Estructura de digitalización de conocimiento

1. Introducción

El presente documento presenta la estructura de digitalización de conocimiento que se ha definido para facilitar el procesamiento y acceso a la información contenida en corpus digitales de lenguas amazónicas de escasos recursos. La estructura descrita representa la manera en que se deberían almacenar los datos relevantes dentro de un corpus para que se apliquen adecuadamente métodos de acceso y procesamiento.

Cada una de las secciones describe las clases y sus atributos con su respectiva descripción.

2. Corpus

La clase Corpus agrupa los elementos relevantes del corpus y su metadato asociado, permitiendo un manejo centralizado de la información. Se espera que el corpus esté compuesto por oraciones, ya que no se admite el procesamiento de textos debido a la falta de un tokenizador específico para oraciones.

- root_directory: Directorio raíz desde donde se lee el archivo del corpus.
- encoding: Tipo de codificación utilizada para leer el corpus.
- lengua_principal: Idioma principal del corpus (por defecto, None).
- n_entries: Número de entradas en el corpus.
- entries: Lista de objetos de tipo CorpusEntry.
- text_column: Nombre de la columna que contiene el texto principal del corpus.
- file_column: Nombre de la columna que contiene el nombre del archivo asociado a las entradas.
- pos_column: Nombre de la columna que contiene las etiquetas de parte del discurso (POS tags).

- mb_column: Nombre de la columna que contiene la separación morfológica.
- id_column: Nombre de la columna que contiene los identificadores de las entradas.

3. Corpus Entry

La clase CorpusEntry representa una entrada específica dentro del corpus, que puede ser una oración o una palabra.

- file: Archivo al que pertenece la entrada.
- text: Texto de la entrada.
- words: Lista de objetos de tipo WordEntry
- id: Identificador único de la entrada

4. Word entry

La clase WordEntry almacena la información de cada palabra en una entrada del corpus.

- word: La palabra en sí, o el signo de puntuación correspondiente.
- mb: Lista de morfemas correspondientes a la palabra.
- pos: Lista de etiquetas POS para cada morfema.

5. Multilingual corpus

Esta clase hereda de Corpus y permite gestionar corpus multilingües, aceptando la integración de corpus paralelos.

- languages: Lista de lenguas adicionales además de la principal.
- gloss_columns: Diccionario de lenguas con sus respectivas columnas de glosas.
- ft_columns: Diccionario de lenguas con sus respectivas columnas de traducción libre.

6. **Multilingual corpus entry**

MultilingualCorpusEntry hereda de CorpusEntry y permite gestionar entradas de corpus paralelos.

- ft: Diccionario que contiene las traducciones libres en las lenguas adicionales.

7. **Multilingual Word entry**

Esta clase extiende WordEntry y permite almacenar glosas para palabras en diferentes lenguas.

- gloss: Diccionario de lenguas adicionales con sus glosas correspondientes.