UNIVERSIDAD TÉCNICA PARTICULAR DE LOJA

Sistemas basados en conocimiento

Integrantes:

- Elvis Ayala
- Fernando León

Tutor: Ing. Janneth Chicaiza Espinosa

Tema: Informe - Clustering de publicaciones Covid 19

Introducción

Dada la gran cantidad de publicaciones científicas relacionadas a la pandemia por la que está atravesando actualmente el mundo al año 2020 por el virus denominado como coronavirus y a la rápida propagación de este, se ha decidido trabajar con estas publicaciones en las cuales se encuentra información importante acerca de estudios realizados por científicos y laboratorios alrededor del mundo, y debido a la necesidad y a la gran importancia que tiene acceder a este tipo de información para los investigadores y científicos relacionados con este tema, se ha decidido facilitar de cierta forma el acceso a esta información por medio de la agrupación de estas publicaciones con la finalidad de que se tenga una mejor relación entre publicaciones para que la búsqueda resulte con un menor grado de complejidad.

Pre-procesamiento de datos

Para la exploración y limpieza de archivos se usó python y algunas librerías necesarias para la exploración como pandas, numpy, json, etc. Los archivos se encuentran en formato CSV, en este formato de archivos lo que se hizo es utilizar herramientas como Excel y LibreOffice para la limpieza de datos por medio de fórmulas y por la funcionalidad de búsqueda y reemplazo que brindan estas 2 herramientas.

Luego se sacó los datos necesarios que contiene el archivo CSV acorde con el modelo conceptual que se planteó, ahora actualizados al modelo ontológico propuesto en “Protege”, se aplicaron algunas funciones necesarias para obtener los datos como autores, organización, locación, abstract, el tipo de recurso bibliográfico. Una vez obtenidos y limpiados los datos del CSV se procedió a su transformación por medio de Jena.

Especificación de Uri

En este se definen las **URI's** a utilizar en el proyecto, para lo cual se van a reutilizar los principales vocabularios a usar en la ontología. Asi como tambien a crear una **URI** base y sus diferentes derivaciones para poder describir los diferentes datos de la ontología.

URI's reutilizadas:

Recurso	Uri	Descripción
DoCO	http://purl.org/spar/doco	en una ontología que proporciona un vocabulario estructurado escrito de componentes del documento, tanto estructurales, como retóricos.
PRO	http://purl.org/spar/pro/	es una ontología para la caracterización de los roles de los agentes: personas, organismos corporativos y agentes computacionales en el proceso de publicación.
Foaf	http://xmlns.com/foaf/0.1/	Esta especificación describe el lenguaje FOAF, definido como un diccionario de propiedades y clases con nombre utilizando la tecnología RDF de W3C.
DCMI	http://purl.org/dc/terms/	Estos términos están destinados a usarse en combinación con términos de metadatos de otros vocabularios compatibles en el contexto de los perfiles de aplicación.
FRAPO	http://purl.org/cerif/frapo	es una ontología para describir la información administrativa de proyectos de investigación, por ejemplo, solicitudes de subvenciones, organismos de financiación, socios de proyectos, etc.
FABIO	http://purl.org/spar/fabio	es una ontología para registrar y publicar en la Web Semántica descripciones de entidades que están publicadas o potencialmente publicables, y que contienen o son referidas por referencias bibliográficas, o entidades utilizadas para definir tales referencias bibliográficas.

Diseño de Uri

Para el diseño de la uri vamos a tomar una estructura base y a partir de esta crear otras para cada recurso del modelo ontológico.

Uri base	http://utpl.edu.ec/sbc/COVID19publications/ontology/
Recurso	**Uri**
catalog	http://utpl.edu.ec/sbc/COVID19publications/ontology/catalog/
dataset	http://utpl.edu.ec/sbc/COVID19publications/ontology/dataset
publication	http://utpl.edu.ec/sbc/COVID19publicationsontology//publication
person	http://utpl.edu.ec/sbc/COVID19publications/ontology/person
author	http://utpl.edu.ec/sbc/COVID19publications/ontology/person/author
organization	http://utpl.edu.ec/sbc/COVID19publications/ontology/organization
journal	http://utpl.edu.ec/sbc/COVID19publications/ontology/journal

Definición de la licencia

Las licencias Creative Commons: Luego de analizar cada una de las licencias propuestas se definió usar esta debido a que permite adaptar y compartir la información siempre y cuando se de atribución a los autores de la misma.

Fuentes de Datos Actualizadas

Nombre de la fuente de datos	COVID-19 Open Research Dataset Challenge	
Proveedor	Kaggle	https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge/data
Última actualización:	09/06/2020	

Nombre de la fuente de datos	COVID-19 Open Research Dataset Challenge	
Proveedor	bioRxiv	https://www.biorxiv.org/search/covid%252

		B19
Última actualización:	01/06/2020	

Tareas de limpieza

Tabla resumen de datos recolectados

ORGANIZACIÓN: biorxiv - medrxiv - PMC (PubMed central)

CATÁLOGO: Kaggle

DATASET: COVID-19 Open Research Dataset Challenge (CORD-19)

CADA DATO DEL DATASET SERIA UNA PUBLICACIÓN

Clase	Instancias
BibliographicResource	10498
Catalog	1
Dataset	1
Organization	3
Person	17.534
Journal	10498
License	10498

Transformación de datos RDF

Para realizar la transformación de los datos de un archivo csv a RDF utilizaremos la librería de Java Jena la cual nos servirá a modelar nuestro esquema de datos rdf con respecto a nuestro modelo ontológico.

Debido al preprocesamiento de los datos, con su conversión y tratamiento de archivos json a csv, la transformación de datos está en una etapa inicial.

