Visualización de Información Texto

Daniela Opitz
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**.

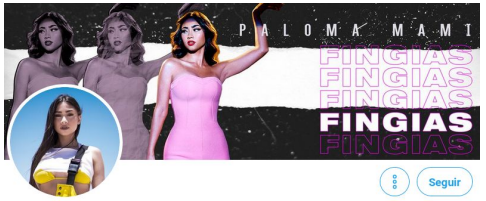
¿Letras, palabras, frases, párrafos, documentos?

¿Temas?¿Semántica del contenido?

¿Gramática?

¿Analizaremos un documento o varios documentos?

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

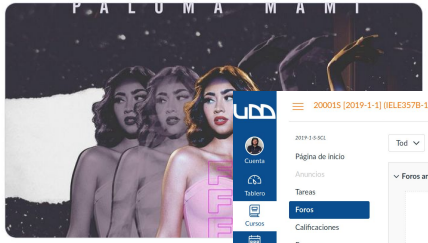


Seguir

Paloma Mami
@palomamamicl
palomamamimusic.com Se unió el enero de 2019
26 Siguiendo **65,2 K** seguidores
Maya, Jeanette Orozco y 18 más de las cuentas que sigues siguen a este usuario

Tweets | Tweets y respuestas | Multimedia | Me gusta

Tweet fijado
Paloma Mami @palomamamicl · 22 mar.
FINGIAS smarturl.it/Fingias/youtube



654 2,4 K 9,7 K

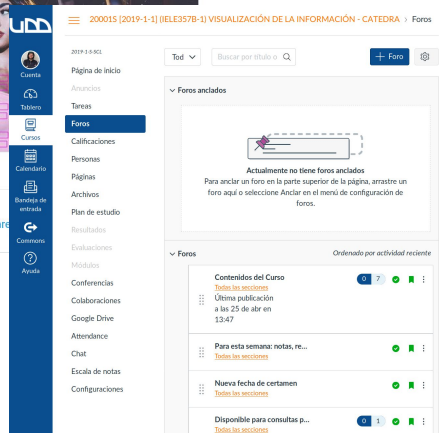
Paloma Mami @palomamamicl · 26 abr.
Gracias @Spotify por tener Fingias en #bailare
❗❗❗

WIKIPEDIA

Universidad del Desarrollo
universidad privada chilena

La **Universidad del Desarrollo** es una universidad privada autónoma en Chile Concepción y en Santiago, específicamente en las comunas de Las Condes

Universidad del Desarrollo	
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

