Herramientas para la investigación digital. Introducción práctica al CorText Manager. Parte II.

Los artículos científicos como fuentes de información.

VII Escuela Doctoral ESOCITE

UNAM - IPN
2021

¿**Artículos científicos** como **fuente** de **información**?

Los **artículos** ocupan un lugar importante en el **ciclo** de **producción** de **conocimiento** científico.

Un artículo científico se considera un **indicador importante** de la producción de investigación científica (pero no el único).

El "conocimiento certificado" es el que ha sido sometido a la **crítica colegiada** y ha **resistido** sus **objeciones** (Callon et al., 1995)

El ciclo de producción de los conocimientos certificados (Callon, Courtial y Penan 1995)

¿Para qué se usa?

Algunos ejes de trabajo frecuente

ELSEVIER

Research Policy 26 (1997) 1–18

research policy

Journal

What is research collaboration? [1]

Título del artículo

J. Sylvan Katz, Ben R. Martin *

Autores: colaboración científica

Science Policy and Research Evaluation Group, ESRC Centre for Science, Technology, Energy ad Environment Policy, Science Policy Research Unit, University of Sussex, Falmer, Brighton BN1 9RF, UK

Direcciones

Accepted 11 January 1995

Fecha de publicación

Abstract

Although there have been many previous studies of research collaboration, comparatively little attention has been given to the concept of 'collaboration' or to the adequacy of attempting to measure it through co-authorship. In this paper, we distinguish between collaboration at different levels and show that inter-institutional and international collaboration need not necessarily involve inter-individual collaboration. We also show that co-authorship is no more than a partial indicator of collaboration. Lastly, we argue for a more symmetrical approach in comparing the costs of collaboration with the undoubted benefits when considering policies towards research collaboration. © 1997 Elsevier Science B.V. All rights reserved.

Resumen de contenido (Abstract)

¿Para qué se usa?

Algunos ejes de trabajo frecuente

References

[1] C. Balog, 1979/80, Multiple authorship and author collaboration in agricultural research publications, Journal of Research Communication Studies 2, 159–169.

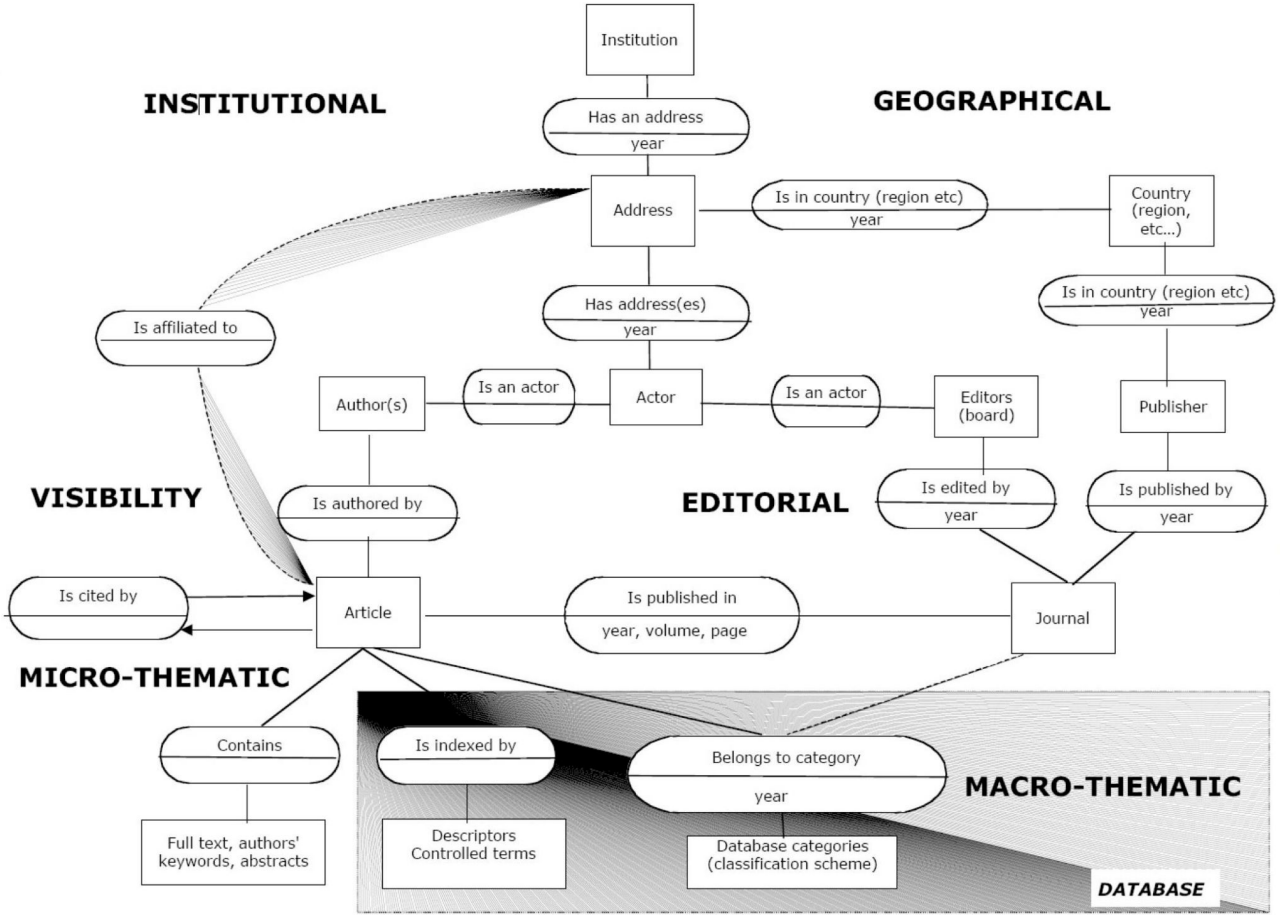
[2] D.de B. Beaver and R. Rosen, 1978, Studies in scientific collaboration: Part I—The professional origins of scientific co-authorship, Scientometrics 1, 65–84.

