pitney bowes

Spectrum Technology Platform

Version 12.0

Guide Information Extraction

Table des matières

1 - Introduction

In this section

Module Information Extraction

Le module Information Extraction fournit des fonctionnalités avancées de traitement de texte et d'extraction des informations du texte d'entrée en langage naturel.

Fonctions fournies

Text Categorization	Vous permet d'affecter des catégories personnalisées à un texte non structuré. Cela est possible une fois que vous avez formé un *modèle de catégorisation de texte* à l'aide de l'utilitaire Administration. Cette fonction peut servir à indexer des rapports de soins de santé de patients, à classer des documents par domaines et sous-domaines et à catégoriser des courriers électroniques en SPAM et non SPAM, entre autres applications.
Entity Extraction	Vous permet de former des modèles pour extraire des entités de données non structurées. Le module est livré avec certaines *entités préexistantes*. Si nécessaire, définissez des entités personnalisées à l'aide du type de modèle *CustomEntity*. Une fois que vous avez créé et formé des modèles spécifiques au domaine, vous pouvez extraire des entités en fonction du modèle que vous avez formé.
Relationship Extraction	Vous permet d'identifier le type de relation liant une paire d'entités dans tout texte d'entrée en langage naturel.

Langues prises en charge

Pour tous les stages du module Information Extraction, la version actuelle prend en charge les fonctionnalités d'extraction des informations du texte d'entrée en langue *anglaise* uniquement.

Remarque : Le stage **Entity Extractor**, outre l'anglais, prend en charge ces langues dans la phase *bêta* :

es	Espagnol (Mexique)
fr	Français
de	Allemand
pt	Portugais (Brésil)

Important : Ces langues *bêta* sont disponibles uniquement en cas de *Custom Entity* et non pour les entités préexistantes.

Sécurité du modèle

Des autorisations de sécurité doivent être accordées dans **Management Console** pour exécuter différentes fonctions via Information Extraction :

- Des autorisations d'affichage sont nécessaires pour catégoriser ou répertorier le modèle.
- Des autorisations de modification sont nécessaires pour reformer ou importer le modèle (si le modèle existe déjà).
- Des autorisations de création sont nécessaires pour importer ou former le modèle.
- Des autorisations de suppression sont nécessaires pour supprimer le modèle.

2 - Text Categorization

In this section

Introduction à la catégorisation de texte

La catégorisation de texte, également appelée classification de texte, est le processus consistant à affecter des catégories personnalisées au contenu non structuré ou au texte en clair, tel que des courriers électroniques, des articles d'actualité et des commentaires, en fonction de la quantité de contenu correspondant à ladite catégorie. La catégorisation peut être effectuée en fonction du sujet, de l'auteur, de la date ou encore de quasiment tout système de classification de votre choix.

Vous pouvez créer votre propre élément de catégorisation en formant un modèle d'élément de catégorisation avec vos données et vos catégories. Le formateur analyse les données et stocke les informations qu'il obtient dans le processus de formation. Il analyse le contenu et détermine la catégorie à laquelle il appartient.

La fonctionnalité de catégorisation de texte utilise un processus de catégorisation de texte statistique. Elle applique des méthodes d'apprentissage machine pour apprendre des règles de classification automatique basées sur des documents de formation libellés par l'homme.

Étant donné que vous pouvez appliquer la catégorisation de votre choix, vous devez d'abord « former » votre modèle à « apprendre » les catégories. Ensuite, vous pouvez utiliser ce modèle dans le stage **Text Categorizer** pour catégoriser vos données non structurées.

Spectrum™ Technology Platform utilise les commandes de l'utilitaire Administration pour gérer les modèles de catégorisation de texte. Pour obtenir une description de ces commandes, reportez-vous à la section **Commandes de l'utilitaire Administration** à la page 30.

Préparation des données

La première étape de l'utilisation de la catégorisation de texte est la préparation de votre fichier d'entrée et de votre fichier de test. Pour ce faire, vous devez structurer les données sous forme de valeurs séparées par des onglets dans les deux fichiers. Les détails des fichiers doivent se présenter au format suivant :

- Codage UFT-8
- Données séparées par des onglets dans deux colonnes, où la première colonne contient le nom de catégorie (par exemple : « Patient » ou « Fournisseur ») et la deuxième colonne dispose des données pour chaque catégorie (comme dans l'exemple ci-dessous)

Vos données devraient prendre la forme suivante :

```
Patient     John Smith dob04181963 224 Main St. Atl GA 30311
Provider    Mark Johnson M.D. NPI5489512047 412 Washington Atl GA 30301
```

Configuration des options

Cela implique la création d'un fichier `Options de formation` contenant des informations sur votre modèle et les options à appliquer pour la formation du modèle. Ce fichier doit être au format XML avec un codage UFT-8 et inclure les fonctions de formation requises et l'en-tête suivants :

En-tête du fichier Options de formation

L'en-tête mentionne les détails du modèle, son type et le chemin d'accès aux fichiers de test et d'entrée.

- `modelName` : nom du modèle
- `modelType` : type du modèle (c'est-à-dire `TC`, ce qui signifie Text Categorization dans ce cas)
- `modelDescription` : description du modèle
- `inputFilePath` : emplacement du fichier d'entrée utilisé pour la formation du modèle
- `testFilePath` : emplacement du fichier de test

  Remarque :

  Le fichier de test vérifie l'efficacité d'un modèle. Il détermine le comportement du modèle personnalisé avec différents paramètres de formation. Comme bonne pratique, il est conseillé d'utiliser différents fichiers d'entrée et de test pour former et évaluer vos modèles personnalisés.

`algorithm` : algorithme Machine Learning utilisé pour la formation du modèle (par défaut, `MaxEnt`)

Fonctions de formation

Il s'agit des fonctions de formation que vous pouvez utiliser pour créer une catégorie.

Remarque : Si vous utilisez plusieurs fonctions, elles peuvent être placées dans n'importe quel ordre dans le fichier.

- **Lingustic feature** : pour spécifier les propriétés de langue
  - `Stemming` : réduit les termes à leur souche, ou racine. Par exemple, « insurer » (assureur), « insured » (assuré) et « insures » (assure) peuvent être réduits à la racine « insure » (assurer).

```
<trainingFeature>
  <featureName>Stemming</featureName>
</trainingFeature>
```

- **Keyword features** : pour définir la liste des mots clés
  - `IgnoreWords` : également appelés stopwords. Cette fonction filtre les termes courants qui n'ont aucun effet sur la catégorisation, comme « the » (le), « and » (et) et « but » (mais). Ces termes

doivent être uniquement séparés par des virgules et non pas par des espaces. Vous pouvez également utiliser la clé `Append` avec cette fonction, qui, lorsqu'elle est définie sur « True », est ajoutée à la liste de stopwords existante.

```
<trainingFeature>
  <featureName>IgnoreWords</featureName>
  <featureParams>
   <entry>
    <key>WordList</key>
    <value>
     and,the,for,with,still,tri,rep,cust,keep,get,req,call
    </value>
   </entry>
   <entry>
    <key>Append</key>
    <value>True</value>
   </entry>
  </featureParams>
 </trainingFeature>
```

- `CategoryKeywords` : identifie une catégorie pour une liste de mots clés appartenant à différentes listes personnalisées. Par exemple, Weekdays de la liste `CategoryKeywords` contient les `mots clés` Monday, Tuesday, Wednesday, Thursday et Friday.