Almacenamiento

Para el almacenamiento de los datos RDF nosotros decidimos usar GraphDB debido a que nos resultó más amigable al momento de la subida de datos, sobretodo por las

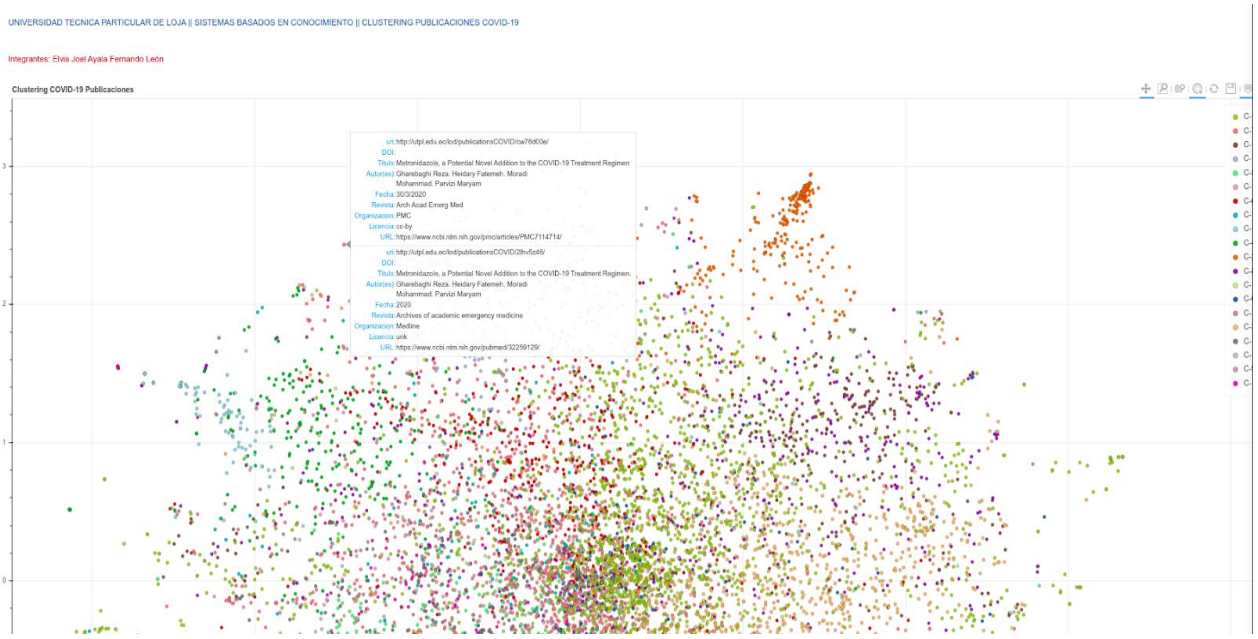
funcionalidades que este presta que hacen que su manejo sea simple, además de su interprete para realizar consultas SPARQL y la manera en que este las presenta al usuario.

Aplicación

Objetivos

- Aplicar técnicas necesarias para el procesamiento de datos
- Realizar clustering por medio de K Means con las publicaciones proporcionadas
- Visualizar la información de publicaciones relacionadas con la enfermedad COVID-19

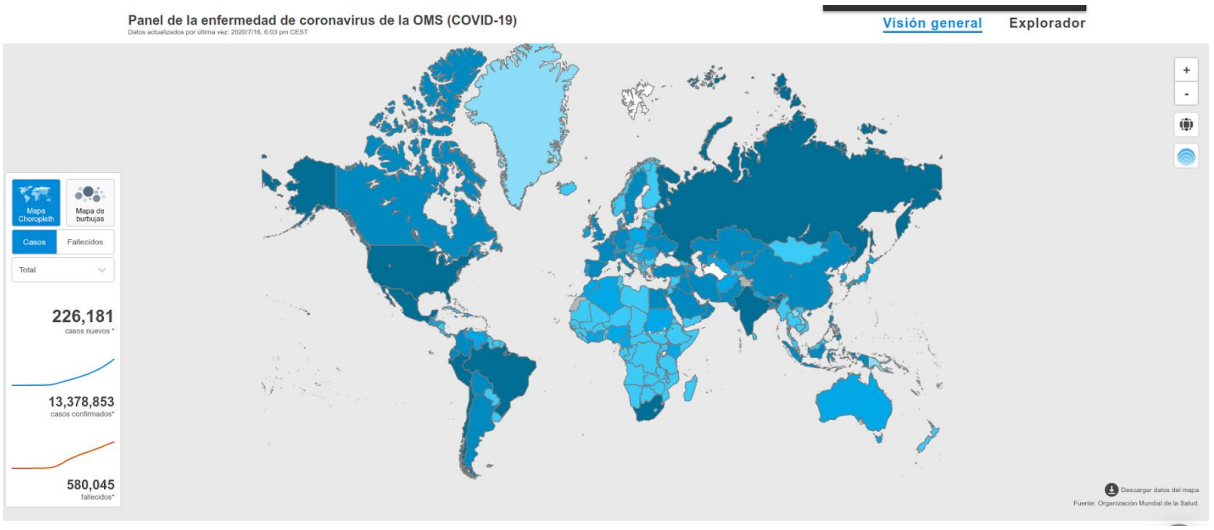
Pantalla Principal



Trabajos Relacionados

Clustering de contagiados por Covid-19 - Organización Mundial de la Salud

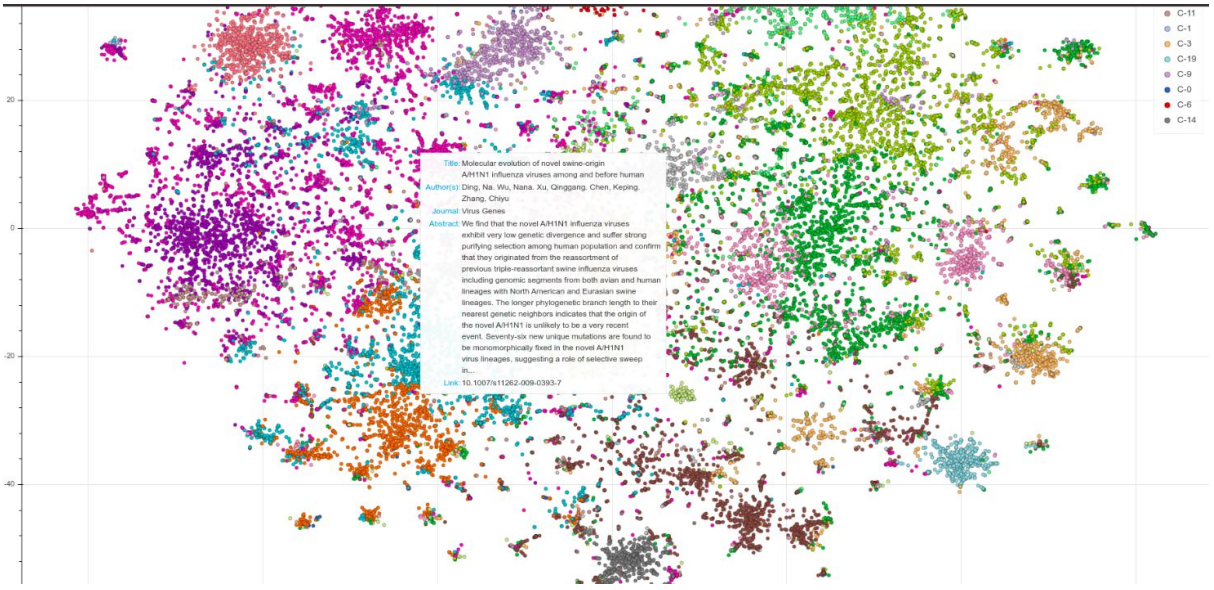
Muestra datos acerca de la enfermedad en todo el mundo, número de contagios, como casos confirmados, número de muertes alrededor del mundo, y casos y muertes por cada país. Asi como tambien la situacion en cada región del mundo