[3] D.de B. Beaver and R. Rosen, 1979, Studies in scientific collaboration: Part II—Scientific co-authorship, research productivity and visibility in the French scientific elite, 1799–1830, Scientometrics 1, 133–149.

[4] D.de B. Beaver and R. Rosen, 1979, Studies in scientific collaboration: Part III—Professionalization and the natural history of modern scientific co-authorship, Scientometrics 1, 231–245.

Citas y referencias del artículo: fuentes científicas del artículo

El **artículo científico** como fuente de información



Basado en
Zitt y Bassecoulard (2004)

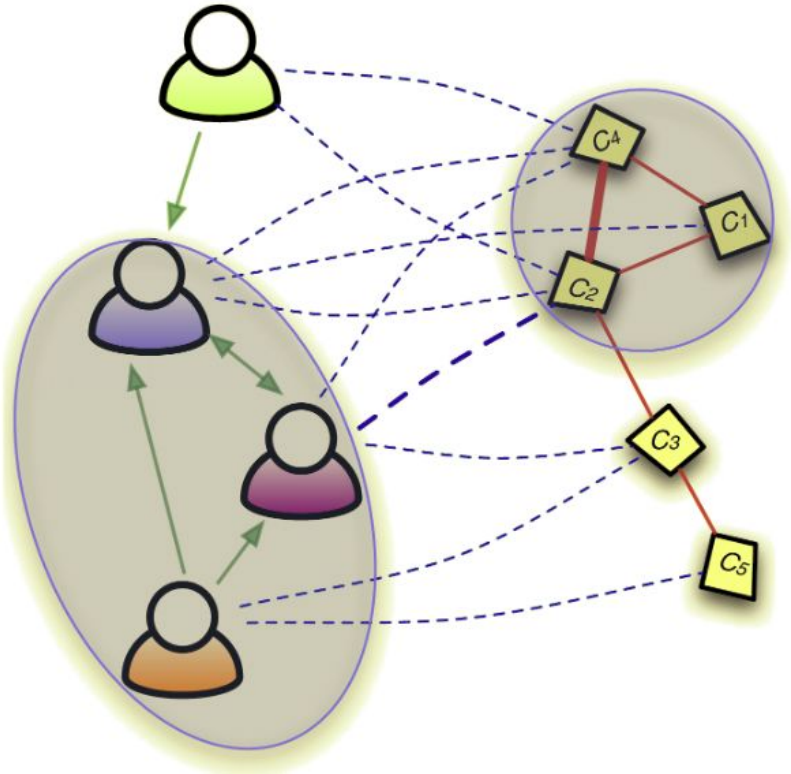
El **minado de información relacional** en publicaciones científicas. *Dimensiones y ejemplos*.

	Micro / Individual (1-100 registros)	Meso / Localizada (101-10,000 registros)	Macro / Global (10,000 < registros)
Análisis estadístico / Perfilado	*Persona individual y perfil experto*	Laboratorios, centros de I+D, universidades, dominios de investigación o regiones	Todo el trabajo de la NSF, NIH, de un país, etc.
Análisis temporal (Cuándo)	*Portafolio de financiamiento de un individuo*	Emergencia de tópicos en 20 años de los "Proceedings of the National Academy of Sciences" (PNAS)	113 años de investigación en Física
Análisis geoespacial (Dónde)	*Trayectoria y carrera de un investigador*	Mapeo del perfil intelectual de una región (provincia o Estado)	Orientación geográfica de los PNAS
Análisis semántico (Qué)	*Conocimiento básico desde el que se soporta un grant para fondos*	Flujos de conocimiento en la investigación química	Mapas temáticos de los fondos otorgados por el NIH
Análisis de Redes (Con quiénes)	Red de directores de proyectos de una sola persona	Redes de co-autorías	Competencias centrales de NIH

(Börner, 2010: 272)