  Cette fonction peut éventuellement spécifier si la correspondance doit être sensible à la casse. Si cela est utilisé, la valeur par défaut est `true`.

```
<trainingFeature>
 <featureName>CategoryKeywords</featureName>
 <featureParams>
  <entry>
   <key>Weekdays</key>
                    <!-- List of weekdays -->
   <value>Monday,Tuesday,Wednesday,Thursday,Friday</value>
  </entry>
  <entry>
   <key>WeekendDays</key>
                    <!-- List of weekend days -->
   <value>Saturday,Sunday</value>
  </entry>
  <entry>
   <key>CaseSensitive</key>
                    <value>True</value>
  </entry>
 </featureParams>
</trainingFeature>
```

- `KeyWords` : recherche les termes que vous avez spécifiés comme appartenant à une liste personnalisée, comme *DaysOfWeek* ou *Month*. Peut également éventuellement indiquer si la

correspondance doit être sensible à la casse ; si cette option est utilisée, la valeur par défaut est `true`.

```
<trainingFeature>
 <featureName>KeyWords</featureName>
 <featureParams>
  <entry>
   <key>KeyWordList</key>
   <value>Monday,Tuesday</value>
  </entry>
  <entry>
   <key>CaseSensitive</key>
   <value>False</value>
  </entry>
 </featureParams>
</trainingFeature>
```

- **Lexical Feature** : pour spécifier les propriétés des lexèmes
  - `NGram` : recherche une partie d'une chaîne plus longue, avec « n », représentant le nombre de termes à rechercher. Par exemple, si vous recherchez l'expression « to be or not to be » (être ou ne pas être), vous pouvez rechercher un unigram « to » ou « be », un bigram « to be » ou « or not », ou un trigram « to be or » ou « not to be».

```
<trainingFeature>
  <featureName>NGram</featureName>
  <featureParams>
   <entry>
    <key>Count</key>
    <value>3</value>
   </entry>
  </featureParams>
</trainingFeature>
```

Voici un exemple de fichier d'options de formation :

```
<trainingOptions>
 <modelName>modelone</modelName>
 <modelType>TC</modelType>
 <modelDescription>modelOne</modelDescription>

<inputFilePath>C:/SpectrumIE/textclassification/train_Input.csv</inputFilePath>


<testFilePath>C:/SpectrumIE/textclassification/train_Test.txt</testFilePath>

    <algorithm>SVM</algorithm>

 <trainingFeatures>
```

```
<!-- Keyword features -->
<trainingFeature>
 <featureName>IgnoreWords</featureName>
 <featureParams>
  <entry>
   <key>WordList</key>
   <value>
    and,the,for,with,still,tri,rep,cust,keep,get,req,call
   </value>
  </entry>
  <entry>
   <key>Append</key>
   <value>True</value>
  </entry>
 </featureParams>
</trainingFeature>

<trainingFeature>
 <featureName>CategoryKeywords</featureName>
 <featureParams>
  <entry>
   <key>Category1/key>
   <value>CategoryKeyword1,CategoryKeyword2</value>
  </entry>
  <entry>
   <key>Category2/key>
   <value>CategoryKeyword3,CategoryKeyword4</value>
  </entry>
            </featureParams>
</trainingFeature>

<trainingFeature>
           <featureName>KeyWords</featureName>
      <featureParams>
  <entry>
   <key>KeyWordList</key>
   <value>
    jam,misfeed,install,help,mechanical,failure,jam,pc,connection
   </value>
  </entry>
 </featureParams>
</trainingFeature>

<!-- Linguistic feature -->
<trainingFeature>
 <featureName>Stemming</featureName>
</trainingFeature>

<!-- Lexical feature -->
<trainingFeature>
 <featureName>NGram</featureName>
 <featureParams>
  <entry>
```

```
    <key>Count</key>
    <value>3</value>
   </entry>
  </featureParams>
</trainingFeature>

</trainingFeatures>
</trainingOptions>
```

Formation du modèle

Après avoir créé un fichier d'options, vous devez former votre modèle à découvrir les relations potentiellement prédictives. Vous pouvez le faire en appliquant les méthodes de Machine Learning. Spectrum™ Technology Platform utilise la commande **iemodel train** à la page 37 CLI pour former un modèle. Après la formation du modèle, vous pouvez l'utiliser dans la catégorisation.

Évaluation du modèle

Vous pouvez souhaiter tester votre modèle après la formation, pour garantir que le fichier d'options de formation est correct et que les catégories sont attribuées comme prévu.

Vous pouvez tester le modèle à l'aide de la commande **iemodel trainAndevaluate model** à la page 37 CLI.

Catégorisation de texte

1. Créez un flux de données incluant un stage source tel que **Read from File** ou **Read from XML**, le stage **Text Categorizer** et un stage de collecteur de données tel que **Write to File** ou **Write to XML**.
2. Dans le stage source, pointe vers votre fichier d'entrée.
3. Dans le stage **Text Categorizer**, sélectionnez le modèle dans le champ **Nom de l'élément de catégorisation**. Il s'agit du modèle que vous avez formé dans la phase de catégorisation de texte. Pour plus d'informations sur la formation d'un modèle, reportez-vous à la section **Formation du modèle** à la page 12.

4. Dans le champ **Nombre de catégories**, sélectionnez le nombre de niveaux de correspondance de catégorie à inclure dans la sortie. Par exemple, la correspondance la plus proche ou la correspondance la plus proche plus la deuxième correspondance proche.

   Remarque : La valeur maximale correspond au nombre de catégories différentes indiqué lors de la formation du modèle.

5. Cliquez sur **OK**.
6. Exécutez le job.

3 - Entity Extraction

In this section

Introduction

L'extraction d'entités est le processus d'identification et d'extraction d'entités de données non structurées. Vous pouvez utiliser les entités préexistantes jointes au stage **Entity Extractor**, ou vous pouvez former un modèle pour qu'il extraie des entités personnalisées.

Entités préexistantes

Les entités préexistantes sont celles fournies avec le module **Information Extraction**.

Pour obtenir une liste des entités préexistantes, ouvrez le stage **Entity Extractor**, cochez la case **Neutralisation des options système par défaut avec les valeurs suivantes** et cliquez sur **Ajout rapide**. La liste des entités s'affiche dans la section **Sélectionner des entités**.

- *Person*
- *Address*
- *ProperNouns*
- *ISBN*
- *CreditCard*
- *ZipCode*
- *WebAddress*
- *Mention*
- *HashTag*
- *SSN*
- *Phone*
- *Email*
- *Date*
- *Location*
- *Organization*

Suivez les étapes restantes de cette section pur extraire ces types d'entités de vos données.

Extraction d'entités pré-existantes

1. Créez un flux de données incluant un stage source **Read from Documents**, un stage **Entity Extractor** et un stage de collecteur de données tel que **Write to File** ou **Write to XML**.

2. Dans le stage source, pointe vers votre fichier d'entrée.
3. Dans le stage **Entity Extractor**, sélectionnez les entités en fonction des données que vous souhaitez extraire du fichier d'entrée. Par exemple, si vous souhaitez sélectionner les noms de toutes les personnes et les adresses du fichier, sélectionnez les entités *Address* et *Person*.

   Remarque : *Address* et *Person* sont les entités par défaut. Pour extraire les données en fonction d'une autre entité, cochez la case **Neutralisation des options système par défaut avec les valeurs suivantes**, puis cliquez sur **Ajout rapide**. La liste des entités s'affiche dans la section **Sélectionner des entités**.

4. Pour obtenir la fréquence dans le fichier d'entrée des données associées aux entités spécifiées, cochez la case **Nombre d'entités de sortie**.
5. Cliquez sur **OK**.
6. Exécutez le job.

Entités personnalisées

Tout comme avec les entités pré-existantes, vous pouvez également former des modèles pour qu'ils récupèrent des entités personnalisées. Ces entités peuvent appartenir à n'importe quel domaine et peuvent être de tout type. Par exemple, vous pouvez utiliser un texte médical pour extraire une liste de diagnostics ou de produits pharmaceutiques. Le processus d'extraction des entités personnalisées inclut :

1. Préparation des données : préparation du fichier d'entrée et du fichier de test
2. Configuration des options : création d'un fichier d'options de formation contenant des informations sur le modèle et les options à appliquer lors de sa formation
3. Formation du modèle
4. Extraction des entités

Lorsque vous effectuez correctement toutes ces étapes, le nouveau type d'entité est ajouté à la liste du stage **Entity Extractor** et vous pouvez l'utiliser pour extraire les détails d'un fichier non structuré.

Préparation des données pour les entités personnalisées

La première étape de création d'entités personnalisées est la préparation de votre fichier d'entrée et de votre fichier de test. La fonction d'entités personnalisées nécessite que les entités figurant dans ces fichiers soient encadrées par le magicWord que vous indiquez dans votre fichier d'options de formation (opération expliquée dans la rubrique suivante).

Imaginons que vous extrayiez des diagnoses de données non structurées figurant de votre fichier d'entrée et que vous ayez désigné le magicWord *DIAGNOSIS* dans votre fichier d'options de formation. Chaque fois que le nom d'une maladie ou d'un état s'affiche dans le texte, le terme est encadré à l'aide de ce magicWord, comme suit :

```
The term diagnostic criteria designates the specific combination of
signs, symptoms, and test results that the clinician uses to attempt
to determine the correct diagnosis. Some examples of diagnostic
criteria, also known as clinical case definitions, are: Amsterdam
criteria for DIAGNOSIShereditary nonpolyposis colorectal cancerDIAGNOSIS
McDonald criteria for DIAGNOSISmultiple sclerosisDIAGNOSIS ACR criteria

for DIAGNOSISsystemic lupus erythematosusDIAGNOSIS Centor criteria for

DIAGNOSISstrep throatDIAGNOSIS.
```

Pour plus d'informations sur l'identification du magicWord, reportez-vous à la rubrique suivante.

