Culturomics con NGram & Wikipedia

Cursos de doctorado:
Herramientas computacionales para el análisis y visualización de datos en Ciencias Sociales y Humanidades

Daniel Torres-Salinas
Universidad de Granada

Escuela Internacional de Posgrado
Actividades Formativas 2024

Índice

1. Breve introducción a la culturomics **2**
2. Inside Books Ngram Viewer **10**
2. Herramientas Wikipedia **17**
3. Ejercicios prácticos **24**
4. Bibliografía básica **26**

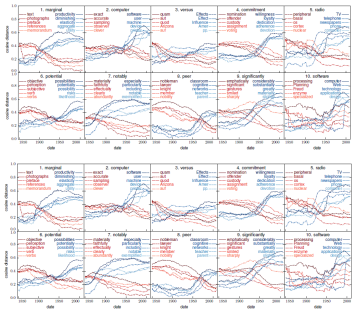
Resumen

El curso aborda dos herramientas digitales abiertas y gratuitas fundamentales en el ámbito de la investigación digital: Google Books Ngram Viewer y los paquetes de análisis de estadísticas de Wikipedia. Mediante estos recursos, se propone introducir y enseñar a utilizar dichas plataformas para rastrear la evolución del lenguaje, así como la evolución e impacto de diversas tendencias culturales, la popularidad y atención que reciben determinados temas académicos o populares. Se busca responder a múltiples preguntas de investigación mediante el análisis cuantitativo de grandes volúmenes de datos textuales, adoptando como marco teórico la perspectiva de la culturonomía. En resumen, este curso ofrece una inmersión profunda en el uso de herramientas digitales avanzadas para el análisis textual y cultural, destacando su relevancia en la investigación contemporánea y en la comprensión de las dinámicas culturales a lo largo del tiempo.

Profesor: Daniel Torres Salinas
Organiza: Escuela Internacional de Posgrado - UGR, Actividades formativas
Duración: 4 Horas - Contenidos teóricos y prácticos
doi: 10.5281/zenodo.10842119
Versión: 1. Publicado el 20/03/2024

1. Breve introducción a la culturomics

Cinco características muy básicas de la culturomics



1

Puede definirse como una disciplina que **utiliza métodos cuantitativos** para analizar **vastas cantidades** de **datos culturales digitales**, con el fin de explorar **patrones**, **tendencias** y **evoluciones** en el ámbito cultural a lo largo del tiempo

2

Se apoya en la premisa de que **los libros y otros materiales digitales son un reflejo de los cambios culturales y lingüísticos** de la sociedad. Al analizar millones de documentos digitalizados, la culturonomía busca i**dentificar tendencias que sen reflejo de los cambios en la sociedad, la política, la economía**

3

Amplía las fronteras de la investigación cultural e **introduce nuevas metodologías** para el estudio de la evolución cultural, permitiendo un **análisis más sistemático y empírico** que era anteriormente difícil o **imposible de realizar con métodos convencionales**.

4

Interdisciplinariedad: existe **una gran intersección de la culturonomía con otras disciplinas científicas**, como la sociología, la historia y la ciencia política, para estudios culturales más integrados. Asimismo mantiene **estrecha relación la Ciencia de Datos**

5

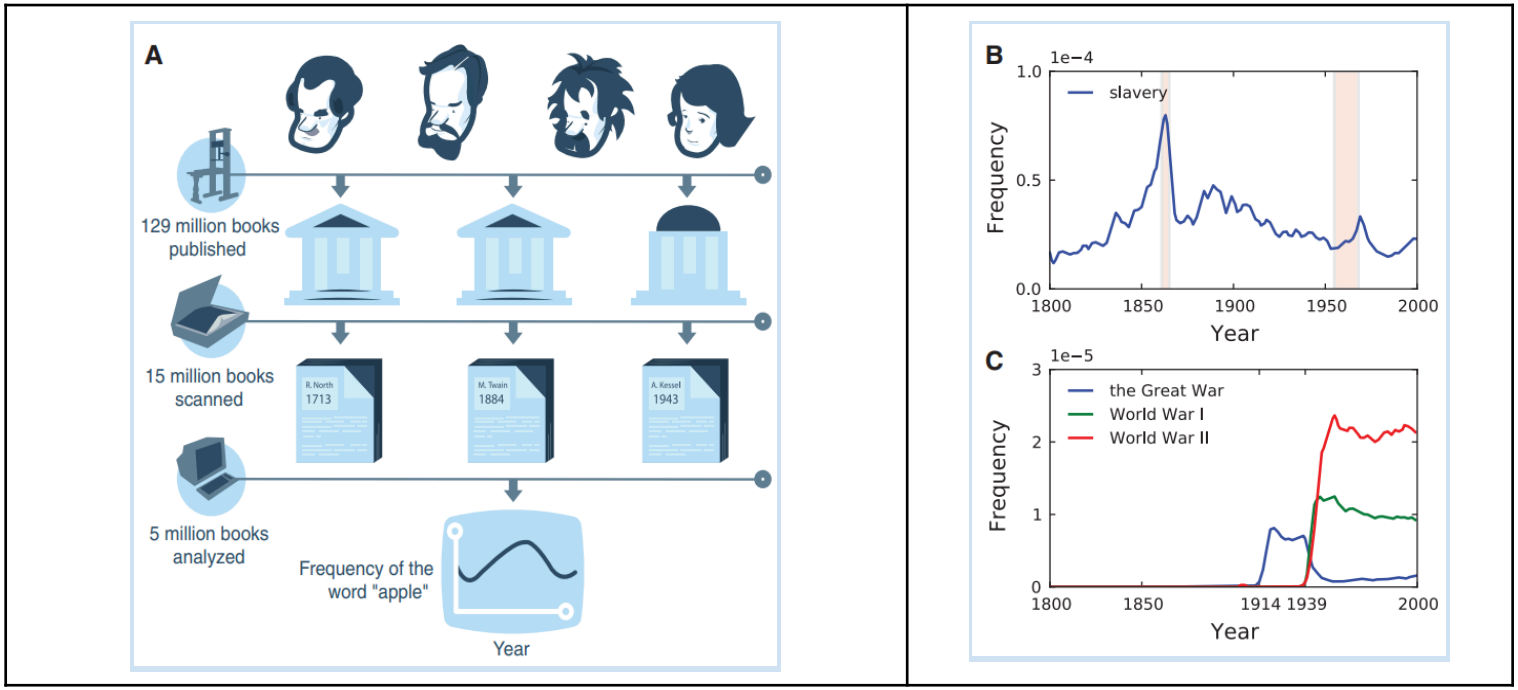
Existe un campo que se ha denominado como la **Culturomics** a partir de este paper publicado en la revista Science “Quantitative Analysis of Culture Using Millions of Digitized Books” (Michel et al., 2011)