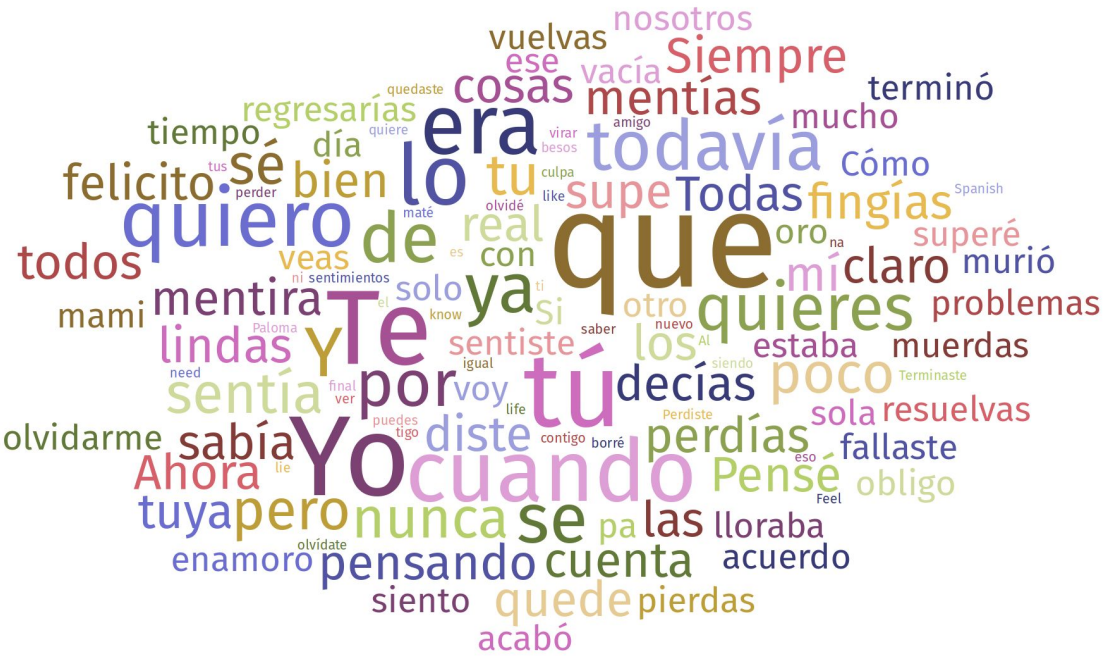
Word Cloud: Frecuencias

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/



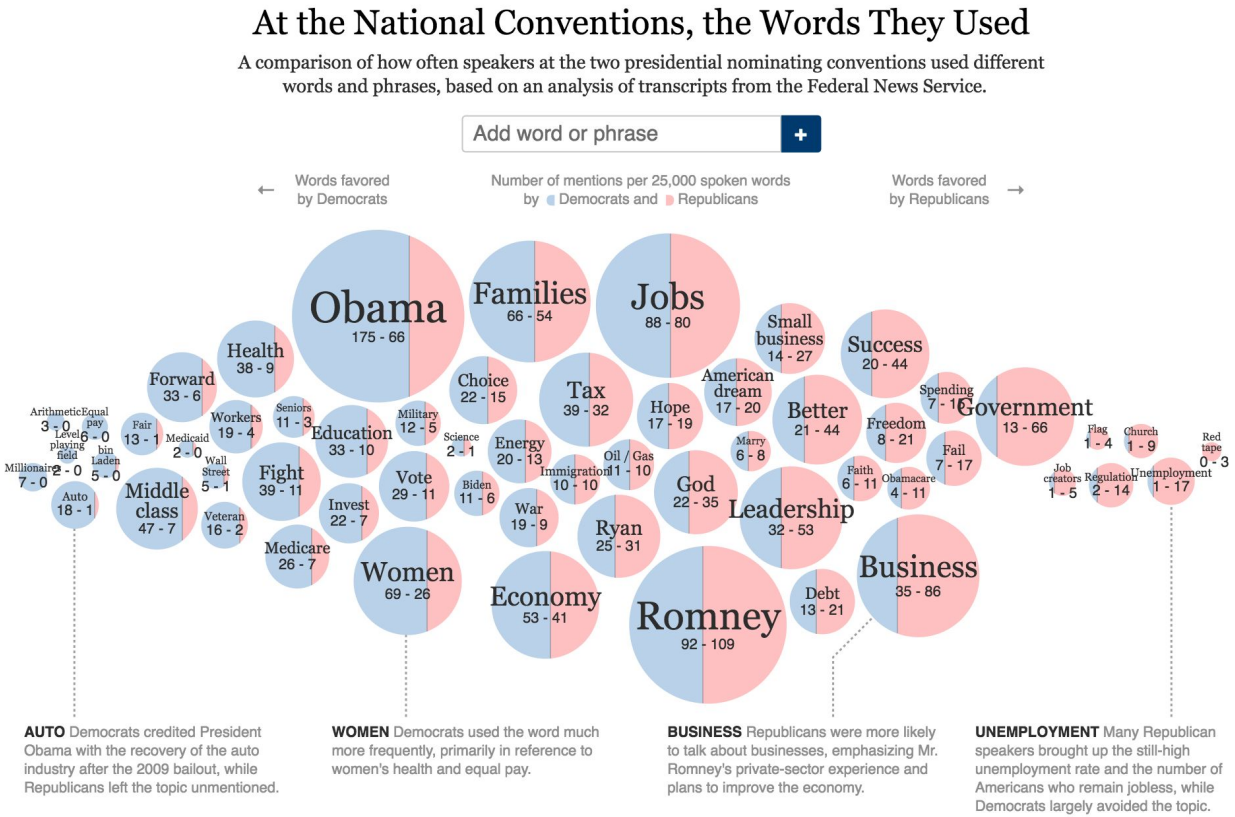
Documento: Fingías de Paloma Mami

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.



