

# The Programming Historian en español

---



## Análisis de corpus con Voyant Tools (/es/lecciones/analisis-voyant-tools)

Silvia Gutiérrez De la Torre  (<https://orcid.org/0000-0001-8717-2291>).

En este tutorial se aprenderá cómo organizar y analizar un conjunto de textos con Voyant-Tools.

 Revisado por pares (<https://github.com/programminghistorian/ph-submissions/issues/211>).

 CC-BY 4.0 (<https://creativecommons.org/licenses/by/4.0/deed.en>).

 Apoyar PH (<https://programminghistorian.org/es/apoyanos#donaciones>).

### editado por

- Jennifer Isasi

### revisado por

- Daniela Ávido
- Jennifer Isasi

### publicado

2019-04-20

### modificado

2020-05-12

### dificultad

Bajo

## Contenidos

- Análisis de corpus con Voyant Tools
  - Análisis de corpus
  - Qué aprenderás en este tutorial
  - Creando un corpus en texto plano
    - 1. Buscar textos
    - 2. Copiar en editor de texto plano
    - 3. Guardar archivo
      - En Windows:
      - En Mac:
      - En Linux
  - Cargar el corpus
  - Explorando el corpus
    - Sumario de los documentos: características básicas de tu conjunto de textos
      - Número de textos, palabras y palabras únicas
        - Actividad
      - Extensión de documentos
        - Actividad 2
      - Densidad del vocabulario
        - Actividad 3
      - Palabras por oración
        - Actividad 4
    - Cirrus y sumario: frecuencias y filtros de palabras vacías
      - Frecuencias sin filtro
        - Actividad 5
      - Palabras vacías
        - Actividad 6
      - Frecuencias con palabras vacías filtradas
        - Actividad 7
    - Términos
      - Frecuencia normalizada
      - Asimetría estadística
      - Palabras diferenciadas
        - Actividad 8
    - Palabras en contexto
      - Actividad 9
      - Exportando las tablas
  - Respuestas a las actividades
    - Actividad 1
    - Actividad 2
    - Actividad 3

- [Actividad 4](#)
- [Actividad 5](#)
- [Bibliografía](#)
- [Notas al pie](#)

## Análisis de corpus con Voyant Tools

En este tutorial se aprenderá cómo organizar un conjunto de textos para la investigación; es decir, se aprenderán los pasos básicos de la creación de un corpus. También se aprenderán las métricas principales del análisis cuantitativo de textos. Para este fin, se enseñará a usar una plataforma que no requiere instalación (sólo conexión a Internet): [Voyant Tools](https://voyant-tools.org/?lang=es) (<https://voyant-tools.org/?lang=es>). (Sinclair y Rockwell, 2016). Este tutorial está pensado como un primer paso en una serie cada vez más compleja de métodos de la lingüística de corpus. En este sentido, podría considerarse este texto como una de las opciones para el análisis de corpus que puedes encontrar en PH (ver por ejemplo: "[Análisis de corpus con Antconc \(/es/lecciones/analisis-de-corpus-con-antconc\)](/es/lecciones/analisis-de-corpus-con-antconc)").

### Análisis de corpus

El análisis de corpus es un tipo de [análisis de contenido](#) (<http://vocabularios.caicyt.gov.ar/portal/index.php?task=fetchTerm&arg=26&v=42>) que permite hacer comparaciones a gran escala sobre un conjunto de textos o corpus.

Desde el inicio de la informática, tanto lingüistas computacionales como especialistas de la [recuperación de la información](http://vocabularios.caicyt.gov.ar/portal/?task=fetchTerm&arg=178&v=42) (<http://vocabularios.caicyt.gov.ar/portal/?task=fetchTerm&arg=178&v=42>) han creado y utilizado software para apreciar patrones que no son evidentes en una lectura tradicional o bien para corroborar hipótesis que intuían al leer ciertos textos pero que requerían de trabajos laboriosos, costosos y mecánicos. Por ejemplo, para obtener los patrones de uso y decaimiento de ciertos términos en una época dada era necesario contratar a personas que revisaran manualmente un texto y anotaran cuántas veces aparecía el término buscado. Muy pronto, al observar las capacidades de "contar" que tenían las computadoras, estos especialistas no tardaron en escribir programas que facilitaran la tarea de crear listas de frecuencias o tablas de concordancia (es decir, tablas con los contextos izquierdos y derechos de un término). El programa que aprenderás a usar en este tutorial, se inscribe en este contexto.

### Qué aprenderás en este tutorial

Voyant Tools es una herramienta basada en Web que no requiere de la instalación de ningún tipo de software especializado pues funciona en cualquier equipo con conexión a Internet.

Como se ha dicho en este otro [tutorial \(/es/lecciones/analisis-de-corpus-con-antconc\)](/es/lecciones/analisis-de-corpus-con-antconc), esta herramienta es una buena puerta de entrada a otros métodos más complejos.

Al finalizar este tutorial, tendrás la capacidad de:

- Armar un corpus en texto plano

- Cargar tu corpus en Voyant Tools
- Entender y aplicar diferentes técnicas de segmentación de corpus
- Identificar características básicas de tu conjunto de textos:
  - Extensión de los documentos subidos
  - Densidad léxica (llamada densidad de vocabulario en la plataforma)
  - Promedio de palabras por oración
- Leer y entender diferentes estadísticas sobre los vocablos: frecuencia absoluta, frecuencia normalizada, asimetría estadística y palabras diferenciadas
- Buscar palabras clave en contexto y exportar los datos y las visualizaciones en diferentes formatos (csv, png, html)

## Creando un corpus en texto plano🔗

Si bien VoyantTools puede trabajar con muchos tipos de formato (HTML, XML, PDF, RTF, y MS Word), en este tutorial utilizaremos el texto plano (.txt). El texto plano tienen tres ventajas fundamentales: no tiene ningún tipo de formato adicional, no requiere un programa especial y tampoco conocimiento extra. Los pasos para crear un corpus en texto plano son:

### 1. Buscar textos🔗

Lo primero que debes hacer es buscar la información que te interesa. Para este tutorial, [Riva Quiroga](https://twitter.com/rivaquioga) (<https://twitter.com/rivaquioga>) y yo preparamos un corpus de los discursos anuales de presidentes de Argentina, Chile, Colombia, México y Perú<sup>1</sup> entre 2006 y 2010, es decir dos años antes y después de la crisis económica de 2008. Este corpus ha sido liberado con una licencia [Creative Commons CC BY 4.0](https://creativecommons.org/licenses/by/4.0/deed.es) (<https://creativecommons.org/licenses/by/4.0/deed.es>) y puedes usarlo siempre y cuando cites la fuente usando el siguiente identificador:

DOI [10.5281/zenodo.2547051](https://zenodo.org/record/2547051#.XE9pc1z0mUk) (<https://zenodo.org/record/2547051#.XE9pc1z0mUk>).

### 2. Copiar en editor de texto plano🔗

Una vez localizada la información, el segundo paso es copiar el texto que te interesa desde la primera palabra dicha hasta la última y guardarla en un editor de texto sin formato. Por ejemplo:

- en Windows podría guardarse en [Bloc de Notas](https://web.archive.org/web/20091013225307/http://windows.microsoft.com/en-us/windows-vista/Notepad-frequently-asked-questions) (<https://web.archive.org/web/20091013225307/http://windows.microsoft.com/en-us/windows-vista/Notepad-frequently-asked-questions>).
- en Mac, en [TextEdit](https://support.apple.com/es-mx/guide/textedit/welcome/mac) (<https://support.apple.com/es-mx/guide/textedit/welcome/mac>);
- y en Linux, en [Gedit](https://wiki.gnome.org/Apps/Gedit) (<https://wiki.gnome.org/Apps/Gedit>).