Elementos metodológicos para el análisis cuantitativo de la cultura: Culture Analytics y Culturomics				
^^**Corpus Digitalizado Masivo**: grandes bases de datos de textos digitalizados, como libros, periódicos, revistas y otros materiales escritos, que sirven como fuente	****Herramientas de Procesamiento de Lenguaje Natural (PLN)**: Empleo de tecnologías y algoritmos de PLN para analizar, entender y extraer información de los textos en el corpus.	****Análisis de N-Gramas**: Estudio de secuencias de palabras (n-gramas) para identificar y seguir la frecuencia de aparición de términos específicos a lo largo del tiempo.	^^**Técnicas Estadísticas y Matemática**s: Aplicación de métodos estadísticos para analizar y interpretar los patrones encontrados en los datos, permitiendo identificar tendencias y correlaciones.	^^**Visualización de Datos**: Creación de representaciones gráficas de los resultados del análisis, como gráficos de tendencias temporales, mapas de calor y redes semánticas, facilitando la interpretación y comprensión de los hallazgos.
^^**Estudios Longitudinales**: Realización de análisis a lo largo de extensos periodos temporales, permitiendo observar la evolución de la cultura y del lenguaje a través de décadas o incluso siglos.	^^**Acceso abierto y Reproducibilidad**: Promoción del acceso abierto a los datos y herramientas utilizados, así como la transparencia en los métodos de análisis, para facilitar la verificación y replicación de los estudios.	^^**Ética y privacidad**: Consideración de aspectos éticos relacionados con el uso de datos digitalizados y el respeto a la privacidad y derechos de autor de los materiales analizados.	^^**Interpretación Crítica**: se enfatiza la importancia de una interpretación crítica y contextual de los resultados, teniendo en cuenta los factores socioculturales, históricos y políticos relevantes.	^^**Preocupación por la representatividad y el sesgo**:Existe una preocupación compartida por desarrollar metodologías que permitan interpretaciones justas y equitativas de los datos culturales.

** Técnicas específicas de las cuturonomicas que se centra en el lenguaje >> N-Gram

^^El resto son características compartidas con la Culture Analytics >> Wikipedia

Enfoque metodológico de la propuesta original de la culturomics de Michel et al. 2011

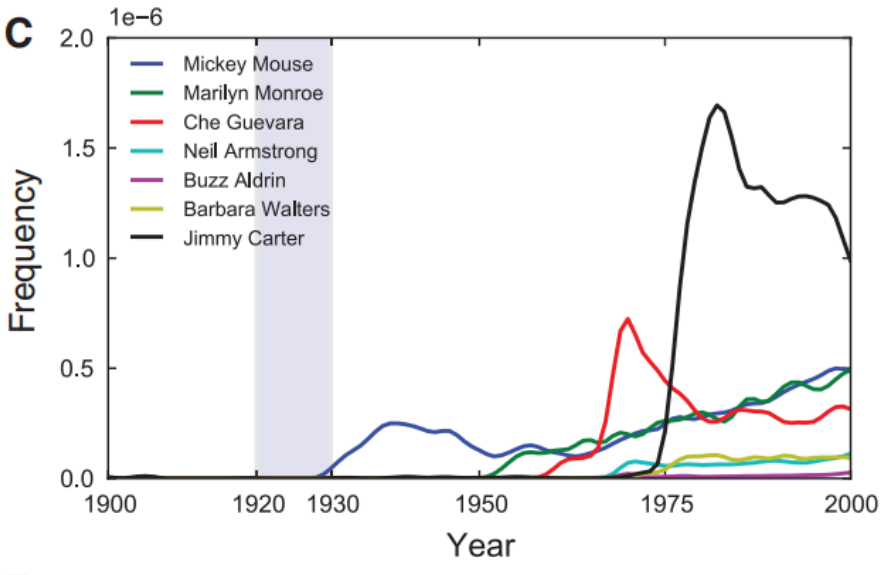


Los análisis culturónomicos estudian millones de libros al mismo tiempo. (A) Fila superior: aproximadamente 129 millones de ediciones de libros han sido publicadas desde el advenimiento de la imprenta. Segunda fila: Bibliotecas y editoriales proveen libros a Google para su digitalización (medio izquierda). Más de 15 millones de libros han sido digitalizados. Tercera fila: Cada libro se asocia con metadatos. Cinco millones de libros son seleccionados para análisis computacional. . (B) Frecuencia de uso de "esclavitud". La Guerra Civil (1861–1865) y el movimiento por los derechos civiles (1955–1968) se destacan en rojo. El número en la parte superior izquierda (1e-4 = 10–4) es la unidad de frecuencia. (C) Frecuencia de uso a lo largo del tiempo para "la Gran Guerra" (azul), "Primera Guerra Mundial" (verde) y "Segunda Guerra Mundial" (rojo).

Las dos herramientas fundamentales	
Google Books	Google Books Ngram Viewer
Google Books constituye una infraestructura fundamental para la culturonomía, albergando una **amplia colección de textos digitalizados que abarcan una multiplicidad de disciplinas, épocas y lenguas**. Este proyecto, iniciado por Google, ha facilitado la transición de obras literarias, documentos históricos y tratados científicos al **dominio digital, democratizando el acceso al conocimiento y ofreciendo una plataforma inigualable para la investigación académica**. La contribución de Google Books a la culturonomía radica en su capacidad para **suministrar un corpus textual extenso, necesario para el análisis cuantitativo de patrones a gran escal**a.	a herramienta analítica de gran valor **permitiendo el rastreo y la visualización de la frecuencia de n-gramas** (secuencias de palabras) dentro del extenso corpus de Google Books. Esta herramienta posibilita investigaciones sobre la evolución del lenguaje y las fluctuaciones en la prevalencia de conceptos culturales, ofreciendo una **perspectiva cuantitativa sobre las transformaciones en el discurso y las prácticas discursiva**s a lo largo del tiempo. **Varios corpus. Varios idiomas** La simplicidad de su interfaz, combinada con la profundidad de su base de datos, convierte a **NGram Viewer en un instrumento esencial para el análisis empírico, además es gratuito**
Google Books Now - Google Books Classics - El proyecto	Ngram Viewer

★ Ojo con los datos es antiguo **pero fijaros en 2011** tenían **5,195,769** libros escaneados con las siguientes palabras por idioma": 500 mil millones de palabras.