Mapeando información

Métricas y técnicas de visualización

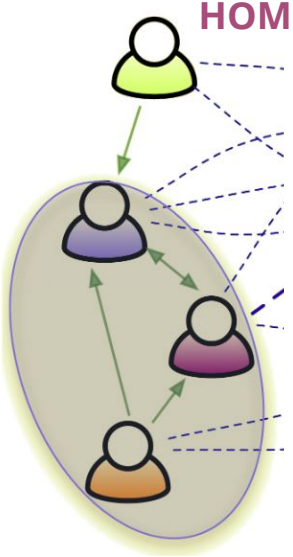


- Los sistemas reales contienen múltiples componentes que interactúan entre sí.
- En este caso, los corpus de documentos pueden contar o ser enriquecidos con múltiples tipos de información.
- La relación entre los documentos puede modelarse de dos formas:
  - Cómo redes homogéneas (sin distinguir tipos de objetos y enlaces en las redes).
  - Cómo redes heterogéneas (distinguiendo los tipos objetos y enlaces).
- El modelado de redes heterogéneas es una oportunidad analítica.

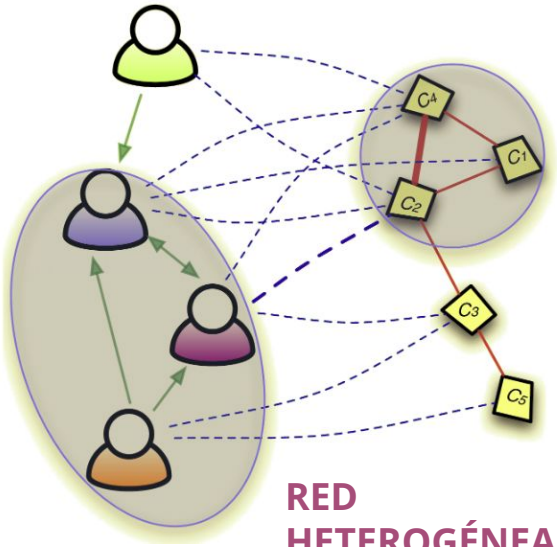
Mapeando información

Métricas y técnicas de visualización

RED HOMOGÉNEA



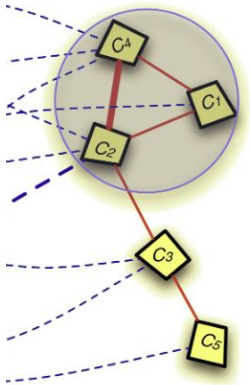
Las personas tiene relaciones: **redes sociológicas**