Situación Global

Covid-19 Literature Clustering

Agrupación de artículos de investigación similares y relacionadas al coronavirus, las publicaciones de temas muy similares comparten una etiqueta y se trazarán una cerca de la otra.



Herramientas utilizadas

GraphDB

GraphDB es una base de datos de gráficos semánticos que sirve a las organizaciones para almacenar, organizar y gestionar contenido en forma de datos inteligentes semánticamente enriquecidos.

Jena

Apache Jena es un framework Java que sirve para construir aplicaciones basadas en ontologías. Jena se desarrolló en HP Labs en el 2000, en 2009 HP cedió el proyecto a la fundación Apache que decidió adoptarlo en noviembre de 2010.

Java

Java es un lenguaje de programación y una plataforma informática que sirve para desarrollar aplicaciones y sitios web. Java es rápido, seguro y fiable para las personas que le den uso a este lenguaje de programación.

Protégé

Protégé es una herramienta que permite crear, editar ontologías y un sistema de adquisicion de conocimiento. Esta herramienta es de código abierto.

Python 3.7.32

Python es un lenguaje de programación que se basa en la legibilidad y claridad de su código, se utiliza mayormente para el desarrollo de aplicaciones y para el procesamiento de datos.

Jupyter Notebook

Jupyter Notebook es una herramienta que nos permite crear y compartir documentos con código en vivo, ecuaciones, visualizaciones y texto explicativo. es una referencia a los tres lenguajes de programación principales soportados por Jupyter, que son Julia, Python y R, y también un homenaje a los cuadernos de Galileo que registran el descubrimiento de los satélites de Júpiter.

Bokeh

Bokeh es una biblioteca de visualización interactiva de Python. EL objetivo de esta herramienta es proporcionar una visualización de datos de forma elegante y concisa compuesta de gráficos novedosos.

Consultas SPARQL

1) Cuales son todas las Publicaciones que hay?

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX dcterms: <http://purl.org/dc/terms/>
SELECT DISTINCT *
WHERE{
   ?Publicaciones rdf:type dcterms:BibliographicResource .
}
```

	Publicaciones
1	http://utpl.edu.ec/lod/publicationsCOVID/nryv9t4i/
2	http://utpl.edu.ec/lod/publicationsCOVID/68eogdyc/
3	http://utpl.edu.ec/lod/publicationsCOVID/rj1uy1ju/
4	http://utpl.edu.ec/lod/publicationsCOVID/w3fsxg90/
5	http://utpl.edu.ec/lod/publicationsCOVID/uy9l9aj9/
6	http://utpl.edu.ec/lod/publicationsCOVID/64y43o0y/
7	http://utpl.edu.ec/lod/publicationsCOVID/wlhzsyp9/
8	http://utpl.edu.ec/lod/publicationsCOVID/a1ha0hx4/
9	http://utpl.edu.ec/lod/publicationsCOVID/t32w8qfb/
10	http://utpl.edu.ec/lod/publicationsCOVID/4j3dwgy3/
11	http://utpl.edu.ec/lod/publicationsCOVID/rly2ydn5/
12	http://utpl.edu.ec/lod/publicationsCOVID/jtsxpjhc/
13	http://utpl.edu.ec/lod/publicationsCOVID/dyuw4kbm/
14	http://utpl.edu.ec/lod/publicationsCOVID/11hi1jel/
15	http://utpl.edu.ec/lod/publicationsCOVID/d9ytgvs2/
16	http://utpl.edu.ec/lod/publicationsCOVID/rkf971xn/
17	http://utpl.edu.ec/lod/publicationsCOVID/lj582swv/

2) Cuál es el título y autores de la publicación?

```
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX data: <http://utpl.edu.ec/lod/publicationsCOVID#>
SELECT ?TituloPublicacion ?NombreAutor
WHERE{
        ?Publicaciones rdf:type dcterms:BibliographicResource .
        ?Publicaciones dcterms:title ?TituloPublicacion .
        ?Publicaciones data:authors ?NombreAutor .
}
```

	TituloPublicacion	NombreAutor
1	"Integrated Scheduling of Information Services and Logistics Flows in the Omnichannel System"	"Ivanov Dmitry; Sokolov Boris"
2	"Low rate of bacterial co-infection in patients with COVID-19"	"Adler Hugh; Ball Robert; Fisher Michael; Mortimer Kalani; Vardhan Madhur S"
3	"Effect of Heat inactivation on Real-Time Reverse Transcription PCR of the SARS-COV-2 Detection"	"liu y.; cao z.; chen m.; zhong Y.; luo y.; shi g.; Xiang H.; Luo J.; Zhou H."
4	"Risk Factors Associated with Clinical Outcomes in 323 COVID-19 Patients in Wuhan, China"	"Hu Ling; Chen Shaoqiu; Fu Yuanyuan; Gao Zitong; Long Hui; Ren Hong;wei; Zuo Yi; Li Huan; Wang Jie; Xv Qing;bang; Yu Wen;xiong; Liu Jia; Shao Chen; Hao Jun;jie; Wang Chuan;zhen; Ma Yao; Wang Zhanwei; Yanagihara Richard; Wang Jian;ming; Deng Youping"
5	"Development of diabetes in Chinese with the metabolic syndrome: a 6-year prospective study."	"Cheung Bernard M Y; Wat Nelson M S; Man Yu Bun; Tam Sidney; Thomas G Neil; Leung Gabriel M; Cheng Chun Ho; Woo Jean; Janus Edward D; Lau Chu Pak; Lam Tai Hing; Lam Karen S L"
6	"Identification of neutralizing human monoclonal antibodies from Italian Covid-19 convalescent patients"	"Andreano Emanuele; Nicastri Emanuele; Paciello Ida; Pileri Piero; Manganaro Noemi; Piccini Giulia; Manenti Alessandro; Pantano Elisa; Kabanova Anna; Troisi Marco; Vacca Fabiola; Cardamone Dario; De Santi Concetta; Agrati Chiara; Capobianchi Maria Rosaria; Castilletti Concetta; Emiliozzi Arianna; Fabbiani Massimiliano; Montagnani Francesca; Montomoli Emanuele; Sala Claudia; Ippolito Giuseppe; Rappuoli Rino"

3) Cuál es la organización en la que está publicado el artículo?