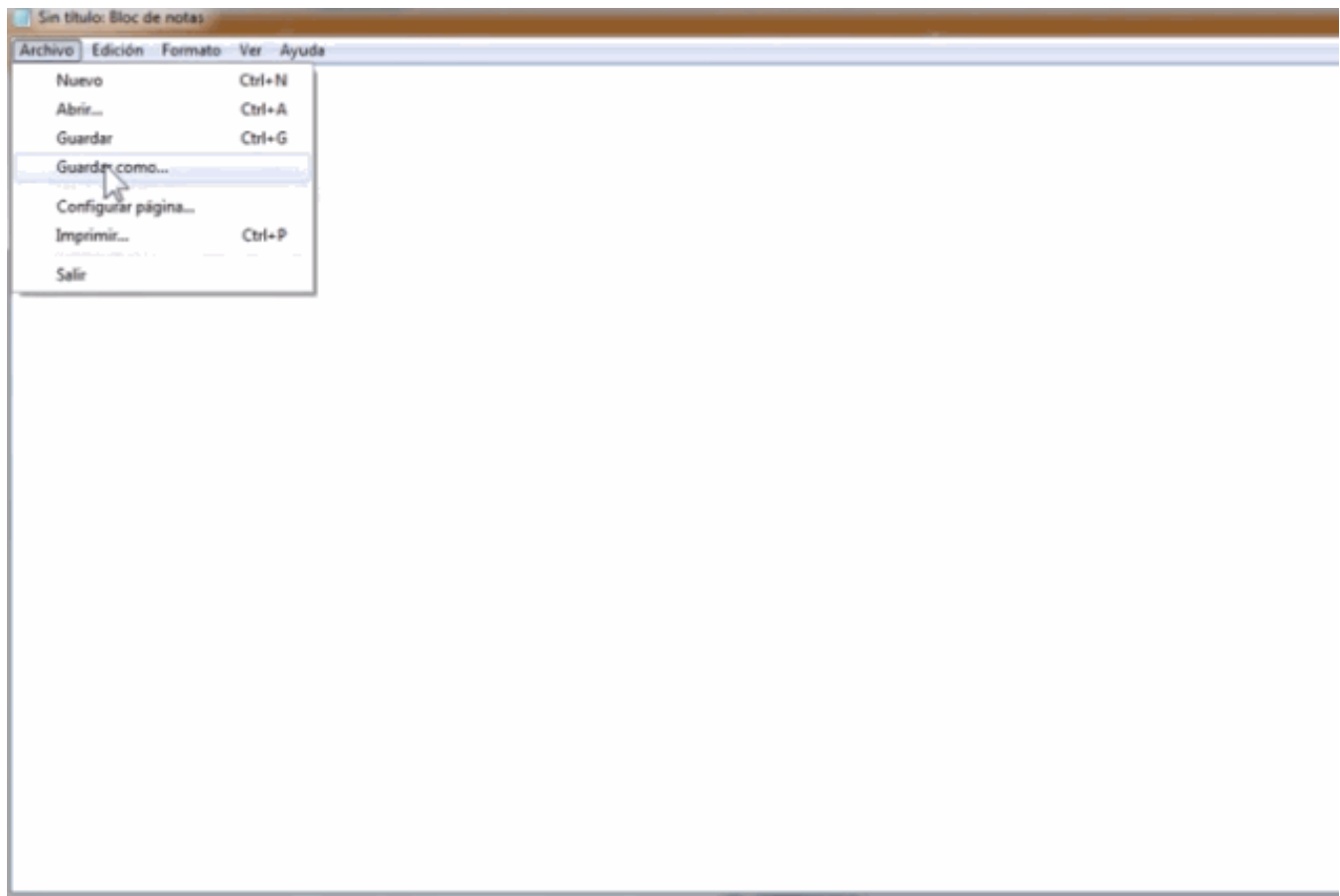
### 3. Guardar archivo🔗

Cuando guardes el texto debes considerar tres cosas esenciales:

Lo primero es **guardar tus textos en UTF-8**, que es un formato de codificación estándar para el español y otros idiomas.

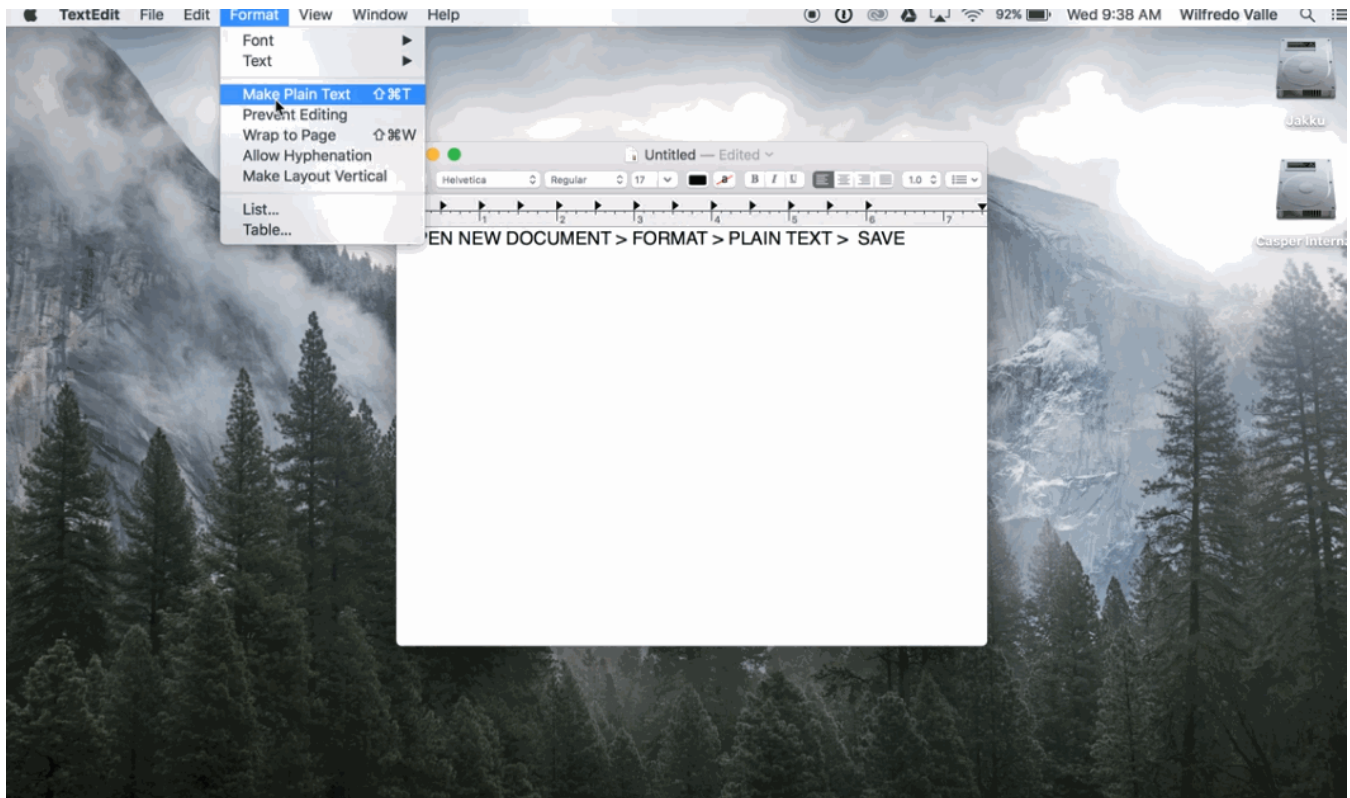
**¿Qué es utf-8?** Si bien en nuestra pantalla vemos que al teclear una "É" aparece una "É"; para una computadora "É" es una serie de ceros y unos que son interpretados en imagen dependiendo del "traductor" o "codificador" que se esté usando. El codificador que contiene códigos binarios para todos los caracteres que se usan en el español es UTF-8. Siguiendo con el ejemplo "11000011", es una cadena de ocho bits –es decir, **ocho** espacios de información– que en UTF-**8** son interpretados como "É"

### En Windows:🔗



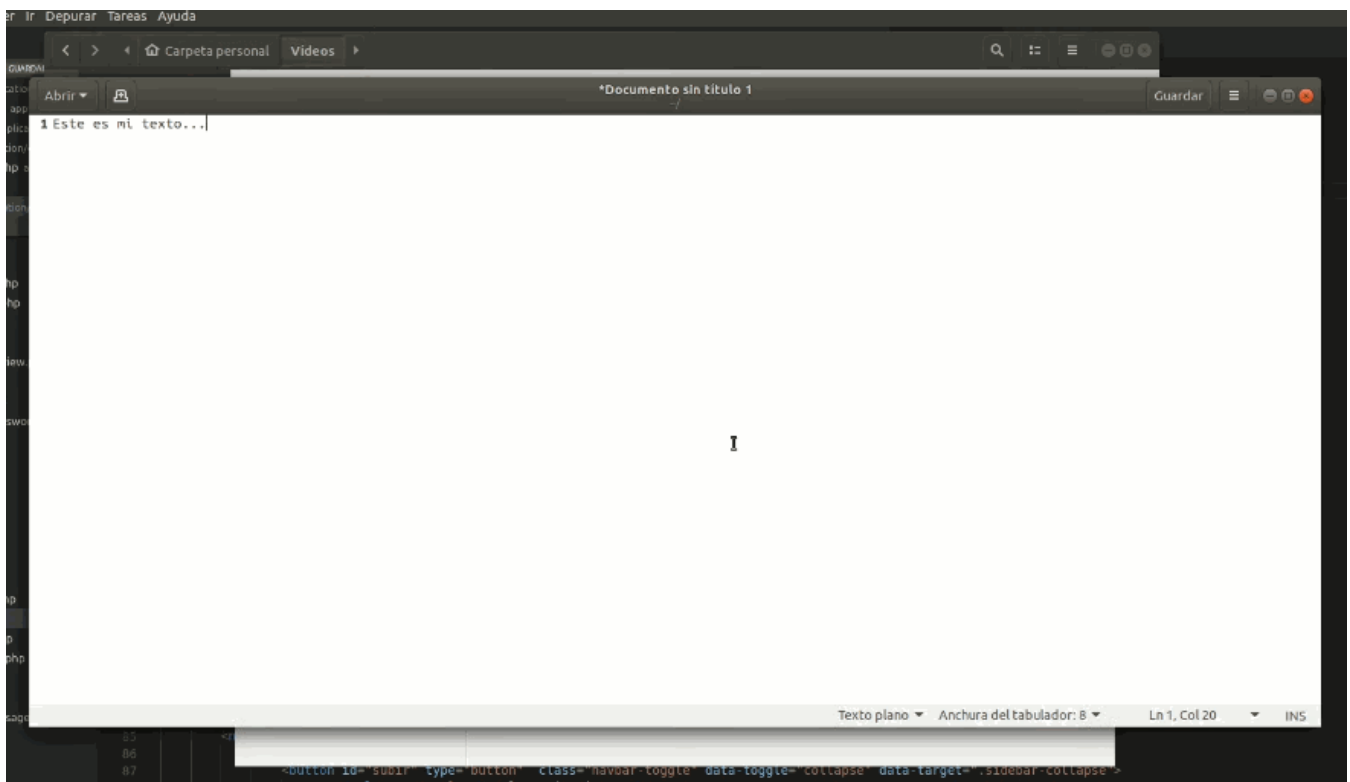
Guardar en UTF-8 en Windows: 1) Abrir Bloc de Notas, 2) Después de pegar o escribir el texto, dar clic en 'Guardar como' 3) En la ventana de 'codificación' seleccionar 'UTF-8' 4) Elegir nombre de archivo y guardar como .txt (Torresblanca, 2014)

### En Mac:🔗



Guardar en UTF-8 en Mac: 1) Abrir TextEdit 2) Pegar el texto que se desea guardar 3) Convertir a texto plano (opción en el menú de 'Formato') 4) Al guardar, seleccionar el encoding 'UTF-8' (Creative Corner, 2016)

## En Linux



Guardar en UTF-8 en Ubuntu: 1) Abrir Gedit 2) Después de pegar el texto, al guardar, seleccionar 'UTF-8' en la ventana de 'Codificación de caracteres'

La segunda es que **el nombre de tu archivo no debe contener acentos ni espacios**, esto asegurará que pueda ser abierto en otros sistemas operativos

**¿Por qué evitar acentos y espacios en los nombres de archivo?** Por razones similares a el inciso anterior, un archivo que se llame Ébano.txt no siempre será entendido de forma correcta por todos los sistemas operativos pues varios tienen otro codificador por defecto. Muchos usan ASCII, por ejemplo, que sólo tiene siete bits de manera que el último bit (1) de "11000011" es interpretado como el inicio del siguiente carácter y se descuadra la interpretación.

