

# Visualización de Información

## Texto

Daniela Opitz

[dopitz@udd.cl](mailto:dopitz@udd.cl)

Data Science Institute, Universidad del Desarrollo  
Edición 2023

# Texto:

## Datos No Estructurados

A diferencia de los data sets que hemos visto hasta ahora, el texto no tiene una estructura clara. Sin embargo, posiblemente gran parte de la información disponible **es texto**.

¿Letras, palabras, frases, párrafos, documentos?

¿Temas? ¿Semántica del contenido?

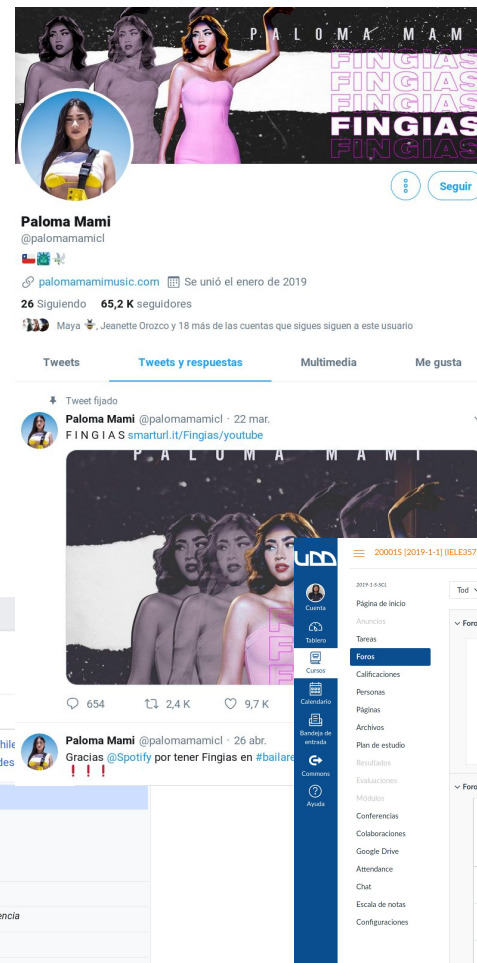
¿Gramática?

¿Analizaremos un documento o varios documentos?

¿Un conjunto de documentos (corpus) o varios? ¿Distintos idiomas?



WIKIPEDIA	
Universidad del Desarrollo	
universidad privada chilena	
✎	
La <b>Universidad del Desarrollo</b> es una <b>universidad</b> privada autónoma en Chile <b>Concepción</b> y en <b>Santiago</b> , específicamente en las comunas de <b>Las Condes</b>	
Universidad del Desarrollo	
 Universidad del Desarrollo Universidad de Excelencia	
Sigla	UDD
Lema	Universidad de Excelencia
Tipo	Privada
Fundación	1990



# ¿Para qué visualizar texto?

**Entender** *lo que contiene un documento o conjunto de documentos (corpus).*

**Agrupar** *documentos distintos dentro de una misma categoría de acuerdo a su similitud.*

**Comparar** *y medir qué diferencia un texto o colección de documentos de otro(a).*

**Medir la evolución** *en el tiempo de un texto de una colección de documentos.*

**Correlacionar** *patrones en el texto con los de otros data sets, por ej., con los de una red social.*

# ¿Qué es lo que se visualiza?

No siempre se visualiza el texto directamente. Usualmente se utiliza un **modelo de lenguaje**:

- Frecuencia de términos (*tokens*), usualmente palabras
- Secuencias de palabras, bolsas de palabras (*bag of words*)
- Componentes latentes (*latent semantic analysis*), modelamiento de tópicos (*topic modeling*), etc.

Frecuencias, Secuencias, Estructura Gramatical

## Word Cloud: Frecuencias

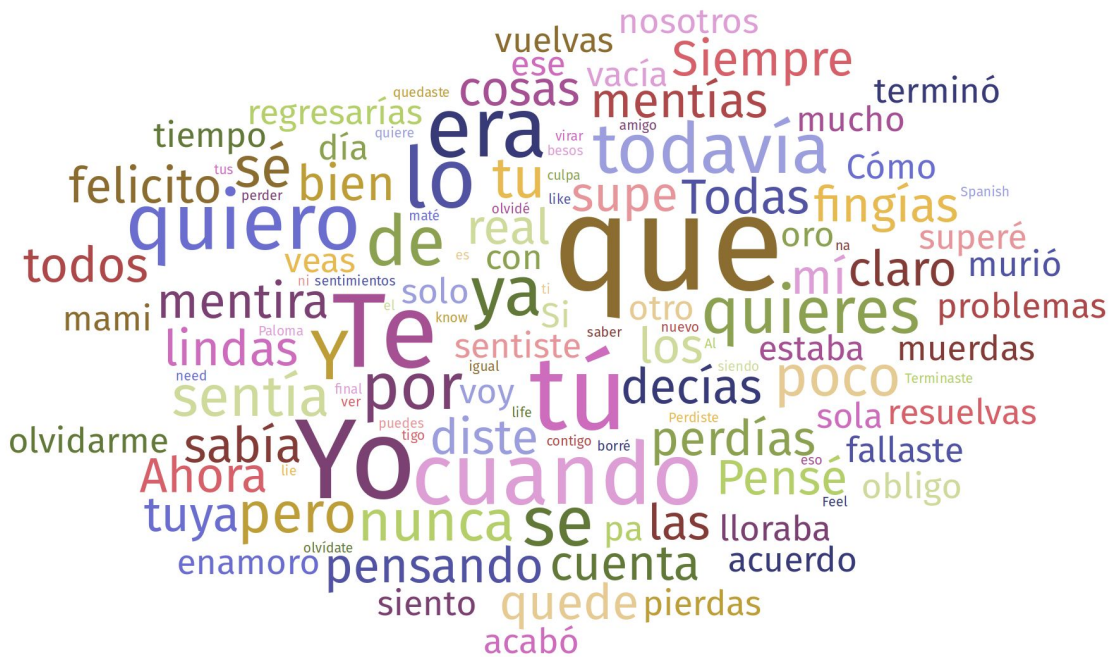
Cada palabra en el texto o corpus se grafica con un tamaño proporcional a su frecuencia (cantidad de apariciones).

Quizás uno de los tipos de visualización más popular. También uno de los más ineficientes en función de los principios de diseño: el canal de área utilizado para graficar la frecuencia dificulta comparaciones, tanto por percepción como al largo de las palabras.

