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 Modelos de Tópicos

En esta sección explicaremos el procedimiento seguido para la generación de los modelos de tópicos, así como para la validación del número de tópicos. Comenzaremos explicando cómo se lleva a cabo la generación del corpus de entrenamiento, y posteriormente se analizará la validación de los parámetros del modelo, y el formato de los ficheros de salida generados.

Lematización y Generación del Corpus de datos para Mallet

Una primera tarea a llevar a cabo para la extracción de tópicos consiste en al homogeneización de los textos asociados al abstract de cada artículo que compone el dataset extendido. Para ello se han realizado las siguientes tareas:

1. Lematización: permite agrupar términos con el mismo significado eliminando información morfológica como pueden ser el número o el género. La lematización se ha llevado a cabo utilizando una versión dockerizada de IXA pipes (<https://github.com/librairy/nlp>). Durante la lematización se presta atención también al *Part of Speech* de cada término, conservando únicamente nombres, adjetivos y verbos.
2. Eliminación de stopwords: Además de stopwords genéricas de lengua inglesa, se han añadido 2.032 stopwords adicionales que son términos de escaso valor semántico para este corpus documental (e.g., *paper, research, agriculture, etc.).* La obtención de estos términos se ha realizado a partir de un modelo de tópicos inicial, de manera que se garantiza la eliminación de aquellos términos que tienen un impacto más significativo en los tópicos obtenidos.
3. Aplicación de equivalencias: Nuevamente, mediante una supervisión manual se han añadido 32 términos equivalentes. Se trata tanto de acrónimos que preferimos visualizar en mayúsculas, como algunas variantes morfológicas que no han sido corregidas adecuadamente por el lematizador (e.g.: breed/breeding).
4. Tokenización: Usando la siguiente expresión regular: [\p{L}\p{N}][\p{L}\p{N}\p{P}]\*\p{L}
5. Filtrado de términos que aparecen en menos de 10 documentos o en más del 60% de los artículos del corpus. De esta manera, conseguimos un vocabulario más reducido que conserva únicamente términos que permiten discriminar temáticas (se eliminan aquellos que aparecen en casi todos los documentos), y eliminamos además algunas erratas.

Tras realizar el filtrado de términos el vocabulario final consta de 15.329 términos.

1. Importación de documentos al formato de Mallet, eliminando aquellos documentos que constan de menos de 15 lemas (salvo los del dataset base, que se conservan todos).

Tras filtrar los documentos con un número de lemas insuficiente, el dataset de entrenamiento consta de 35.322 documetnos.

Validación del Número de Tópicos

Se ha llevado a cabo el entrenamiento con un número variable del número de tópicos, manteniendo constante otros hiperparámetros como son el valor de *alpha*, que define el nivel de dispersión a priori de la representación de los documentos, y el intervalo de reestimación de hiperparámetros. Sin embargo, estos datos suelen mostrar una influencia limitada en la calidad de los tópicos extraídos, siendo el número de tópicos el parámetro más relevante.

Para seleccionar un número adecuado de tópicos, usamos un indicador de coherencia, cuya evolución con el número de tópicos se muestra en la siguiente figura:



La figura muestra que un número de tópicos entre 40 y 250 aproximadamente proporciona valores elevados de la coherencia. De cara a una mayor interpretabilidad se ha seleccionado el modelo con 40 tópicos; sin embargo, una evaluación más precisa podría obtenerse a partir de una evaluación manual subjetiva de los tópicos y/o la validación de los resultados finales, una vez mapeada la salida del modelo de tópicos a las categorías de EuroSciVoc.

La descripción de los tópicos a partir de los términos más relevantes de cada tópico se ha incluido en la siguiente tabla, en la que se indica además, junto a cada tópico, su importancia relativa en el modelo.

0 0.0185 rat effect level study mouse exposure liver dose serum day control result significant decrease treatment compare diet mg/Kg concentration plasma

1 0.01062 patient result fracture treatment elbow study pain surgery month score conclusion complication injury bone mm follow-up treat evaluate perform total

2 0.01689 insect volatile aphid host female pest male control adult larva effect egg feed study test result pheromone compare trap parasitoid

3 0.02841 stress salt tolerance drought growth nacl na level abiotic water temperature effect study treatment high abiotic\_stress result cold mm root

4 0.01725 soil root biomass microbial effect carbon decomposition nitrogen litter change nutrient rhizosphere study addition total isotope respiration decrease result affect

5 0.01561 temperature ph study acid effect starch result treatment higher pef concentration fermentation oil decrease sample storage milk investigate min stability

6 0.01943 genetic genotype marker study analysis accession allele locus breed qtl polymorphism rs variation chromosome snp map association region result find

