1. Formato de los archivos

Para que corran los modelos de análisis de texto es crucial que los archivos crudos vengan en el formato adecuado. Originalmente, se había propuesto la siguiente estructura:

- `general`: Carpeta general
  - `Libro 1`: Carpeta de un libro particular
    - `pdf`: Carpeta con páginas en PDF
      - `hoja 1.pdf`
      - `hoja 2.pdf`
      - ...
    - `jpg`: Carpeta con páginas en JPG
      - `hoja 1.jpg`
      - `hoja 2.jpg`
      - ...
    - `txt`: Carpeta con páginas en texto
      - `full-txt/libro completo.txt`: Carpeta con un solo texto correspondiente al libro completo
      - `hoja 1.jpg`
      - `hoja 2.jpg`
      - ...
  - `Libro 2`: Carpeta con otro libro particular
  - ...
- `models`: Carpeta con los modelos (para uso interno)
- `results`: Carpeta con los resultados (HTMLs, XMLs, JSONs, etc)

Sin embargo, después de la primera etapa se decidió utilizar la estructura elegida por la UNAM, que divide los archivos por tipo primero y luego por libro. A continuación presentamos el nuevo formato con más detalle:

- `general`: Carpeta general
  - `pdf`: Carpeta con libros en PDF
    - `Libro 1`: Carpeta con PDFs de un libro particular
      - `hoja 1.pdf`

        - ◇ `hoja 2.pdf`
        - ◇ ...
    - ○ ...
- `jpg`: Carpeta con libros en JPG
    - ○ `Libro 1`: Carpeta con JPGs de un libro particular
        - ◇ `hoja 1.jpg`
        - ◇ `hoja 2.jpg`
        - ◇ ...
    - ○ ...
- `txt`: Carpeta con libros *completos* en texto (no se utilizaron las páginas sueltas)
    - ○ `idioma 1`: Carpeta con libros de un idioma en particular
        - ◇ `Libro 1.txt`: Texto de un libro completo
        - ◇ `Libro 2.txt`
        - ◇ ...
    - ○ ...
- $\texttt{pdf}_{50mb}$: *Carpeta con libros en PDF*
    - `Libro 1`: Carpeta con PDFs de un libro particular, en archivos de 50 MB
        - ○ `parte 1.pdf`
        - ○ `parte 2.pdf`
        - ○ ...
    - ...

- `meta`: Carpeta con un archivo de metadatos por libro (adentro tiene su idioma)
- `models`: Carpeta con los modelos (para uso interno)
- `results`: Carpeta con los resultados (HTMLs, XMLs, JSONs, etc)
- `extracts`: Carpeta con un JSON por libro que incluye los extractos del texto

Una de las ventajas de este formato es que es fácil concentrar el uso de un tipo de archivo. Como los modelos usualmente utilizan únicamente un tipo de datos, esta estructura es particularmente útil. En el caso particular

de la carpeta de los textos se tiene el beneficio adicional de que la estructura aporta metadatos sobre los libros (el idioma, por ejemplo, está implícito).

A pesar de todo eso, en el momento de la instalación del software se encontró que en el servidor los archivos estaban en el formato que se había contemplado en un inicio. Esto presentó un problema importante porque para correr el proceso en todos los libros se requeriría copiarlos o moverlos al formato correcto. Dado el tamaño de la colección, no fue viable hacerlo, salvo para una muestra pequeña (aunque no trivial) de libros, con el fin de corroborar que el programa funcionara sin problemas.

Como se prefirió la estructura inicial, en el futuro se deberá cambiar el formato de entrada del análisis de texto. Para ello se tendrá que cambiar por completo la forma en la que se explota la estructura de carpetas, lo que tendrá impacto tanto en el proceso de `luigi` como en diversos módulos y funciones del proceso, ya que la decisión de qué se debe correr y qué no se obtiene directamente de cómo están organizadas las carpetas. Además de cambiar la lectura de los datos, habrá algunas decisiones arquitectónicas que se deberá tomar. Por ejemplo, se tendrá que decidir cómo y dónde incluir los idiomas dentro de la estructura, dónde incorporar los archivos de metadatos, etc. Una propuesta posible sería la estructura original, pero con los siguientes cambios:

- En la carpeta de un libro en particular `Libro i`:
  - `idioma libro i/Libro i.txt`: Texto completo del libro
  - `Libro i.meta`: Archivo de metadatos del libro
- `general/meta`: Carpeta con metadatos (listas de archivos, listas de idiomas, etc)
  - `Idioma 1`: Metadatos de un idioma particular
  - ...

Por supuesto, habrá que hacer un análisis detallado para ver si todas estas propuestas son viables y si no presentarán dificultades inherentes a la estructura.