Problema: palabras más frecuentes no son informativas.

Hagan las suyas en

<https://www.jasondavies.com/wordcloud/>



Documento: Fingías de Paloma Mami

## At the National Conventions, the Words They Used

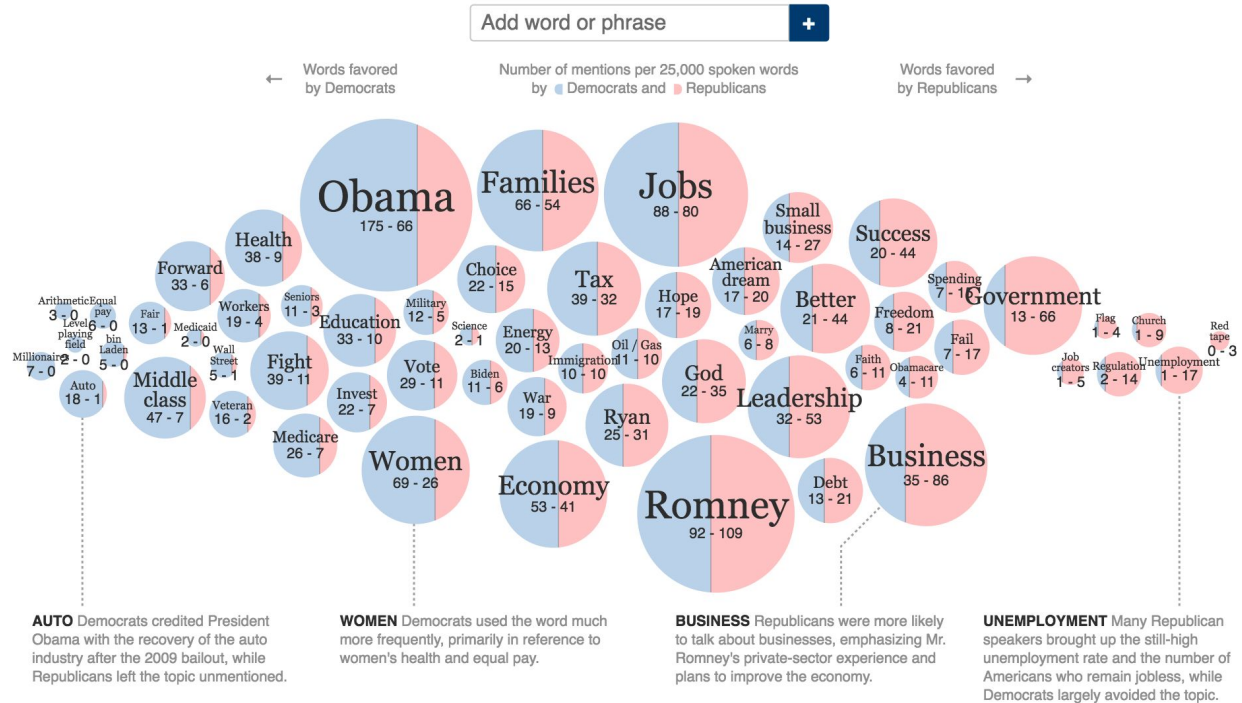
A comparison of how often speakers at the two presidential nominating conventions used different words and phrases, based on an analysis of transcripts from the Federal News Service.

### Bubble Clouds

Alternativa a las word clouds. Resuelve algunas de sus limitaciones, y permite visualizar otros atributos del dataset.

Estos gráficos son agradables estéticamente y son fáciles de entender.

Es posible asignarle un significado a la posición de cada burbuja. En el ejemplo, el eje x codifica la asociación de cada palabra con los partidos políticos en los Estados Unidos.