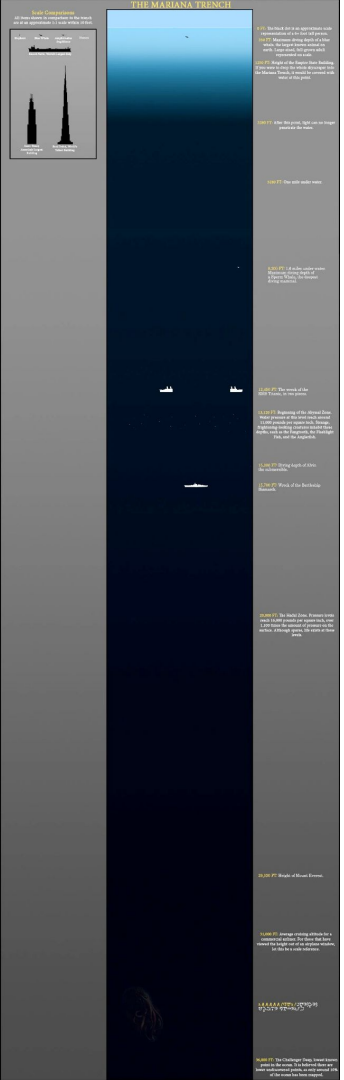
RED HETEROGÉNEA

Las personas producen textos: **redes heterogéneas socio-semánticas**

RED HOMOGÉNEA



Los textos tienen relaciones entre conceptos: **redes semánticas**



Detalle

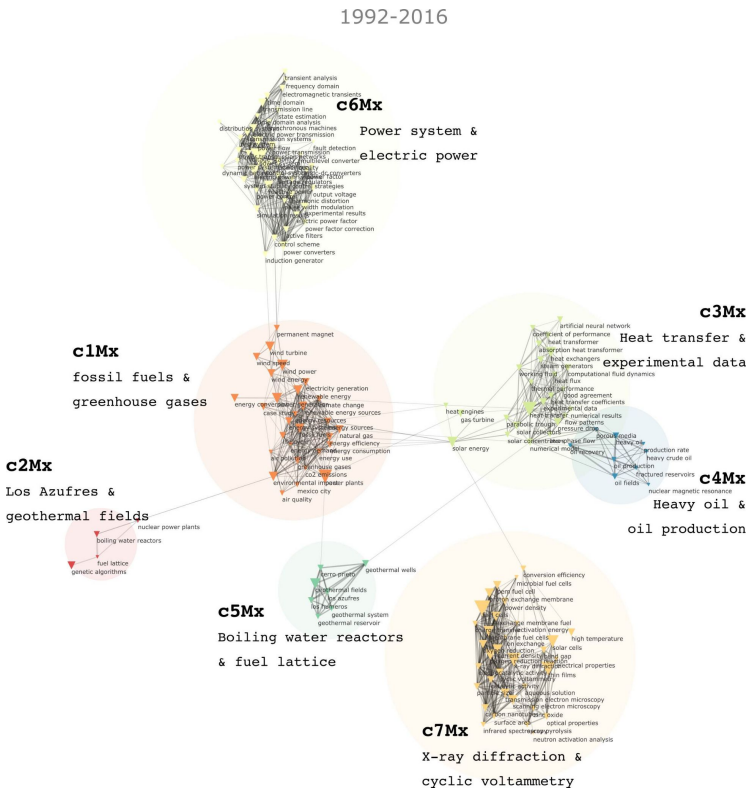
¿Qué tan profundo ir con nuestro análisis?

- Como cuando exploramos el mar, depende de qué querramos ver:
  - A distintas profundidades podemos ver criaturas y comportamientos diferentes.
  - Cuando más profundo vamos, más oscuro es el mar.
  - Dependiendo de la profundidad, ajustamos nuestros instrumentos.
- El nivel de detalle se puede ajustar en distintos momentos (recuperación, procesamiento y análisis-visualización)
- Podemos pensar tres niveles (análisis-visualización)
  - Macro: vista general, agrupamientos destacados, estructura general.
  - Meso: foco en un área de la red.
  - Micro: detalle sobre un tema analíticamente relevante.

Infografía de la Fosa de las Marianas, en el Océano Pacífico.

Macro

Mirada general del fenómeno estudiado

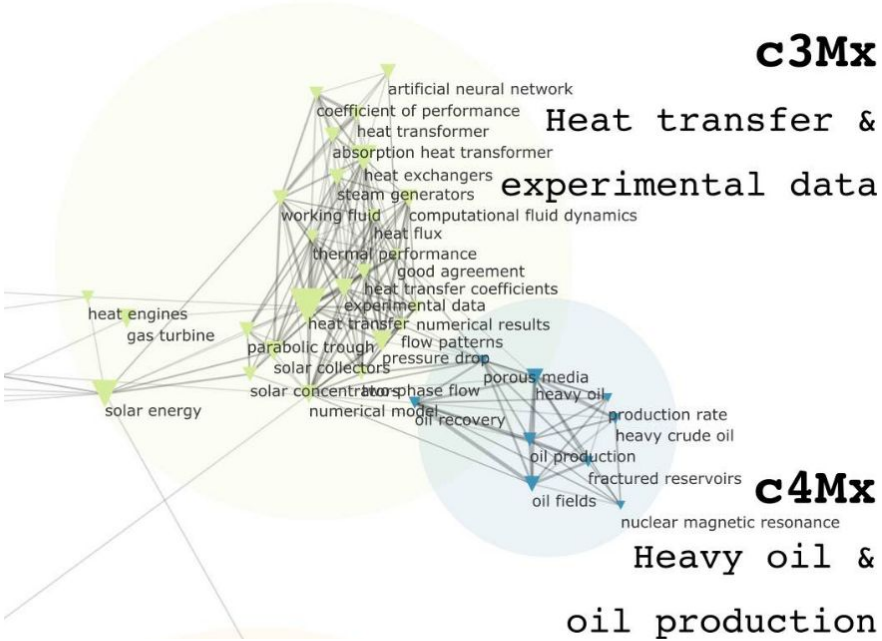


- Tomemos un ejemplo: la producción mexicana en energías renovables entre 1992 y 2016 en SCOPUS.
- Una red homogénea de expresiones semánticas.
- Métricas que nos interesan:
  - Números de clústeres
  - Densidad del gráfico y sus clústeres
  - Tamaño de los clústeres
  - Posiciones centrales o periféricas

Especificaciones: Análisis de co-ocurrencia de términos en documentos, umbral de 150 términos más destacados, algoritmo de clasificación Lovain (resolución 1)