```
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
SELECT DISTINCT ?Publicaciones ?Organizacion ?NombreOrg
WHERE{
   ?Publicaciones rdf:type dcterms:BibliographicResource ;
   dcterms:publisher ?Organizacion .
   ?Organizacion foaf:name ?NombreOrg .
}
```

	Publicaciones	Organizacion	NombreOrg
1	http://utpl.edu.ec/lod/publicationsCOVID/nryv9t4i/	http://utpl.edu.ec/lod/publicationsCOVID/nryv9t4i/organization/PMC	"PMC"
2	http://utpl.edu.ec/lod/publicationsCOVID/68eogdyc/	http://utpl.edu.ec/lod/publicationsCOVID/68eogdyc/organization/PN	"PMC"
3	http://utpl.edu.ec/lod/publicationsCOVID/rj1uy1ju/	http://utpl.edu.ec/lod/publicationsCOVID/rj1uy1ju/organization/MedF	"MedRxiv"
4	http://utpl.edu.ec/lod/publicationsCOVID/w3fsxg90/	http://utpl.edu.ec/lod/publicationsCOVID/w3fsxg90/organization/Me	"MedRxiv"
5	http://utpl.edu.ec/lod/publicationsCOVID/uy9l9aj9/	http://utpl.edu.ec/lod/publicationsCOVID/uy9l9aj9/organization/Med	"Medline"
6	http://utpl.edu.ec/lod/publicationsCOVID/64y43o0y/	http://utpl.edu.ec/lod/publicationsCOVID/64y43o0y/organization/Bi	"BioRxiv"
7	http://utpl.edu.ec/lod/publicationsCOVID/wlhzsyp9/	http://utpl.edu.ec/lod/publicationsCOVID/wlhzsyp9/organization/PM	"PMC"

4) **En qué Revista se publicaron y cual es el título de la Publicación?**

```
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
SELECT ?Revista ?Titulo
WHERE{
        ?Publicaciones rdf:type dcterms:BibliographicResource .
        ?Publicaciones dcterms:title ?Titulo .
        ?Publicaciones dbo:AcademicJournal ?journal .
        ?journal foaf:name ?Revista .
} ORDER BY(?Titulo)
```

	Revista	Titulo
1	"RoFo : Fortschritte auf dem Gebiete der Rontgenstrahlen und der Nuklearmedizin"	""
2	"Acta medica Okayama"	""Hook and roll technique"" using an articulating hook cautery to provide a critical view during single-in cision laparoscopic cholecystectomy.
3	"Current radiology reports"	""Imaging Evaluation of Collaterals in the Brain: Physiology and Clinical Translation""
4	"Sales Excellence"	""Immer bereit sein, die Extrameile zu gehen""
5	"The New Zealand medical journal"	""Just say no"'--reducing the use of antibiotics for colds, bronchitis and sinusitis.
6	""	""No test is better than a bad test"": Impact of diagnostic uncertainty in mass testing on the spread of Covid-19
7	"return"	""Panic first!""
8	"Cardiovascular revascularization medicine : including molecular interventions"	""Subintimal external crush"" technique for a ""balloon uncrossable"" chronic total occlusion.
9	"InFo Neurologie"	""Videosprechstunden sind nicht f�r alle die L�sung""
10	"&#x000e4"	""Viele Dinge bleiben zu bedenken""
11	""	""Virus hunting""using radial distance weighted discrimination

5) Nombre de Revista con artículos que contengan en su Título la palabra "pandemic"

```
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
SELECT ?Publicaciones ?Titulo ?Revista ?NombreRevista
WHERE{
        ?Publicaciones rdf:type dcterms:BibliographicResource .
        ?Publicaciones dcterms:title ?Titulo .
        ?Publicaciones dbo:AcademicJournal ?Revista .
        FILTER regex(?Titulo, "pandemic", "i") .
        OPTIONAL{
                ?Revista foaf:name ?NombreRevista .
         }
} ORDER BY DESC(?NombreRevista)
```

	Publicaciones	Titulo	Revista	NombreRevista
1	http://utpl.edu.ec/lod/publicationsCOVID/0yqyclxk/	"Epitope-based chimeric peptide vaccine design against S, M and E proteins of SARS-CoV-2 etiologic agent of global pandemic COVID-19: an in silico approach"	http://utpl.edu.ec/lod/publicationsCOVID/0yqyclxk/	"bioRxiv"
2	http://utpl.edu.ec/lod/publicationsCOVID/3b5jk6o3	"Emerging phylogenetic structure of the SARS-CoV-2 pandemic"	http://utpl.edu.ec/lod/publicationsCOVID/3b5jk6o3	"bioRxiv"
3	http://utpl.edu.ec/lod/publicationsCOVID/431ksdno	"Coronavirus, as a source of pandemic pathogens"	http://utpl.edu.ec/lod/publicationsCOVID/431ksdno	"bioRxiv"
4	http://utpl.edu.ec/lod/publicationsCOVID/62xsu7qa	"Epigenetic regulator miRNA pattern differences among SARS-CoV, SARS-CoV-2 and SARS-CoV-2 world-wide isolates delineated the mystery behind the epic pathogenicity and distinct clinical characteristics of pandemic COVID-19"	http://utpl.edu.ec/lod/publicationsCOVID/62xsu7qa	"bioRxiv"
5	http://utpl.edu.ec/lod/publicationsCOVID/691ybkan	"Coronavirus activates a stem cell-mediated defense mechanism that reactivates dormant tuberculosis: implications in COVID-19 pandemic"	http://utpl.edu.ec/lod/publicationsCOVID/691ybkan	"bioRxiv"

6) Cual es el DOI y el título de las publicaciones?