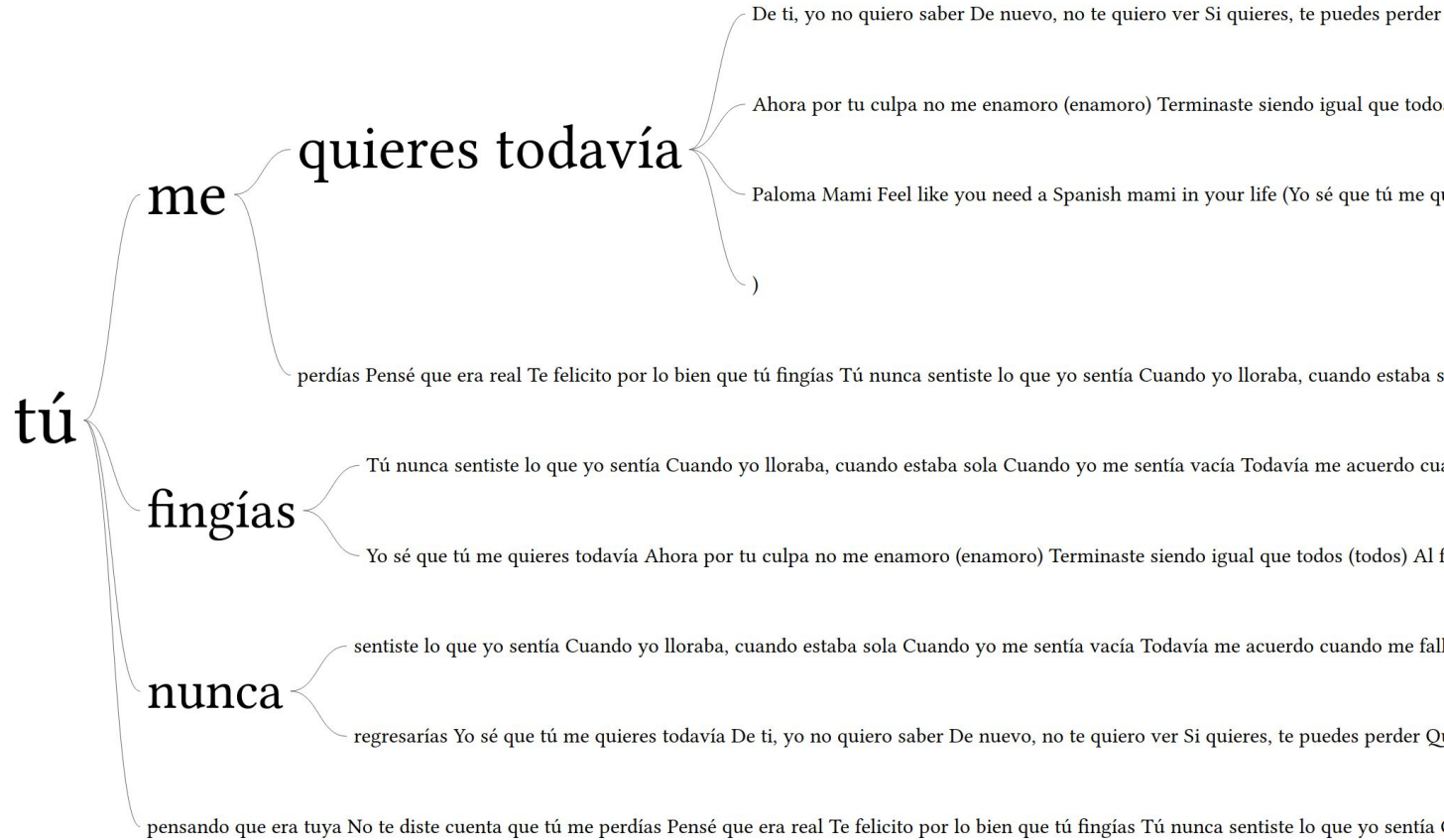
# Word Tree: Secuencias

Esta técnica visualiza la estructura secuencial en el texto (misma canción de Paloma Mami), creando árboles de texto, donde los nodos son palabras o secuencias de palabras.

Hagan los suyos en:

<https://www.jasondavies.com/wordtree>

Wattenberg, M., & Viégas, F. B. (2008). The word tree, an interactive visual concordance. *IEEE transactions on visualization and computer graphics*, 14(6), 1221-1228.

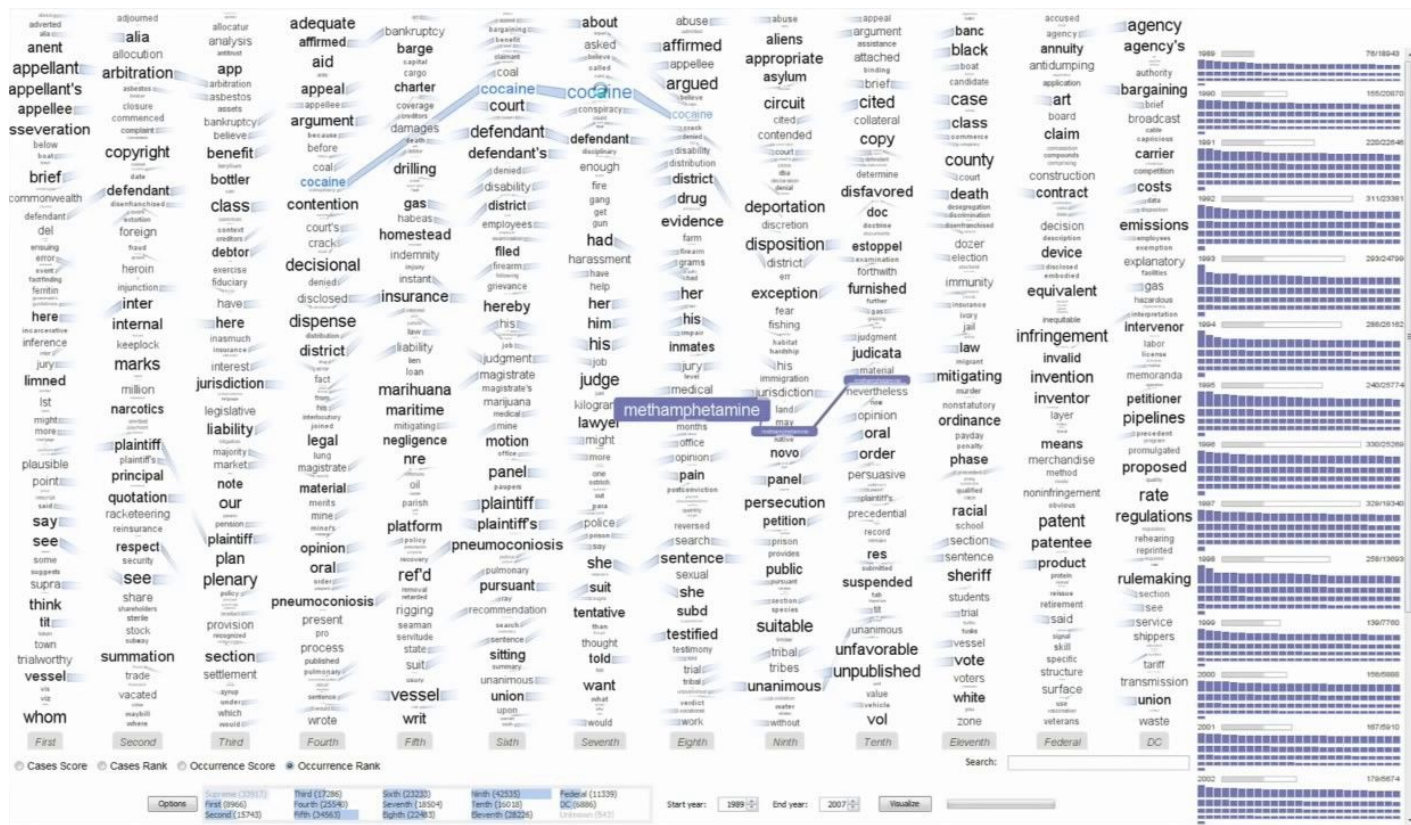




# Parallel Tag Clouds

Pueden existir distintas facetas en un corpus, y la distribución del texto puede ser distinta en cada una de ellas. Facetas incluyen temáticas, tiempo de publicación, entre otras.

Esta visualización muestra para cada faceta la distribución de la frecuencia o relevancia de palabras, y al mismo tiempo, cómo esa relevancia varía a lo largo de las facetas.



Collins, C., Viegas, F. B., & Wattenberg, M. (2009, October). Parallel tag clouds to explore and analyze faceted text corpora. In *2009 IEEE Symposium on Visual Analytics Science and Technology* (pp. 91-98). IEEE.

