## Qu'est-ce que le format CONLL ?

Le format CoNLL est  l'un des formats de structure de données le plus couramment utilisé dans le domaine de traitement automatique des langues, et il existe plusieurs versions dans son évolution, par exemple CoNLL-2000, CoNLL-2002, CoNLL-2003, CoNLL-X, CoNLL-U etc. Dans ce cas, on aborde principalement la première version CoNLL-2000 et la version CoNLL-U qu' on utilise le plus souvent.

Aujourd'hui, on utilise principalement le format CoNLL-U qui est encodées dans des fichiers de texte brut avec trois types de lignes :

1. Lignes de mots contenant l'annotation d'un mot/token dans 10 champs séparés par des caractères de tabulation simples ; voir ci-dessous.
2. Lignes vides marquant les limites de la phrase.
3. Lignes de commentaires commençant par un dièse (#).

## Quel est le rôle du format CONLL ?

Le format CONLL peut être utilisé dans de nombreuses tâches de traitement du langage naturel, telles que la reconnaissance d'entités nommées, l'étiquetage morpho-syntaxique, l'analyse syntaxique, etc. En raison de sa simplicité et de son utilisation répandue, le format CONLL est devenu l'un des formats d'annotation les plus couramment utilisés dans la recherche en traitement du langage naturel.

## Les variantes du format CONLL

1. CoNLL-2000

CONLL-2000 a été conçu pour les tâches de marquage morpho-syntaxique et de découpage en blocs (chunking). Il y a trois colonnes, une colonne pour les mots, et deux colonnes pour les étiquettes de parties du discours.

Exemple :

| | | | | | |
|---|---|---|---|---|---|
| Le | chat | mange | une | souris | . |
| DET | NOUN | VERB | DET | NOUN | PUNCT |
| B-NP | I-NP | B-VP | B-NP | INP | O |

L'étiquette B-NP signifie le début de NP (begin NP) et L'étiquette I-NP signifie l' intérieur de NP(inside NP).

2. CoNLL-U

CoNLL-U est une nouvelle version qui est basée sur CoNLL-X. Il existe principalement des différences ci-dessous par rapport à CoNLL-X.

| | |
|---|---|
| éléments de colonne | Le format CoNLL-X représente chaque colonne comme une information d'étiquetage, tandis que le format CoNLL-U inclut plusieurs attributs d'information d'étiquetage dans chaque colonne. |

| | |
|---|---|
| extensibilité | Le format CoNLL-X n'est pas conçu pour être extensible, tandis que le format CoNLL-U est extensible et permet facilement l'ajout de nouvelles informations. |
| gestion des caractères Unicode | Le format CoNLL-X n’arrive pas à gérer les caractères Unicode, tandis que le format CoNLL-U utilise l'encodage UTF-8 pour les gérer. |
| désignation des attributs | Les noms d'attributs dans le format CoNLL-X sont basés sur les tâches et les ensembles de données, tandis que ceux dans le format CoNLL-U sont génériques et indépendants des tâches et des ensembles de données. |

Donc, le format CoNLL-U est plus extensible et étendu que le format CoNLL-X. Et CoNLL-U contient les champs suivants:

| | |
|---|---|
| ID | Indice du mot, nombre entier commençant à 1 pour chaque nouvelle phrase ; peut être une plage pour les mots multiples ; peut être un nombre décimal pour les nœuds vides. |
| FORM | Forme du mot ou symbole de ponctuation. |
| LEMMA | Lemme ou racine de la forme du mot. |
| UPOS | Universal POS tags |
| XPOS | Language-specific POS; underscore si non disponible. |
| FEATS | Liste des caractéristiques morphologiques provenant de l'inventaire universel des caractéristiques ou d'une extension spécifique à une langue définie ; soulignement si non disponible. |
| HEAD | Tête du mot courant, qui est soit une valeur de ID, soit zéro (0). |
| DEPREL | Relation de dépendance universelle avec le HEAD (racine si HEAD = 0) ou un sous-type de celui-ci défini et spécifique à la langue. |
| DEPS | Graphe de dépendance amélioré sous la forme d'une liste de paires tête-décharge. |
| MISC | Toute autre annotation. |

Les champs doivent en outre répondre aux contraintes suivantes :

1. Les champs ne doivent pas être vides.

2. Les champs autres que FORM, LEMMA et MISC ne doivent pas contenir de caractères d'espace.

3. Le tiret bas (_) est utilisé pour indiquer des valeurs non spécifiées dans tous les champs, sauf ID.