Meso

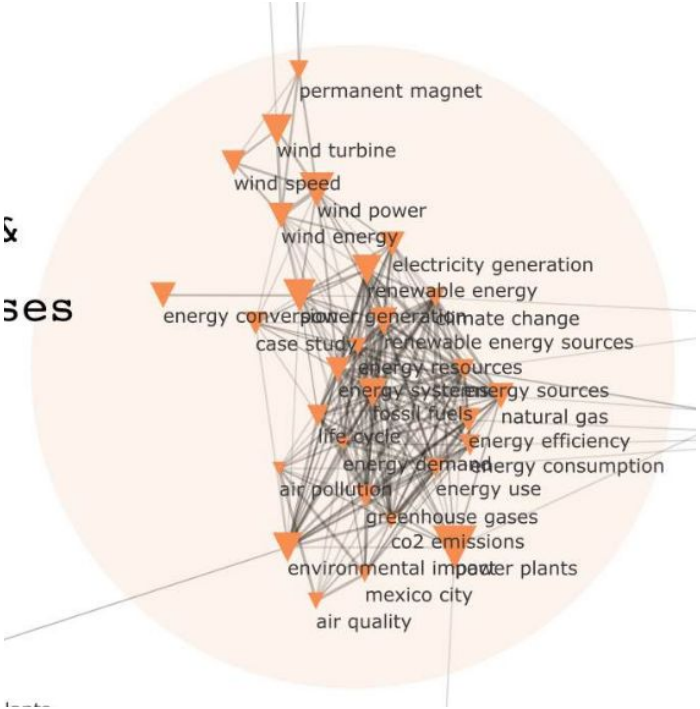
Mirada a una sección del mapa construído



- Posición relativa de un grupo de clústeres.
- Puntos de conexión con el resto del mapa.
- Espacios compartidos entre los grupos temáticos en los que nos interesamos.

Micro

Detalle sobre un grupo temático



- Interpretación endogámica de la composición de los clústeres.
- Atención a los tamaños y posiciones relativas de los nodos.
- Métricas que nos interesan:
  - Centralidad de los nodos
  - Peso de los nodos en el clúster
  - Composición de los clústeres (nodos, enlaces)

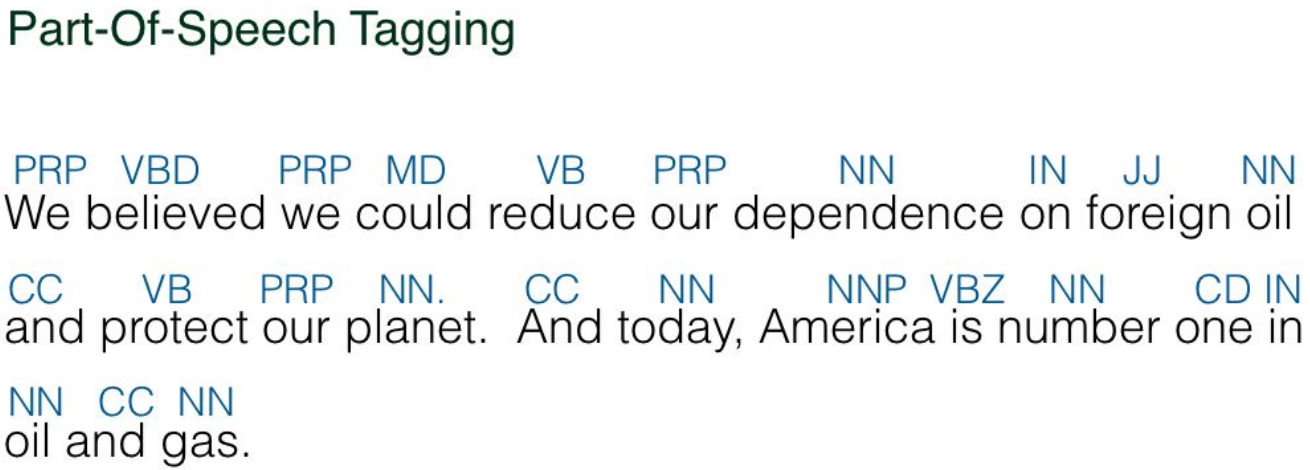
Extracción de información

Cómo procesamos el texto para generar información útil

- Podemos procesar la información textual para extraer información de acuerdo a dos lógicas:
  - Inductiva: queremos detectar regularidades y luego interpretarlas.
  - Deductiva: tenemos una hipótesis y queremos probarla.
- Podemos ocupar en nuestro flujo de trabajo distintos momentos en los que intercalemos abordajes desde ambas perspectivas.
- En el caso de un abordaje **inductivo** se vuelven importantes los algoritmos y parámetros para la detección de términos.
- En el caso de un abordaje **deductivo** se vuelven importantes los procesos de depuración, selección y validación de listas de palabras o *diccionarios*.