Temáticas (*topics*) y Reducción Dimensional

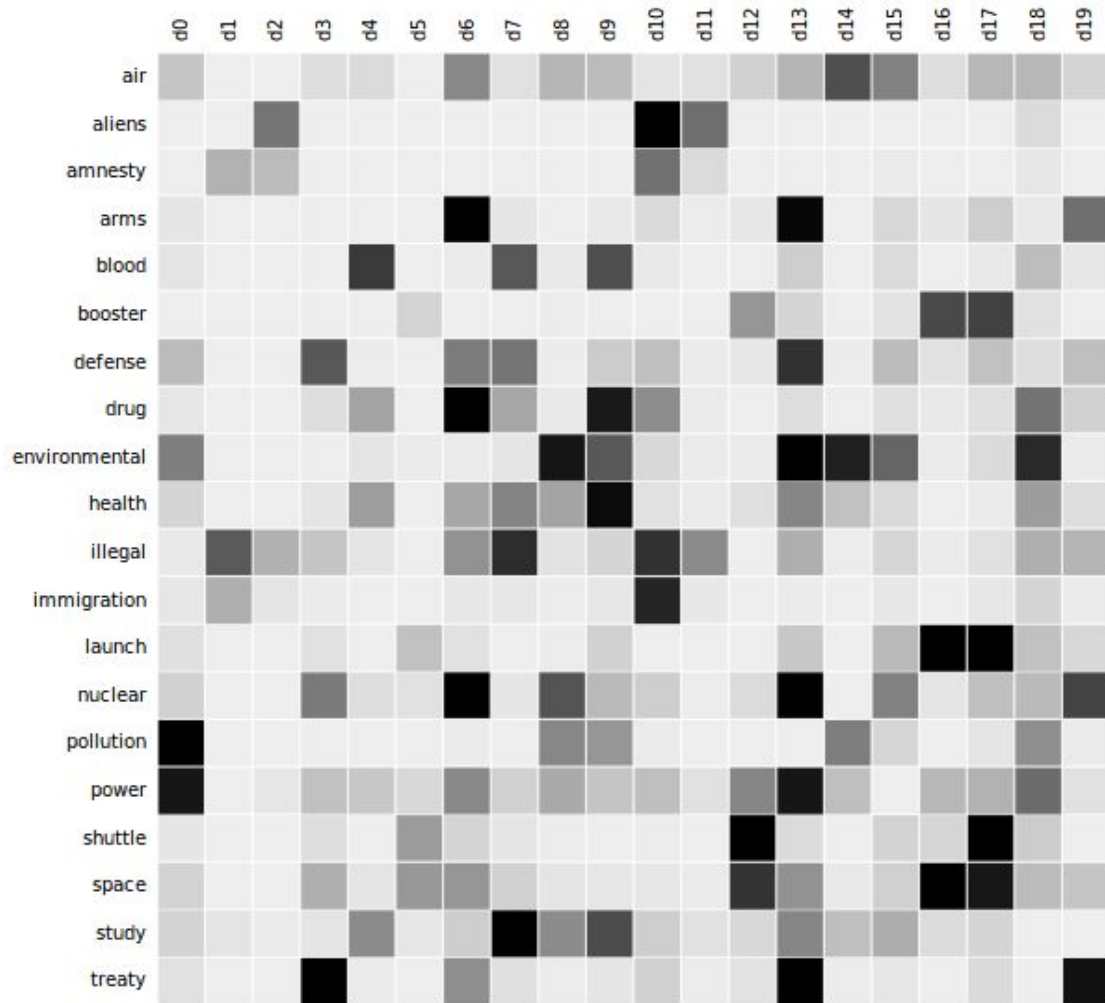
# Estructura de Datos:

## Document-Term Matrices (DTM)

¿Cómo representar una colección de documentos? Una manera de hacerlo es a través de un enfoque conocido como **bag of words** (*bolsa de palabras*). En este esquema, una colección se puede expresar como una **matriz**: en la imagen, cada columna es un documento, y cada fila es una palabra. *Cada celda contiene la cantidad de veces que aparece una palabra en un documento.*

Ventaja: al ser una matriz, podemos utilizar métodos de álgebra lineal y cálculo para trabajar con el texto.

Desventaja: esta representación pierde información, no se sabe el orden en el que aparecieron las palabras en el documento.



# Topic Modeling

¿Cuántas palabras puede tener una colección de documentos? ¿Basta describir relaciones entre documentos y palabras para entender un corpus? **Topic Modeling** busca encontrar cuáles son los *temas* o *tópicos* o *dimensiones latentes* en los documentos de un corpus.

Existen técnicas como *Latent Semantic Analysis*, *Latent Dirichlet Allocation*, y *Non-Negative Matrix Factorization*, entre otras. Todas funcionan así: **se define (o elige) un número  $k$  de dimensiones, y se encuentra una representación matricial de documentos asociados a temas, y palabras asociadas a temas.**

(otra interpretación es una **reducción dimensional**)

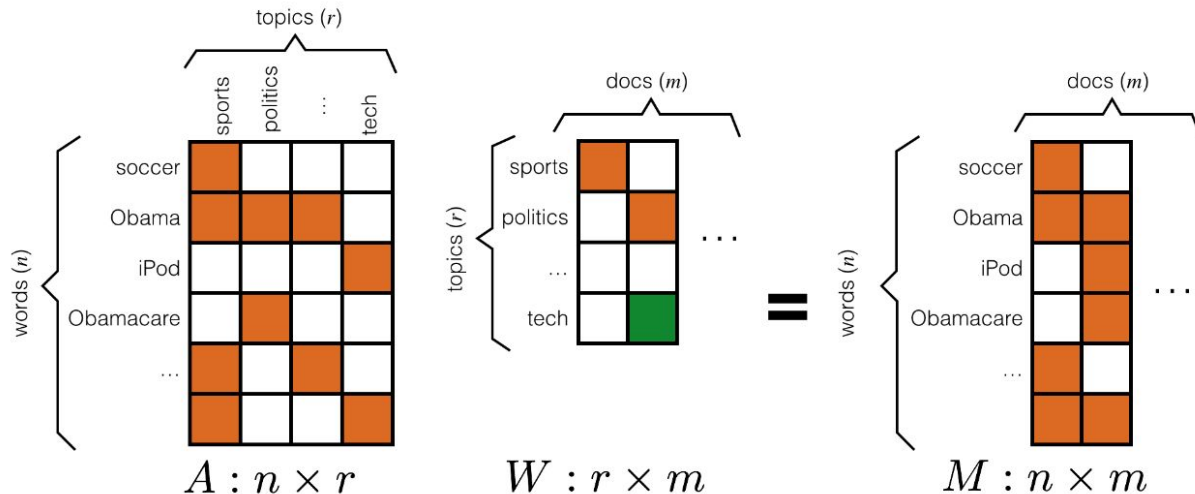
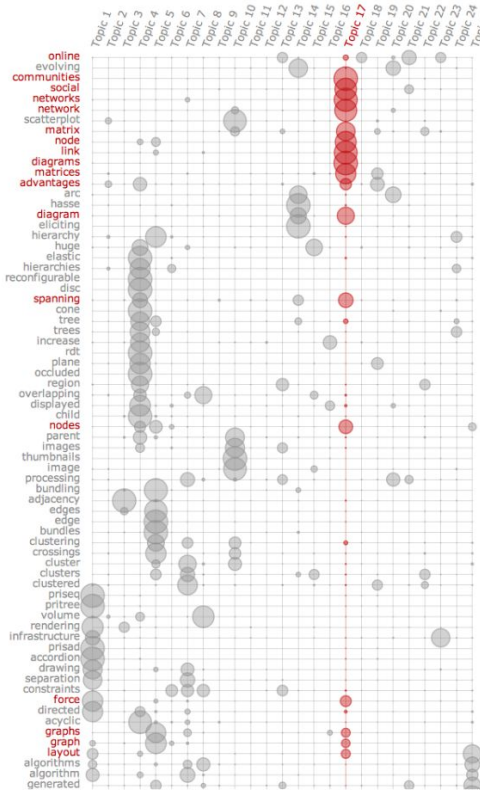