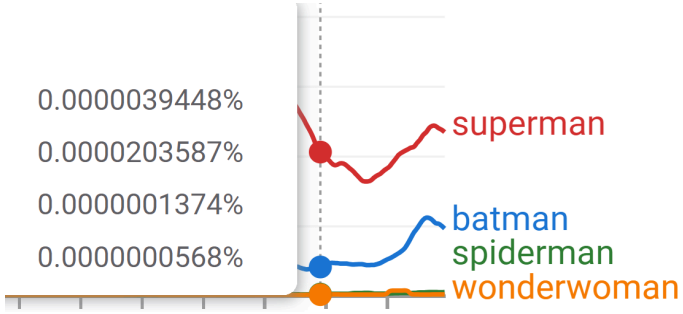
- inglés (361 mil millones)
- francés (45 mil millones)
- español (45 mil millones)
- alemán (37 mil millones)
- chino (13 mil millones)
- ruso (35 mil millones)
- hebreo (2 mil millones).



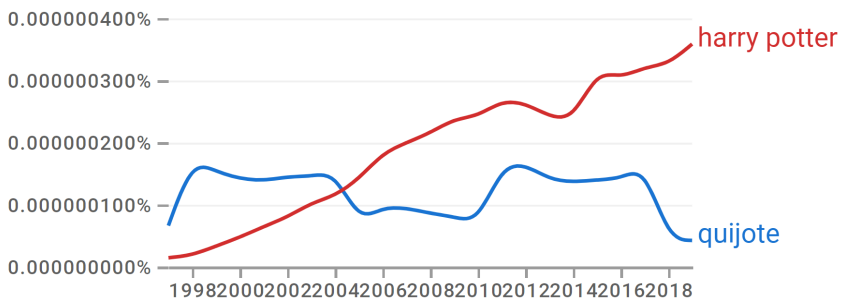
A lo largo del trabajo podemos ver algunos ejemplos como por ejemplo este de la popularidad de algunos personajes históricos de EEUU

Veamos otros ejemplos directos de Michel et al.

A continuación veamos otros ejemplos de todo tipo que he recopilado y he inventado 🎈

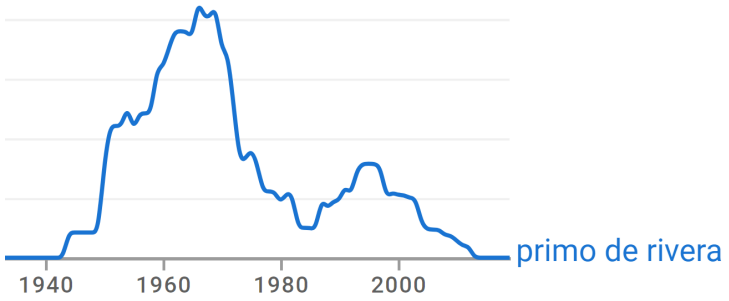


Superman: clásico eterno



El declive de Don Quijote en el mundo anglosajón

- 🦸Cultura popular: superhéroes famosos
- 🧙 **El quijote Vs Harry Potter**
  - Aquí en el resultado en el Corpus Inglés
  - Aquí el resultado en el Corpus Español
- 🥤 En la alimentación: lucha entre dos refrescos

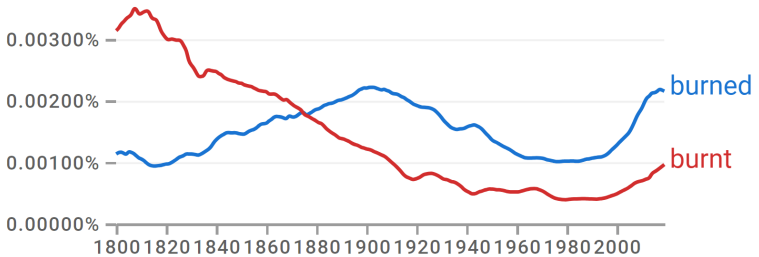


A quién le interesa ahora este tío

Un autor inmortal en la su cima

★ **En la censura y política**: Marc Chagall o Herman HessePrimo de Rivera en España
★ **Tendencias académicas**: ¿Cuáles son las tendencias actuales en traducción?
★ **Movimientos**: tenemos un tipo que analiza el marxismo
★ **Medios**: Otro que analiza cuatro medios de comunicación
★ **Ciudades:** para ver las ciudades más literarias
★ **Términos**: para ver el uso de un término en un contexto político social

★ Según los autores "*Culturomics has profound consequences for the study of language, lexicography, and grammar*" (Véase figura 2 del artículo)



Uso de **Burnt** or **Burned**

- Más ejemplos lingüísticos
  - Backwards compatibility o backward compatibility
  - video game o computer game

- Otros experimentos y ejercicios similares
  - GoogleWriting Tools: Ngram Viewer and Define
  - La alternativa francesa 🇫🇷 Gallicagram
  - Think Outside the box 📦

★ Aquí podemos encontrar más ejemplos de todo lo que hemos visto
★ Aquí muy importante, errores y limitaciones: Pitfalls

2. Inside Books Ngram Viewer

★ Los Corpus tradicionales ¿cómo funcionan y cuáles son sus elementos?

Uso de corpus y diccionarios combinatorios en la traducción

Mariana Orozco-Jutorán

Corpus	Autor	Volumen	Fuentes	Periodo	Distribución	Características
Corpus del Español (CdE) ➢ Abrir enlace	Mark Davies, Brigham Young University (EEUU)	Ca. 7,2 mil millones de palabras	Prensa digital	2012-2019 (ampliaciones periódicas)	78% América 22% España	Búsqueda y comparación por países y en el tiempo; el corpus más actual y amplio
Corpus del Español del Siglo XXI (CORPES) ➢ Abrir enlace	RAE	Ca. 286 millones de formas	Textos escritos y orales: ficción, no ficción, prensa, Internet	2001-2016 (ampliaciones periódicas)	70% América 30% España	
Corpus de Referencia del Español Actual (CREA) ➢ Abrir enlace	RAE	Ca. 160 millones de formas	Textos escritos y orales: ficción, no ficción, prensa y medios de comunicación	1975-2000	50% América 50% España	Amplia variedad de géneros y temáticas; búsqueda por países, temáticas, medios y autores.
Corpus Diacrónico del Español (CORDE) ➢ Abrir enlace	RAE	Ca. 200 millones de formas	Textos escritos (prosa y verso): ficción, no ficción, prensa	Hasta 1974	26% América 74% España	

