Revisión Metadata

#Borrador Revision Metadata
Lista sin orden de importancia

1. Chequear que los links funcionen tanto en episodios como en los contenidos (Conectar vpn de ser necesario)

2. Buscar inconsistencias en 'Title' y 'CleanTitle'.
Hay cosas que considero que se ven con el ojo y no con algun metodo automatizado.
Desde el Studio 3T pueden ayudarse mucho con las queries y hacer busquedas como:
- 'Season' (en el/los idioma/s que corresponda/n)
- Parentesis, corchetes, ":", "-", etc
- Titulos repetidos (comparar la metadata entre si para ver si son realmente repetidos o revisar tmb deeplinks)
- titulos con 'Ep' (por episodio) o parecidos, (Capitulo, Episode, etc)
- 'Live'
De los ejemplos dados, no necesariamente quiere decir que sean titulos que hayan que limpiar y/o esten mal. Son palabras generales que pueden implicar que se deba hacer una limpieza en los titulos para aplicar en el scraping de la plataforma (no en el script ya hecho aparte replace.py)

Asimismo en CleanTitle buscar casos que se pudieron haber aplicado o titulos que terminen quedando vacios o sin sentido en comparacion al Title (sobre todo con los regex) por ejemplo donde el titulo sea "Caperucita: la pelicula" y el CleanTitle haya quedado como "Caperucita: ".

3. Confirmar modelos de negocio y revisar que los episodios tengan los mismos que las series.
Todos los modelos de negocio que figuren en los episodios deben también figurar para las series incluso si hay al menos 1 episodio con un package diferente.

4. Mirar e indagar sobre patrones que se repitan o parezca que se repitan mucho en todos los fields.
Por ejemplo: si notan que todos los 'Year' son 2020. o todos los generos son 'Action', o todos los títulos empiezan con x letras/palabras.

5. Duraciones que puedan no estar en minutos

Por ej si ven un solo contenido que tenga 1000 minutos, o 9472 , puede que este realmente indicando el tiempo en segundos y no en mins como debiera

6. Metadata que se guarda en listas (Genre, Directors, Cast, Providers):
Chequear que no haya nombres raros que no parezcan pertenecer a la lista correspondiente
Chequear que no haya nombres repetidos, esten mal escritos o tengan espacios entre item e item.
Que no sean un solo string que contengan todos los datos en vez de estar separados correctamente
Tambien a veces se pueden reconocer que hay directores en cast (porque sabes que son directores o conoces la pelicula) que no tengan tmp tags html

7. Chequear links de imagenes

8. Sinopsis, buscar:
- tags de html o escapes
- que tengan en ciertos casos promociones de la plataforma que no correspondan
- fechas que no correspondan a la sinopsis
- Sinopsis repetidas en distintos contenidos

9. Metadata que figure en episodios pero no en series (en la serie es donde mas importa por el match) como año, director, cast.

10. Rating que no sea el Parental Rating sino el de puntaje del contenido

11. OriginalTitle, estos casos vienen de la plataforma pero no siempre son correctos o estan bien. Confirmar que son Original titles.
Tmb hay que buscar patrones o ver si agregan algun dato que no corresponda.
Por ej: 'Pepito, 2020'.

14. Revision de Season y Episode (numbers) en los episodios.
Ver que esten bien estructurados
Gralmente no deberian haber casos repetidos en una serie con season y/o episode igual
Hemos visto casos particulares en que desde la plataforma venian repetidos pero con diferente idioma, ej en ingles y en español. Estos casos se dejan asi.

15. Observar si figura algun titulo que no corresponda a la plataforma. Por ejemplo: si la plataforma es alemana y se encuentra metadata en español. Pueden ser casos que realmente figuren asi en la plataforma o que el scraping se este revisando con browser o parametro de otro idioma

16. Hay cosas particulares que podemos notar en plataformas de ciertos paises, ahora el unico ejemplo que se me ocurre es que muchas plataformas de Francia aclaran el idioma en el titulo, como VF u otro. Seguramente esto este en el replace, pero puede que sigan quedando titulos asi. Este punto se puede explayar mejor.

17. BuyPrice & RentPrice :
Posibles errores:
- que sean iguales entre si
- que RentPrice sea mayor que BuyPrice
Lo mismo relacionandolo con las definiciones, que SD por ejemplo sea mas caro que 4K.

17. Definiciones repetidas

18. Currencies que no correspondan al pais scrapeado

20. Types , chequear. El upload no admite cosas diferentes a movie o serie, pero podemos por ejemplo mirar/saber si hay mayoria de pelis o series y que la plataforma no muestre lo mismo

21. Episodios que puedan estar figurando como pelis o series.

22. Cuando comparamos entre varios paises tmb deberiamos observar si en gral traen los mismos datos si no se informo que era el mismo contenido,
- Idiomas que esten mal entre los diferentes paises de la plataforma
-

PLUS: (mas relacionada posiblemente al script o a la plataforma)

1. Metadata *importante* (Year, Directors, Episode, Season, etc) que aparezca como null se puede chequear en la plataforma (ej año, director, n° episode/season). A veces no lo conseguimos en el scraping inicial que hicimos pero descubrimos que la plataforma lo muestra.

2. Ids que sean hashes o no sean necesariamente un id unico que venga de la plataforma. Chequear su viabilidad (ej a veces ponemos un slug pero no es la mejor opcion o puede haber repetidos que sean contenidos diferentes)

3. Lo ideal seria corroborar en la plataforma al menos varios datos que sean visibles y que sean los mismos o tengan coherencia con lo scrapeado.
Ej. En la plataforma vemos que un contenido indica como año '2013' y en el api de donde sacamos ese dato dice '2019'. Sobre todo puede pasar con los episodios o series esto.