Configuration des options des entités personnalisées

Cela implique la création d'un fichier `Options de formation` contenant des informations sur votre modèle et les options à appliquer pour la formation du modèle. Ce fichier doit être au format XML avec un codage UFT-8 et inclure les fonctions de formation requises et l'en-tête suivants :

En-tête du fichier Options de formation

L'en-tête mentionne des détails sur le modèle, le chemin d'accès aux fichiers de test et d'entrée et le mot clé d'annotation des entités personnalisées.

- `modelName` : nom du modèle personnalisé
- `modelType` : type du modèle personnalisé (c'est-à-dire *CustomEntity*).
- `modelDescription` : description du modèle personnalisé
- `inputFilePath` : chemin d'accès au fichier balisé utilisé pour la formation du modèle (fichier d'entrée)
- `testFilePath` : chemin d'accès au fichier utilisé pour tester le modèle
- `magicWord` : mot clé utilisé pour annoter les entités personnalisées
- `language` : langue utilisée dans le texte.

  Remarque : L'anglais est pris en charge. L'allemand, l'espagnol, le français et le néerlandais sont en phase bêta.

Fonctions de formation

Vous pouvez utiliser ces fonctions de formation pour créer des entités personnalisées.

- **Linguistic features** : pour spécifier les propriétés de langue

- `POSTagger` : balisage pour identifier des parties du texte, comme des noms, des pronoms, des adjectifs et des verbes.

```
<trainingFeature>
   <featureName>POSTagger</featureName>
</trainingFeature>
```

- **Orthographic features** : pour spécifier les propriétés structurelles
  - `CaseIdentifier` : détermine si les entités personnalisées sont tout en majuscules, en minuscules, ou en une combinaison des deux.

```
<trainingFeature>
 <featureName>CaseIdentifier</featureName>
</trainingFeature>
```

  - `NumericIdentifier` : détermine si les entités personnalisées sont numériques ou alphanumériques.

```
<trainingFeature>
 <featureName>NumericIdentifier</featureName>
</trainingFeature>
```

  - `1st2ndIdentifier` : détermine si les entités personnalisées sont des nombres ordinaux tels que 1er, 2e, 3e, etc.

```
<trainingFeature>
 <featureName>1st2ndIdentifier</featureName>
</trainingFeature>
```

  - `PatternMatcher` : met des termes en correspondance par rapport à un ou plusieurs modèles via des expressions régulières. Lorsque plusieurs expressions sont fournies, inclut la condition de jointure `AND` pour toutes les expressions ou `OR` (valeur par défaut) pour toute expression.

```
<trainingFeature>
 <featureName>PatternMatcher</featureName>
  <featureParams>
   <entry>
    <key>RegEx1</key>
    <value>b[aeiou]t</value>
   </entry>
   <entry>
    <key>RegEx2</key>
    <value>b[xyz]t</value>
   </entry>
   <entry>
    <key>JoinCondition</key>
    <value>AND</value>
   </entry>
```

```
  </featureParams>
</trainingFeature>
```

- **Keyword features** : pour définir la liste des mots clés
  - `CategoryKeywords` : identifie une catégorie pour une liste de mots clés appartenant à différentes listes personnalisées. Par exemple, Weekdays de la liste `CategoryKeywords` contient les `mots clés` Monday, Tuesday, Wednesday, Thursday et Friday.

    Cette fonction peut éventuellement spécifier si la correspondance doit être sensible à la casse. Si cela est utilisé, la valeur par défaut est `true`.

```
<trainingFeature>
 <featureName>CategoryKeywords</featureName>
 <featureParams>
  <entry>
   <key>Weekdays</key>
                    <!-- List of weekdays -->
   <value>Monday,Tuesday,Wednesday,Thursday,Friday</value>
  </entry>
  <entry>
   <key>WeekendDays</key>
                    <!-- List of weekend days -->
   <value>Saturday,Sunday</value>
  </entry>
  <entry>
   <key>CaseSensitive</key>
                    <value>True</value>
  </entry>
 </featureParams>
</trainingFeature>
```

  - `KeyWords` : recherche les termes que vous avez spécifiés comme appartenant à une liste personnalisée, comme *DaysOfWeek* ou *Month*. Peut également éventuellement indiquer si la correspondance doit être sensible à la casse ; si cette option est utilisée, la valeur par défaut est `true`.

```
<trainingFeature>
 <featureName>KeyWords</featureName>
 <featureParams>
  <entry>
   <key>KeyWordList</key>
   <value>Monday,Tuesday</value>
  </entry>
  <entry>
   <key>CaseSensitive</key>
   <value>False</value>
  </entry>
 </featureParams>
</trainingFeature>
```

- `Substring` : extrait une partie d'une chaîne tel que spécifié dans les paramètres. Peut également être utilisé pour l'extraction de préfixe et de suffixe.
  - `StartLocation` : à gauche ou à droite. Position d'extraction de la sous-chaîne. La valeur par défaut est `Left`.
  - `StartPosition` : position de départ de la sous-chaîne. La valeur par défaut est `0`.
  - `EndPosition` : position d'arrivée de la sous-chaîne. La valeur par défaut est `3`.
  - `MinLength` : longueur minimale du terme auquel cette fonction doit s'appliquer. La valeur par défaut est `3`.

```
<trainingFeature>
 <featureName>Substring</featureName>
  <featureParams>
   <entry>
    <key>StartLocation</key>
   </entry>
   <entry>
    <key>StartPosition</key>
    <value>1</value>
   </entry>
   <entry>
    <key>EndPosition</key>
    <value>4</value>
   </entry>
   <entry>
    <key>MinLength</key>
  </featureParams>
</trainingFeature>
```

- **Lexical Features** : pour spécifier les propriétés des lexèmes
  - `FeatureWindow` : indique la fenêtre de génération de la fonction

```
<trainingFeature>
 <featureName>FeatureWindow</featureName>
 <!-- Number of preceding tokens used to create the feature set.
Default is 3 -->
  <entry>
   <key>Before</key>
   <value>1</value>
  </entry>
 <!-- Number of succeeding tokens used to create the feature set.
Default is 3 -->
  <entry>
   <key>After</key>
   <value>2</value>
  </entry>
</trainingFeature>
```

Voici un exemple complet de fichier d'options de formation d'entités personnalisées :

```
<trainingOptions>
 <modelName>CustomModel</modelName>
 <modelType>CustomEntity</modelType>
 <modelDescription>CustomDiagnosesModel</modelDescription>

<inputFilePath>C:/SpectrumIE/custom_model/Custom_Input.csv</inputFilePath>

 <testFilePath>C:/SpectrumIE/custom_model/Custom_Test.txt</testFilePath>

       <magicWord>DIAGNOSIS</magicWord>
       <language>English</language>

      <trainingFeatures>

 <!-- Lexical features-->
 <trainingFeature>
  <featureName>FeatureWindow</featureName>
  <featureParams>
   <entry>
    <key>Before</key>
    <value>1</value>
   </entry>
   <entry>
    <key>After</key>
    <value>2</value>
   </entry>
  </featureParams>
 </trainingFeature>

 <!-- Orthographic features-->
 <trainingFeature>
  <featureName>CaseIdentifier</featureName>
      </trainingFeature>

 <trainingFeature>
  <featureName>NumericIdentifier</featureName>
 </trainingFeature>
 </trainingFeatures>
 </trainingOptions>
```

Formation du modèle des entités personnalisées

Après avoir créé un fichier d'options, vous devez former votre modèle à identifier les entités personnalisées. Spectrum™ Technology Platform effectue cette opération avec la commande **iemodel train** à la page 37 CLI. Un modèle formé est utilisé pour extraire des entités personnalisées.

Évaluation du modèle des entités personnalisées

Vous pouvez souhaiter tester votre modèle après la formation, pour garantir que le fichier d'options de formation est correct et que les entités sont extraites comme prévu. Pour tester votre modèle, utilisez la commande **iemodel trainAndevaluate model** à la page 37 CLI.

Extraction d'entités personnalisées

L'entité personnalisée formée, qui est désormais disponible dans la liste d'entités du stage **Entity Extractor**, peut être utilisée pour extraire des informations pertinentes de vos données non structurées.

Pour connaître les étapes d'extraction des entités préexistantes, reportez-vous à la section **Extraction d'entités pré-existantes** à la page 15.

4 - Relationship Extraction

In this section

Introduction

Relationship Extraction vous permet d'identifier les types de relation entre une paire d'entités identifiée dans le contenu source. Il analyse le contenu en langage naturel du document source et identifie les types de relation existants entre toutes les paires d'entités identifiées.