La tercera es **integrar metadatos de contexto (v.g. fecha, género, autor, origen) en el nombre del archivo** que te permitan partir tu corpus según diferentes criterios y también leer mejor los resultados. Para este tutorial hemos nombrado los archivos con el año del discurso presidencial, el código del país ([ISO 3166-1 alfa-2](https://es.wikipedia.org/wiki/ISO_3166-1#C%C3%B3digos_oficialmente_asignados) ([https://es.wikipedia.org/wiki/ISO\\_3166-1#C%C3%B3digos\\_oficialmente\\_asignados](https://es.wikipedia.org/wiki/ISO_3166-1#C%C3%B3digos_oficialmente_asignados))) y el apellido de quien profirió el discurso.

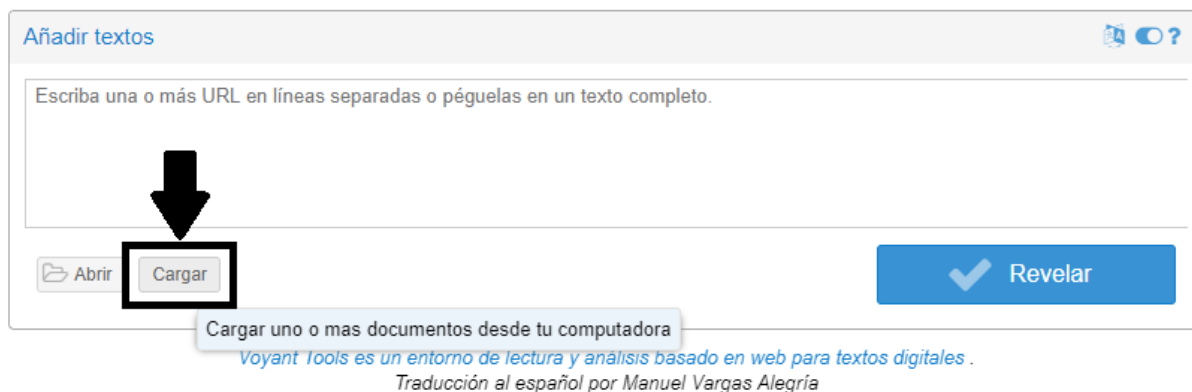
2007\_mx\_calderon.txt ([https://github.com/corpusenespanol/discursos-presidenciales/blob/master/mexico/2007\\_mx\\_calderon.txt](https://github.com/corpusenespanol/discursos-presidenciales/blob/master/mexico/2007_mx_calderon.txt)) tiene el año del discurso dividido con un guión bajo, el código de dos letras del país (México = mx) y el apellido del presidente que dictó el discurso, Calderón, (sin acentos ni ñes)

## Cargar el corpus

En la página de entrada de Voyant Tools encontrarás cuatro opciones sencillas para cargar textos.<sup>2</sup> Las dos primeras opciones están en el cuadro blanco. En este cuadro puedes pegar directamente un texto que hayas copiado de algún lugar; o bien, pegar direcciones web –separadas por comas– de los sitios en donde se encuentren los textos que quieres analizar. Una tercera opción es dar clic en “Abrir” y seleccionar alguno de los dos corpus que Voyant tiene precargados (las obras de Shakespeare o las novelas de Austen: ambos en inglés).

Por último, está la opción que usaremos en este tutorial, en la que puedes cargar directamente los documentos que tengas en tu computadora. En este caso subiremos el corpus completo ([./assets/analisis-voyant-tools/corpus\\_presidentes.zip](https://assets/analisis-voyant-tools/corpus_presidentes.zip)) de discursos presidenciales.

Para cargar los materiales pulsa sobre el icono que dice “Cargar”, abre tu explorador de archivos y, dejando presionada la tecla ‘Shift’ selecciona todos los archivos que desees analizar.



Cargar documentos

## Explorando el corpus

Una vez cargados todos los archivos llegarás a la 'interfaz' ('skin') que tiene cinco herramientas por defecto. A continuación, una breve explicación de cada una de estas herramientas:

- Cirrus: nube de palabras que muestra los términos más frecuentes





Lector: espacio para la revisión y lectura de los textos completos con una gráfica de barras que indica la cantidad de texto que tiene cada documento

**Lector** ○ TérminosBerry ?

### 2006\_ar\_kircher

Señor Vicepresidente de la Nación; señor presidente provisional del Senado; señor presidente de la Cámara de Diputados; señores gobernadores; señores ministros del Poder Ejecutivo Nacional; señores jefes del Estado Mayor Conjunto y de los Estados Mayores Generales de las Fuerzas Armadas; señores legisladores; miembros del Cuerpo Diplomático; señoras y señores: vengo a dejar inauguradas las sesiones del Honorable Congreso de la Nación como lo dispone el inciso 8 del artículo 99 de la Constitución de la Nación Argentina, en esta ocasión prevista por nuestra Ley Fundamental para que en mi carácter de Presidente dé cuenta ante la Asamblea Legislativa del estado de la Nación, repasando lo que hasta aquí hemos recorrido, verificar lo que estamos haciendo y marcar los rumbos que debemos seguir, de eso se trata.

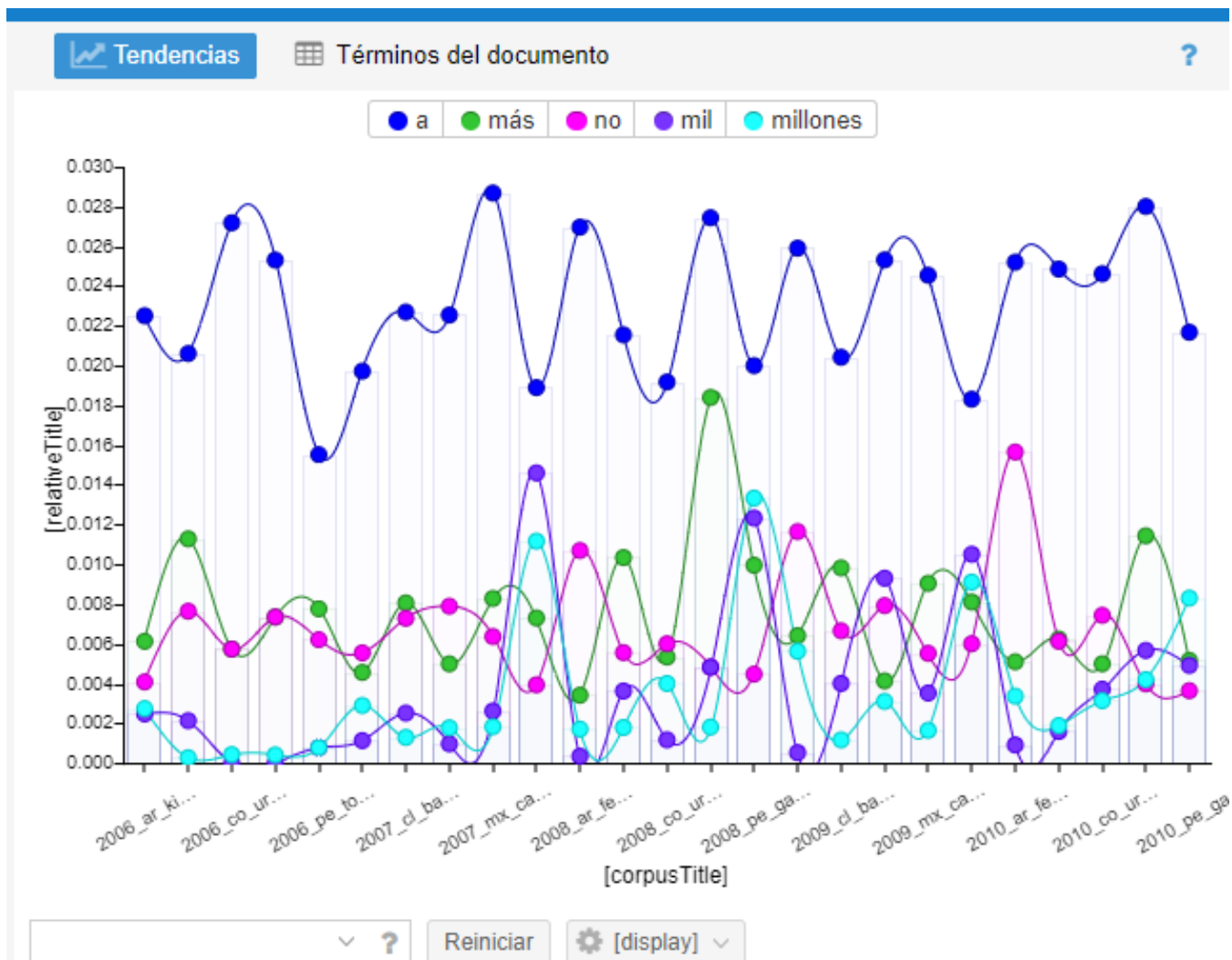
Es preciso siempre recordar de qué situación venimos; vamos de a poco superando con esfuerzo lo que constituyó la peor crisis de nuestra historia; vamos escalando peldaño a peldaño lo que ha sido y todavía es el calvario de la Argentina. Venimos del infierno intentando todavía salir de él, por eso debemos actuar con memoria. Debemos repasar los hechos que marcan con toda contundencia a veces cuánto hemos avanzado, otras veces cuánto nos falta recorrer y otras tantas cuánto cuesta reconstruir lo que ha sido destruido.