Imagen por Catalin Voss, Stanford.

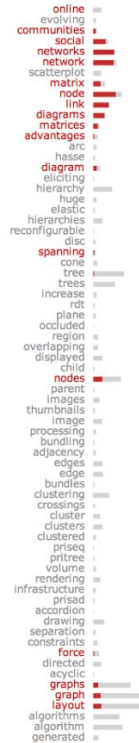
# Termite

Termite es otra aplicación para explorar y evaluar los resultados de topic modeling.

En Termine, el foco está en la matriz de vocabulario y tópicos. La visualización permite reordenar las filas y columnas de la matriz, ver la distribución de palabras, y también explorar los documentos.



Word Frequency



Representative Documents

A Comparison of the Readability of Graphs Using Node-Link and Matrix-Based Representations Mohammad Ghoniem Jean-Daniel Fekete Philippe Castagliola
Using Multilevel Call Matrices in Large Software Projects Frank van Ham
Improving the Readability of Clustered Social Networks using Node Duplication Nathalie Henry Anastasia Bezerianos Jean-Daniel Fekete
MatrixExplorer: a Dual-Representation System to Explore Social Networks Nathalie Henry Jean-Daniel Fekete
<b>NodeTrix: a Hybrid Visualization of Social Networks</b> Nathalie Henry Jean-Daniel Fekete Michael J. McGuffin The need to visualize large social networks is growing as hardware capabilities make analyzing large networks feasible and many new data sets become available. Unfortunately, the visualizations in existing systems do not satisfactorily resolve the basic dilemma of being readable both for the global structure of the network and also for detailed analysis of local communities. To address this problem, we present NodeTrix, a hybrid representation for networks that combines the advantages of two traditional representations: node-link diagrams are used to show the global structure of a network, while arbitrary portions of the network can be shown as adjacency matrices to better support the analysis of communities. A key contribution is a set of interaction techniques. These allow analysts to create a NodeTrix visualization by dragging selections to and from node-link and matrix forms, and to flexibly manipulate the NodeTrix representation to explore the dataset and create meaningful summary visualizations of their findings. Finally, we present a case study applying NodeTrix to the analysis of the InfoVis 2004 coauthorship dataset to illustrate the capabilities of NodeTrix as both an exploration tool and an effective means of communicating results.
Visualizing Causal Semantics using Animations Nivedita R. Kadaba Pourang P. Irani Jason Leboe
<b>Balancing Systematic and Flexible Exploration of Social Networks</b> Adam Perer Ben Shneiderman Social network analysis (SNA) has emerged as a powerful method for understanding the importance of relationships in networks. However, interactive exploration of networks is currently challenging because: (1) it is difficult to find patterns and comprehend the structure of networks with many nodes and links, and (2) current systems are often a medley of statistical methods and overwhelming visual output which leaves many analysts uncertain about how to explore in an orderly manner. This results in exploration that is largely opportunistic. Our contributions are techniques to help structural analysts understand social networks more effectively. We present SocialAction, a system that uses attribute ranking and coordinated views to help users systematically examine numerous SNA measures. Users can (1) flexibly iterate through visualizations of measures to gain an overview, filter nodes, and find outliers; (2) aggregate networks using link structure, find cohesive subgroups, and focus on communities of interest; and (3) untangle networks by viewing different link types separately, or find patterns across different link types using a matrix overview. For each operation, a stable node layout is maintained in the network visualization so users can make comparisons. SocialAction offers analysts a strategy beyond opportunism, as it provides systematic, yet flexible, techniques for exploring social networks.
Causality Visualization Using Animated Growing Polygons Niklas Elmqvist Philippos Tsigas
SpicyNodes: Radial Layout Authoring for the General Public Michael Douma Grzegorz Ligierko Ovidiu Ancuta Pavel Gritsai Sean Liu

Chuang, J., Manning, C. D., & Heer, J. (2012, May). Termite: Visualization techniques for assessing textual topic models. In *Proceedings of the international working conference on advanced visual interfaces* (pp. 74-77). ACM.