Cómo se procesa el texto

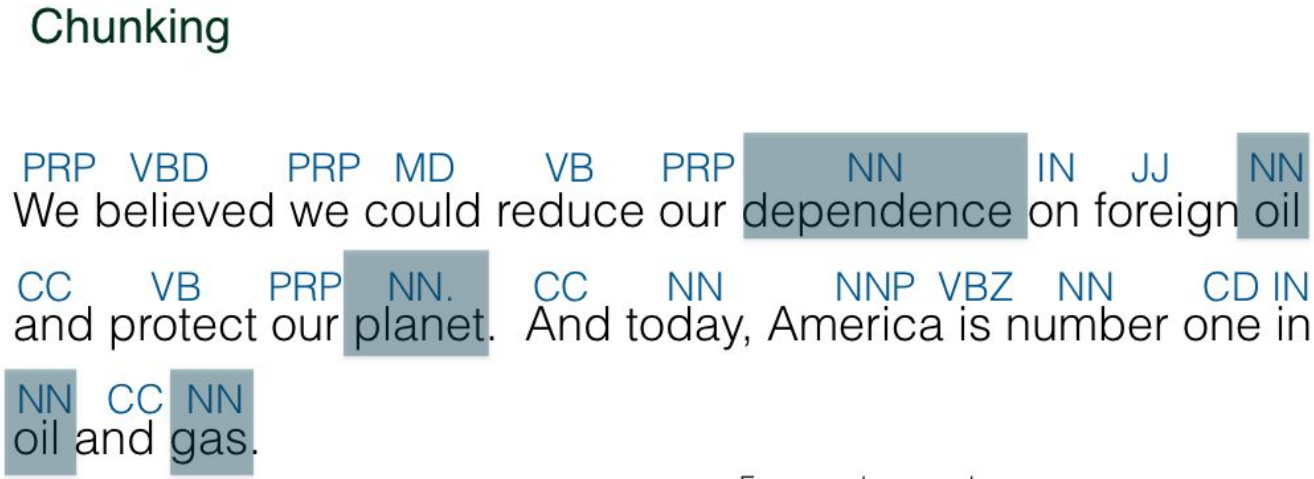
Qué leen las computadoras



- La computadora detecta qué tipo de entidades de lenguaje hay en el texto.
- Verbos, sustantivos, etc.

Cómo se procesa el texto

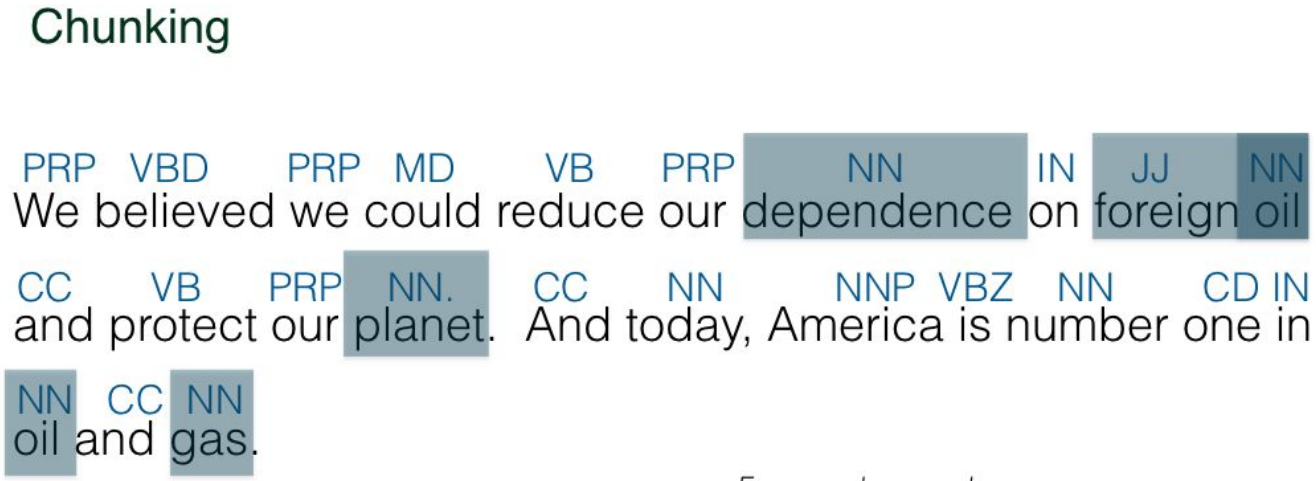
Qué leen las computadoras



- Luego de identificar los tipos de palabras, se enfoca el procesamiento en un tipo.
- En este caso, el foco está en identificar las frases nominales o *'noun phrases'*

Cómo se procesa el texto

Qué leen las computadoras



Extracted noun phrases:

- *dependence*
- *planet*
- *oil*
- *gas*
- *foreign oil*

- Luego busca qué otras entidades hay cercanas

Cómo se procesa el texto

Qué leen las computadoras

Chunking

PRP VBD PRP MD VB PRP NN IN JJ NN
We believed we could reduce our dependence on foreign oil

CC VB PRP NN. CC NN NNP VBZ NN CD IN
and protect our planet. And today, America is number one in

NN CC NN
oil and gas.

Extracted noun phrases:

- *dependence*
- *planet*
- *oil*
- *gas*
- *foreign oil*
- *dependence on foreign oil*
- *oil and gas*

- Las palabras se agrupan en trozos o ‘chunks’.
- En este caso, el foco es en las frases nominales o *‘noun phrases’*

Cómo se procesa el texto

Qué leen las computadoras

Stemming and Standardizing

PRP VBD PRP MD VB PRP NN IN JJ NN
We believed we could reduce our dependence on foreign oil

CC VB PRP NN. CC NN NNP VBZ NN CD IN
and protect our planet. And today, America is number one in

NN CC NN
oil and gas.