Types d'entité pris en charge

Actuellement, les types d'entité pris en charge pour l'extraction de relations sont les suivants :

- *Person*
- *Organization*
- *Location*

Relationship Types

RelationshipType	Entity1 Type	Entity2 Type	Relationships Covered
AffiliatedWith	*Person*	*Organization*	Indique toute relation professionnelle ou académique entre les entités *Person* et *Organization*. La relation peut être l'une de ces relations ou toute autre relation similaire : • *Person* Étudie ou a étudié à*Organization* • *Person* Travaille ou a travaillé avec*Organization* • *Person* S'est vu offrir un poste chez *Organization* **Remarque :** Il s'agit d'une liste indicative des relations couvertes par ce type. Par exemple, `James has studied from the University of Toronto and works at ABC Corp.` Ici, deux relations peuvent être analysées : Entity1 = `James`, RelationshipType = `AffiliatedWith`, Entity2 = `University of Toronto` Entity1 = `James`, RelationshipType = `AffiliatedWith`, Entity2 =`ABC Corp`

RelationshipType	Entity1 Type	Entity2 Type	Relationships Covered
LivesIn	*Person*	*Location*	Indique une relation entre les entités *Person* et *Location*. La relation peut être l'une de ces relations : • *Person* Réside ou a résidé à *Location* • *Person* Transféré à *Location* • *Person* Est né à *Location* • *Person* Est décédé à *Location* **Remarque :** Il s'agit d'une liste indicative des relations couvertes par ce type. Par exemple, `John Jamison, a National Weather Service meteorologist in Galveston, reported that a massive hurricane was about to hit the East Coast the next day.` Entity1 =`John Jamison`, RelationshipType = *LivesIn*, Entity2 = `Galveston`
OrgBasedIn	*Organization*	*Location*	Indique qu'au moins un des bureaux de *Organization* se trouve à *Location*. *Location* peut être une filiale, un bureau de développement, le siège social, etc. Par exemple, `HSBC Holdings Plc. is headquartered in London, United Kingdom.` Ici, deux relations peuvent être analysées : Entity1 = `HSBC Holdings Plc.`, RelationshipType = *OrgBasedIn*, Entity2 =`London` Entity1 = `HSBC Holdings Plc.`, RelationshipType = *OrgBasedIn*, Entity2 = `United States of America`

RelationshipType	Entity1 Type	Entity2 Type	Relationships Covered
LocatedIn	*Location*	*Location*	Indique la relation entre deux endroits différents, où l'une des entités est géographiquement contenue dans l'autre. **Exemple 1 :** `Canberra is the capital of Australia.` Ici, Entity1 = `Canberra`, RelationshipType = *LocatedIn*, Entity2 =`Australia` **Exemple 2 :** `India has as its capital New Delhi.` Ici, Entity1 = `India`, RelationshipType = *LocatedIn*, Entity2 = `New Delhi`
Negative	*Person* *Organization* *Location*	*Organization* *Location*	Indique qu'aucun des types de relation ci-dessus n'a pu être analysé entre les deux entités correspondantes. Par exemple, `New Delhi and New York are good places to live in.` Lors de l'analyse de ce texte d'entrée, aucun des types de relation pris en charge n'est analysé entre aucune paire d'entités identifiées. Par conséquent, la relation peut être divisée en types de relation *Negative* entre les entités identifiées : Entity1 = `New Delhi`, RelationshipType = *Negative*, Entity2 = `New York`

Remarque : Vous pouvez connecter un stage **Splitter** à la sortie du stage **Relationship Extractor** pour extraire les types de relation identifiés et les paires correspondantes d'entités liées par la relation. Le stage splitter convertit la sortie hiérarchique de ce stage en une sortie de texte.

Exemple

En cas de texte d'entrée complexe, plusieurs combinaisons de types de relation possibles peuvent être analysées pour la même phrase.

Par exemple,

```
James McCarthy has settled in New York, United States as director of ABC
Technologies.
```

Lorsque le stage **Relationship Extractor** analyse ce texte d'entrée à l'aide des types de relation sélectionnés dans les options de stage, les relations trouvées sont les suivantes :

Relationship 1	Entity1 = `James McCarthy`, Entity1 Type = *Person*, RelationshipType = *LivesIn*, Entity2 = `New York`, Entity2 Type = *Location*
Relationship 2	Entity1 = `James McCarthy`, Entity1 Type = *Person*, RelationshipType = *AffiliatedWith*, Entity2 = `ABC Technologies`, Entity2 Type = *Organization*
Relationship 3	Entity1 = `ABC Technologies`, Entity1 Type = *Organization*, RelationshipType = *OrgBasedIn*, Entity2 = `United States`, Entity2 Type = *Location*
Relationship 4	Entity1 = `ABC Technologies`, Entity1 Type = *Organization*, RelationshipType = *OrgBasedIn*, Entity2 = `New York`, Entity2 Type = *Location*
Relationship 5	Entity1 = `James McCarthy`, Entity1 Type = *Person*, RelationshipType = *LivesIn*, Entity2 = `United States`, Entity2 Type = *Location*
Relationship 6	Entity1 = `New York`, Entity1 Type = *Location*, RelationshipType = *LocatedIn*, Entity2 = `United States`, Entity2 Type = *Location*

5 - Commandes de l'utilitaire Administration

In this section

Commandes de l'utilitaire Administration

Les commandes de l'utilitaire Administration servent à gérer, évaluer et former les modèles du module Information Extraction.

iemodel delete

La commande `iemodel delete` renvoie une liste de tous les modèles du module Information Extraction.

Utilisation

`iemodel delete --n`*modelName*

Requis	Argument	Description
Oui	`--n` *modelName*	Indique le nom du modèle à supprimer. Les chemins de répertoire que vous indiquez ici sont relatifs à l'emplacement dans lequel vous exécutez l'utilitaire Administration.

Exemple

Cet exemple supprime le modèle appelé « MyModel ».

```
iemodel delete --n MyModel
```

iemodel evaluate model

La commande `iemodel evaluate` évalue un modèle du module Information Extraction précédemment formé.

Utilisation

`iemodel evaluate model--n`*modelName*`--t`*testFileName*`--o`*outputFileName*`--c`*categoryCount*`--d`*trueOrfalse*

Obligatoire	Argument	Description
Oui	`--n` *modelName*	Indique le nom et l'emplacement du modèle à évaluer. Les chemins de répertoire que vous indiquez ici sont relatifs à l'emplacement dans lequel vous exécutez l'utilitaire Administration.
Oui	`--t` *testFileName*	Indique le nom et l'emplacement du fichier de test utilisé pour évaluer le modèle.
Non	`--o` *outputFileName*	Indique le nom et l'emplacement du fichier de sortie qui va stocker les résultats de l'évaluation.
Non	`--c` *categoryCount*	Spécifie le nombre de catégories du modèle ; doit être une valeur numérique. **Remarque :** S'applique uniquement au modèle Text Classification.
Non	`--d` *trueOrfalse*	Spécifie s'il faut afficher une table avec une analyse détaillée des entités ; la valeur doit être `true` ou `false`, comme suit : **true** Les résultats d'évaluation détaillés sont requis. **false** Les résultats d'évaluation détaillés ne sont pas requis. La valeur par défaut est `false`. La table *Résultats d'évaluation du modèle*, et *Matrice de confusion* avec ses colonnes, comme décrit ci-dessous, affichent les nombres par entité. **Remarque :** Si la commande est exécutée sans cet argument, ou avec la valeur d'argument `false`, la table *Résultats d'évaluation du modèle* et *Matrice de confusion* ne sont pas affichées. Seules les *Statistiques d'évaluation du modèle* sont affichées.

Sortie

Statistiques d'évaluation du modèle

L'exécution de cette commande affiche ces statistiques d'évaluation sous forme de tableau :

- **Précision** : il s'agit d'une mesure de l'exactitude. La précision définit la proportion d'uplets correctement identifiés.
- **Rappel** : il s'agit d'une mesure de l'exhaustivité des résultats. Rappel peut être défini sous forme de fraction d'instances pertinentes récupérées.

- **Mesure F1** : il s'agit de la mesure de la précision d'un test. Le calcul du score F1 prend en compte à la fois la précision et le rappel du test. Il peut être interprété comme la moyenne pondérée de la précision et du rappel, où un score F1 de valeur 1 est le meilleur score et un score F1 de valeur 0 est le pire score.
- **Exactitude** : cette option mesure le degré d'exactitude des résultats. Elle définit la proximité de la valeur mesurée par rapport à la valeur connue.