<http://users.cecs.anu.edu.au/~lan.Wood/termite/50-run1-LIWCvocab/public.html/>

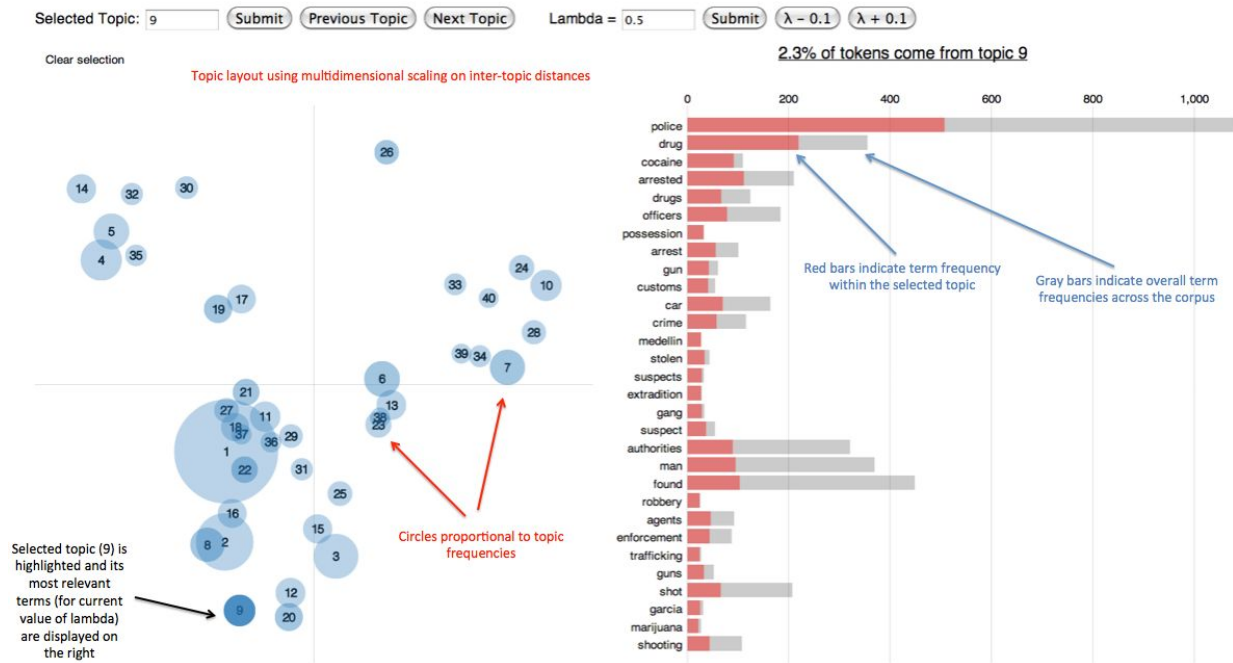
# LDA Vis

Ahora bien, los *tópicos* no siempre son interpretables. En ocasiones tienen coherencia algebraica, pero no tienen significado para nosotros.

El software LDA Vis utiliza una visualización interactiva compuesta para que podamos explorar el espacio de *topics* de un corpus.

A la derecha muestra cada tópico como una burbuja, con una posición calculada utilizando reducción dimensional.

A la izquierda muestra para un tópico específico su distribución de palabras más relevantes (y la comparación con la asociación global de esas palabras).



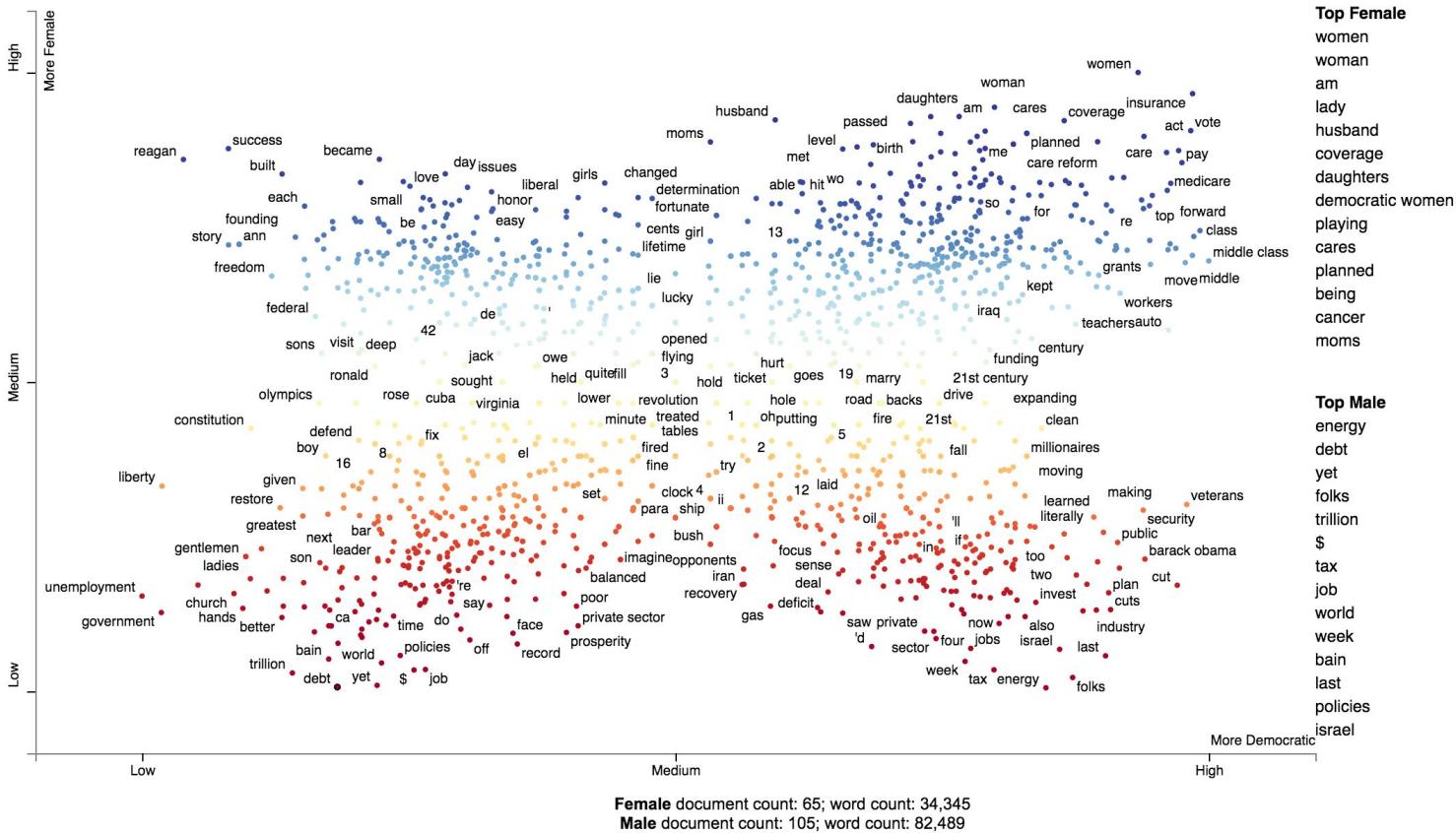
Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces* (pp. 63-70).



# ScatterText

ScatterText es una visualización que pone los términos más **relevantes** de una colección (¡no necesariamente los más frecuentes!). Esa relevancia se puede calcular sobre la matriz DT.

En el ejemplo, el eje Y codifica la asociación en el vocabulario de un corpus de política hacia hombres y mujeres, y el eje x hacia republicanos y demócratas en los EEUU.