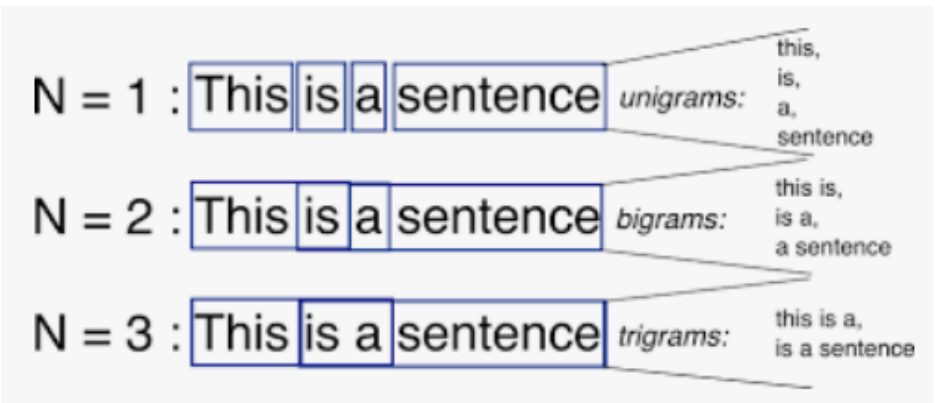
Ya hemos visto que NGram es una aplicación lanzada por Google en 2010, en este caso nos encontramos ante un corpus masivo de carácter multilingüe

★ "*Al examinar los libros de forma colectiva, Google puede procesar el texto y proporcionar la repetición de la aparición de palabras basada en datos estadísticos. Con la herramienta de búsqueda Google Ngram Viewer, puede buscar a través de estos enormes datos estadísticos de manera rápida y efectiva. Al comparar la popularidad relativa de las palabras, puede establecer cómo cambia el idioma y la cultura a lo largo del tiempo. Ngram puede hacer mucho más que simplemente informar la frecuencia de las palabras*" (Fuente)

Pero ¿Qué es un **Ngram**?

Son elementos de una muestra de texto

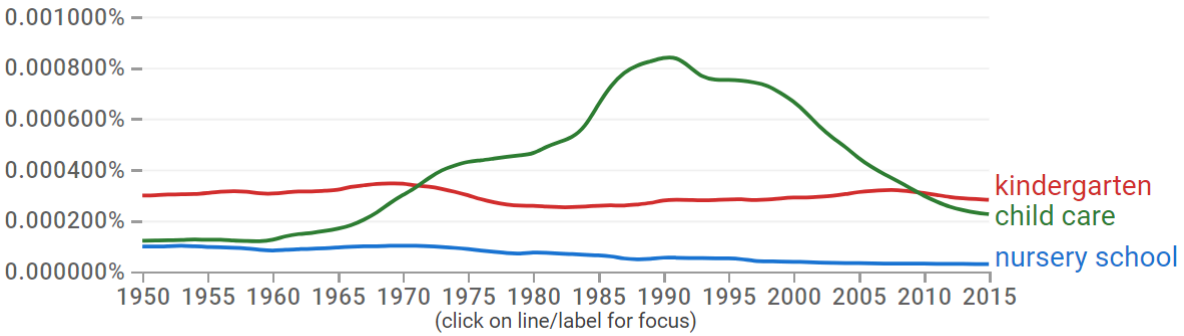
Los N-Grams, siendo secuencias de 'n' ítems (que pueden ser fonemas, sílabas, letras, palabras o pares de palabras) extraídos de un texto o corpus de habla. El N-Grama puede estar compuesto por **grandes bloques de palabras o por una única palabra** o Token. Tenemos dos tipos de **tokens: words and nonwords.** Los N-Gramas se utilizan en el procesamiento del lenguaje natural como una forma de **predecir el texto**. Dependiendo del número de elementos se puede denominar **unigrama, bigrama o trigrama**



Ejemplo básico del concepto de NGram

★ **Comparado unigramas y bigramas.**

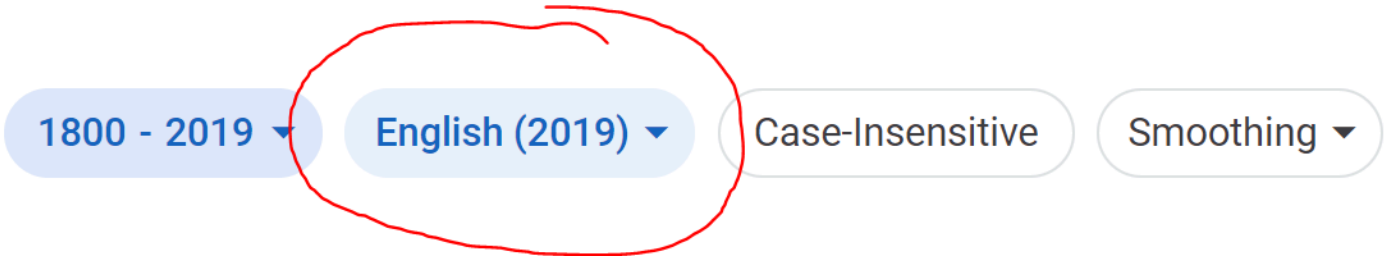
En el siguiente Ejemplo en el siguiente gráfica vemos las frecuencia de tres términos concretos (Kindergarten, Child Care y Nursery School) durante un período de tiempo (1950 y 2015). Claramente se ve el uso de los tres términos. En este caso Kindergarten es un unigram y Chil Care y Nursery School son Bigramas



NGram es una colección de Corpus de diferente naturaleza de varios idiomas y períodos

En inglés, chino (simplificado), francés, alemán, hebreo, italiano, ruso o español. Corpora especializados en inglés, tales como inglés americano, inglés británico, inglés de ficción e "inglés un millón. Todos los corpus se generaron en julio de 2009, julio de 2012 y febrero de 2020; actualizan estos corpus a medida que continúa el escaneo de libros