```
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX fabio: <http://purl.org/spar/fabio/>
SELECT DISTINCT ?Publicaciones ?Titulo_Publicacion ?DOI
WHERE{
   ?Publicaciones rdf:type dcterms:BibliographicResource .
   ?Publicaciones dcterms:title ?Titulo_Publicacion ;
   fabio:doi ?DOI .
}
```

	Publicaciones	Titulo_Publicacion	DOI
1	http://utpl.edu.ec/lod/publicationsCOVID/nryv9t4i/	"Integrated Scheduling of Information Services and Logistics Flows in the Omnichannel System"	"10.1007/978-3-030-43177-8_7"
2	http://utpl.edu.ec/lod/publicationsCOVID/68eogdyc/	"Low rate of bacterial co-infection in patients with COVID-19"	"10.1016/s2666-5247(20)30036-7"
3	http://utpl.edu.ec/lod/publicationsCOVID/rj1uy1ju/	"Effect of Heat inactivation on Real-Time Reverse Transcription PCR of the SARS-COV-2 Detection"	"10.1101/2020.05.19.20101469"
4	http://utpl.edu.ec/lod/publicationsCOVID/w3fsxg90/	"Risk Factors Associated with Clinical Outcomes in 323 COVID-19 Pa tients in Wuhan, China"	"10.1101/2020.03.25.20037721"
5	http://utpl.edu.ec/lod/publicationsCOVID/uy9l9aj9/	"Development of diabetes in Chinese with the metabolic syndrome: a 6-year prospective study."	""
6	http://utpl.edu.ec/lod/publicationsCOVID/64y43o0y/	"Identification of neutralizing human monoclonal antibodies from It alian Covid-19 convalescent patients"	"10.1101/2020.05.05.078154"
7	http://utpl.edu.ec/lod/publicationsCOVID/wlhzsyp9/	"Breekt nood de wet?"	"10.1007/s12459-020-0309-0"

7) Título de Publicación, Nombre de la Revista y el Tipo de Licencia cc-by

```
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX dbo: <http://dbpedia.org/ontology/>
SELECT DISTINCT ?Publicaciones ?Titulo_Publicacion ?Revista ?NombreRevista ?Licencia
?Tipo_Licencia
WHERE{
   ?Publicaciones rdf:type dcterms:BibliographicResource .
   ?Publicaciones dcterms:title ?Titulo_Publicacion ;
   dcterms:license ?Licencia .
   ?Licencia foaf:name ?Tipo_Licencia .
   ?Publicaciones dbo:AcademicJournal  ?Revista .
   ?Revista foaf:name ?NombreRevista .
   FILTER regex(?Tipo_Licencia, "cc-by", "i") .
} ORDER BY(?NombreRevista)
```

	Publicaciones	Titulo_Publicacion	Revista	NombreRevista	Licencia	Tipo_Licencia
1	http://utpl.edu.ec/lod/publication	"Assessment of spontaneous breathing during pressure controlled ventilation with superimposed spontaneous breathing using respiratory flow signal analysis"	http://utpl.edu.ec/lod/publications	"J Clin Monit Comput"	http://utpl.edu.ec/lod/publications/by	"cc-by"
2	http://utpl.edu.ec/lod/publication	"ACUTE HEPATITIS ASSOCIATED WITH MOUSE LEUKEMIA : V. THE NEUROTROPIC PROPERTIES OF THE CAUSAL VIRUS"	http://utpl.edu.ec/lod/publications	"J Exp Med"	http://utpl.edu.ec/lod/publications/by-nc-sa	"cc-by-nc-sa"
3	http://utpl.edu.ec/lod/publication	"Letter from the (un)seen virus: (post)humanist perspective in corona times"	http://utpl.edu.ec/lod/publications	"Soc Anthropol"	http://utpl.edu.ec/lod/publications/by-nc-nd	"cc-by-nc-nd"
4	http://utpl.edu.ec/lod/publication	"Authors reply"	http://utpl.edu.ec/lod/publications	"Lung India"	http://utpl.edu.ec/lod/publications/by	"cc-by"
5	http://utpl.edu.ec/lod/publication	"SURE EU Capacity for Stabilising Employment and Incomes in the Pandemic"	http://utpl.edu.ec/lod/publications	"Inter Econ"	http://utpl.edu.ec/lod/publications/by	"cc-by"
6	http://utpl.edu.ec/lod/publication	"Instructions for Authors"	http://utpl.edu.ec/lod/publications	"Chin Med J (Engl)"	http://utpl.edu.ec/lod/publications/by-nc-sa	"cc-by-nc-sa"

8) Muestra Publicaciones, DOI ,Titulo de Publicación , Nombre de Autor Nombre de Revista Nombre Organizacion, Tipo de Licencia, URL

```
PREFIX dcterms: <http://purl.org/dc/terms/>
```

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX fabio: <http://purl.org/spar/fabio/>
SELECT DISTINCT ?Publicaciones ?DOI ?Titulo_Publicacion ?Autor ?NombreAutor ?Revista
?NombreRevista ?Organizacion ?NombreOrganizacion ?Licencia ?Tipo_Licencia ?URL
WHERE{
   ?Publicaciones rdf:type dcterms:BibliographicResource .
   ?Publicaciones dcterms:title ?Titulo_Publicacion ;
   dcterms:license ?Licencia .
   ?Publicaciones dcterms:creator ?Autor .
   ?Autor foaf:name ?NombreAutor .
   ?Publicaciones dcterms:publisher ?Organizacion .
   ?Organizacion foaf:name ?NombreOrganizacion .
   ?Publicaciones fabio:doi ?DOI .
   ?Publicaciones fabio:hasURL ?URL .
   ?Licencia foaf:name ?Tipo_Licencia .
   ?Publicaciones dbo:AcademicJournal  ?Revista .
         OPTIONAL {
          ?Revista foaf:name ?NombreRevista
         }
} ORDER BY(?Titulo_Publicacion)
```