← → [?] ?

Lector

- Tendencias: gráfico de distribución que muestra los términos en todo el corpus (o dentro de un documento cuando sólo se carga uno)



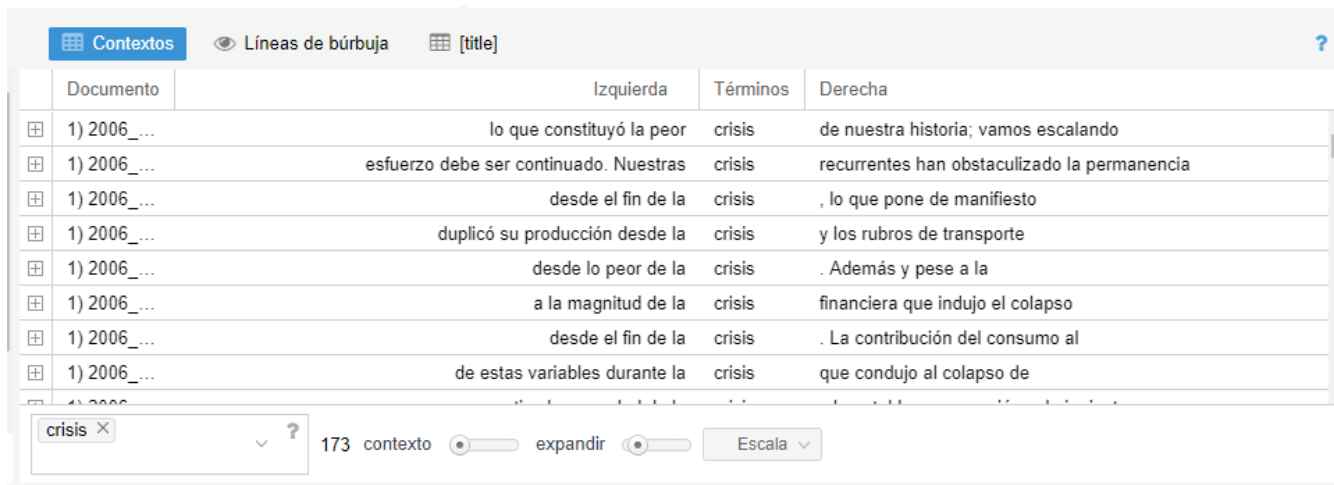
## Tendencias

- Sumario: proporciona una visión general de ciertas estadísticas textuales del corpus actual



## Sumario

- Contextos: concordancia que muestra cada ocurrencia de una palabra clave con un poco de contexto circundante



Documento	Izquierda	Términos	Derecha
1) 2006_...	lo que constituyó la peor	crisis	de nuestra historia; vamos escalando
1) 2006_...	esfuerzo debe ser continuado. Nuestras	crisis	recurrentes han obstaculizado la permanencia
1) 2006_...	desde el fin de la	crisis	, lo que pone de manifiesto
1) 2006_...	duplicó su producción desde la	crisis	y los rubros de transporte
1) 2006_...	desde lo peor de la	crisis	. Además y pese a la
1) 2006_...	a la magnitud de la	crisis	financiera que indujo el colapso
1) 2006_...	desde el fin de la	crisis	. La contribución del consumo al
1) 2006_...	de estas variables durante la	crisis	que condujo al colapso de

crisis x 173 contexto expandir Escala

## Contextos

### Sumario de los documentos: características básicas de tu conjunto de textos

Una de las ventanas más informativas de Voyant es la del sumario. Aquí obtenemos una vista de pájaro sobre algunas estadísticas de nuestro corpus por lo que funciona como un buen punto de partida. En las siguientes secciones obtendrás una explicación de las diferentes medidas que aparecen en esta ventana.

### Número de textos, palabras y palabras únicas

La primera frase que leemos se ve algo como esto:

Este corpus tiene 25 documentos con 261,032 total de palabras y 18,550 formulario de palabra única. Creado hace 8 horas atrás [el texto es producto de una traducción semi-automática del inglés y por eso se lee raro]

De entrada con esta información sabemos exactamente cuántos documentos distintos fueron cargados (25); cuántas palabras hay en total (261,032); y cuántas palabras únicas existen (18,550).

En las siguientes líneas encontrarás nueve actividades que pueden ser resueltas en grupos o individualmente. Cinco de ellas tienen respuestas al final del texto para servir de guía. Las últimas cuatro están abiertas a la reflexión/discusión de quienes las lleven a cabo

## Actividad

Si nuestro corpus estuviera compuesto de dos documentos; uno que dijera: "tengo hambre"; y otro que dijera: "tengo sueño". ¿Qué información aparecería en la primera línea del sumario? Completa:

Este corpus tiene \_ documentos con un total de palabras de \_ y \_ palabras únicas.

## Extensión de documentos

Lo segundo que vemos es la sección de "extensión del documento". Ahí aparece lo siguiente:

- Más largo: 2008 cl bachelet (<https://voyant-tools.org/?corpus=b6f0e2c5ee1bc9b644ffda6b86a93740&panels=cirrus,reader,trends,summary,contexts#>) (20702); 2007 ar kircher (<https://voyant-tools.org/?corpus=b6f0e2c5ee1bc9b644ffda6b86a93740&panels=cirrus,reader,trends,summary,contexts#>) (20390); 2006 ar kircher (<https://voyant-tools.org/?corpus=b6f0e2c5ee1bc9b644ffda6b86a93740&panels=cirrus,reader,trends,summary,contexts#>) (18619); 2010 cl pinera (<https://voyant-tools.org/?corpus=b6f0e2c5ee1bc9b644ffda6b86a93740&panels=cirrus,reader,trends,summary,contexts#>) (16982); 2007 cl bachelet (<https://voyant-tools.org/?corpus=b6f0e2c5ee1bc9b644ffda6b86a93740&panels=cirrus,reader,trends,summary,contexts#>) (15514)
- Más corto: 2006 pe toledo (<https://voyant-tools.org/?corpus=b6f0e2c5ee1bc9b644ffda6b86a93740&panels=cirrus,reader,trends,summary,contexts#>) (1289); 2006 mx fox (<https://voyant-tools.org/?corpus=b6f0e2c5ee1bc9b644ffda6b86a93740&panels=cirrus,reader,trends,summary,contexts#>) (2450); 2008 mx calderon (<https://voyant-tools.org/?corpus=b6f0e2c5ee1bc9b644ffda6b86a93740&panels=cirrus,reader,trends,summary,contexts#>) (3317); 2006 co uribe (<https://voyant-tools.org/?corpus=b6f0e2c5ee1bc9b644ffda6b86a93740&panels=cirrus,reader,trends,summary,contexts#>) (4709); 2009 co uribe (<https://voyant-tools.org/?corpus=b6f0e2c5ee1bc9b644ffda6b86a93740&panels=cirrus,reader,trends,summary,contexts#>) (5807)

## Actividad 2

1. ¿Qué podemos concluir sobre los textos más largos y los más cortos considerando los metadatos en el nombre del archivo (año, país, presidente)?
2. ¿Para qué nos sirve saber la longitud de los textos?

## Densidad del vocabulario

La densidad de vocabulario se mide dividiendo el número de palabras únicas entre el número de palabras totales. Entre más cercano a uno es el índice de densidad quiere decir que el vocabulario tiene mayor variedad de palabras, es decir, que es más denso.

## Actividad 3

## 1) Calcula la densidad de las siguientes estrofas, compara y comenta:

- Estrofa 1. De "Hombres necios que acusáis" de Sor Juana Inés de la Cruz

¿Qué humor puede ser más raro que el que, falto de consejo, él mismo empaña  
el espejo, y siente que no esté claro?