Extracted classes of Noun Phrases:

- *dependence on foreign oil: {dependence on foreign oil ; foreign oil dependence}*
- *oil and gas: {oil and gas; gas and oil}*
- *planet: {planet, planets}*
- *etc.*

- Podemos parametrizar para elegir qué tipo de procesamiento.
- En este caso, el foco es en las frases nominales o *'noun phrases'*

Cómo se procesa el texto

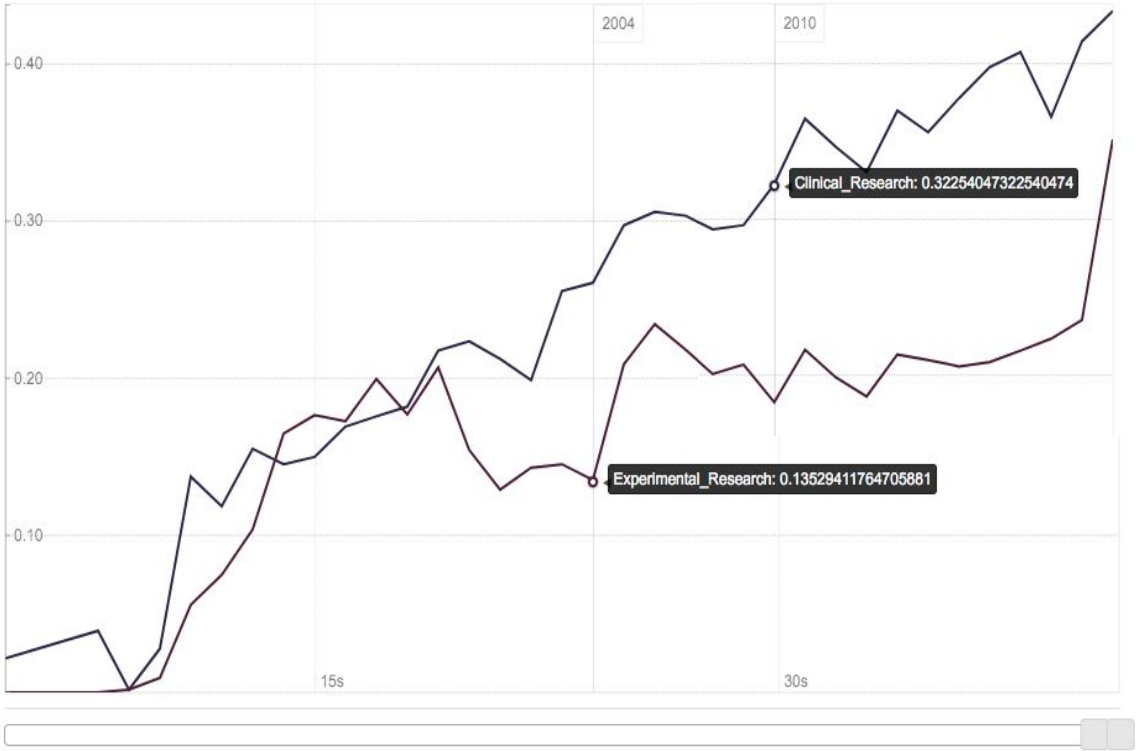
Un ejemplo de resultados

label	forms
abundance	abundance
acres	acres|&lacre
aliens	aliens|&lalien
armed forces	armed forces|&larmed force|&lforce of arms|&larms and force
business men	business men|&lbusiness man|&lmen of business|&lman of business|&lbusinessman|&lbusinessmen
children	children|&lchild
commerce and navigation	commerce and navigation|&lnavigation and commerce|&lnavigation or commerce
crews	crews|&lcrew
crime	crime|&lcrimes
crisis	crisis|&lcrises
crops	crops|&lcrop
cruisers	cruisers|&lcruiser
democracy	democracy|&ldemocracies
diplomatic relations	diplomatic relations
farm products	farm products|&lproducts of the farm|&lproducts of farm
great importance	great importance|&lgreater importance|&lgreatest importance
health care costs	**health care costs|&lcost of health care|&lhealth care cost**
income tax	income tax|&ltax the income|&ltax on the income
peace and freedom	peace and freedom|&lfreedom and peace|&lpeace with freedom
property rights	property rights|&lproperty right|&lrights of property|&lright of property|&lrights and property

- Estos términos detectados pueden ser depurados:
  - eliminando ruido
  - dándoles categorías manualmente
- Esta es buena materia prima para la construcción de diccionarios.