Résultats d'évaluation du modèle

Si la commande est exécutée avec l'argument `--d true`, le nombre de correspondances de toutes les entités est affiché dans un tableau. Les colonnes de la table sont les suivantes :

Input Count	Nombre d'occurrences de l'entité dans les données d'entrée.
Mismatch Count	Nombre de fois où la correspondance d'entités a échoué.
Match Count	Nombre de fois où la correspondance d'entités a réussi.

Matrice de confusion

La *Matrice de Confusion* (illustrée ci-dessous) permet de visualiser les performances d'un algorithme. Elle illustre les performances d'un modèle de classification.

La colonne représente les instances d'une classe prédite, tandis que la ligne représente les instances d'une classe réelle. Certains des termes associés à la matrice de confusion sont les suivants :

Réel	Nombre d'occurrences de l'entité dans la classe réelle.
Prédit	Nombre d'occurrences de l'entité dans la classe prédite.
TP	**True Positive :** Nombre d'occurrences de l'entité prédites comme positives et, en réalité, également vraies.
TN	**True Negative :** Nombre d'occurrences de l'entité prédites comme négatives mais, en réalité, vraies.
FP	**False Positive :** Nombre d'occurrences de l'entité prédites comme positives mais, en réalité, fausses.
FN	**False Negative :** Nombre d'occurrences de l'entité prédites comme négatives et, en réalité, également fausses.

Exemple

Cet exemple :

- Évalue le modèle appelé « MyModel ».
- Utilise un fichier de test appelé « ModelTestFile » au même emplacement.
- Enregistre la sortie de l'évaluation dans un fichier appelé « MyModelTestOutput ».
- Indique un nombre de catégories 4.
- Indique qu'une analyse détaillée de l'évaluation est requise.

```
iemodel evaluate model --n MyModel --t
C:\Spectrum\IEModels\ModelTestFile --o
C:\Spectrum\IEModels\MyModelTestOutput --c 4 --d true
```

iemodel evaluate train_model

La commande `iemodel evaluate train_model` évalue et forme un modèle existant du **module Information Extraction**. Cette fonction ne peut pas être effectuée sur un nouveau modèle.

> **Remarque :** Pour obtenir de meilleurs résultats d'évaluation et de formation d'un **module Information Extraction** existant, utilisez cette commande : `iemodel trainAndevaluate model`. Pour des informations plus détaillées, reportez-vous à la section **iemodel trainAndevaluate model** à la page 37.

Utilisation

`iemodel evaluate train_model --f` *trainingOptionsFile* `--u` *trueOrFalse* `--o` *outputFileName* `--c` *categoryCount* `--d` *trueOrfalse*

Obligatoire	Argument	Description
Oui	`--f` *trainingOptionsFile*	Indique le nom et l'emplacement du fichier d'options de formation utilisé pour former le modèle. Les chemins de répertoire que vous indiquez ici sont relatifs à l'emplacement dans lequel vous exécutez l'utilitaire Administration.
Non	`--u` *overWriteIfExists*	Spécifie s'il faut remplacer le modèle formé existant (le cas échéant). *TrueOrFalse* est l'un des éléments suivants : **Vrai** Remplace le modèle existant. **false** Ne remplace pas le modèle existant.
Non	`--o` *outputFileName*	Indique le nom et l'emplacement du fichier de sortie qui va stocker les résultats de l'évaluation.
Non	`--c` *categoryCount*	Spécifie le nombre de catégories du modèle ; doit être une valeur numérique. **Remarque :** S'applique uniquement au modèle Text Classification.
Non	`--d` *trueOrfalse*	Spécifie s'il faut afficher une table avec une analyse détaillée des entités ; la valeur doit être `true` ou `false`, comme suit : **true** Les résultats d'évaluation détaillés sont requis.

Obligatoire	Argument	Description
		false Les résultats d'évaluation détaillés ne sont pas requis. La valeur par défaut est `false`. La table *Résultats d'évaluation du modèle*, avec ses colonnes, comme décrit ci-dessous, affiche le nombre par entité. **Remarque :** Si la commande est exécutée sans cet argument, ou avec la valeur d'argument `false`, la table Résultats d'évaluation du modèle n'est pas affichée. Seules les Statistiques d'évaluation du modèle sont affichées.

Sortie

Statistiques d'évaluation du modèle

L'exécution de cette commande affiche ces statistiques d'évaluation sous forme de tableau :

- Précision
- Rappel
- Mesure F1

Résultats d'évaluation du modèle

Si la commande est exécutée avec l'argument `--d true`, le nombre de correspondances de toutes les entités est affiché dans un tableau. Les colonnes de la table sont les suivantes :

Input Count	Nombre d'occurrences de l'entité dans les données d'entrée.
Mismatch Count	Nombre de fois où la correspondance d'entités a échoué.
Match Count	Nombre de fois où la correspondance d'entités a réussi.

Exemple

Cet exemple :

- Utilise un fichier d'options de formation appelé « ModelTrainingFile » qui se trouve dans « C:\Spectrum\IEModels ».
- Remplace tout fichier de sortie existant portant le même nom.
- Enregistre la sortie de l'évaluation dans un fichier appelé « MyModelTestOutput ».
- Indique un nombre de catégories 4.
- Indique qu'une analyse détaillée de l'évaluation est requise.

```
iemodel evaluate train_model --f
C:\Spectrum\IEModels\ModelTrainingFile --u true --o
C:\Spectrum\IEModels\MyModelTestOutput --c 4 --d true
```

iemodel export

La commande `iemodel export` exporte un modèle du module Information Extraction et ses métadonnées.

Utilisation

`iemodel export --n`*modelName* `--o`*outputDirectory*

Requis	Argument	Description
Oui	`--n` *modelName*	Indique le nom du modèle à exporter. Les chemins de répertoire que vous indiquez ici sont relatifs à l'emplacement dans lequel vous exécutez l'utilitaire Administration.
Oui	`--o` *outputDirectory*	Indique l'emplacement du dossier qui contiendra le modèle exporté et ses métadonnées.

Exemple

Cet exemple exporte un modèle nommé`MyModel` qui place la sortie dans un dossier appelé « MyModelExport », qui se trouve dans « C:\Spectrum\IEModels\MyModelExport ».

```
iemodel export --n MyModel --o
C:\Spectrum\IEModels\MyModelExport
```

iemodel import

La commande `iemodel import` importe un modèle de module Information Extraction et ses métadonnées.

Utilisation

`iemodel import --n`*modelName* `--o` *inputDirectory*`--u` *trueOrFalse*

Requis	Argument	Description
Oui	`--n` *modelName*	Indique le nom du modèle à importer. Les chemins de répertoire que vous indiquez ici sont relatifs à l'emplacement dans lequel vous exécutez l'utilitaire Administration.
Oui	`--o` *inputDirectory*	Indique l'emplacement du dossier qui contiendra le modèle importé et ses métadonnées.
Non	`--u` *overWriteIfExists*	Spécifie s'il faut remplacer le modèle existant (le cas échéant). *TrueOrFalse* est l'un des éléments suivants : **vrai** Remplace le modèle existant. **faux** Ne remplace pas le modèle existant.

Exemple

Cet exemple importe un modèle nomméMyModel qui stocke le modèle dans un dossier appelé « MyModelExport », qui se trouve dans « C:\Spectrum\IEModels\MyModelExport ». Il remplace également tout modèle existant portant le même nom.

```
iemodel import --n MyModel --o
C:\Spectrum\IEModels\MyModelExport --u true
```

iemodel list

La commande `iemodel list` renvoie une liste de tous les modèles du module Information Extraction.

Utilisation

```
iemodel list
```

Exemple

Cet exemple répertorie tous les modèles.

```
iemodel list
```

iemodel train

La commande `iemodel train` forme un modèle du module Information Extraction. Elle appelle votre fichier d'options de formation, qui pointe vers votre fichier d'entrée et applique les options que vous avez spécifiées.

Utilisation

`iemodel train --f`*trainingOptionsFile* `--u`*trueOrFalse*

Obligatoire	Argument	Description
Oui	`--f` *trainingOptionsFile*	Indique le nom et l'emplacement du fichier d'options de formation utilisé pour former le modèle. Les chemins de répertoire que vous indiquez ici sont relatifs à l'emplacement dans lequel vous exécutez l'utilitaire Administration.
Non	`--u` *trueOrFalse*	Indique si le modèle existant portant le même nom (le cas échéant) doit être remplacé, où *TrueOrFalse* est l'une des valeurs suivantes : **true** Remplace le modèle existant. **false** Ne remplace pas le modèle existant.