- Estrofa 2. De "Despacito" de Erika Ender, Luis Fonsi y Daddy Yankee

Pasito a pasito, suave suavecito Nos vamos pegando poquito a poquito Cuando  
tú me besas con esa destreza Veo que eres malicia con delicadeza

## 2) Lee los datos de densidad léxica de los documentos de nuestro corpus, ¿qué te dicen?

- Más alto: 2006\_pe\_toledo (<https://voyant-tools.org/?corpus=b6f0e2c5ee1bc9b644ffda6b86a93740&panels=cirrus,reader,trends,summary,contexts#>) (0.404); 2006\_co\_uribe (<https://voyant-tools.org/?corpus=b6f0e2c5ee1bc9b644ffda6b86a93740&panels=cirrus,reader,trends,summary,contexts#>) (0.340); 2009\_co\_uribe (<https://voyant-tools.org/?corpus=b6f0e2c5ee1bc9b644ffda6b86a93740&panels=cirrus,reader,trends,summary,contexts#>) (0.336); 2008\_co\_uribe (<https://voyant-tools.org/?corpus=b6f0e2c5ee1bc9b644ffda6b86a93740&panels=cirrus,reader,trends,summary,contexts#>) (0.334); 2006\_mx\_fox (<https://voyant-tools.org/?corpus=b6f0e2c5ee1bc9b644ffda6b86a93740&panels=cirrus,reader,trends,summary,contexts#>) (0.328)
- Más bajo: 2008\_cl\_bachelet (<https://voyant-tools.org/?corpus=b6f0e2c5ee1bc9b644ffda6b86a93740&panels=cirrus,reader,trends,summary,contexts#>) (0.192); 2007\_mx\_calderon (<https://voyant-tools.org/?corpus=b6f0e2c5ee1bc9b644ffda6b86a93740&panels=cirrus,reader,trends,summary,contexts#>) (0.192); 2007\_ar\_kircher (<https://voyant-tools.org/?corpus=b6f0e2c5ee1bc9b644ffda6b86a93740&panels=cirrus,reader,trends,summary,contexts#>) (0.206); 2007\_pe\_garcia (<https://voyant-tools.org/?corpus=b6f0e2c5ee1bc9b644ffda6b86a93740&panels=cirrus,reader,trends,summary,contexts#>) (0.214); 2010\_ar\_fernandez (<https://voyant-tools.org/?corpus=b6f0e2c5ee1bc9b644ffda6b86a93740&panels=cirrus,reader,trends,summary,contexts#>) (0.217)

### 3) Compáralos con la información sobre su extensión, ¿qué notas?

#### Palabras por oración

La forma en que Voyant calcula la longitud de las oraciones debe considerarse muy aproximada, especialmente por lo complicado que es distinguir entre el final de una abreviatura y el de una oración o de otros usos de la puntuación (por ejemplo, en algunos casos un punto y coma marca el límite entre oraciones). El análisis de las oraciones es realizado por una plantilla con instrucciones o 'clase' del lenguaje de programación Java que se llama BreakIterator (<https://docs.oracle.com/javase/tutorial/i18n/text/about.html>).

### Actividad 4

1) Observa las estadísticas de palabras por oración (ppo) y contesta: ¿qué patrón o patrones puedes observar si consideras el índice de "ppo" y los metadatos de país, presidente y año contenidos en el nombre del documento?

2) Da clic sobre los nombre de algunos documentos que te interesen por su índice de "ppo". Dirige tu mirada a la ventana de "Lector" y lee algunas líneas, ¿leer el texto original agrega información nueva a tu lectura de los datos? Comenta por qué.

#### Cirrus y sumario: frecuencias y filtros de palabras vacías

Ya que tenemos una idea de algunas características globales de nuestros documentos, es momento de que empecemos con las características de los términos en nuestro corpus y uno de los puntos de entrada más comunes es entender qué significa analizar un texto a partir de sus frecuencias.

#### Frecuencias sin filtro

El primer aspecto con el que vamos a trabajar es con el de **frecuencia bruta** y para esto utilizaremos la ventana de Cirrus.

### Actividad 5

1) ¿Qué palabras son las más frecuente en el corpus?

2) ¿Qué nos dicen estas palabras del corpus?, ¿son significativas todas?

**Tip** pasa el mouse sobre las palabras para obtener sus frecuencias derecho

#### Palabras vacías

La importancia no es un valor intrínseco y dependerá siempre de nuestros intereses. Justo por eso Voyant ofrece la opción de filtrar ciertas palabras. Un procedimiento común para obtener palabras relevantes es el de filtrar las unidades léxicas gramaticales o *palabras vacías*: artículos,

preposiciones, interjecciones, pronombres, etc. (Peña y Peña, 2015).

## Actividad 6

- 1) ¿Qué palabras vacías están en la nube de palabras?
- 2) ¿Cuáles eliminarías y por qué?

Voyant tiene ya cargada una lista de *stop words* o palabras vacías del español; no obstante, nosotros podemos editarla de la siguiente manera: 1) Colocamos nuestro cursor en el superior derecho de la ventana de Cirrus y damos clic sobre el icono que parece un interruptor.



- 2) Aparecerá una ventana con diferentes opciones, seleccionamos la primera "Editar lista"





4) Una vez que hayamos añadido las palabras que deseamos filtrar damos clic en "salvar" (sic).

**Cuidado:** por defecto está seleccionada una caja que dice “Aplicar a todo”; si ésta se deja seleccionada el filtrado de palabras afectará las métricas de todas las otras herramientas. Es muy importante que documentes tus decisiones. Una buena práctica es guardar la lista de palabras vacías en un archivo de texto (.txt) Para este tutorial hemos creado una [lista de palabras para filtrar \(/assets/analisis-voyant-tools/stopwords-es.txt\)](/assets/analisis-voyant-tools/stopwords-es.txt), y la puedes usar si así lo quieres, sólo recuerda que esto afectará tus resultados. Por ejemplo: en la lista de palabras filtradas incluí “todas” y “todos”, habrá personas para las que estas palabras podrían ser interesantes dado que muestran que “todos” es mucho más utilizado que “todas” y esto podría darnos pistas sobre el uso de lenguaje incluyente.

## Frecuencias con palabras vacías filtradas🔗

Volvamos entonces a esta sección del sumario. Como dijimos en el inicio anterior las palabras filtradas afectan otros campos de Voyant. En este caso, si dejaste seleccionada la caja de “Aplicar a todo”, en la lista que aparece debajo de la leyenda: **Palabra más frecuente en el corpus**, se mostrarán las palabras que se repiten más **sin contar** aquéllas que fueron filtradas. En mi caso, muestra:

[social \(https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#\)](https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#). (437); [nacional \(https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#\)](https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#). (427); [nuestro \(https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#\)](https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#). (393); [inversión \(https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#\)](https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#). (376); [ley \(https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#\)](https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#). (369)

## Actividad 7🔗

1. Reflexiona sobre estas palabras y piensa qué información te proporcionan y cómo se distingue esta información de la que obtienes viendo la nube de palabras.
2. Si estás en un grupo discute las diferencias de tus resultados con los de los demás

## Términos🔗

Si bien las frecuencias pueden decirnos algo sobre nuestros textos, existen muchas variables que pueden hacer que estos números sean poco significativos. En los siguientes apartados se explicarán diferentes estadísticas que pueden obtenerse en la pestaña o solapa de “Términos” que está a la izquierda del botón de “Cirrus” en la disposición default de Voyant.

## Frecuencia normalizada

En el apartado anterior hemos observado la “frecuencia bruta” de las palabras. Sin embargo, si tuviéramos un corpus de seis palabras y otro de 3,000 palabras, las frecuencias brutas son poco informativas. Tres palabras en un corpus de seis palabras representa 50% del total, tres palabras en un corpus de 6,000 representan el 0.1% del total. Para evitar la sobre-representación de un término, los lingüistas han ideado otra medida que se llama: “frecuencia relativa normalizada”. Ésta se calcula de la siguiente manera: Frecuencia Bruta \* 1,000,000 / Número total de palabras. Analicemos un verso como ejemplo. Tomemos la frase: “pero mi corazón dice que no, dice que no”, que tiene ocho palabras en total. Si calculamos su frecuencia bruta y relativa tenemos que:

### **palabra frecuencia bruta frecuencia normalizada**

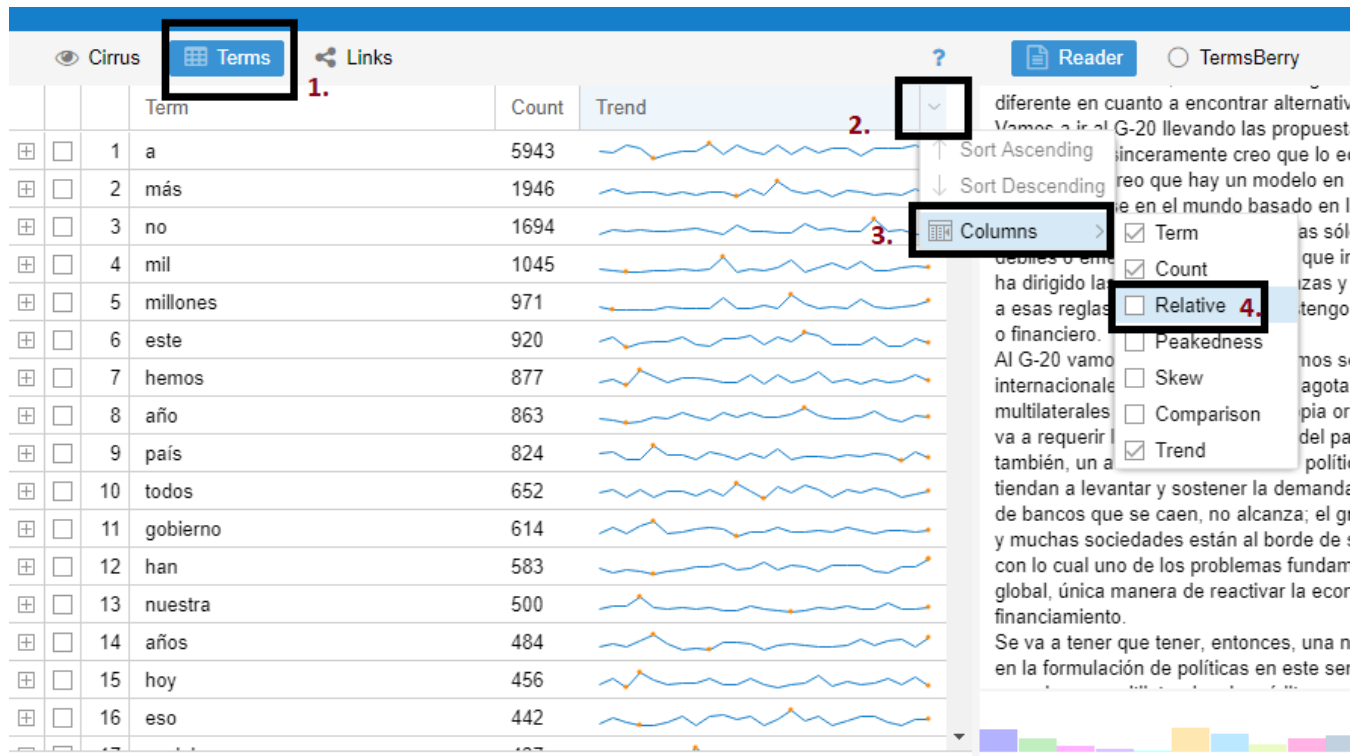
corazón 1  $1 * 1,000,000 / 8 = 125,000$

dice 2  $2 * 1,000,000 / 8 = 111,000$

¿Cuál es la ventaja de esto? Que si tuviéramos un corpus en el que la palabra corazón tuviera la misma proporción, por ejemplo 1,000 ocurrencias entre 8,000 palabras; si bien la frecuencia bruta es muy distinta, la frecuencia normalizada sería la misma, pues  $1,000 * 1,000,000 / 8,000$  también es 125,000.

