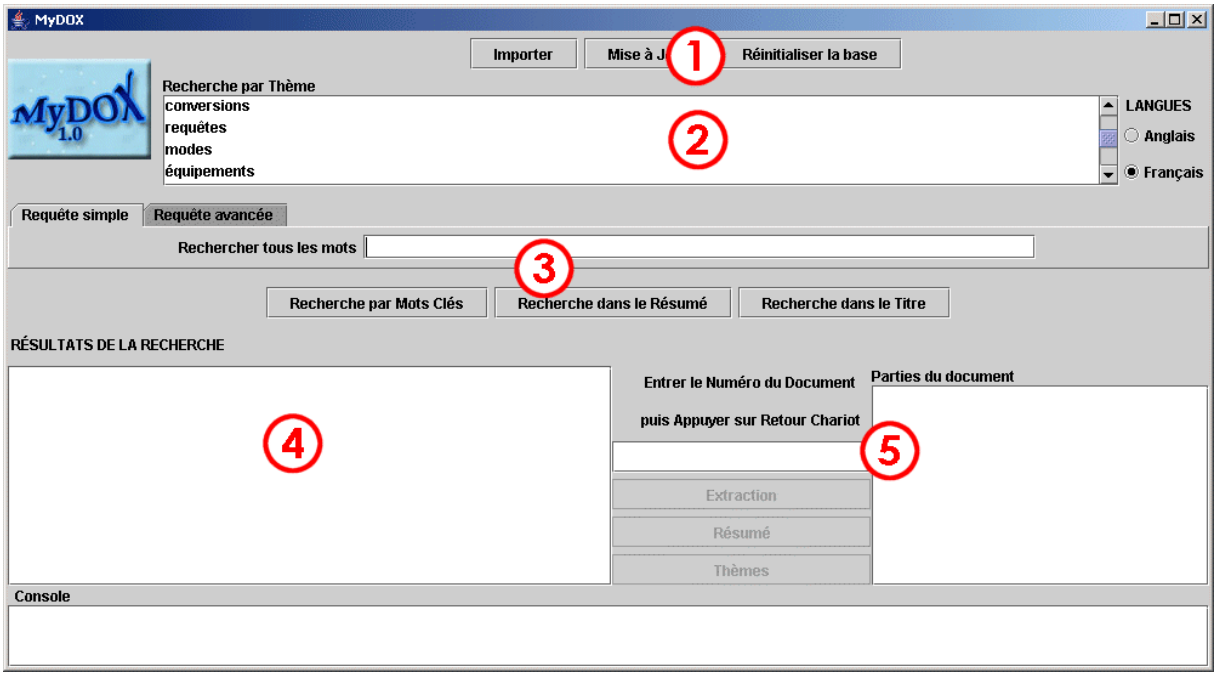
II. Description générale de MyDOX

1°) But du logiciel

MyDOX est un logiciel d'aide à la classification de documents hétérogènes. Il permet d'importer divers types de fichiers contenant du texte, en particulier dans la version 1.0, les fichiers .txt et .mail (texte brut), les fichiers .rtf et documents Word (.doc), les .pdf, et enfin .xml, .html et .htm. Seul le contenu textuel des fichiers est stocké et indexé.

Une fois les documents importés, une liste des thèmes contenus dans l'ensemble des documents est affichée. Il est également possible d'effectuer des recherches sur le contenu, d'obtenir des extraits ou les résumés des documents.

2°) Description globale de l'interface



L'interface de MyDOX

L'interface se décompose en 5 parties. En haut, la zone 1 correspond à la gestion de la base de documents. En dessous, la zone 2 donne un aperçu du contenu de la base. Ensuite en 3 vient la zone de requête. Les résultats des requêtes s'affichent en 4. La dernière zone en bas à droite, 5, permet d'obtenir plus d'information sur un document en particulier (résumé, thèmes, extraits).