Los diferentes corpus incluidos en Ngram



American English 2019	eng_us_2019	**English One Million**	eng_1m_2009	**Spanish 2019**	spa_2019
American English 2012	eng_us_2012	**Chinese 2019**	chi_sim_2019	**Spanish 2012**	spa_2012
American English 2009	eng_us_2009	**Chinese 2012**	chi_sim_2012	**Spanish 2009**	spa_2009
British English 2019	eng_gb_2019	**Chinese 2009**	chi_sim_2009	**Russian 2019**	rus_2019
British English 2012	eng_gb_2012	**French 2019**	fre_2019	**Russian 2012**	rus_2012
British English 2009	eng_gb_2009	**French 2012**	fre_2012	**Russian 2009**	rus_2009
English 2019	eng_2019	**French 2009**	fre_2009	**Italian 2019**	ita_2019
English 2012	eng_2012	**German 2019**	ger_2019	**Italian 2012**	ita_2012
English 2009	eng_2009	**German 2012**	ger_2012		
English Fiction 2019	eng_fiction_2019	**German 2009**	ger_2009		
English Fiction 2012	eng_fiction_2012	**Hebrew 2019**	heb_2019		
English Fiction 2009	eng_fiction_2009	**Hebrew 2012**	heb_2012		
		Hebrew 2009	heb_2009		

★ Opciones y técnicas de búsqueda en NGram 🔍

- Con la opción de ortografía case-sensitive (que compara el uso exacto de las mayúsculas y minúsculas), esto es especialmente importantes para nombres propios y puede dar resultado totalmente diferentes. Aquí tenemos un ejemplo

💡 Si nos situamos en sobre la línea de un N-Gram y pulsamos el botón derecho veremos solo ese Ngram

- Podemos buscar por ejemplo con la **Wildcard** en lugar de una palabra y te pone un ranking de la 10 repeticiones más comunes
  - Aquí por ejemplo con “Forest *”
  - Aquí por ejemplo con “Igualdad *”

💡 En la zona inferior tenemos enlaces al lugar exacto donde se ha extraído la palabra, veremos la palabra con un enlace a diferentes períodos temporales que nos remiten a Google Books, una vez que vemos los resultados de un determinado N-Gram incluido podemos acceder al libro 💡 Si nos situamos en sobre la línea de un N-Gram y pulsamos el botón derecho veremos solo ese Ngram

- **Una inflexión** es la modificación de una palabra para representar varias categorías gramaticales como aspecto, caso, género, modo, número, persona, tiempo y voz. Puede buscarlos agregando _INF a un ngram.
  - Por ejemplo con "Search_INF for Information" me busca search y searching

💡 puedes combinar sin problemas Wildcard e Inflexiones

Operadores

Partes de las Oración. Si una palabra incluye muchas partes del discurso, puede agregar los operadores de texto para ser más específicos al final de los ngrams

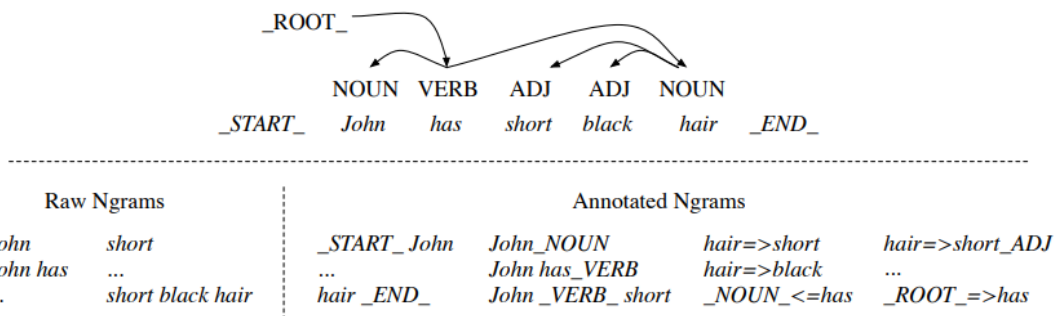
- Podemos indicarle **si queremos buscar un verbo, nombre, un adverbio**, etc..
  - Aquí puedes encontrar todas las opciones de búsqueda
  - Un ejemplo: Play puede ser nombre (play_NOUN) y verbo (play_VERB)

- Operadores **Aritméticos**
  - **"+" y "-" para sumar o restar resultados**
  - Por ejemplo comparativa entre walkman, ipod y spotify

- **Dependencia**: a veces ayuda pensar en las palabras en términos de dependencias en lugar de patrones. Para eso, Ngram Viewer proporciona r**elaciones de dependencia con el operador =>**
  - Por ejemplo todas las veces que aparece el término house con charming
  - o todas las veces que house aparece con charming, lovely o enchanting

- **Combinaciones avanzadas**
  - Todo se puede combinar quiero buscar todos los nombres que aparecen dependiendo del término Pay, utilizamos tres de los operadores que hemos empleado. Otros ejemplos combinación el tipo y la dependencia

★ Ahora podemos comprender mejor como se trabaja con la metodología Ngram. Cuando se coge una frase las frases se convierten en Ngrams que se etiquetan.



Fuente

2. Herramientas Wikipedia

Naturaleza Colaborativa: Permite que cualquier usuario, desde cualquier parte del mundo, pueda editar casi cualquier artículo en cualquier momento. Esto fomenta un esfuerzo colectivo en el que el conocimiento es construido y mejorado por voluntarios. Datos de los Wikipedians Tipos de Usuarios en Wikipedia Ranking de usuarios ES	**Proceso de Revisión y Discusión**: Cada artículo en Wikipedia tiene asociada una página de "discusión" donde los editores pueden debatir mejoras, correcciones y contenido controvertido. Este proceso democrático asegura un consenso colectivo en torno al conocimiento compartido. Ver discusión en Durruti Ver historial en Golpe de Tejero	**Ediciones en Varios Idiomas:** Wikipedia está disponible en más de 300 idiomas, lo que facilita el acceso global al conocimiento y permite la inclusión de perspectivas culturales diversas. Número de Wikipedias y Usuarios activos en cada de ellas
Modelo de Contenido Libre: Wikipedia opera bajo una licencia de contenido libre, específicamente la Licencia de Documentación Libre de GNU y Creative Commons Attribution-ShareAlike, lo que permite que su contenido sea utilizado y distribuido libremente, siempre que se otorgue el debido crédito y se comparta de manera similar.	**Sistema de Referencias y Citas**: Los artículos en Wikipedia deben estar bien referenciados con fuentes fiables. Esto es crucial para asegurar la veracidad y fiabilidad de la información presentada. Ejemplos de citación de artículos científicos en Wikipedia (Torres-Salinas et al., 2019)	**Funcionalidades Complementarias**: Wikipedia ofrece diversas herramientas y funcionalidades, como la posibilidad de crear libros a partir de artículos seleccionados, portales temáticos, resúmenes de noticias y acceso a bases de datos relacionadas como Wikimedia Commons Portales / Portal Al-Andalus Portales más famosos en ES Wikimedia