Veamos cómo funciona esto en Voyant Tools:

1. En la sección de Cirrus (la nube de palabras), damos clic sobre ‘Terms’ o ‘Términos’. Esto abrirá una tabla que por defecto tiene tres columnas: Términos (con la lista de palabras en los documentos, sin las filtradas), Contar (con la ‘frecuencia bruta o neta’ de cada término) y Tendencia (con una gráfica de la distribución de una palabra tomando su frecuencia relativa). Para obtener información sobre la frecuencia relativa de un término, en la barra de los nombres de columna, en el extremo derecho, se da clic sobre el triángulo que ofrece más opciones y en ‘Columnas’ se selecciona la opción ‘Relativo’ como se muestra en la imagen a continuación:



Frecuencia relativa

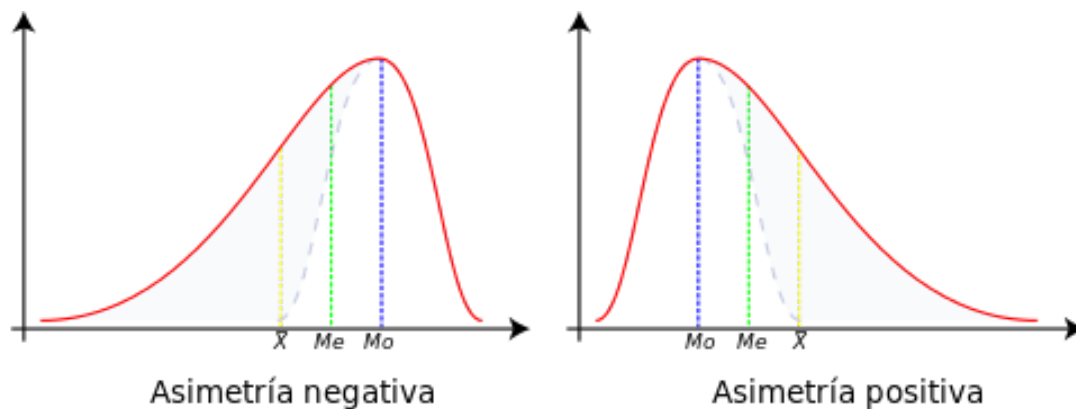
1. Si ordenas las columnas en orden descendiente como lo harías en un programa de hojas de cálculo, observarás que el orden de la frecuencia bruta ('Contar') y la frecuencia relativa ('Relativo') el orden es el mismo. ¿Para qué nos sirve entonces esta medida? Para cuando comparamos diferentes corpus. Un corpus es un conjunto de textos con algo en común. En este caso, Voyant está interpretando todos los discursos como un solo corpus. Si quisiéramos que cada país fuera un corpus distinto, tendríamos que guardar nuestro texto en una tabla, en HTML o en XML, donde los metadatos estuvieran expresados en columnas (en el caso de la tabla) o en etiquetas (en el caso de HTML o XML).<sup>3</sup>

## Asimetría estadística

Aunque la frecuencia relativa no sirve para entender la distribución de nuestro corpus, existe una medida que sí nos da información sobre qué tan constante es un término a lo largo de nuestros documentos: la asimetría estadística.

Esta medida nos da una idea de la distribución de probabilidad de una variable sin tener que hacer su representación gráfica. La forma en que se calcula es observando las desviaciones de una frecuencia con respecto a la media, para obtener si son mayores las que ocurren a la derecha de la media (asimetría negativa) que las de la izquierda (asimetría positiva). Entre más cercano a cero sea el grado de la asimetría estadística, significa que la distribución de ese término es más regular (es decir que ocurre con una media muy similar en todos los documentos). Algo que no es muy intuitivo es que si un término tiene una asimetría estadística con **números positivos** significan que ese término está **por debajo** de la media, y entre más grande el número más asimétrico es el

término (es decir, que ocurre muchísimo en un documento pero que casi no ocurre en el corpus). Los **números negativos**, por el contrario, indican que ese término tiende a estar **por arriba** de la media.



Asimetría estadística

Para obtener esta medida en Voyant, tenemos que repetir los pasos que hicimos para obtener la frecuencia relativa, pero esta vez seleccionar "Oblicuidad" ("Skew"). Esta medida nos permite observar entonces, que la palabra "crisis" por ejemplo, a pesar de tener una alta frecuencia, no sólo no tiene una frecuencia constante a lo largo del corpus, sino que ésta tiende a estar por debajo de la media pues su asimetría estadística es positiva (1.9).

### Palabras diferenciadas

Como tal vez ya sospechas, la información más interesante generalmente no se encuentra dentro de las palabras más frecuentes, pues éstas tienden a ser también las más evidentes. En el campo de la recuperación de la información se han inventado otras medidas que permiten ubicar los términos que hacen que un documento se distinga de otro. Una de las medidas más usadas se llama tf-idf (del inglés *term frequency – inverse document frequency*). Esta medida busca expresar numéricamente qué tan relevante es un documento en una colección determinada; es decir, en una colección de textos sobre "manzanas" la palabra manzana puede ocurrir muchas veces, pero no nos dicen nada nuevo sobre la colección, por lo que no queremos saber la frecuencia bruta de las palabras (*term frequency*, frecuencia de término) pero sopesarla en qué tan única o común es en la colección dada (*inverse document frequency*, frecuencia inversa de documento).

En Voyant el tf-idf se calcula de la siguiente manera

(<https://twitter.com/VoyantTools/status/1025458748574326784>):

Frecuencia Bruta (tf) / Número de Palabras (N) \* log10 ( Número de Documentos / Número de veces que aparece el término en los documentos)

$$tfidf_{t,D} = \left( \frac{tf_{t,d}}{N_i} \right) \cdot \log_{10} \frac{|D|}{|\{d \in D : t \in d\}|}$$

Fórmula de TF-IDF

## Actividad 8

Observa las **palabras diferenciadas (comparado con el resto del corpus)** de cada uno de los documentos y anota qué hipótesis puedes derivar de ellas

1. 2006 ar kircher (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>):  
[uruguay](https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#) (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (12),  
[2004](https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#) (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (13), [2005](https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#)  
<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (31), [plata](https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#)  
<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (7), [inclusión](https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#)  
<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (16).
2. 2006 cl bachelet (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>):  
[innovación](https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#) (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (15),  
[rodrigo](https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#) (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (8),  
[alegremente](https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#) (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (4),  
[barrios](https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#) (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (9), [cobre](https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#)  
<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (10).
3. 2006 co uribe (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>):  
[tutela](https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#) (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (5),  
[reelección](https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#) (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (6),  
[regalías](https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#) (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (7), [iva](https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#)  
<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (6), [publicación](https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#)  
<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (5).
4. 2006 mx fox (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>):  
[atenta](https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#) (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (5), [apego](https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#)  
<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (5), [federalismo](https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#)  
<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (3), [intransigencia](https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#)  
<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (2), [fundamento](https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#)  
<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (3).
5. 2006 pe toledo (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>):  
[entrego](https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#) (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (5), [señor](https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#)  
<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (14), [señora](https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#)  
<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (5), [amigo](https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#)  
<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (5), [tracemos](https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#)  
<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (2).

6. 2007 ar kircher (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>): 2006 (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (65), mercosur (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (12), uruguay (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (9), provincias (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (16), interanual (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (5).
7. 2007 cl bachelet (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>): macrozona (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (7), deudores (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (12), cuna (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (9), subvención (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (10), pesimismo (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (4).
8. 2007 co uribe (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>): guerrilla (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (10), sindicalistas (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (7), paramilitares (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (8), inversionista (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (10), despeje (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (7).
9. 2007 mx calderon (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>): igualar (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (9), transformar (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (19), tortilla (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (4), acuíferos (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (4), miseria (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (10).
10. 2007 pe garcia (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>): huancavelica (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (9), redistribución (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (10), callao (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (8), 407 (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (4), lima (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (7).
11. 2008 ar fernandez (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>): abordar (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (17), capítulo (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (12), presupone (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (5), lesa (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (8), articular (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (5).
12. 2008 cl bachelet (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>): desafío (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (18), mirada (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (10), aprobamos (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (6),

- adulto (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (6), diez (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (11).
13. 2008 co uribe (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>): ecopetrol (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (6), revaluación (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (4), juegos (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (4), desatrasar (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (3), billones (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (6).
  14. 2008 mx calderon (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>): cártel (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (5), noches (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (3), mexicanas (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (6), controlaba (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (3), federales (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (6).
  15. 2008 pe garcia (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>): poblados (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (11), kilómetros (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (52), lima (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (11), carreteras (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (21), mineros (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (4).
  16. 2009 ar fernandez (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>): sosteniendo (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (7), dirigencia (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (5), coparticipación (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (6), catamarca (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (7), pbi (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (9).
  17. 2009 cl bachelet (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>): sello (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (5), fortalecidos (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (5), crisis (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (48), gente (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (24), aplauzo (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (4).
  18. 2009 co uribe (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>): colombia (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (20), calzada (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (6), contributivo (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (5), desplazados (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (6), notificado (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (3).
  19. 2009 mx calderon (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>): federal (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>).