III. Utilisation de MyDOX

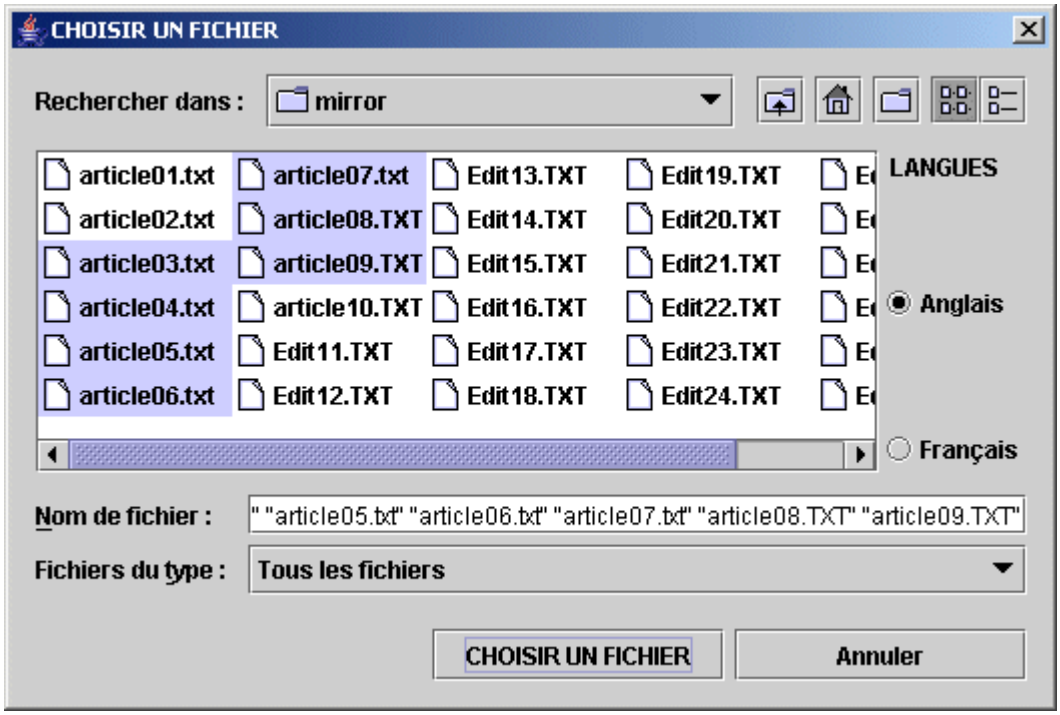
1°) Gestion de la base de documents

a) Importation de documents

Les types de fichiers acceptés sont

- txt (texte brut)
- mail (texte brut)
- rtf (texte formaté)
- doc (Word)
- pdf (Portable Document Format)
- xml
- html
- htm

Pour importer un nouveau fichier, cliquez sur Importer, ce qui a pour effet d'afficher la boite de dialogue d'importation.



La boite de dialogue d'importation

Cette boite de dialogue permet de sélectionner plusieurs documents à la fois.

☞ Attention : La sélection de la langue d'importation (à droite de la boite) est très importante. Elle conditionne la façon dont le résumé et les thèmes sont générés, ainsi que la liste des mots qui sont ignorés par l'indexation.

Cliquez sur CHOISIR UN FICHIER, et les fichiers sont importés dans la base. Selon le type de fichiers cela peut prendre plusieurs secondes (en particulier les gros fichiers binaires).

b) Mise à jour de documents

Lorsqu'un fichier a été modifié sur le disque, il faut le réimporter dans la base pour mettre à jour les index de recherche. Pour cela cliquez sur le bouton Mettre à jour.

La boite de dialogue d'importation s'ouvre, et s'utilise de la même façon que pour importer un document.

c) Réinitialisation de la base

Pour supprimer tous les documents de la base, cliquez sur Réinitialiser la base.

d) Mais quelle est donc la différence entre 'Importer' et 'Mettre à Jour' ?

Le fonctionnement global est le même, en particulier on peut utiliser Mettre à Jour pour insérer un document pour la première fois.