	Publicaciones	DOI	Titulo_Publicacion	Autores	Nombre_Revista	Nombre_Organizacion	Tipo_Licencia	URL	Fecha_Publicacion
1	http://utpl.edu.ec/lod/	"10.1055/a-1149-3625"	""	"Antoch Gerald; Urbach Horst; Mentzel Hans;Joachim; Reimer Peter; Weber Werner; Wujciak Detlef"	"RoFo : Fortschritte auf dem Gebiete der Rontgenstrahlen und der Nuklearmedizin"	"Medline"	"unk"	"https://doi.org/10.1055/a-1149-3625"	"1/4/2020"
2	http://utpl.edu.ec/lod/	""	""Hook and roll technique"" using an articulating hook cautery to provide a critical view during single-incision laparoscopic cholecystectomy.	"Idani Hitoshi; Nakano Kanyu; Asami Shinya; Kubota Tetsushi; Komoto Satoshi; Kurose Yohei; Kubo Shinichiro; Nojima Hiroki; Hioki Katsuyoshi; Kin Hitoshi; Takakura Norihisa"	"Acta medica Okayama"	"Medline"	"unk"	"https://www.ncbi.nlm.nih.gov/pubmed/23970325/"	"2013"
3	http://utpl.edu.ec/lod/	""	""Imaging Evaluation of Collaterals in the Brain: Physiology and Clinical Translation""	"Sheth Sunil A; Liebeskind David S"	"Current radiology reports"	"Medline"	"unk"	"https://www.ncbi.nlm.nih.gov/pubmed/25478305/"	"2014"
4	http://utpl.edu.ec/lod/	"10.1007/s35141-020-0329-3"	""Immer bereit sein, die Extrameile zu gehen""	"Bjeli? Dr. Aleksandar"	"Sales Excellence"	"PMC"	"no-cc"	"https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7282888/"	"10/6/2020"
5	http://utpl.edu.ec/lod/	""	""Just say no""--reducing the use of antibiotics for colds, bronchitis and sinusitis.	"Thomas M G; Arroll B"	"The New Zealand medical journal"	"Medline"	"unk"	"https://www.ncbi.nlm.nih.gov/pubmed/10935570/"	"2000"
6	http://utpl.edu.ec/lod/	"10.1101/2020.04.16.20067884"	""No test is better than a bad test"": Impact of diagnostic uncertainty in mass testing on the spread of Covid	"Gray Nicholas; Calleja Dominic; Wimbush Alex; Miralles;Dolz Enrique; Gray Ander; De-Angelis Marco; Pers	""	"MedRxiv"	"medrxiv"	"http://medrxiv.org/cgi/content/short/2020.04.16.20067884v1?rss=1"	"22/4/2020"

Implementación de Aplicación

Para crear la aplicación se ha optado por la herramienta jupyter notebook, que nos permite crear cuadernos de lenguaje python por bloques. Los bloques más importantes del código se listan a continuación.

- **Carga de datos**

  Para la carga de datos utilizamos la biblioteca SPARQLWrapper de python que nos permite conectarnos a un endpoint remoto y ejecutar consultas saparql

In [3]:
```
import pandas as pd
from SPARQLWrapper import SPARQLWrapper, JSON
#NLP
from spacy.lang.en.stop_words import STOP_WORDS
# import en_core_sci_lg  # model downloaded in previous step
```

In [6]:
```
sparql = SPARQLWrapper("http://localhost:7200/repositories/publications")
sparql.setQuery("""
    PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
    PREFIX foaf: <http://xmlns.com/foaf/0.1/>
    PREFIX dcterms: <http://purl.org/dc/terms/>
    PREFIX dc: <http://purl.org/dc/elements/1.1/>
    PREFIX fabio: <http://purl.org/spar/fabio/>
    PREFIX data: <http://utpl.edu.ec/lod/publicationsCOVID#>
    PREFIX dbo: <http://dbpedia.org/ontology/>
    SELECT ?Publicaciones ?DOI ?Titulo_Publicacion ?Autores ?Nombre_Revista ?Nombre_Organizacion
    WHERE{
        ?Publicaciones rdf:type dcterms:BibliographicResource .
        ?Publicaciones dcterms:title ?Titulo_Publicacion ;
        dcterms:license ?Licencia .
        ?Licencia foaf:name ?Tipo_Licencia .
        ?Publicaciones fabio:doi ?DOI .
        ?Publicaciones data:authors ?Autores .
        ?Publicaciones dcterms:publisher ?Organizacion .
        ?Organizacion foaf:name ?Nombre_Organizacion .
        ?Publicaciones fabio:hasURL ?URL .
        ?Publicaciones dbo:AcademicJournal  ?Revista .
        ?Revista foaf:name ?Nombre_Revista .
        ?Publicaciones dcterms:date ?Fecha_Publicacion .
    } ORDER BY(?Titulo_Publicacion)
""")
sparql.setReturnFormat(JSON)
results = sparql.query().convert()
```

La consulta insertada nos devolverá todas las publicaciones alojadas en el repositorio de datos rdf GraphDB, con sus respectivos atributos.

- **Cargar los datos en un dataframe**

  Se cargan los datos extraídos de la consulta sparql en un dataframe para que posteriormente sean tratados.

In [8]:
```
for idx, res in enumerate(results["results"]["bindings"]):
    #print(f'Processing index: {idx} of {len(results["results"]["bindings"])}')

    dict_['Uri_Pb'].append(res['Publicaciones']['value'])
    dict_['Doi'].append(res["DOI"]["value"])
    dict_['Titulo'].append(res["Titulo_Publicacion"]["value"])
    try:
        # if more than one author
        autores = res["Autores"]["value"]
        authors = autores.split(';')
        if len(authors) > 2:
            # if more than 2 authors, take them all with html tag breaks in between
            dict_['Autores'].append(get_breaks('. '.join(authors), 40))
        else:
            # authors will fit in plot
            dict_['Autores'].append(". ".join(authors))
    except Exception as e:
        # if only one author - or Null valie
        dict_['Autores'].append(res["Autores"]["value"])

    dict_['Fecha'].append(res["Fecha_Publicacion"]["value"])
    dict_['Revista'].append(res["Nombre_Revista"]["value"])
    dict_['Organizacion'].append(res["Nombre_Organizacion"]["value"])
    dict_['Licencia'].append(res["Tipo_Licencia"]["value"])
    dict_['URL'].append(res["URL"]["value"])

    df_covid = pd.DataFrame(dict_, columns=['Uri_Pb', 'Doi', 'Titulo','Autores', 'Fech
    df_covid.info()
    print('-------------')
    df_covid.head()
```

- **Preprocesamiento de los datos**

  Instalamos e importamos un modelo de spacy para poder trabajar con procesamiento de lenguaje natural

In [91]:
```
from IPython.utils import io
with io.capture_output() as captured:
    !pip install https://s3-us-west-2.amazonaws.com/ai2-s2-scispacy/releases/v0.2.4/en_core_sci_lg-0.2.4.tar.gz
```

In [10]:
```
import spacy
from spacy.lang.en.stop_words import STOP_WORDS
import en_core_sci_lg  # model downloaded in previous step
```

Eliminamos palabras vacías (Stop words) que actúan como ruido e interfieren en la clusterización de los datos.

