Agriculture Dataset

# Enriquecimiento de datos de los artículos del dataset

Vamos a analizar en primer lugar los artículos proporcionados por la Universidad de Murcia.

Identificador PMID (para búsqueda en Semantic Scholar)

Se han traducido los PMC de los 127 artículos proporcionados a su identificador PMID que permite una búsqueda más sencilla en nuestra base de datos de Semantic Scholar. A pesar de que Europe PMC proporciona un CSV que podría permitir dicha “traducción”, hemos comprobado que para varios de los artículos del dataset proporcionado no se incluye el identificador PMID correspondiente. Por dicho motivo, hemos recurrido al servicio online: <https://www.ncbi.nlm.nih.gov/pmc/utils/idconv/v1.0>

De los 127 artículos del dataset, únicamente hay uno que no se resuelve correctamente, y que de hecho parece tener una temática un poco desviada de la propuesta (agricultura). (PMC4395563). Por este motivo, la resolución del PMID de dicho artículo se ha realizado de forma manual:

df.loc[df.PMCID == "PMC4392563", ['PMID', 'DOI']] = '25697273', '10.1007/s11751-015-0211-9'

Descarga del texto completo de los artículos

Para su posible uso tanto para la extracción de entidades como para la extracción de tópicos, hemos descargado el texto completo tanto en formato PDF como XML. En ambos casos, la descarga se ha realizado directamente desde EuropePMC, ya que la API REST de artículos permite la descarga completa del fichero en formato XML. El PDF no está disponible a través de la API REST, pero la dirección asociada a cada artículo sigue un formato normalizado, por lo que los artículos en dicho formato han podido descargarse fácilmente mediante un proceso de crawling.

* Descarga en formato PDF: Se dispone de PDF asociado a todos los artículos salvo los 5 siguientes: PMC4844397, PMC5521873, PMC5620588, PMC5669304 y PMC6472519
* Descarga en formato XML: Se dispone de XML asociado a todos los artículos salvo el PMC6472519.

Descarga de las anotaciones de los artículos

Europe PMC dispone de una API de anotaciones que permite acceder a entidades anotadas por terceros, tanto sobre el abstract del artículo como sobre el texto completo. Se han descargado dichas anotaciones para su potencial uso tanto por Everis como por parte de la UC3M para la extracción de tópicos. Para ello se considerará la extracción de tópicos a partir únicamente de esas entidades anotadas, lo que puede simplificar considerablemente la limpieza del vocabulario y mejorar la calidad de los tópicos obtenidos, al obtenerse palabras basadas en términos más relevantes.

La siguiente figura muestra el tipo de entidades incluidas en las anotaciones de los artículos, y que pueden descargarse de manera conjunta en formato JSON a partir de la API disponible en Europe PMC.

A screenshot of a cell phone

Description automatically generated

En cuanto al formato de las anotaciones, se incluye un ejemplo que muestra la obtención de una entidad de tipo “Gene Ontology” proporcionada directamente por Europe PMC. Se puede observar que dicha anotación (la 1.181 del artículo) se proporciona contextualizada (prefix/postfix), y se incluye tanto la entidad precisa (PTP1B) como el tipo de entidad de la que se trata (Gene\_Proteins).

A picture containing bird

Description automatically generated

Cabe notar que la precisión de las anotaciones está fuera del propósito de Europe PMC por lo que resultar posible notar un elevado número de fallos que sugiere mostrar cierta cautela a la hora de usarlas al menos de manera exclusiva. A modo de ejemplo, obsérvense las anotaciones 1154 a 1157 para el mismo artículo. A pesar de que las entidades obtenidas pueden resultar de utilidad podemos comprobar que:

* El tipo de entidad es erróneo en 3 de las 4 anotaciones
* Se pierden palabras que podrían resultar relevantes para un modelado de tópicos (diabetic).

A screenshot of a cell phone

Description automatically generated

# Ampliación del dataset para entrenamiento de los modelos de tópicos

Vamos a estudiar distintos procedimientos de ampliación del dataset, con el objetivo de generar un conjunto de artículos más extenso sobre el que poder aplicar con garantías el algoritmo de extracción automática de tópicos. En concreto, se han implementado los dos siguientes métodos de extensión:

* Incorporar las referencias de los 127 artículos del dataset base
* Incorporar todos los artículos co-autorizados por los autores de los 127 artículos del dataset base

El motivo por el que se han seleccionado estas estrategias es por resultar de sencilla implementación a partir de la información disponible en Semantic Scholar, además de proporcionar la representación más completa a la que podemos tener acceso de la producción científica de los autores del conjunto de entrenamiento. Otras estrategias posibles que no se han implementado, pero podrían explorarse en caso necesario, serían ampliar con artículos de revistas seleccionadas (por ejemplo a partir del catálogo de la *National Agriculture Library*[[1]](#footnote-1)) o utilizar alguna medida de distancias (como BM25) para seleccionar artículos similares a los del dataset base.

Sobre las dos estrategias implementadas, hay que tomar una serie de precauciones a la hora de construir el dataset ampliado:

* No incluir artículos repetidos, ya que muchos artículos pueden aparecer múltiples veces en las referencias y/o autorías
* Verificar que los artículos dispongan de Abstract, necesario para poder hacer el análisis de tópicos.
* Verificar que se dispone de PMID. Aunque no es imprescindible, permite cruzar con la base de datos de EuropePMC y descargar las anotaciones asociadas a los artículos.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Artículos únicos | Con PMID | Con Abstract | Con PMID y Abstract |
| Autorías (847  autores únicos) | 54.102 | 22.114 | 43.390 | 21.610 |
| Referencias | 9.234 | 4.877 | 8.688 | 4.833 |

De acuerdo a los números que se presentan en la tabla de arriba, queda patente que un elevado porcentaje de los artículos seleccionados disponen de un abstract, siendo la disponibilidad de un PMID una condición más restrictiva. Por este motivo, se pueden considerar dos criterios diferentes para generar un dataset extendido:

**El número total de artículos en el dataset extendido es de 51.042 artículos cuando se seleccionan artículos con abstract, y de 25.792 si se seleccionan artículos con abstract y PMID.**

Distribución de los artículos por año de publicación

En primer lugar vamos a analizar la distribución de los artículos del dataset base y ampliado según el año de publicación. El siguiente diagrama de barras muestra que los artículos facilitados han sido publicados entre 2012 y 2019, aumentando el número de artículos disponibles según el año de publicación.