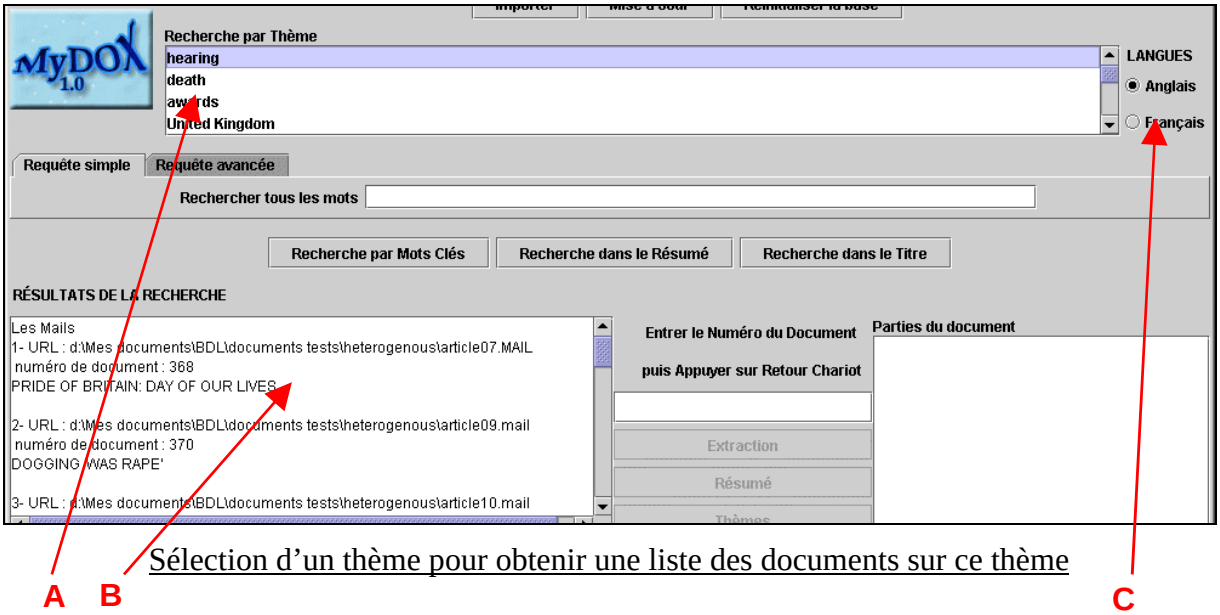
La différence se situe lorsque le document se trouve déjà dans la base (le document est identifié par son chemin complet sur le disque). Importer bloque l'importation alors que Mettre à Jour réimporte le document.

2°) Le contenu de la base – Les thèmes

a) Que sont les thèmes ?

Lors de l'importation, 10 thèmes (au maximum) sont déterminés pour chaque document. Le calcul de ces termes est fait à partir d'une base de connaissance et bien sûr il est essentiel d'avoir sélectionné la bonne langue du document lors de l'importation.

b) La fenêtre des thèmes



Sélection d'un thème pour obtenir une liste des documents sur ce thème

La fenêtre des thèmes (A) permet de sélectionner un thème pour obtenir dans la fenêtre de résultats (B) la liste des documents portant sur ce sujet.

☞ Les thèmes anglais sont séparés des thèmes français. Vous pouvez sélectionner la langue avec le sélecteur de langue (C).

☞ Dans la fenêtre de résultats, il y a une distinction entre les mails (fichiers avec extension .mail) et les autres documents.

3°) Les requêtes

On distingue deux modes de requêtes, les requêtes simples (recherche sur un ensemble de mots, type Google) et les requêtes avancées (qui nécessitent une connaissance de la syntaxe de la fonction `CONTAINS` d'Oracle).

a) Requêtes simples

Sélectionnez l'onglet Requête simple, s'il n'est pas déjà sélectionné.