Exemple

Cet exemple forme un modèle figurant dans le fichier *TrainingOptions.xml* qui est stocké sur le lecteur `C:` et remplace tout modèle existant portant le même nom.

```
iemodel train --f c:/TrainingOptions.xml --u true
```

iemodel trainAndevaluate model

La commande `iemodel trainAndevaluate model` évalue et forme un nouveau modèle, ainsi que d'un modèle existant. S'il s'agit d'un modèle existant, vous devez le remplacer par le modèle récemment formé à l'aide de « true » pour l'argument--u de la commande.

Cette commande appelle votre fichier d'options de formation et fournit un fichier de sortie facultatif avec les résultats de l'évaluation, si vous décidez de produire ce fichier.

Utilisation

```
iemodel trainAndevaluate
model--ftrainingOptionsFile--utrueOrFalse--ooutputFileName--ccategoryCount--dtrueOrfalse
```

Obligatoire	Argument	Description
Oui	`--f` *trainingOptionsFile*	Indique le nom et l'emplacement du fichier d'options de formation utilisé pour former le modèle. Les chemins de répertoire que vous indiquez ici sont relatifs à l'emplacement dans lequel vous exécutez l'utilitaire Administration.
Non	`--u` *overWriteIfExists*	Spécifie s'il faut remplacer le modèle formé existant (le cas échéant). **Vrai** Remplace le modèle existant. **false** Ne remplace pas le modèle existant.
Non	`--o` *outputFileName*	Indique le nom et l'emplacement du fichier de sortie qui va stocker les résultats de l'évaluation.
Non	`--c` *categoryCount*	Spécifie le nombre de catégories du modèle ; doit être une valeur numérique. **Remarque :** S'applique uniquement au modèle Text Classification.
Non	`--d` *trueOrfalse*	Spécifie s'il faut afficher une table avec une analyse détaillée des entités ; la valeur doit être `true` ou `false`, comme suit : **true** Les résultats d'évaluation détaillés sont requis. **false** Les résultats d'évaluation détaillés ne sont pas requis. La valeur par défaut est `false`. La table *Résultats d'évaluation du modèle*, et *Matrice de confusion* avec ses colonnes, comme décrit ci-dessous, affichent les nombres par entité. **Remarque :** Si la commande est exécutée sans cet argument, ou avec la valeur d'argument `false`, la table *Résultats d'évaluation du modèle* et *Matrice de confusion* ne sont pas affichées. Seules les *Statistiques d'évaluation du modèle* sont affichées.

Sortie

Statistiques d'évaluation du modèle

L'exécution de cette commande affiche ces statistiques d'évaluation sous forme de tableau :

- **Précision** : il s'agit d'une mesure de l'exactitude. La précision définit la proportion d'uplets correctement identifiés.
- **Rappel** : il s'agit d'une mesure de l'exhaustivité des résultats. Rappel peut être défini sous forme de fraction d'instances pertinentes récupérées.
- **Mesure F1** : il s'agit de la mesure de la précision d'un test. Le calcul du score F1 prend en compte à la fois la précision et le rappel du test. Il peut être interprété comme la moyenne pondérée de la précision et du rappel, où un score F1 de valeur 1 est le meilleur score et un score F1 de valeur 0 est le pire score.
- **Exactitude** : cette option mesure le degré d'exactitude des résultats. Elle définit la proximité de la valeur mesurée par rapport à la valeur connue.

Résultats d'évaluation du modèle

Si la commande est exécutée avec l'argument `--d true`, le nombre de correspondances de toutes les entités est affiché dans un tableau. Les colonnes de la table sont les suivantes :

Input Count	Nombre d'occurrences de l'entité dans les données d'entrée.
Mismatch Count	Nombre de fois où la correspondance d'entités a échoué.
Match Count	Nombre de fois où la correspondance d'entités a réussi.

Matrice de confusion

La *Matrice de Confusion* (illustrée ci-dessous) permet de visualiser les performances d'un algorithme. Elle illustre les performances d'un modèle de classification.

La colonne représente les instances d'une classe prédite, tandis que la ligne représente les instances d'une classe réelle. Certains des termes associés à la matrice de confusion sont les suivants :

Réel	Nombre d'occurrences de l'entité dans la classe réelle.
Prédit	Nombre d'occurrences de l'entité dans la classe prédite.
TP	**True Positive :** Nombre d'occurrences de l'entité prédites comme positives et, en réalité, également vraies.
TN	**True Negative :** Nombre d'occurrences de l'entité prédites comme négatives mais, en réalité, vraies.
FP	**False Positive :** Nombre d'occurrences de l'entité prédites comme positives mais, en réalité, fausses.
FN	**False Negative :** Nombre d'occurrences de l'entité prédites comme négatives et, en réalité, également fausses.

Exemple

Cet exemple :

- Utilise un fichier d'options de formation appelé « ModelTrainingFile » qui se trouve dans « C:\Spectrum\IEModels ».
- Remplace tout fichier de sortie existant portant le même nom.
- Enregistre la sortie de l'évaluation dans un fichier appelé « MyModelTestOutput ».
- Indique un nombre de catégories 4.
- Indique qu'une analyse détaillée de l'évaluation est requise.

```
iemodel trainAndevaluate model --f
C:\Spectrum\IEModels\ModelTrainingFile --u true --o
C:\Spectrum\IEModels\MyModelTestOutput --c 4 --d true
```

6 - Référence aux stages

In this section

Composants du module Information Extraction

Le module Information Extraction inclut ces stages.

- **Read From Documents** : lit les données d'entrée non structurées de différents formats de fichier et en extrait le contenu.
- **Entity Extractor** : extrait les entités telles que les noms et les adresses de données non structurées transmises sous forme de chaînes.
- **Text Categorizer** : affecte des catégories personnalisées à un contenu non structuré ou à un texte en clair (tel que des courriers électroniques, des articles d'actualité et des commentaires) en fonction de la quantité de contenu correspondant à ladite catégorie.
- **Relationship Extractor** : extrait les relations entre les entités.

Read from Documents

Read from Documents est un stage source qui lit les données d'entrée non structurées de différents formats de fichier et en extrait le contenu. Il peut s'agir de documents juridiques, de retours/commentaires de clients, de revues de produits, d'articles d'actualités, de blogs, de réseaux sociaux, etc. Read from Documents extrait également les champs de métadonnées comme l'auteur et la date de création. Une fois les données extraites, elles peuvent être utilisées pour différents types de traitement, comme, entre autres, l'extraction d'entités et la manipulation de chaînes. Il est également possible d'utiliser les données pour générer des index de recherche à des fins de recherche de texte non structuré.

Remarque : Chaque document est considéré comme un enregistrement de ce stage.

Entrée

L'entrée de Read from Documents est un seul fichier ou dossier. Ce stage prend en charge les types de fichier suivants :

- Text
- PDF
- Microsoft Outlook
- Microsoft Word
- HTML

Read from Documents réalise trois types d'extraction :

- Document : utilisation du document tout entier
- Page : utilisation d'une page spécifique d'un document
- Selective : utilisation d'une partie sélectionnée d'un document
- Bookmarks : utilisation des signets d'un document PDF

Read from Documents fait partie du module Information Extraction.

Options

Onglet de propriétés du fichier

Le tableau suivant répertorie les options qui contrôlent le type d'informations renvoyé par Read from Documents.

Tableau 1 : Options de ReadfromDocuments

Option	Description
Server name	Indique le nom du serveur Spectrum Technology Platform en cours d'utilisation.
File/folder name	Chemin d'accès et nom du document ou du dossier source. Si vous souhaitez pointer vers un dossier, utilisez un astérisque comme caractère générique (« * ») pour sélectionner tous les fichiers du dossier. Si vous souhaitez pointer vers plusieurs fichiers du même type figurant dans un dossier, utilisez le caractère générique suivi de l'extension de fichier (« *.pdf »).
File type	Type de fichier du document source, qui sera automatiquement sélectionné une fois que vous sélectionnez une source : • Texte • PDF • Microsoft Outlook • Microsoft Word • HTML

Option	Description	
Extraction type	**Documentation**	Utilise le document tout entier.
	Page	Utilise une page spécifique d'un document.
	Selection	Utilise une portion sélectionnée d'un document.
	Bookmarks	Utilise les signets d'un document PDF.
Page selection	Uniquement avec le type d'extraction Page. Sélectionne toutes les pages ou une plage de pages.	
Selected extraction	Uniquement avec le type d'extraction Selection. Indique le type de recherche.	
Specify text	Uniquement avec le type d'extraction Selection. Indique le texte à rechercher.	
Exclude start text	Uniquement avec le type d'extraction Selection et l'option Start text. Omet la chaîne saisie des données renvoyées.	
Specify end text	Uniquement avec le type d'extraction Selection. Indique le texte de fin à rechercher.	
Exclude end text	Uniquement avec le type d'extraction Selection. Omet la chaîne saisie de la fin des données renvoyées.	
Selection return	Uniquement avec le type d'extraction Selection. Indique le nombre de paragraphes à renvoyer pour chaque résultat. Par exemple, si vous sélectionnez « 2 », les données renvoyées pour chaque résultat incluent le paragraphe contenant le résultat plus le paragraphe suivant, pour un total de deux paragraphes. La valeur par défaut est 1. Non valide lorsque le texte de fin est indiqué.	