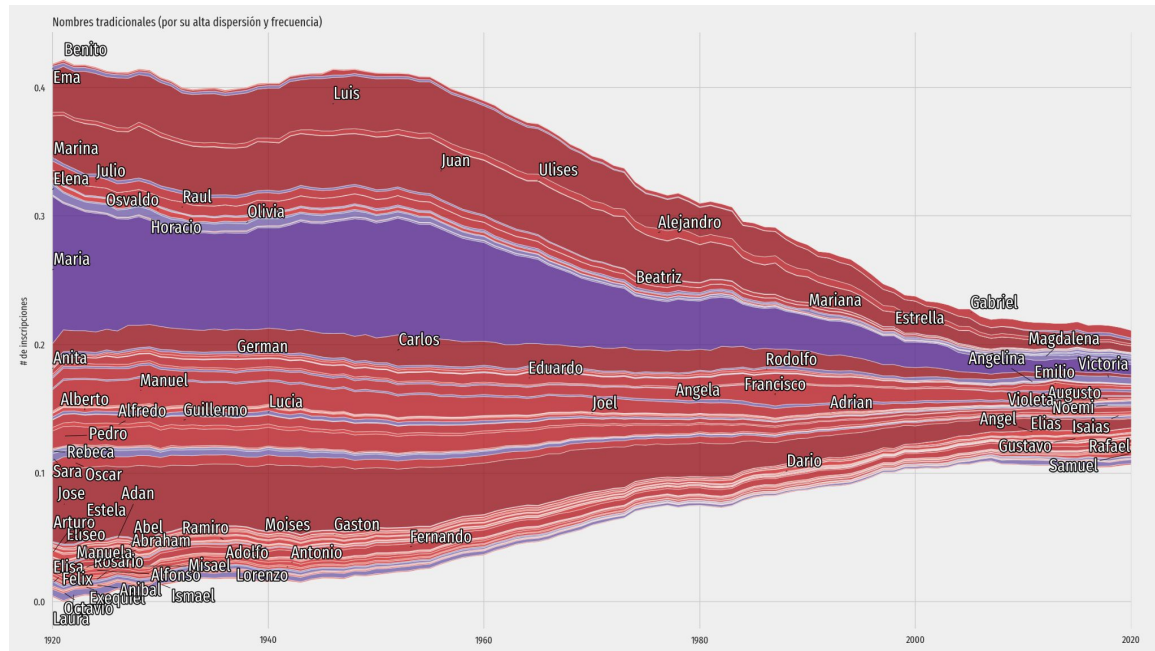
Kessler, J. (2017). Scattertext: a Browser-Based Tool for Visualizing how Corpora Differ. *Proceedings of ACL 2017, System Demonstrations*, 85-90.  
Código: <https://github.com/jasonkessler/scattertext>

# StreamGraphs

Generalización de visualizaciones apiladas. Se enfatiza la continuidad horizontal versus ítemes verticales.

Datos:

- llave categórica (por ej., artista, película, nombre, etc.),
- 1 atributo ordinal o cuantitativo (por ej., número de apariciones)





# Reducción Dimensional con UMAP

La DTM (document-term matrix) puede ser reducida dimensionalmente. Así, sobre la reducción dimensional se pueden ejecutar algoritmos de clustering, que permiten encontrar grupos de documentos similares entre sí.

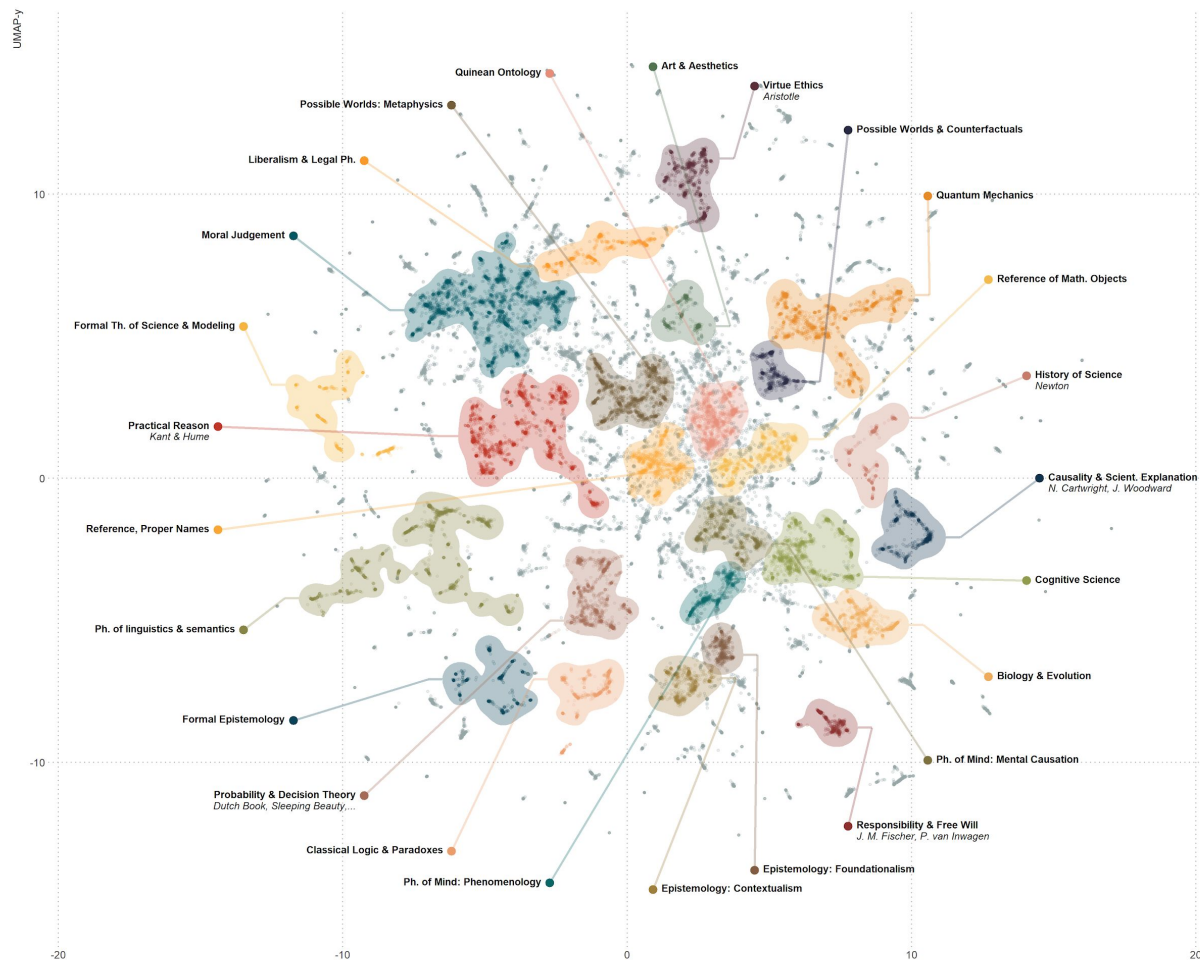
Esta visualización muestra un corpus de documentos filosóficos, para mostrar “la estructura de la Filosofía reciente”. Para ello, primero aplica UMAP sobre la DTM; luego el clustering HDBSCAN; y finalmente un algoritmo de posicionamiento de etiquetas.