In [11]:
```
import string

punctuations = string.punctuation
stopwords = list(STOP_WORDS)
stopwords[:10]
```

Out[11]:
```
['should',
 'everywhere',
 'do',
 'no',
 'than',
 'under',
 'have',
 'move',
 'such',
 'it']
```

In [14]:
```
custom_stop_words = [
    'doi', 'preprint', 'copyright', 'peer', 'reviewed', 'org', 'https', 'et', 'al', 'author', 'figure',
    'rights', 'reserved', 'permission', 'used', 'using', 'biorxiv', 'medrxiv', 'license', 'fig', 'fig.',
    'al.', 'Elsevier', 'PMC', 'CZI', 'www'
]

for w in custom_stop_words:
    if w not in stopwords:
        stopwords.append(w)
```

In [15]:
```
parser = en_core_sci_lg.load(disable=["tagger", "ner"])
parser.max_length = 7000000

def spacy_tokenizer(sentence):
    mytokens = parser(sentence)
    mytokens = [ word.lemma_.lower().strip() if word.lemma_ != "-PRON-" else word.lower_ for word in mytokens ]
    mytokens = [ word for word in mytokens if word not in stopwords and word not in punctuations ]
    mytokens = " ".join([i for i in mytokens])
    return mytokens
```

Ejecutamos el modelo y procesamos los datos de los tópicos, para hallar similitudes entre los tópicos

In [15]:
```
parser = en_core_sci_lg.load(disable=["tagger", "ner"])
parser.max_length = 7000000

def spacy_tokenizer(sentence):
    mytokens = parser(sentence)
    mytokens = [ word.lemma_.lower().strip() if word.lemma_ != "-PRON-" else word.lower_ for word in mytokens
    mytokens = [ word for word in mytokens if word not in stopwords and word not in punctuations ]
    mytokens = " ".join([i for i in mytokens])
    return mytokens
```

In [16]:
```
from tqdm import tqdm
tqdm.pandas()
df_covid["processed_text"] = df_covid["Titulo"].progress_apply(spacy_tokenizer)
```

```
100%|██████████| 10798/10798 [00:30<00:00, 356.97it/s]
```

- **Vectorización**

Luego de preprocesar los datos, se tienen que convertirlos a un formato que nuestros algoritmos puedan manejar. Para este propósito usaremos tf-idf. Este convierte nuestros datos con formato de cadena en una medida de la importancia de cada palabra para la instancia fuera de la literatura en su conjunto.

In [17]:
```
from sklearn.feature_extraction.text import TfidfVectorizer
def vectorize(text, maxx_features):

    vectorizer = TfidfVectorizer(max_features=maxx_features)
    X = vectorizer.fit_transform(text)
    return X
```

In [18]:
```
text = df_covid["processed_text"].values
X = vectorize(text, 2 ** 12)
X.shape
```

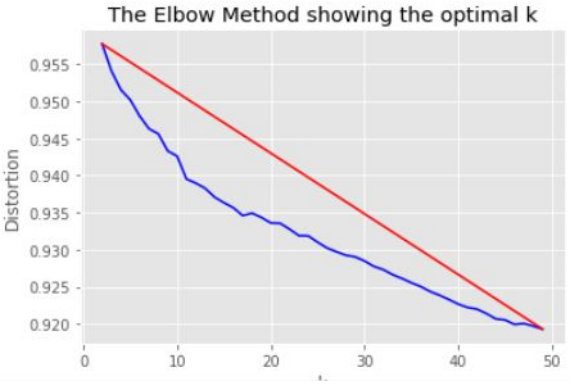
- **Aplicación de Clustering**

  Procedemos a realizar el cálculo de las distancias entre los nodos para determinar los centroides de los grupos, para esto usamos la fórmula euclidean.

In [20]:
```
import numpy as np
from sklearn.cluster import KMeans
from sklearn import metrics
from scipy.spatial.distance import cdist

# run kmeans with many different k
distortions = []
K = range(2, 50)
for k in K:
    k_means = KMeans(n_clusters=k, random_state=42).fit(X_reduced)
    k_means.fit(X_reduced)
    distortions.append(sum(np.min(cdist(X_reduced, k_means.cluster_centers_, 'euclidean'),
    #print('Found distortion for {} clusters'.format(k))
```

Para encontrar el mejor valor de k, es decir el número de clusters que existirán en los cuales se tienen que integrar cada publicación utilizamos el método de codo, para generar una gráfica que nos indique el número óptimo de clusters, en este caso el número óptimo está entre 15 y 25, por lo que tomamos el valor intermedio; K=20.

In [21]:
```
import matplotlib.pyplot as plt
plt.style.use('ggplot')

X_line = [K[0], K[-1]]
Y_line = [distortions[0], distortions[-1]]

# Plot the elbow
plt.plot(K, distortions, 'b-')
plt.plot(X_line, Y_line, 'r')
plt.xlabel('k')
plt.ylabel('Distortion')
plt.title('The Elbow Method showing the optimal k')
plt.show()
```



Aplicamos el clustering de los datos con la función predict y le usamos el número óptimo de k.

```
In [22]: k = 20
         kmeans = KMeans(n_clusters=k, random_state=42)
         y_pred = kmeans.fit_predict(X_reduced)
         df_covid['y'] = y_pred
```

Generamos el plot (gráfico) de los datos agrupados con etiquetas de colores.

```
In [25]: %matplotlib inline
         from matplotlib import pyplot as plt
         import seaborn as sns

         # sns settings
         sns.set(rc={'figure.figsize':(15,15)})

         # colors
         palette = sns.hls_palette(20, l=.4, s=.9)

         # plot
         sns.scatterplot(X_embedded[:,0], X_embedded[:,1], hue=y_pred, legend='ful
         plt.title('t-SNE with Kmeans Labels')
         plt.savefig("improved_cluster_tsne.png")
         plt.show()
```

- **Aplicación de LDA para el modelado de tópicos**

  A través del modelado de temas descubriremos cuáles son los términos más importantes para cada grupo. Esto agregará más significado al clúster al proporcionar palabras clave para identificar rápidamente los temas del clúster. Para el modelado de temas, usamos LDA (Asignación de Dirichlet Latente). En LDA, cada documento puede describirse mediante una distribución de temas y cada tema puede describirse mediante una distribución de palabras.