- [corpus=77227f21c006f5ef083d820d77667627#](https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#)). (27), [organizado](https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#) (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (10), [cambiar](https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#) (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (13), [propongo](https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#) (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (8), [policiacos](https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#) (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (4).
20. [2009](https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#) [pe garcia](https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#) (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>): [lima](https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#) (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (11), [1,500](https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#) (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (6), [tingo](https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#) (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (4), [pampas](https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#) (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (4), [desorden](https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#) (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (6).

## Palabras en contexto🔗

El proyecto con el que algunas historias dan por inauguradas las Humanidades Digitales es el *Index Thomisticus*, una concordancia de la obra de Tomás de Aquino liderada por el filólogo y religioso Roberto Busa (Hockey, 2004), en la que participaron decenas de mujeres en la codificación (Terras, 2013). Este proyecto que tomó años en completarse, es una función integrada en Voyant Tools: en la esquina inferior derecha, en la ventana de "Contextos" es posible hacer consultas de las concordancias izquierdas y derechas de términos específicos.

La tabla que vemos tiene las siguientes columnas predeterminadas:

1. **Documento:** aquí aparece el nombre del documento en el que ocurre(n) la(s) palabra(s) clave(s) de la consulta
2. **Izquierda:** contexto izquierdo de la palabra clave (este puede ser modificado para abarcar más palabras o menos y si se da clic sobre la celda, ésta se expande para mostrar más contexto)
3. **Términos:** palabra(s) clave(s) de la consulta
4. **Derecha:** contexto derecho

Se puede añadir la columna **Posición** que indica el lugar en el documento en el que se encuentra el término consultado:

Contextos						Líneas de burbuja		[title]			
Documento	Izquierda	Términos	Derecha	Posición ↑							
11) 2008_ar_fe...	una sociedad desequilibrada, c...	crisis	, de una Argentina volátil, de	101							
16) 2009_ar_fe...	luego, en marco de graves	crisis	, pero en la mayoría de	115							
16) 2009_ar_fe...	mayoría de los casos eran	crisis	provocadas en nuestro propio								
11) 2008_ar_fe...	24 años, antes de la	crisis	, la Argentina había tenido 9								
16) 2009_ar_fe...	vez, los coletazos de alguna	crisis	muy focalizada o localizada que								
1) 2006_ar_ki...	lo que constituyó la peor	crisis	de nuestra historia; vamos esc...								
25) 2010_pe_ga...	Y SUPERAR LA MÁS GRAVE	crisis	ECONÓMICA DE LOS ÚLTIM...								
1) 2006_ar_ki...	esfuerzo debe ser continuado. ...	crisis	recurrentes han obstaculizado ...	242							
19) 2009_mx_ca...	nuestra sociedad. Primero. Vivi...	crisis	económica mundial más grave...	251							

crisis

173 contexto

expandir

Escala

Agregar columna de posición

**Consulta avanzada** Voyant permite el uso de comodines para buscar variaciones de una palabra. Estas son algunas de las combinaciones

- **famili \*** : esta consulta arrojará todas las palabras que empiecen con el prefijo "famili" (familias, familiares, familiar, familia)
- **\* ción**: términos que terminan con el sufijo "ción" (contaminación, militarización, fabricación)
- **pobreza, desigualdad**: puedes buscar más de un término separándolos por comas
- **"contra la pobreza"**: buscar la frase exacta
- **"pobreza extrema" ~ 5**: buscar los términos dentro de las comillas, el orden no importa, y pueden haber hasta 5 palabras de por medio (esa condición regresaría frases como "la extrema desigualdad y la pobreza" donde se encuentra la palabra "pobreza" y "extrema")

## Actividad 9

1. Busca el uso de algún término que te parezca interesante, utiliza alguna de las estrategias de la consulta avanzada
2. Ordena las filas usando las diferentes columnas (Documento, Izquierda, Derecha y Posición): ¿qué conclusiones puedes derivar sobre tus términos utilizando la información de estas columnas?

**Cuidado:** el orden de las palabras en la columna "Izquierda" es inverso; es decir, de derecha a izquierda desde la palabra clave.

## Exportando las tablas🔗

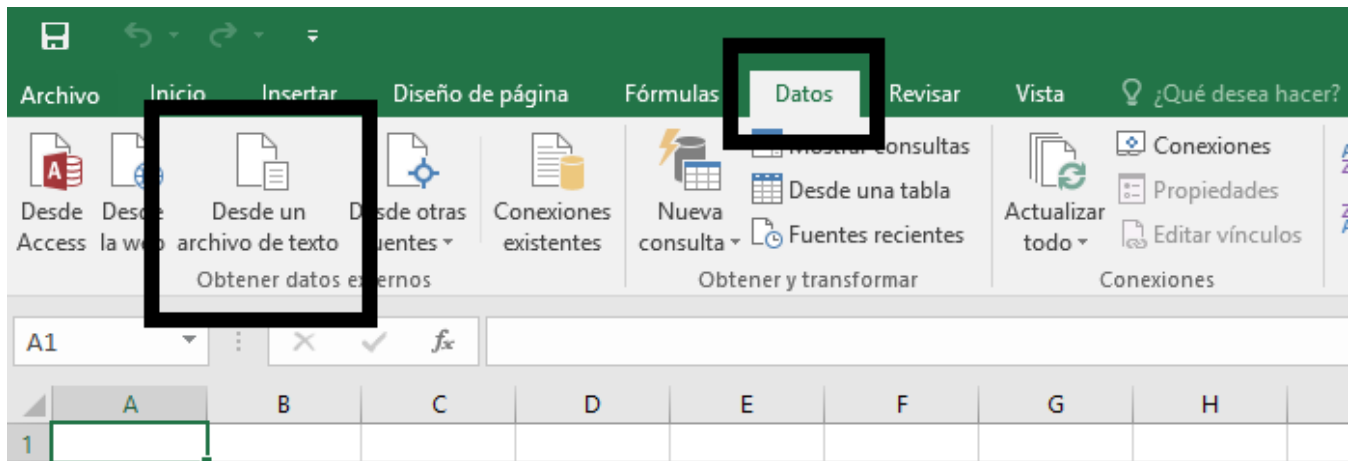
Para exportar los datos se da clic en el cuadro con flecha que aparece cuando pasas el cursor sobre la esquina derecha de "Contextos". En seguida se selecciona la opción "Exportar datos actuales" y se da clic sobre la última opción **Export all available data as tab separated values (text)grid**.

Eso lleva a una página donde están separados los campos por un tabulador:

DocIndex	Position	Left	Middle	Right
0	147	lo que constituyó la peor	crisis	de nuestra historia; vamos escalando
0	242	esfuerzo debe ser continuado. Nuestras	crisis	recurrentes han obstaculizado la permanencia
0	804	desde el fin de la	crisis	, lo que pone de manifiesto
0	1082	duplicó su producción desde la	crisis	y los rubros de transporte
0	1113	desde lo peor de la	crisis	. Además y pese a la
0	1122	a la magnitud de la	crisis	financiera que indujo el colapso
0	1211	desde el fin de la	crisis	. La contribución del consumo al
0	1270	de estas variables durante la	crisis	que condujo al colapso de
0	1309	perspectiva la gravedad de la	crisis	y la notable recuperación subsiguiente
0	1376	modo, luego de que la	crisis	provocara una contracción en el
0	2431	desde el fin de la	crisis	en un 70 y 50
0	3100	e indigencia, que durante la	crisis	habían alcanzado valores inéditos para
0	3369	así la probabilidad de nuevas	crisis	macroeconómicas financieras, que son las
0	3402	de la salida de la	crisis	, que fue haber mantenido un
1	3999	mal administrados que terminaron en	crisis	. Nuestro continente es abundante en
5	557	confirma que, luego de la	crisis	económica y financiera más grave
5	3460	que la Argentina vivía una	crisis	energética. Lo que yo digo
5	3646	cuando hay peligro de una	crisis	energética algunos medios lo titulan
5	5342	a la vulnerabilidad de cualquier	crisis	internacional". Pagamos 10.200, 10.300 millones
5	6007	afronta el pago de la	crisis	, ha ganado autonomía cancelando la
5	8684	tejido social destruido por la	crisis	. En el terreno del desarrollo
5	8722	del Producto Bruto en la	crisis	se necesitan tres del ciclo
5	10469	conflictivo derivado de la profunda	crisis	de la que venimos, el
5	18193	de lo profundo de la	crisis	, como está claro que debemos
5	18209	económica, sin caer en otra	crisis	, obteniendo la sustentabilidad. Argentina necesitaba
5	18278	nos condenaban a las recurrentes	crisis	. Argentina puede crecer y redistribuir
5	18361	de la Argentina de la	crisis	recurrente que estamos dejando atrás
5	18400	trabajo y discutir salarios, las	crisis	facilitan la concentración y el
5	18412	de los poderosos. En las	crisis	, ejecutando fenomenales ajustes, los sectores
5	18440	distribución del ingreso. En esas	crisis	la experiencia ha enseñado que
5	18481	internacional del país. En las	crisis	logran medidas que no pueden
6	1457	manejados y que terminaron en	crisis	. En Chile, sin ir más
8	2350	a volver a vivir otra	crisis	económica en el país. Mi
8	2462	estas turbulencias hubieran generado una	crisis	económica. Las reservas del Banco
10	101	una sociedad desequilibrada, con fuertes	crisis	, de una Argentina volátil, de
10	131	24 años, antes de la	crisis	, la Argentina había tenido 9
10	1041	un fuerte escudo contra las	crisis	internacionales que en otras oportunidades
11	1209	se ha transformado en una	crisis	financiera global. Las consecuencias de
11	1859	sólo los efectos de la	crisis	internacional, sino del otro gran
11	6916	bajas emisiones. Hablo de la	crisis	de los alimentos, que está
12	2663	a aquellas de vísperas de	crisis	. Que hay esfuerzos distintos al
12	5142	la clave para resistir la	crisis	de la economía global que
14	320	de millones. Ahora es la	crisis	mundial, es el alza brutal
14	1049	corregirse. Y, sin embargo, con	crisis	externa y con defectos al
14	1354	contra un agravamiento de la	crisis	mundial, porque podría ser que
14	1968	Y, así, cuando pase la	crisis	mundial de los precios, tendremos
15	115	luego, en marco de graves	crisis	, pero en la mayoría de
15	124	mayoría de los casos eran	crisis	provocadas en nuestro propio país
15	137	vez, los coletazos de alguna	crisis	muy focalizada o localizada que
15	419	fueron los causantes de esta	crisis	puedan tener la capacidad intelectual
15	512	que hay un modelo en	crisis	que tiene que ver también
15	900	que no han provocado la	crisis	, emergen y se trasladan hacia
15	999	modelos nacionales si realmente	la	crisis se prolonga en el tiempo
15	1032	de nuevo e inédito esta	crisis	: que emergiendo de los países
15	1073	encuentra a los argentinos esta	crisis	sin precedentes a escala global
15	1080	precedentes a escala global? Esta	crisis	nos encuentra en nuestro sexto
15	1961	total y fundamental de la	crisis	de lo que se denominaba
15	3044	preguntarme: ¿qué pasaría si esta	crisis	a nivel mundial hubiera encontrado
15	3092	pagando los costos de esta	crisis	en otro momento? Creo que
15	3227	hay ajustes y que hay	crisis	, ha podido decirle a sus
15	4663	en que nos toma la	crisis	, nos debe dar la necesidad
15	5886	actitud diferente frente a esta	crisis	inédita a nivel mundial, que
15	5952	años que, de prolongarse la	crisis	tal cual como se preanuncia
15	7154	mercado interno frente a la	crisis	del sector externo, adquiere también
15	7299	mira y advierte que la	crisis	no está en que no
15	7316	o cual receta económica, la	crisis	es como siempre han sido
15	7324	siempre han sido las grandes	crisis	que marcaron los cambios en
15	7332	los cambios en la humanidad,	crisis	de las ideas. Estamos ante
15	7339	las ideas. Estamos ante la	crisis	de un sistema de ideas
16	409	Somos el país azotado por	crisis	económicas que se han ensañado
16	409	la crisis económica de cada	crisis	que destruye las fundaciones de