Los cinco pilares de WIkipedia («Wikipedia en español», 2024)

Es una enciclopedia
Busca un «punto de vista neutral»
Es de contenido libre
Sigue unas reglas de etiqueta
No tiene normas fijas

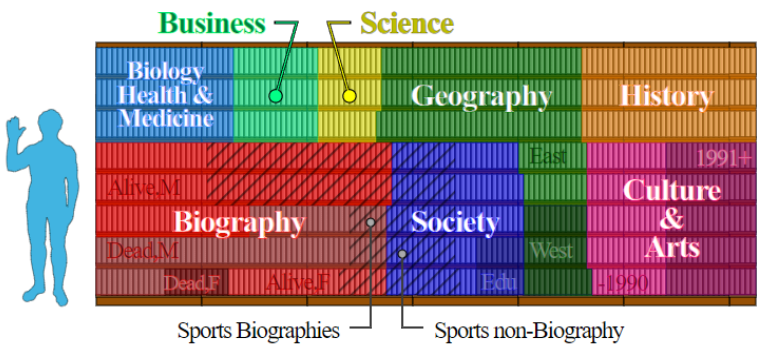
★ **Algunas estadísticas globales de wikipedia**

En esta páginas podemos consultar diferentes aspectos /

- Tasa de crecimiento / Número de palabras / Número de páginas

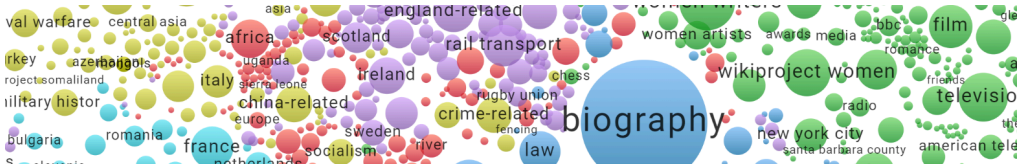
Wikipedia - Estadísticas / Estadística de Wikimedia

★ **Contenidos en Wikipedia**



Aquí vemos la distribución de temas de la wikipedia en inglés y como se distribuyen

En un trabajo intentamos clasificar la Wikipedia (Arroyo-Machado et al., 2022)



Aquí puedes verlo de manera interactiva: link a mapa en VoSviewer y además si consultas el artículo podrás ver que áreas están más "academizadas" en función del número de páginas que tienen un mayor porcentajes de referencias a dois e ISBN

★ Evaluación de contenido en Wikipedia («Wikipedia», 2024)

All rated articles by quality and importance						
Quality	Importance					
	Top	High	Mid	Low	???	Total
FA	1,550	2,443	2,361	1,875	182	**8,411**
FL	170	632	722	673	109	**2,306**
A	354	680	782	576	83	**2,475**
GA	3,132	7,099	14,376	18,791	1,727	**45,125**
B	16,497	31,961	52,645	64,619	19,571	**185,293**
C	16,458	52,677	131,285	291,769	83,039	**575,228**
Start	18,482	91,486	410,070	1,560,947	394,042	**2,475,027**
Stub	4,234	31,566	277,968	2,764,351	760,623	**3,838,742**
List	4,708	16,643	52,566	185,321	65,928	**325,166**

- **FA (Artículo Destacado)**: Máxima calidad, cumple con los criterios más estrictos de Wikipedia.
- **FL (Lista Destacada)**: Listas que cumplen criterios similares a los artículos destacados.
- **A: Excelente calidad**, aunque no alcanza el estándar de artículo destacado.
- **GA (Buen Artículo)**: Artículos que son buenos pero no están tan refinados como los artículos destacados.
- **B: Artículos que son decentes**, pero necesitan trabajo adicional para alcanzar un nivel superior.
- **C: Artículos con un nivel básico** de información, pero con contenido y estructura que requieren mejora.
- **Start (Inicio)**: Artículos en etapas iniciales de desarrollo, a menudo con deficiencias.
- **Stub (Esbozo):** Artículos muy cortos o incompletos.
- **List (Lista):** Artículos que son principalmente listas.

★ **¿Cómo podemos utilizar wikipedia en investigación?**

Desarrollo del contenido a lo largo del Tiempo: Observar cómo han crecido y evolucionado los artículos desde su creación. Esto puede dar luces sobre el interés en ciertos temas y cómo ha cambiado con el tiempo.	**Popularidad de temas:** Mediante el seguimiento de las vistas de página y la actividad de edición en los artículos, se pueden identificar tendencias en la popularidad de ciertos temas.	**Cobertura geográfica y cultural:** Analizar cómo la cobertura de temas varía por regiones y culturas, lo que podría indicar sesgos o deficiencias en la representación del conocimiento global.	**Actividad de edición durante eventos:** Examinar la actividad de edición durante eventos importantes (como elecciones, desastres naturales, lanzamientos de productos, etc.) podría mostrar cómo la información se expande y se corrige en tiempo real.
Polarización y consenso: Estudiar la discusión y la edición de temas controvertidos podría revelar información sobre polarización en la sociedad, así como sobre cómo se llega al consenso en la comunidad de Wikipedia.	**Enlaces externos y referencias:** El análisis de las referencias y los enlaces externos puede ofrecer una perspectiva sobre las fuentes de información más utilizadas y confiables en diferentes campos del conocimiento.	**Demografía de los editores:** Comprender quiénes son los editores de Wikipedia, su ubicación, y la diversidad de su demografía puede iluminar aspectos de inclusión y sesgo en la creación de contenido.	**Cambios en las políticas de edición:** Observar cómo las políticas de Wikipedia han cambiado y cómo esto afecta la creación y edición de contenido.

★ **Ejemplos de contenido popular**
The top 50 Report - 2023
Wikipedia:Today's featured article/Most viewed
Buscar en Wikimedia en la preguntas artículos más vistos

★ **Herramientas: Pageviews**
La página "Pageviews Analysis" de Wikimedia Toolforge proporciona herramientas analíticas para evaluar y comparar las vistas de página de los artículos de Wikipedia. Los usuarios pueden examinar datos de uso diario o mensual, filtrar por tipo de dispositivo y tipo de usuario (humano o bot), y obtener visualizaciones como gráficos de líneas o barras. Es una herramienta útil para entender la popularidad de diferentes páginas de Wikipedia y observar tendencias a lo largo del tiempo.