Cómo se procesa el texto

Un ejemplo de resultados



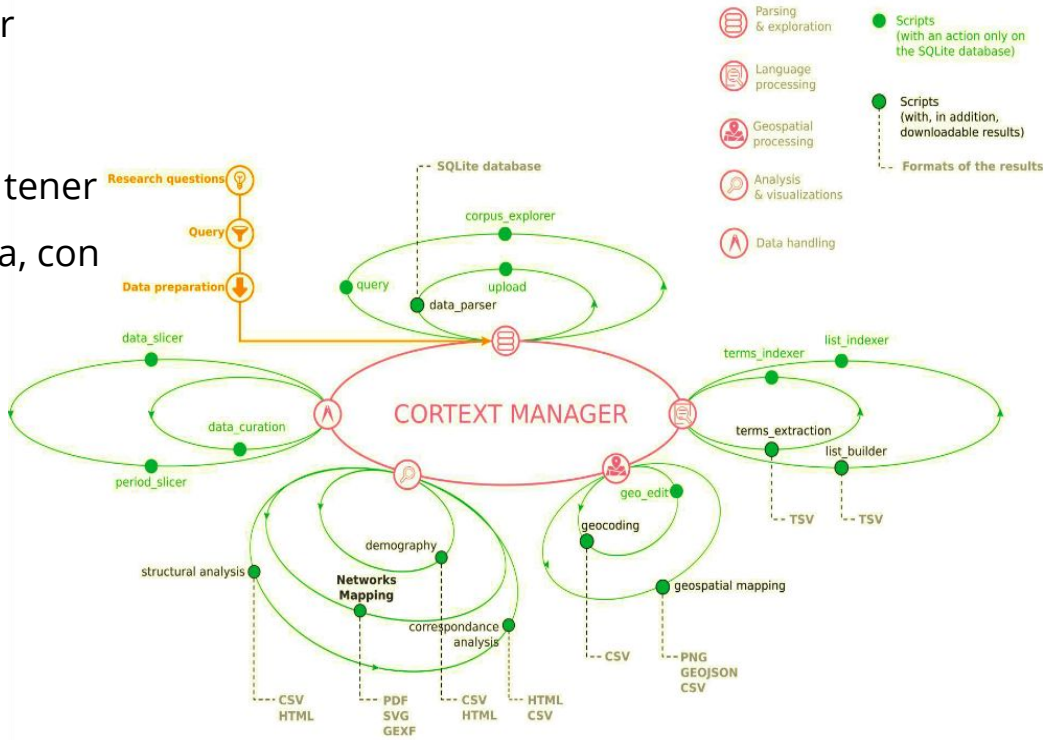
Ejemplo:

Evolución de la investigación clínica y experimental en ciencias biomédicas en México entre 1989 y 2020.

- Categoría 1: documentos mencionando términos asociados a investigación clínica (rats, gene expressions, DNA, etc.)
- Categoría 2: documentos mencionando términos asociados a investigación experimental (cases, patients, control group, etc.)

Cómo seguimos

1. Creen una cuenta en CorText Manager https://managerv2.cortext.net/
2. Nos vemos el **jueves 29 de julio** para tener una sesión práctica con la herramienta, con énfasis en:
   a. Procesamiento de lenguaje
   b. Manejo de información
   c. Visualización y análisis



Cómo seguimos

1. Creen una cuenta en CorText Manager
   https://managerv2.cortext.net/

2. Descarguen las bases de datos para trabajar

   Nos vemos el **jueves 29 de julio** a las **8 AM de Ciudad de México**

   8 AM en Bogotá, Lima, Quito;
   9 AM en Santiago de Chile;
   10 AM en Río de Janeiro, Buenos Aires, Montevideo

   ¿Dudas sobre el horario?
   Pueden revisar el de su ciudad en: https://www.worldtimebuddy.com/

Referencias utilizadas

Berry, D. M. (Ed.). (2012). *Understanding digital humanities*. Palgrave Macmillan.

Börner, K. (2010). *Atlas of science: Visualizing what we know*. The MIT Press.

Callon, M., Courtial, J.-P., Penan, H., & Callon, M. (1995). *Cienciometría: La medición de la actividad científica ; de la bibliometría a la vigilancia tecnológica*. Ed. TREA.

Cointet, J.-P. (2017). *La cartographie des traces textuelles comme méthodologie d'enquête en sciences sociales*. ÉCOLE NORMALE SUPÉRIEURE.

de Solla Price, D. J. (1963). *Little science, big science*. Columbia University Press, New York.

Garfield, E. (1955). Citation Indexes for Science. *Science*, *122*(3159), 108–111. https://doi.org/10.1126/science.123.3185.61-a

Moretti, F. (2013). *Distant reading*. Verso.

Rogers, R. (2013). *Digital methods*. The MIT Press.

Salganik, M. J. (2018). *Bit by bit: Social research in the digital age*. Princeton University Press.

Shi, C., Li, Y., Zhang, J., Sun, Y., & Yu, P. S. (2017). A survey of heterogeneous information network analysis. *IEEE Transactions on Knowledge and Data Engineering*, *29*(1), 17–37. https://doi.org/10.1109/TKDE.2016.2598561

Zitt, M., & Bassecoulard, E. (2004). Internationalisation in science in the prism of bibliometric indicators. En *Handbook of quantitative science and technology research* (pp. 407–436). Springer.

Para estar conectadxs
durante el este curso y después

Telegram

https://t.me/joinchat/gbAPSBt0-2g2Mzcx

`cortext_esocite`