Onglet Champs

Cliquez sur **Régénérer** pour définir des champs d'entrée.

Tableau 2 : Options de données de sortie

Option	Description
Attribute Name	Indique l'attribut le plus proche du champ d'entrée. Par exemple, si l'un de vos champs contient des informations de date et que vous l'appelez « Date », l'attribut « Date » est affecté à ce champ. Cette colonne n'est pas modifiable.
Nom	Nom du champ. Cette colonne est modifiable.
Type	Type de données du champ.
Include	Indique les champs à inclure dans un index de recherche.

Sortie

Le stage Read from Documents comporte deux ports sortants. Un port capture les données lues par le stage et renvoyées en fonction des critères saisis. Il peut s'agir de texte en clair ou de métadonnées (comme l'auteur, la langue, la date de création, etc.). Ce port peut être connecté à tout stage capable de lire les données entrantes, comme Write to File ou Write to XML, ainsi qu'à des stages primaires comme Validate Address ou Write to Search Index. Il peut également être connecté au stage Information Extractor si vous souhaitez renvoyer des informations sur certains types d'entité qui se trouvent dans le document. Lorsque vous sélectionnez le type d'extraction Document, la sortie contient des données plates ; lorsque vous sélectionnez le type d'extraction Page ou Selection, la sortie contient des données hiérarchiques.

L'autre port collecte tout enregistrement que le flux de données n'a pas correctement traité. Il s'agit du port d'erreur et les enregistrements arrivant dans le collecteur de données, via ce port, sont considérés comme non conformes. Capturer des enregistrements non conformes peut vous aider à identifier le problème avec ces enregistrements. Lorsque vous associez un collecteur de données au port d'erreur, le fichier de sortie qui en résulte contient tous les champs des enregistrements non conformes. Il contiendra également un champ Raison qui indique la raison pour laquelle l'enregistrement a échoué.

Tableau 3 : Sortie d'Unstructured Reader

Nom du champ	Description/Valeurs valides
Author	Contient généralement le nom de la personne qui a créé ou mis à jour le document. Ces informations font partie des métadonnées du document.
Bookmark	Contient tous les signets du fichier d'entrée au format PDF. Pour les types d'extraction Bookmarks uniquement.
BookmarkNo	Contient tous les signets du fichier d'entrée au format PDF. Pour les types d'extraction Bookmarks uniquement.
ContentLength	Indique la longueur du document. Cette valeur varie suivant le type d'extraction sélectionné : **Document** Nombre de pages du document. **Page** « 1 », pour représenter une seule page de contenu.
Table des matières	Varie en fonction du type d'extraction. Par exemple, les types d'extraction Document sortent le document tout entier sous forme de données plates. Les types d'extraction Page, Selection et Bookmarks sortent des données hiérarchiques.
ContentType	Indique le type de document lu, comme PDF, .txt, etc.
Creator	Contient généralement le nom de la personne qui a créé le document. Ces informations font partie des métadonnées du document.
Date	Indique la date de création ou de la dernière mise à jour du document.
Keywords	Contient tous le mots clés fournis dans les métadonnées du document.
Langue	Indique la langue dans laquelle le document a été rédigé.
NPages	Indique le nombre de pages du document.

Nom du champ	Description/Valeurs valides
PageContents	Contient le contenu de la ou des pages sélectionnées. Pour les types d'extraction Page uniquement.
PageNo	Contient le numéro de page du signet. Pour les types d'extraction Page uniquement.
Parent	Contient le chemin d'accès au signet, similaire au XPath d'un fichier XML. Pour les types d'extraction Bookmarks uniquement.
ResourceName	Indique le nom de fichier du document.
SectionContents	Contient le contenu de la section sélectionnée. Pour les types d'extraction Selection uniquement.
SectionNo	Indique le numéro de cette section au sein du document. Pour les types d'extraction Selection uniquement.
Subject	Contient le sujet du document fourni dans les métadonnées du document.
Titre	Contient le titre du document fourni dans les métadonnées du document.

Entity Extractor

Entity Extractor extrait des entités telles que des noms et des adresses de chaînes de données non structurées (également connues sous le nom de « texte en clair »).

Il est possible que toutes les entités d'un type sélectionné ne soient pas renvoyées, car la précision varie en fonction du type d'entrée. Étant donné que Entity Extractor utilise un traitement de langage naturel, une chaîne contenant une phrase grammaticalement correcte d'un article d'actualité ou d'un blog renverra probablement des noms plus précis qu'une simple liste de noms et de dates.

Entrée

Entity Extractor accepte des chaînes de données non structurées comme entrée. Il peut également utiliser le stage **Read from Documents** comme entrée si vous souhaitez extraire des entités d'un document non structuré. Le stage **Read from Documents** lit le document et renvoie le texte en fonction des paramètres définis par l'utilisateur. **Entity Extractor** extrait les informations requises de ce texte en fonction des entités sélectionnées.

Tableau 4 : Format d'entrée

Nom du champ	Description
PlainText	Chaîne de données non structurée dont vous souhaitez extraire des informations.

Options

Les options Entity Extractor vous permettent de sélectionner des entités dont vous souhaitez extraire des informations de la chaîne d'entrée. Par défaut, vous pouvez extraire des informations via *Person* et *Address* comme types d'entité. Cependant, vous pouvez utiliser la fonction **Ajout rapide** et sélectionner tout ou partie des 15 entités préconfigurées.

Nom de l'option	Description
Neutralisation des options système par défaut avec les valeurs suivantes	Cochez la case pour remplacer les types d'entité par défaut *Address* et *Person*. Lorsque vous cochez la case, le bouton **Ajout rapide** est activé. Cliquez sur ce bouton et sélectionnez les entités dont vous avez besoin pour l'extraction du texte. Les entités sélectionnées sont ajoutées à la liste **Type d'entité**.

Nom de l'option	Description
Type d'entité	Indique le type de données que vous souhaitez extraire de la chaîne non structurée. ***Address*** ***CreditCard*** ***Date*** ***Email*** ***HashTag*** ***ISBN*** ***Location*** ***Mention*** ***Organization*** ***Person*** ***Phone*** ***ProperNouns*** ***SSN*** ***WebAddress*** ***ZipCode***
Nombre d'entités de sortie	Indique s'il faut renvoyer le nombre de fois où une entité donnée s'est retrouvée dans la sortie. **true** Renvoie le nombre des entités retrouvées dans la chaîne non structurée. **false** Ne renvoie pas le nombre des entités retrouvées dans la chaîne non structurée.

Réponse

La sortie de **Entity Extractor** est une liste des entités correspondantes retrouvées dans la chaîne d'entrée. Par exemple, si vous avez sélectionné un type d'entité « Person », la sortie est une liste des noms de personne retrouvés dans la chaîne d'entrée. De même, si vous avez sélectionné un **Type d'entité** « Date », la sortie est une liste des dates retrouvées dans la chaîne d'entrée.

Chaque entité,qu'il s'agisse d'un nom, d'une adresse ou d'une date, est renvoyée une seule fois, même si l'entité apparaît plusieurs fois dans la chaîne d'entrée.

Pour afficher le nombre de fois où l'entité apparaît dans la chaîne d'entrée, vous pouvez sélectionner l'option **Nombre d'entités de sortie** dans la fenêtre Options **Entity Extractor**.

Nom du champ	Description
Text	Texte extrait de la chaîne.
Type	Type d'entité du texte extrait. L'un des éléments suivants : ***Address*** ***CreditCard*** ***Date*** ***Email*** ***HashTag*** ***ISBN*** ***Location*** ***Mention*** ***Organization*** ***Person*** ***Phone*** ***ProperNouns*** ***SSN*** ***WebAddress*** ***ZipCode***
Count	Si l'option de renvoi d'un nombre est activée, ce champ contient le nombre de fois où une entité donnée est apparue dans l'entrée. Par exemple, si vous avez décidé de renvoyer les entités `Name` et que le texte d'entrée contient cinq instances du nom `John`, le nom `John` est inclus une seule fois dans la sortie, avec `Name` comme type d'entité et « 5 » comme nombre de sorties.