Opciones de consulta (I) Visitas (menú superior)

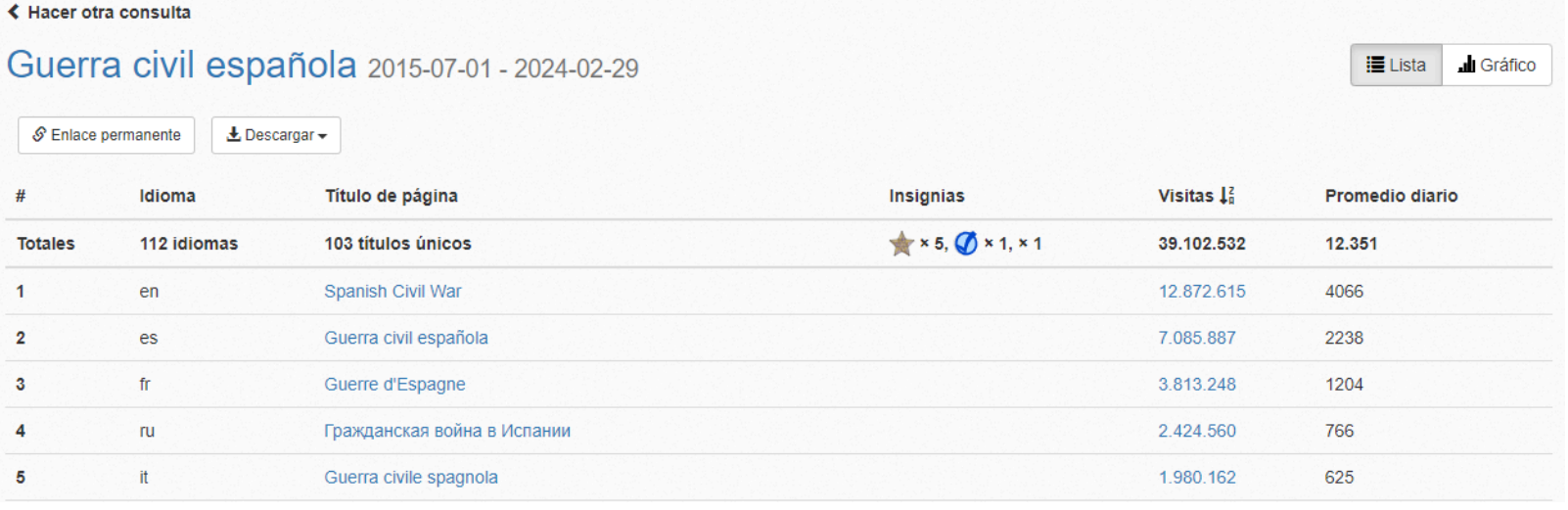
- Filtros que podemos hacer
  - Fechas
  - Tipo de Fecha
  - Proyecto
    - Plataforma Usuario
- Veamos un ejemplo de como funciona analizando la popularidad de los líderes políticos españoles actuales (enlace a la consulta)
- A continuación veamos dos cuestiones
  - Tipos de gráficos que podemos emplear (línea, barra, radial, …)
  - Exportación de los datos (CSV, PNG, JSON)

- Es una herramienta para ver las popularidad de eventos en diferentes contextos, por ejemplo a continuación vamos a ver la popularidad a través de diferentes proyectos de WIkipedia de una acontecimiento histórico como la Guerra Civil Español
  - Guerra Civil Española en es.wikipedia (7.085.887 visitas)
  - Spanish Civil War en en.wikipedia (12.872.615 visitas)
  - Guerre d'Espagne en fr.wikipedia (3.813.visitas)
  - Spanische Bürgerkrieg en de.wikipedia (1.957.041 visitas)

Opciones de consulta (II) Langviews (menú superior)

- Podemos buscar la información de una página en todos los idiomas

Hemos buscado Civil Española y vemos que ocurre en todos los idiomas



Hacer otra consulta

Guerra civil española 2015-07-01 - 2024-02-29

Lista | Gráfico

Enlace permanente | Descargar

#	Idioma	Título de página	Insignias	Visitas	Promedio diario
Totales	112 idiomas	103 títulos únicos	× 5, × 1, × 1	39.102.532	12.351
1	en	Spanish Civil War		12.872.615	4066
2	es	Guerra civil española		7.085.887	2238
3	fr	Guerre d'Espagne		3.813.248	1204
4	ru	Гражданская война в Испании		2.424.560	766
5	it	Guerra civile spagnola		1.980.162	625

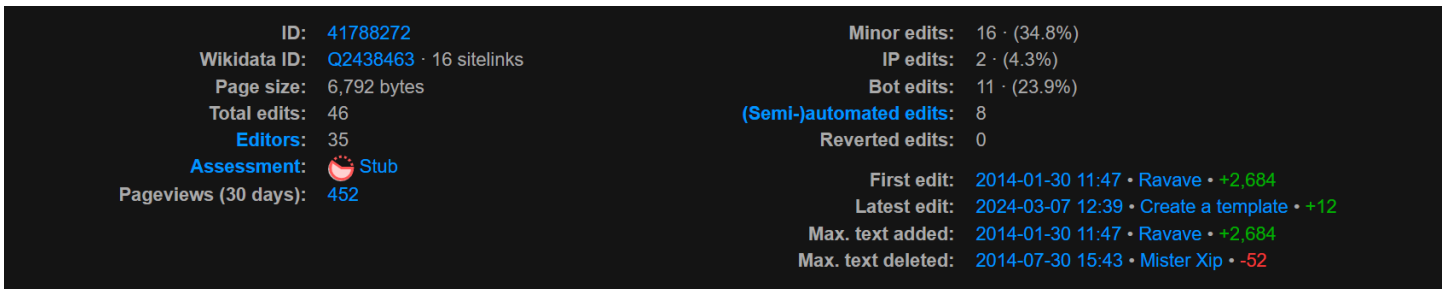
Opciones de consulta (III) Topviews(menú superior)

- Aquí podemos las páginas webs más populares pero es interesante ya que también podemos buscar webs específicas y conocer su posición específica en el ranking de Wikipedia
- En este caso vemos las guerras en España más populares en wikipedia en Febrero de 2024, curiosamente la más consultada es la Primera Guerra Mundial (posición 30 del ranking) frente a la Guerra Civil Española (posición 325 del ranking)

Opciones de consulta (IV) Redirecciones (menú superior)

- Sí bien no es una herramienta de datos, nos permite conocer cual es el control de autoridades o cual es al entrada oficial sobre un tema concreto
- por ejemplo aquí podemos cuál es la entrada exacta que deberíamos consultar para la 11-M y cuales son las entradas que redirigen a la misma