Tapez simplement les mots, séparés par des espaces. Utilisez ensuite l'un des 3 boutons Recherche par Mots-clés Recherche dans le titre Recherche dans le résumé. Le résultat s'affiche immédiatement dans la fenêtre de résultats.

b) Requêtes avancées

Sélectionnez l'onglet Requête avancée, s'il n'est pas déjà sélectionné.

Utilisez la syntaxe d'Oracle pour remplir le champ.

En particulier, les opérateurs `AND` et `OR` sont acceptés.

`Mot1 ~ Mot2` recherche les documents contenant `Mot1` mais pas `Mot2`.

`?malin` retournera les documents contenant des mots qui ressemblent à `malin`.

% remplace un nombre quelconque de caractères

_ remplace un caractère

Utilisez ensuite l'un des 3 boutons Recherche par Mots-clés Recherche dans le titre Recherche dans le résumé. Le résultat s'affiche immédiatement dans la fenêtre de résultats.

4°) Les résultats

Dans la fenêtre de résultats, l'url (emplacement du fichier) s'affiche, ainsi que son numéro d'indexation, la pertinence du résultat, et le titre s'il a été trouvé.

5°) Informations complémentaires sur les documents

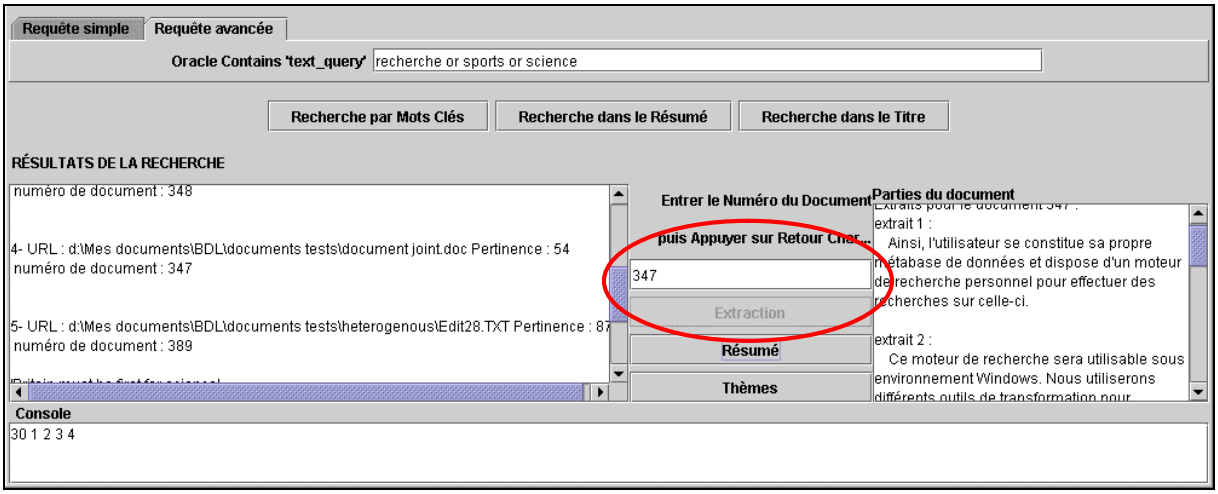
a) Extraits

Après avoir effectué une requête, il est possible d'obtenir la liste des paragraphes d'un document qui vérifient cette requête. Pour cela :

Relevez le numéro d'indexation du document qui vous intéresse (dans la fenêtre résultats)

Entrez ce numéro dans le champ prévu à cet effet, terminez par [Entrée]

Cliquez sur Extraction.



Obtenir les extraits du document vérifiant la requête

☞ Il n'est possible d'obtenir les extraits d'un document qu'après avoir effectué une requête. Les extraits sont en effet déterminés en fonction de la requête effectuée.

b) Résumé

Le résumé est obtenu en tapant un numéro de document dans la zone prévue et en tapant [Entrée] puis en terminant par cliquer sur Résumé.

Le résumé est toujours accessible, pour tous les documents présents dans la base, contrairement aux extraits (voir a).

c) Thèmes

Le fonctionnement est le même que pour les résumés, voir (b).