Fuente (incluye código en R):

<https://homepage.univie.ac.at/noichlm94/poets/structure-of-recent-philosophy-ii/>

## The structure of recent Philosophy

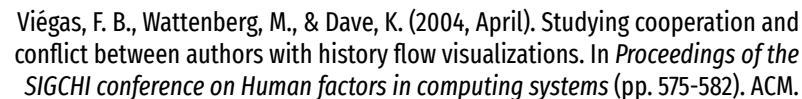
A umap & hdbscan-cluster-analysis of ~ 50000 papers in philosophy that brings out the major groupings of the discipline.



# Evolución y Cambios

Además de las facetas del texto (como la fecha de publicación), existen otros atributos relevantes. Por ejemplo, el historial de modificación de un documento - particularmente si es escrito por más de una persona.

History Flow permite ver el proceso de escritura y construcción de los artículos de Wikipedia, tanto a nivel temporal (cuándo se agregó o quitó contenido) como autoral (quién agregó o quitó contenido).

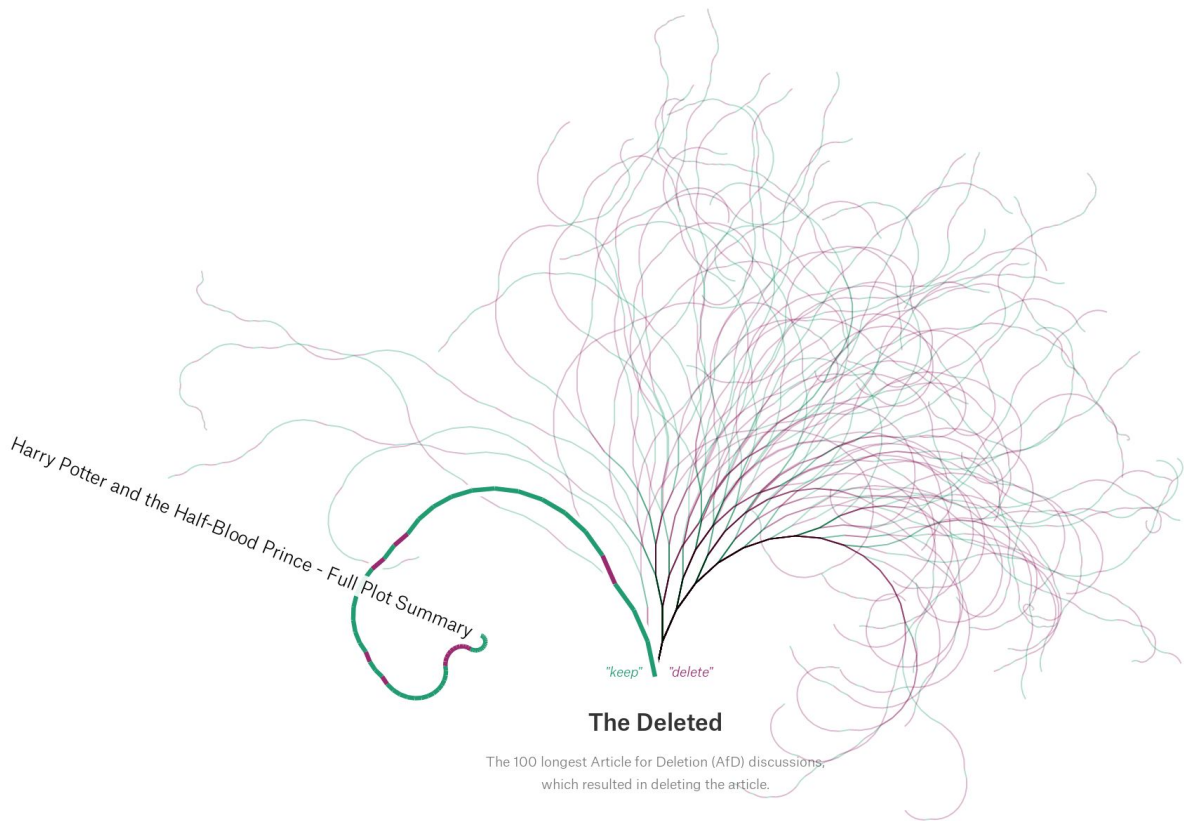


# Notabilia

Notabilia es una visualización orgánica que muestra los procesos de edición en Wikipedia desde el conflicto: se focaliza en los artículos que son marcados para eliminación.

Cada artículo es una línea cuya trayectoria se forma a medida que la discusión decide si el artículo se elimina o se mantiene en la enciclopedia.

Si el artículo recibe votos de mantención, se suma un segmento verde, hacia la izquierda. Si recibe votos de eliminación, se suma un segmento rojo, hacia la derecha. Así, el consenso (o la falta de éste) aparece en la visualización.



Sistema: <http://notabilia.net/> Por [Moritz Stefaner](#) • [Dario Taraborelli](#) • [Giovanni Luca Ciampaglia](#)

# ¿Preguntas?

Hoy nos enfocamos en la visualización de texto en un sentido tradicional.

Dos recursos relevantes en este aspecto son los siguientes:

- [Text Analysis with Visualization](#), capítulo del libro *Search User Interfaces* de Marti Hearst.
- [Text Visualization Browser](#), una colección de vínculos e imágenes de visualización de texto.

Sin embargo, existen otras áreas en las cuales también se usa texto. Una de ellas es la secuenciación genética.

**¿Qué texto (o colección) te gustaría visualizar? ¿Para qué?**