★ **Otras Herramientas: xtools** - Page History



ID:	41788272	Minor edits:	16 · (34.8%)
Wikidata ID:	Q2438463 · 16 sitelinks	IP edits:	2 · (4.3%)
Page size:	6,792 bytes	Bot edits:	11 · (23.9%)
Total edits:	46	(Semi-)automated edits:	8
Editors:	35	Reverted edits:	0
Assessment:	Stub		
Pageviews (30 days):	452	First edit:	2014-01-30 11:47 • Ravave • +2,684
		Latest edit:	2024-03-07 12:39 • Create a template • +12
		Max. text added:	2014-01-30 11:47 • Ravave • +2,684
		Max. text deleted:	2014-07-30 15:43 • Mister Xip • -52

3. Ejercicios prácticos

1. Utilizando el Corpus español me gustaría conocer la evolución en el uso de los términos “ordenador” “computadora” y “pc”. Describe brevemente la evolución
2. ¿Qué término se utiliza más “escritor independiente” o “escritor freelance”
3. Describeme la popularidad de las siguientes escritoras españolas (rosa montero,almudena grandes,rosa chacel,carmen laforet) a lo largo de las décadas (1980-1990, 1990-2000, etc…)
4. Consulta en el Diccionario Panhispánico de dudas la palabra Yincana. A continuación cuatro voces incorrectas del término Yincana. Compara todos los términos en Ngram (Corpes español) y establece un ranking de la más empleada. ¿A partir de de qué fecha se empieza a emplear más el término Yincana? ¿Serías capaz de encontrar una explicación a este fenómeno?
5. En el siguiente artículo June Casagrande utiliza NGram Viewer para ver el uso de una palabra. ¿qué palabra analiza? ¿qué fenómeno o tendencia identifica mediante NGram? ¿a qué conclusión sobre su uso llega sobre la palabra estudiada?
6. Una rápida con el corpus español
    a. ¿Qué se utiliza más Clinex, Klinex?
    b. ¿Qué se utiliza más Clinex, Klinex, Kleenex?
    c. ¿Qué se utiliza más Clinex, Klinex, Kleenex o pañuelo de papel?
7. Inhóspito e inhabitable ¿qué palabra es la qué consideras que deberíamos utilizar y cuándo se produjo el gran cambio en su uso?
8. Según en n-Gram en español 2019 que es más importante para nosotros ¿el dinero, la salud o el amor?
    a. ¿Dime tres nombres que aparezcan/dependan de dinero?
    b. ¿Dime tres verbos que aparezcan/dependan de dinero?
    c. ¿Dime tres adjetivos que aparezcan/dependan de amor?
9. Finalmente realiza el mismo pequeño estudio con el American-English 2019.

10. Explorando las aplicaciones de Ngram

a. Considerando el tema de su tesis doctoral o su campo de especialización, reflexione sobre cómo los N-Grams de Google, que son secuencias de 'n' palabras extraídas de una vasta colección de textos digitalizados, podrían ser aplicados para enriquecer su investigación. Analice qué patrones, tendencias lingüísticas o evoluciones conceptuales podrían descubrirse mediante este enfoque. Diseñe un experimento que utilice N-Grams para explorar aspectos específicos de su temática, como la frecuencia y la evolución de términos clave a lo largo del tiempo. Después, compartiremos y discutiremos colectivamente los resultados y metodologías aplicadas, en un esfuerzo por ampliar nuestra comprensión y aplicabilidad de los N-Grams en el ámbito académico

11. Miniensayo Wikipedia

a. A través de las dos herramientas que hemos visto, Visitas y Page History (incluida en Xtools), compara básicamente los datos de algunos de los temas que proponemos. Recaba la información que consideres oportuna y realiza un miniensayo con alguna gráfica comparativa. Te dejo algunas ideas: podrías analizar la evolución de su popularidad en un período largo de tiempo, la comunidad que existe en torno a ambas (por ejemplo, número de usuarios y usuarios destacados). Podrías asimismo determinar cuál es más discutido a través de las ediciones y otros indicadores, o podrías analizar cuál es más popular en el exterior

b. Temas a escoger una opción

i. Ciencias políticas: Pablo Iglesias VS Santiago Abascal
ii. Patrimonio: Alhambra VS Sagrada Familia
iii. Literatura: Javier Cercas VS Javier Marías
iv. Cultura Popular: Patrulla X (Xmen) VS Avengers
v. Cine: El Padrino VS Centauros del Desierto
vi. Tema libre, si lo consideras prueba con los temas que te interesen

4. Bibliografía básica

Arroyo-Machado, W., Torres-Salinas, D., & Costas, R. (2022, septiembre 7). Where is the science in Wikipedia? Identification and characterization of scientifically supported contents. Zenodo. https://doi.org/10.5281/zenodo.6967465

Bohannon, J. (2011). Google Books, Wikipedia, and the Future of Culturomics. Science, 331(6014), 135-135. https://doi.org/10.1126/science.331.6014.135

Hai-Jew, S. (2014). Exploring the Google books Ngram viewer for «Big Data» text corpus visualizations.

Lin, Y., Michel, J.-B., Aiden, E., Orwant, J., Brockman, W., & Petrov, S. (2012). Syntactic annotations for the Google Books Ngram Corpus. July, 169-174.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Google Books Team, Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., & Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. Science (New York, N.Y.), 331(6014), 176-182. https://doi.org/10.1126/science.1199644

Orozco-Jutorán, M. (s. f.). Ejemplos de uso de corpus monolingües generales para traducir al español.

Torres-Salinas, D., Romero-Frías, E., & Arroyo-Machado, W. (2019). Mapping the backbone of the Humanities through the eyes of Wikipedia. Journal of Informetrics, 13(3), 793-803. https://doi.org/10.1016/j.joi.2019.07.002

Wikipedia en español. (2024). En Wikipedia, la enciclopedia libre. https://es.wikipedia.org/w/index.php?title=Wikipedia_en_espa%C3%B1ol&oldid=158129897

Wikipedia:Content assessment. (2024a). En Wikipedia. https://en.wikipedia.org/w/index.php?title=Wikipedia:Content_assessment&oldid=1210317568

Wikipedia:Estadísticas. (2024b). En Wikipedia, la enciclopedia libre. https://es.wikipedia.org/w/index.php?title=Wikipedia:Estad%C3%ADsticas&oldid=157914952

Zhang, S. (s. f.). The Pitfalls of Using Google Ngram to Study Language. Wired. Recuperado 20 de marzo de 2024, de https://www.wired.com/2015/10/pitfalls-of-studying-language-with-google-ngram/