## Exportar contextos

Selecciona todos los datos (Ctrl+A o Ctrl+E); copíalos (Ctrl+C) y pégalos en una hoja de cálculo (Ctrl+V). Si esto no funciona, guarda los datos como en un editor sencillo de texto como .txt (¡no olvides la codificación UTF-8!) y luego en tu hoja de cálculo importa los datos. En Excel esto se hace en la pestaña de "Datos" y después "Desde un archivo de texto"



Importar datos desde un archivo de textos

## Respuestas a las actividades

### Actividad 1

Este corpus tiene 2 documentos con un total de palabras de 4 y 3 palabras únicas (*tengo, hambre, sueño*)

### Actividad 2

1) Podríamos observar, por ejemplo, que los textos más largos son de dos países: Chile y Argentina, y de tres presidentes distintos: Kirchner, Bachelet y Pinera. Sobre los más cortos podríamos ver que si bien el más corto es de Perú, en realidad los que más aparecen entre los breves son los de México y Colombia.

2) Saber la extensión de nuestros textos nos permite entender la homogeneidad o disparidad de nuestro corpus, así como entender ciertas tendencias (por ejemplo, en qué años tendían a ser más cortos los discursos, en qué momento cambió la extensión, etc.)

### Actividad 3

1) La primera estrofa tiene 23 palabras y 20 son palabras únicas, por lo que  $20/23$  da igual a una densidad de vocabulario de 0.870; en realidad de 0.869 pero Voyant Tools redondea estos números: <https://voyant-tools.org/?corpus=b6b17408eb605cb1477756ce412de78e>. La segunda estrofa tiene 24 palabras y 20 son palabras únicas, por lo que  $20/24$  da igual a una densidad de vocabulario de 0.833: <https://voyant-tools.org/?corpus=366630ce91f54ed3577a0873d601d714>.

Como podemos observar la diferencia entre un verso de Sor Juana Inés de la Cruz y otro compuesto por Érika Ender, Daddy Yankee y Luis Fonsi tienen una diferencia de densidad de 0.037, que no es muy alto. Debemos tener cuidado al interpretar estos números pues sólo son un indicador cuantitativo de la riqueza del vocabulario y no incluye parámetros como la complejidad de la rima o de los términos.

Parece haber una correspondencia entre los discursos más cortos y los más densos, esto es natural pues entre más breve es un texto menos "oportunidad" hay para repetirse. No obstante, esto también podría decirnos algo sobre los estilos de diferentes países o presidentes. Entre menos densidad es más probable que recurran a más recursos retóricos.

## Actividad 4

Estos resultados parecen indicar que la presidenta Kirchner, además de tener los discursos más largos es la que hace frases más largas; sin embargo tenemos que tener cuidado con las conclusiones de este tipo pues se trata de discursos orales en los que la puntuación depende de quien transcribe el texto.

## Actividad 5

1. a (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (5943); más (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (1946); no (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (1694); mil (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (1045); millones (<https://voyant-tools.org/?corpus=77227f21c006f5ef083d820d77667627#>). (971)
2. La primera palabra es una preposición, la segunda un adverbio de comparación y la tercera un adverbio de negación. Estas palabras podrían ser significativas si lo que se busca comprender es el uso de este tipo de palabras funcionales. Sin embargo, si lo que se busca son más bien sustantivos, habrá que hacer un filtrado (ver sección: "Palabras más frecuentes")

## Bibliografía

Hockey, Susan. 2004 "The History of Humanities Computing". *A Companion to Digital Humanities*. Schreibman et al. (editores). Blackwell Publishing Ltd. [doi:10.1002/9780470999875.ch1](https://doi.org/10.1002/9780470999875.ch1) ([doi:10.1002/9780470999875.ch1](https://doi.org/10.1002/9780470999875.ch1)).

Peña, Gilberto Anguiano, y Catalina Naumis Peña. 2015. «Extracción de candidatos a términos de un corpus de la lengua general». *Investigación Bibliotecológica: Archivonomía, Bibliotecología e Información* 29 (67): 19-45. <https://doi.org/10.1016/j.ibbai.2016.02.035> (<https://doi.org/10.1016/j.ibbai.2016.02.035>).

Sinclair, Stéfan and Geoffrey Rockwell, 2016. *Voyant Tools*. Web. <http://voyant-tools.org/> (<http://voyant-tools.org/>).

Terras, Melissa, 2013. "For Ada Lovelace Day – Father Busa's Female Punch Card Operatives". *Melissa Terras' Blog*. Web. <http://melissaterras.blogspot.com/2013/10/for-ada-lovelace-day-father-busas.html> (<https://melissaterras.blogspot.com/2013/10/for-ada-lovelace-day-father-busas.html>).

Este tutorial fue escrito gracias al apoyo de la Academia Británica y preparado durante el Taller de escritura de The Programming Historian en la Universidad de los Andes en Bogotá, Colombia, el del 31 de julio al 3 de agosto de 2018.

## Notas al pie

1. Los textos de Perú fueron recopilados por a [Pamela Sertzen](https://twitter.com/madvivacious) (<https://twitter.com/madvivacious>). ↵
2. Existen formas más complejas para cargar el corpus que [puedes consultar en la documentación en inglés](https://voyant-tools.org/docs/#!/guide/corpuscreator) (<https://voyant-tools.org/docs/#!/guide/corpuscreator>). ↵
3. Para más información, consulta la documentación en inglés. ↵

## Acerca del autor

Silvia Gutiérrez De la Torre es la Bibliotecaria de Humanidades Digitales de El Colegio de México y co-fundadora de RLadiesCDMX (México).  (<https://orcid.org/0000-0001-8717-2291>).

## Cita sugerida

Silvia Gutiérrez De la Torre, "Análisis de corpus con Voyant Tools", *The Programming Historian en español* 3 (2019), <https://doi.org/10.46430/phes0043>.

*The Programming Historian en español* (ISSN: 2517-5769) se publica con una licencia [CC-BY](https://creativecommons.org/licenses/by/4.0/deed.es) (<https://creativecommons.org/licenses/by/4.0/deed.es>).


Este proyecto es administrado por ProgHist Limited, con número de compañía [12192946](https://beta.companieshouse.gov.uk/company/12192946) (<https://beta.companieshouse.gov.uk/company/12192946>).

[ISSN 2397-2068 \(inglés\) \(/\)](#)


[ISSN 2517-5769 \(español\) \(/es\)](#)


[ISSN 2631-9462 \(francés\) \(/fr\)](#)

 [Alojado en GitHub](https://github.com/programminghistorian/jekyll) (<https://github.com/programminghistorian/jekyll>).

 [Última actualización el 15 June 2020](#)

(<https://github.com/programminghistorian/jekyll/commits/gh-pages>).

 [Suscripción a RSS](https://programminghistorian.org/feed.xml) (<https://programminghistorian.org/feed.xml>).

 [Versiones anteriores](https://github.com/programminghistorian/jekyll/commits/gh-pages/es/lecciones/analisis-voyant-tools.md) (<https://github.com/programminghistorian/jekyll/commits/gh-pages/es/lecciones/analisis-voyant-tools.md>).

 [Envíanos tus comentarios \(/es/retroalimentacion\)](#)