In [32]:
```
clusters_lda_data = []

for current_cluster, lda in enumerate(lda_models):
    # print("Current Cluster: " + str(current_cluster))

    if vectorized_data[current_cluster] != None:
        clusters_lda_data.append((lda.fit_transform(vectorized_data[current_cluster])))
```

In [33]:
```
# Functions for printing keywords for each topic
def selected_topics(model, vectorizer, top_n=3):
    current_words = []
    keywords = []

    for idx, topic in enumerate(model.components_):
        words = [(vectorizer.get_feature_names()[i], topic[i]) for i in topic.argsort()[:-
        for word in words:
            if word[0] not in current_words:
                keywords.append(word)
                current_words.append(word[0])

    keywords.sort(key = lambda x: x[1])
    keywords.reverse()
    return_values = []
    for ii in keywords:
        return_values.append(ii[0])
    return return_values
```

In [34]:
```
all_keywords = []
for current_vectorizer, lda in enumerate(lda_models):
    # print("Current Cluster: " + str(current_vectorizer))

    if vectorized_data[current_vectorizer] != None:
        all_keywords.append(selected_topics(lda, vectorizers[current_vectorizer]))
```

- **Generando plot con bokeh**

  Los pasos anteriores nos han dado un agrupamiento y un conjunto de datos de documentos reducidos a dos dimensiones. Al combinar esto con Bokeh, podemos crear una trama interactiva de la literatura. Esto debería organizar los documentos de manera que las publicaciones relacionadas estén muy cerca.

  Primero importamos todas las librerías.

In [405]:
```
# required libraries for plot

from call_backs import input_callback, selected_code  # file with customJS callbacks for bokeh
                                                      # github.com/MaksimEkin/COVID19-Literature-Clustering/bl
import bokeh
from bokeh.models import ColumnDataSource, HoverTool, LinearColorMapper, CustomJS, Slider, TapTool, TextInput
from bokeh.palettes import Category20
from bokeh.transform import linear_cmap, transform
from bokeh.io import output_file, show, output_notebook
from bokeh.plotting import figure
from bokeh.models import RadioButtonGroup, TextInput, Div, Paragraph
from bokeh.layouts import column, widgetbox, row, layout
from bokeh.layouts import column
```

  Para tratar de comprender cuáles pueden ser las similitudes, también hemos realizado modelos de temas en cada grupo de documentos para elegir los términos clave. Los cargamos a continuación

In [406]:
```
os.chdir(main_path)
```

In [407]:
```
import os

topic_path = 'topics.txt'
with open(topic_path) as f:
    topics = f.readlines()
```

Luego generamos el setup principal de bokeh.

In [408]:
```
# show on notebook
output_notebook()
# target labels
y_labels = y_pred

# data sources
source = ColumnDataSource(data=dict(
    x= X_embedded[:,0],
    y= X_embedded[:,1],
    x_backup = X_embedded[:,0],
    y_backup = X_embedded[:,1],
    desc= y_labels,
    uri= df_covid['Uri_Pb'],
    doi= df_covid['Doi'],
    titulo= df_covid['Titulo'],
    autores = df_covid['Autores'],
    fecha = df_covid['Fecha'],
    revista = df_covid['Revista'],
    organizacion = df_covid['Organizacion'],
    licencia = df_covid['Licencia'],
    url = df_covid['URL'],
    labels = ["C-" + str(x) for x in y_labels]
    ))

# hover over information
hover = HoverTool(tooltips=[
    ("uri", "@uri{safe}"),
    ("DOI", "@doi{safe}"),
    ("Titulo", "@titulo{safe}"),
    ("Autor(es)", "@autores{safe}"),
    ("Fecha", "@fecha{safe}"),
    ("Revista", "@revista{safe}"),
    ("Organizacion", "@organizacion{safe}"),
    ("Licencia", "@licencia{safe}"),
    ("URL", "@url{safe}"),
],
point_policy="follow_mouse")

# map colors
mapper = linear_cmap(field_name='desc',
                     palette=Category20[20],
                     low=min(y_labels) ,high=max(y_labels))

# prepare the figure
plot = figure(plot_width=1200, plot_height=850,
           tools=[hover, 'pan', 'wheel_zoom', 'box_zoom', 'reset', 'save', 'tap'],
           title="Clustering COVID-19 Publicaciones",
           toolbar_location="above")

# plot settings
plot.scatter('x', 'y', size=5,
          source=source,
```

Y finalmente nos generará un archivo html en el que podremos visulizar el clustering de datos, y todos sus nodos, para este caso hemos tomado una muestra de 10.000 datos.

Cabe recalcar que en este apartado se ha puesto los bloques de código más relevantes, para consultar todo el notebook puede dirigirse a https://github.com/ejayala2/Covid19_Clustering.

Conclusiones

- El algoritmo K Means es adecuado para realizar clustering por medio de técnicas de procesamiento de lenguaje natural (NLP) y vectorización de los datos.
- Las ontologías en la web semántica son adecuadas para dotar a los ordenadores de la capacidad de estructurar y manejar la información en base a una valoración semántica de sus contenidos, con la finalidad de que la inteligencia artificial pueda entender el significado del contenido.
- Los resultados obtenidos a través de la aplicación desarrollada han permitido generar un mayor grado de relación entre las ontologías y los métodos de agrupamiento como kmeans.
- Jena es una herramienta adecuada para la generación de datos RDF, pero cabe mencionar que existen herramientas que permiten una mayor flexibilidad y simplicidad al momento de generar este tipo de datos.
- GraphDB ha brindado resultados positivos al momento de trabajar con datos RDF, lo que ha permitido adentrarse al campo de las consultas SPARQL, generando un mayor grado de aprendizaje en este ámbito.