7 0.01757 mouse expression effect il rat study cd cell level inhibit result decrease treatment activation reduce protein brain disease suggest inflammation

8 0.0228 antioxidant ros oxidative\_stress stress enzyme decrease level chlorophyll photosynthetic leaf result H2O reduce peroxidase chloroplast glutathione photosynthesis reactive\_oxygen effect fluorescence

9 0.01729 bacteria strain genetic isolate DNA sample plasmid PCR analysis antibiotic study sequence detect assay abundance primer citrus rhizosphere reveal result

10 0.01784 study age child result woman prevalence factor adult conclusion man association female higher symptom compare male 95%\_ci analysis find level

11 0.02338 absorption nanoparticle surface study result high nm concentration sample membrane ion film effect particle acid ph exhibit electrode np investigate

12 0.01723 soil biochar effect treatment organic fertilizer manure straw fraction Kg Ha rice fertilization result study carbon residue total decrease control

13 0.02453 farming maize Ha Kg grain legume weed reduce fertilizer effect soil rice result nutrient higher study seasonal treatment compare low

14 0.01196 expression mir cell tissue tumor level study metastasis invasion patient result protein carcinoma crc progression overexpression mrna inhibit normal hcc

15 0.01562 pathogen genetic acid infection defense bacteria salicylic mutant pr host ppgpp fungus expression tomato protein defence result suggest accumulation immunity

16 0.01975 wheat grain rice cultivar flower barley effect pollen genotype floral high flowering temperature result seed difference reduce higher protein weight

17 0.01469 herbivore acid ja defense ethylene attenuata leaf level signaling jasmonate herbivory insect attack jasmonic accumulation defens signal biosynthesis genetic defence

18 0.02138 root transport cell leaf tissue uptake study cell\_wall phloem stomatal xylem phytolith silicon result wall silica distribution suggest plasma\_membrane aquaporin

19 0.0118 virus infection vaccine antibody antigen strain result test study mouse vaccination hcv assay positive serum hbv sample mab disease influenza

20 0.02757 genetic expression analysis transcript express study protein mrna reveal sequence transcriptome mirna RNA tissue cdna level encode microarray pattern family

21 0.01673 emission rice concentration ch water N2O study nitrogen result reduce soil total sediment seasonal flux removal nh hg methane flood

22 0.02465 seed root growth germination ga soybean shoot cytokinin effect acid concentration treatment level length higher medium control study result day

23 0.02288 genetic sequence family analysis study region phylogenetic genus find mutation evolutionary lineage reveal result suggest clade chromosome bp sequencing DNA

24 0.01403 virus infection potato disease host isolate infected leaf symptom spread pathogen vector tuber infect RNA control incidence tomato transmission mosaic

25 0.02178 study antioxidant acid flavonoid medicinal result medicinal\_plants total fraction isolate phytochemical find alkaloid leaf phenolic test evaluate antimicrobial analysis highest

26 0.02458 mutant genetic protein mutation phenotype expression chloroplast plastid cell result embryo regulate suggest wild-type RNA arabidopsis\_thaliana oocyte encode control affect

27 0.03903 analysis result study parameter measure estimate predict sample prediction image test compare determine evaluate difference accuracy perform correlation distribution analyze

28 0.0208 cd concentration soil zn metal uptake root arsenic mg cu toxicity pb se accumulation effect rice decrease level cadmium shoot

29 0.0276 leaf fruit growth effect concentration tomato treatment level higher water temperature total result study decrease cultivar high control glucosinolate affect

30 0.0311 habitat study landscape forest effect change rodent abundance site result pattern biodiversity tree suggest find grassland ecosystem high density bird

31 0.02277 genetic expression transgenic promoter tobacco express protein transgenic\_plants rice senescence encode result level analysis leaf overexpression control transcription\_factor transcription\_factors wrky

32 0.01759 fungus root bacteria isolate strain host endophytic mycorrhizal symbiosis growth endophyte inoculation study soil colonization plant\_growth effect symbiotic nematode arbuscular

33 0.01575 patient study result treatment 95%\_ci compare level conclusion significant disease serum control meta-analysis higher analysis score ci therapy diagnosis healthy

34 0.02429 protein peptide bind protease inhibit express sequence analysis subunit encode purify recombinant residue proteomic study family kda suggest clone find

35 0.01678 protein kinase ca signaling cell activation receptor phosphorylation regulate activate expression er mapk calcium bind mta inhibit signal suggest pak

36 0.02828 metabolism acid level lipid change enzyme sugar amino\_acids accumulation biosynthesis sucrose carbon analysis result reveal study starch growth nitrogen decrease