Al analizar la distribución para los artículos del dataset ampliado vemos que dicha tendencia “creciente” es de hecho propia de la distribución de artículos en Semantic Scholar (y probablemente del número de publicaciones científicas en general). Se observa además un comportamiento muy similar de las publicaciones con y sin PMID (con Abstract en ambos casos).

Finalmente, al analizar el eje horizontal vemos que existen publicaciones muy antiguas (incluso de 1.805) y algunas con fecha sin definir (9.999 en la figura). Se ha decido conservar en el dataset únicamente los artículos con fecha definida y posterior al año 2.000, lo que supone un muy pequeño decremento en el número de artículos disponibles en el dataset ampliado.



**Tras filtrar por año, el número total de artículos en el dataset extendido es de 47.065 artículos cuando se seleccionan artículos con abstract, y de 24.450 si se seleccionan artículos con abstract y PMID.**

Distribución de los artículos por campo de estudio

Es importante notar que la identificación de autores y referencias en Semantic Scholar es un procedimiento automatizado que por tanto puede estar sujeto a ciertos errores. Por este motivo, conviene realizar cuantas validaciones sean posibles para eliminar algunos artículos susceptibles de haber sido seleccionados de manera incorrecta. A tal fin, disponemos del campo fieldsOfStudy que proporciona la temática principal o temáticas principales del artículo, inferida de forma automática.

La figura adjunta muestra la distribución de los valores en el campo fieldsOfStudy para los 127 artículos que componen la base de datos básica de Semantic Scholar. Podemos comprobar que la totalidad de los artículos pertenecen al campo *Medicine*, y que una buena cantidad de los mismos han sido asignados también al campo *Biology*. Aparecen además otros campos de estudio en mucha menor proporción, y siempre en combinación con la asignación a *Medicine*. Por lo tanto, parece razonable exigir a los artículos del dataset ampliado una temática similar a los del dataset básico.

Realizando un análisis similar sobre el conjunto completo de datos del conjunto extendido obtenemos la distribución temática que se muestra en la siguiente figura



Vemos que, nuevamente las temáticas de *Medicine* y *Biology* son muy mayoritarias. De hecho, la etiqueta *Medicine* está nuevamente presente en la práctica totalidad de los artículos con PMID (24.429 de 24.450), lo que sugiere que existe una muy alta correlación entre los artículos con PMID y la aparición de dicha etiqueta. Sin embargo, un alto número de artículos con abstract (algunos con PMID y otros no) presentan las etiquetas *Biology* o *Chemistry* sin incluir también la de *Medicine* (diferencia de altura entre las barras azul y naranja).

Por lo tanto, hemos decidido conservar los artículos que contienen alguna de las tres etiquetas con alta presencia en la base de datos con PMID: *Medicine, Biology y Chemistry.* La siguiente figura muestra la reducción en la aparición del resto de etiquetas, y muestra que en la mayoría de los casos los artículos que las contienen no forman parte de ninguna de las temáticas seleccionadas como nucleares del dataset.



Por este motivo, decidimos incorporar el filtrado por temáticas para garantizar un dataset extendido más homogéneo. Como resultado, el dataset extendido formado por artículos con y sin PMID se reduce en torno a un 25%, mientras que el dataset extendido formado por artículos con PMID apenas sufre variación en el número de artículos. Destacamos, además, que el primer dataset extendido es aproximadamente un 50% mayor que el segundo. Estos datasets constituyen la selección final sobre la que ejecutaremos los modelos para extracción automática de tópicos.

**Tras filtrar por año, el número total de artículos en el dataset extendido es de 36.543 artículos cuando se seleccionan artículos con abstract, y de 24.448 si se seleccionan artículos con abstract y PMID.**

Distribución Del Número de Artículos por Autor

Para la caracterización de cada uno de los 847 autores incluidos en los 127 papers del dataset base utilizaremos el conjunto de artículos autorizados por cada autor disponibles en el conjunto de entrenamiento. Es de esperar que la fiabilidad de las caracterizaciones así obtenidas sea más precisa cuanto mayor sea el número de artículos disponibles para cada autor. La siguiente figura muestra el número de autores que tienen un número de autorías concretos (eje horizontal). Como es de esperar, al considerar el dataset de papers con PMID el histograma se desplaza hacia la izquierda, es decir, el número promedio de autorías por autor disminuye.



Podemos ver que, a pesar de que hay un número relativamente alto de autores con una, dos o tres autorías, el número de autores con entre 5 y 80 co-autorías es también bastante elevado, lo que suponen condiciones favorables para los algoritmos de caracterización que se utilizarán en nuestra implementación.

# Generación de los Corpus de Entrenamiento

1. Generar corpus / Lematizar
2. Generar un primer modelo
3. Exportador de salidas para el proyecto
4. Limpiar el vocabulario
5. Explorar el número de tópicos
6. Explorar con Termolator / Annotations

1. <https://agricola.nal.usda.gov/jia/> [↑](#footnote-ref-1)