Text Categorizer

Ce stage vous permet d'attribuer des catégories personnalisées à un contenu non structuré ou à un texte en clair (tel que des courriers électroniques, des articles d'actualité et des commentaires) en fonction de la quantité de contenu correspondant qu'il comporte. Le stage répertorie les catégories définies à partir desquelles vous pouvez sélectionner celle dont vous avez besoin pour votre

catégorisation. Cependant, vous devez créer ces catégories en formant un modèle d'élément de catégorisation avec vos données. Pour plus d'informations, voir **Introduction à la catégorisation de texte** à la page 7.

Entrée

Le stage accepte des chaînes de données non structurées comme entrée. Il peut également utiliser le stage **Read from Documents** comme entrée si vous souhaitez catégoriser le texte d'un document non structuré. Le stage **Read from Documents** lit le document et renvoie le texte en fonction des paramètres définis par l'utilisateur. Le document est lu par le stage **Text Categorizer** pour vous fournir la sortie de votre choix.

Tableau 5 : Format d'entrée

Nom du champ	Description
PlainText	Chaîne de données non structurée dont vous souhaitez extraire des informations.

Options

Text Categorizer Options vous permet de sélectionner les paramètres en fonction desquels vous souhaitez classer votre chaîne d'entrée de données. Vous pouvez sélectionner le modèle de catégorisation et le nombre de niveaux de correspondance à appliquer à la sortie. Par exemple, uniquement la correspondance la plus proche ou la correspondance la plus proche plus la deuxième correspondance proche.

Nom de l'option	Description
Neutralisation des options système par défaut avec les valeurs suivantes	Pour remplacer l'option par défaut et sélectionner l'élément de catégorisation dans la liste déroulante **Nom de l'élément de catégorisation**.
Nom de l'élément de catégorisation	Spécifie le modèle à utiliser pour la catégorisation du texte. Cela répertorie tous les modèles que vous avez formés dans la phase de catégorisation du texte. **Remarque :** Pour plus d'informations, reportez-vous à la section **Formation du modèle** à la page 12.

Nom de l'option	Description
Nombre de catégories	Nombre de niveaux de correspondance de catégorie que vous souhaitez appliquer à la sortie. Par exemple, sélectionnez 1 pour afficher uniquement la correspondance la plus proche et 2 pour afficher la correspondance la plus proche plus la deuxième correspondance proche. **Remarque :** La valeur maximale correspond au nombre de classes différentes indiqué lors de la formation du modèle.

Réponse

La sortie répertorie les catégories dans lesquelles le contenu de la chaîne d'entrée est classé et le classement de cette catégorie. Le classement signifie le degré de correspondance du contenu d'entrée par rapport à la catégorie. Par exemple, 1 signifie qu'il s'agit de la correspondance la plus proche de la catégorie et 2 signifie qu'il s'agit de la correspondance la plus proche plus la deuxième correspondance proche.

Nom du champ	Description
Category	Catégorie prédite pour chaque enregistrement du fichier d'entrée.
Rank	Classement des catégories du score le plus élevé au score le plus bas.

Relationship Extractor

Le stage **Relationship Extractor** vous permet d'identifier les types de relation entre les entités identifiées dans le contenu source.

Le stage **Relationship Extractor** identifie :

1. Entity1
2. Entity1 Type
3. Relation Type
4. Entity2

5. Entity2 Type

Important : Le stage tente d'obtenir la plus grande précision possible lorsqu'il identifie les types de relation entre deux entités données du texte d'entrée. Cependant, les relations autres que la relation précise entre deux entités peuvent elles aussi être identifiées via l'analyse de phrases complexes du texte d'entrée.

Entrée

Le stage **Relationship Extractor** accepte les chaînes de données en langage naturel comme entrée et identifie les entités et les types de relation qui existent entre chaque paire d'entités.

Utilisez le stage **Read from Documents** comme stage source si le texte d'entrée provient d'un document non structuré. Le stage **Read from Documents** lit le document et renvoie le texte en fonction des paramètres définis par l'utilisateur.

Le stage **Relationship Extractor** identifie alors toutes les entités et le type de relation entre chaque paire d'entités.

Tableau 6 : Format d'entrée

Nom du champ	Description
PlainText	Chaîne de données non structurée dont vous souhaitez identifier les types de relation existants entre chaque paire d'entités.

Options

Les options du stage **Relationship Extractor** vous permettent de spécifier les types de relation que vous souhaitez identifier dans le texte d'entrée.

Par défaut, les types de relation identifiés sont les suivants :

1. *AffiliatedWith*
2. *LivesIn*
3. *OrgBasedIn*
4. *LocatedIn*

Nom de l'option	Description
Neutralisation des options système par défaut avec les valeurs suivantes	Cochez la case pour remplacer les types de relation par défaut identifiés et spécifier les types de relation que vous souhaitez identifier et extraire du texte d'entrée. Lorsque vous cochez la case, le bouton **Ajout rapide** est activé. Cliquez sur **Ajout rapide** pour sélectionner les types de relation que vous souhaitez identifier dans le texte. Les entités sélectionnées sont ajoutées à la liste **Type de relation**.

Réponse

La sortie de **Relationship Extractor** est une liste des jeux de relations identifiés entre des paires d'entités rencontrés dans la chaîne de sortie.

Par exemple, si, dans les options du stage, vous avez sélectionné les types de relation *LivesIn* et *OrgBasedIn* pour les extraire, la sortie contient une liste de tous les jeux de *Person LivesIn Location* et *Organization OrgBasedIn Location* identifiés dans le texte d'entrée.

Chaque paire d'entités avec son type de relation est répertoriée une seule fois.

Pour chaque jeu d'entités extrait et leur relation, les informations extraites sont les suivantes :

Nom du champ	Description
Entity1	La première entité d'une paire d'entités extraite du texte d'entrée.
Entity1 Type	Le type d'entité de la première entité de la paire d'entités extraite du texte d'entrée. Le type d'entité est l'un des suivants : • *Person* • *Organization* • *Location*
Type	Le type de relation identifié entre Entity1 et Entity2. Pour plus d'informations sur les types de relation, reportez-vous à la section **Relationship Types** à la page 25. **Remarque :** Seuls les types de relation sélectionnés pour l'extraction dans les options de stage sont identifiés et répertoriés.

Nom du champ	Description
Entity2	La deuxième entité d'une paire d'entités extraite du texte d'entrée.
Entity2 Type	Le type d'entité de la deuxième entité de la paire d'entités extraite du texte d'entrée. Le type d'entité est l'un des suivants : • *Person* • *Organization* • *Location*

Notices

Avis USPS®

Pitney Bowes Inc. détient une licence non exclusive pour la publication et la vente de bases de données ZIP + 4® sur des supports optiques et magnétiques. Les marques de commerce suivantes appartiennent à United States Postal Service : CASS, CASS Certified, DPV, eLOT, FASTforward, First-Class Mail, Intelligent Mail, LACS^{Link}, NCOA^{Link}, PAVE, PLANET Code, Postal Service, POSTNET, Post Office, RDI, Suite^{Link} , United States Postal Service, Standard Mail, United States Post Office, USPS, ZIP Code et ZIP + 4. Cette liste de marques de commerce appartenant à U.S. Postal Service n'est pas exhaustive.

Pitney Bowes Inc. détient une licence non exclusive de USPS® pour le traitement NCOA^{Link}®.

Les prix des produits, des options et des services de Pitney Bowes Software ne sont pas établis, contrôlés ni approuvés par USPS® ni par le gouvernement des États-Unis. Lors de l'utilisation de données RDI™ pour déterminer les frais d'expédition de colis, le choix commercial de l'entreprise de distribution de colis à utiliser n'est pas fait par USPS® ni par le gouvernement des État-Unis.

Fournisseur de données et avis associés

Les produits de données contenus sur ce support et utilisés au sein des applications Pitney Bowes Software sont protégés par différentes marques de commerce et par un ou plusieurs des copyrights suivants :

© Copyright United States Postal Service. Tous droits réservés.

© 2014 TomTom. Tous droits réservés. TomTom et le logo TomTom logo sont des marques déposées de TomTom N.V.

© 2016 HERE

Source : INEGI (Instituto Nacional de Estadística y Geografía)

Basées sur les données électroniques © National Land Survey Sweden.

© Copyright United States Census Bureau

© Copyright Nova Marketing Group, Inc.

Des portions de ce programme sont sous © Copyright 1993-2007 de Nova Marketing Group Inc. Tous droits réservés.

© Copyright Second Decimal, LLC

© Copyright Canada Post Corporation

Ce CD-ROM contient des données provenant d'une compilation dont Canada Post Corporation possède le copyright.

© 2007 Claritas, Inc.

3001 Summer Street
Stamford CT 06926-0700
USA

www.pitneybowes.com