**Les problèmes rencontrés lors de l'extraction des données des conventions collectives nationales**

Au cours du développement de la solution permettant de récupérer les données des CCN, plusieurs problèmes ont été rencontrés lors de l'extraction des données :

1. Les sections ne sont pas les mêmes, par exemple pour le taux de la cotisation on le trouve parfois sous forme de tableau et d'autres fois sous forme de paragraphe ou autre.

**Sous forme d’un texte :**

A close-up of a computer code

Description automatically generated

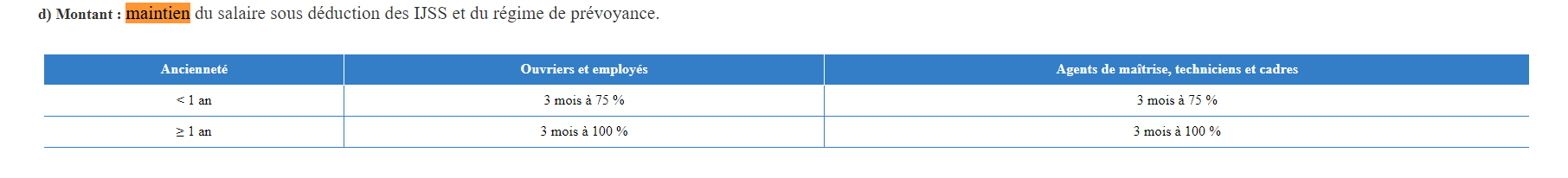
**Sous forme d’un tableau :**

A screenshot of a computer

Description automatically generated

1. La détection des tableaux avec les mots clés n'est pas toujours possible parce que les mots clés qu'on cherche se trouvent parfois dans les entêtes des tableaux et d'autres fois dans des paragraphe au-dessus du tableau

**Titre au-dessous du tableau :**



**Dans l’entête d’un tableau :**

A blue and white rectangular box with white text

Description automatically generated

1. La structure des tableaux n'est pas toujours la même dans les CCN.

**L’une des structures d’un tableau :**

A screenshot of a computer

Description automatically generated

**Autre structure d’un tableau :**

A blue and white rectangular box with white text

Description automatically generated

1. Nous avons essayé d'extraire des données en partant de fichiers PDF au lieu de HTML mais ce n'était pas possible vu que les fichiers PDS sont structurés de manière complexe. En plus, la récupération se fait que sur la page en question, autrement dit si on a un tableau divisé sur deux pages on ne peut récupérer que la partie dans qui se trouve dans la première page détectée.

A screenshot of a computer

Description automatically generated

1. Pour récupérer une partie de texte il faut mettre des conditions de limites pour pouvoir extraire exactement la partie voulue, mais comme déjà évoqué chaque CCN (ou presque) a une structure particulière ce qui nous empêche de mettre des conditions qui peuvent s'appliquer sur toutes les CCN

**Structure d’une page :**

A close up of text

Description automatically generated

**Autre structure d’une page :**

A close up of text

Description automatically generated