37 0.01958 cell inhibit effect expression apoptosis study result treatment assay inhibition growth level tumor decrease differentiation protein cell\_cycle proliferation cell\_lines mouse

38 0.0135 patient tumor diagnosis result lesion treatment study ct imaging PET/CT conclusion disease evaluate undergo month lung detect perform carcinoma image

39 0.02003 enzyme inhibit reaction substrate acid bind study gst derivative biosynthesis synthase catalyze ligand residue result synthesize catalytic inhibition formation ring

Obtención de los Tópicos de los Papers del Dataset Base y de Los Autores entregados

La caracterización de cada artículo o autor de interés del corpus documental se ha volcado en unos ficheros de texto disponibles en cada carpeta asociada a un modelo de tópicos. En todos los casos, se trata de un vector con la misma longitud que el número de tópicos del modelo. En el caso de los artículos, se trata directamente de la representación para dicho documento proporcionada por el algoritmo de extracción de tópicos. En el caso de los autores, se trata del vector promedio de las representación de todos los artículos escritos por el autor en cuestión.

En las dos tablas siguientes se muestran las primeras líneas de dichos ficheros. En el caso de los artículos científicos se incluye el identificador de proyecto, tal y como fue proporcionado por la Unversidad de Murcia. En el caso de los autores, se incluye el nombre del autor junto con el número de artículos asociados a dicho autor (y su correspondiente representación vectorial.

PMC5055509 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.7686644196510315, 0.0, 0.0, 0.2313355803489685, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0

PMC5066492 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.8844794631004333, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.11552053689956665, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0

PMC5095167 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.03900504484772682, 0.0, 0.0, 0.0, 0.0, 0.3882959485054016, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.5726990103721619, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0

Luigia Giuzio 8 0.0, 0.0, 0.0, 0.1182233989238739, 0.0, 0.0035839362535625696, 0.009915398433804512, 0.0, 0.029166322201490402, 0.0, 0.0, 0.028806405141949654, 0.0, 0.056935179978609085, 0.0099061643704772, 0.0, 0.44114452600479126, 0.0, 0.0, 0.0, 0.0, 0.007371360901743174, 0.0, 0.0, 0.0, 0.06431687623262405, 0.0, 0.006368404719978571, 0.0, 0.11073829233646393, 0.0, 0.0012645808747038245, 0.0, 0.0, 0.10180072486400604, 0.0, 0.010458420030772686, 0.0, 0.0, 0.0

Ryota Kawaguchi 4 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.22206296026706696, 0.0, 0.16450275480747223, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.004344842862337828, 0.0, 0.012789078988134861, 0.0, 0.0, 0.0, 0.029724635183811188, 0.0, 0.0, 0.10799359530210495, 0.0, 0.0, 0.00430265860632062, 0.0, 0.0, 0.0, 0.3471006155014038, 0.0, 0.0, 0.10717882215976715

Sandra Pelletier 35 0.000579490268137306, 0.0, 0.0, 0.042322978377342224, 0.00422121025621891, 0.02148803137242794, 0.03204987943172455, 0.0005330310668796301, 0.01643768884241581, 0.0, 0.015339047648012638, 0.002654248382896185, 0.0, 0.0, 0.0, 0.14738938212394714, 0.007481575943529606, 0.0029100936371833086, 0.0971606969833374, 0.0012639821507036686, 0.14650654792785645, 0.0, 0.09140985459089279, 0.00807173177599907, 0.014755317941308022, 0.0031544340308755636, 0.1397145539522171, 0.005137971602380276, 0.003831938374787569, 0.03954064100980759, 0.0034258821979165077, 0.026930879801511765, 0.002312909811735153, 0.0003256032650824636, 0.011915876530110836, 0.017568320035934448, 0.07275331765413284, 0.0, 0.0033061013091355562, 0.017506813630461693

Resulta evidente a la vista de los resultados que la representación de los artículos es claramente más dispersa que la de los autores, lo cual resulta lógico si tenemos en cuenta que la representación de cada autor se obtiene como el promedio de un número (generalmente) alto de vectores correspondientes a artículos diferentes.

La siguiente figura muestra la importancia promedio de cada tópico del modelo de 40 tópicos, al promediar sobre todos los artículos del dataset base, y al hacerlo para cada uno de los autores del dataset.



# Exportación de resultados a las categorías de la Taxonomía EuroSciVoc

Jesús (TBD)

# Posibles extensiones

1. Entrenar modelos con las entidades obtenidas por Everis
2. Entrenar modelos con las entidades reconocidas por EuropePMC
3. Entrenar modelos con el dataset amplio filtrando los abstract con las entidades reconocidas (EuropePMC)

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