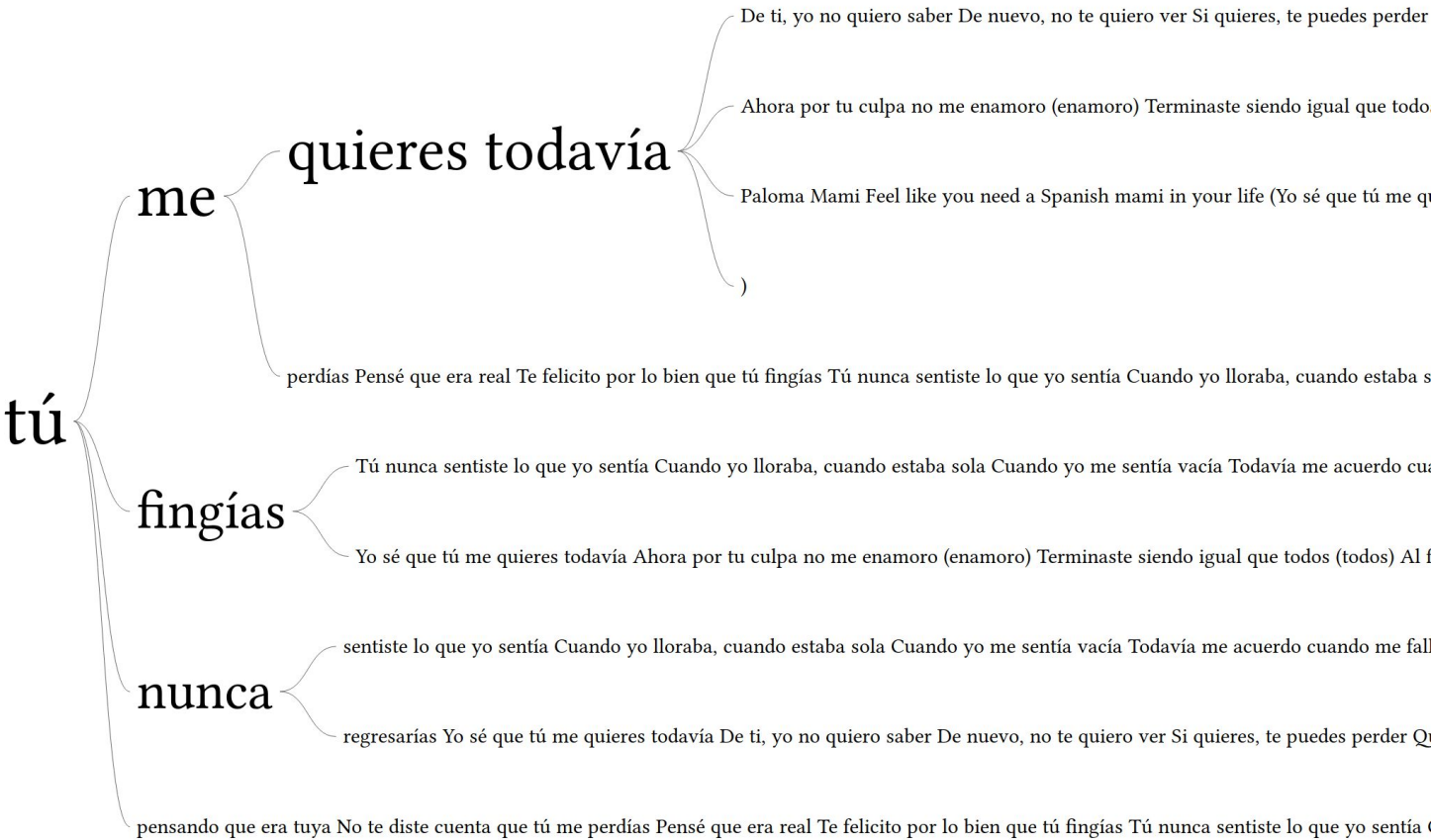
New York Times

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

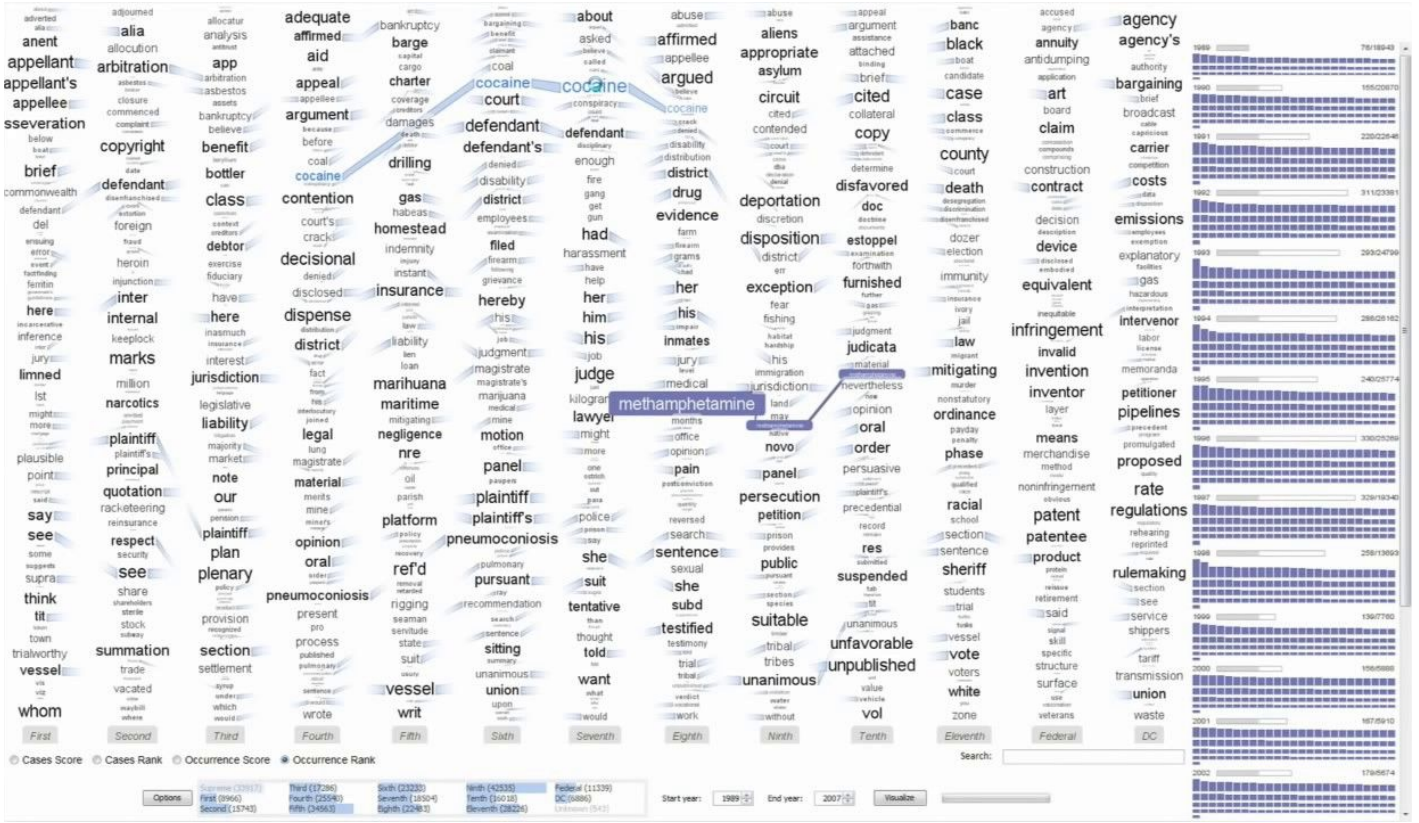
Wattenberg, M., & Viégas, F. B. (2008). The word tree, an interactive visual concordance. *IEEE transactions on visualization and computer graphics*, *14*(6), 1221-1228.



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.

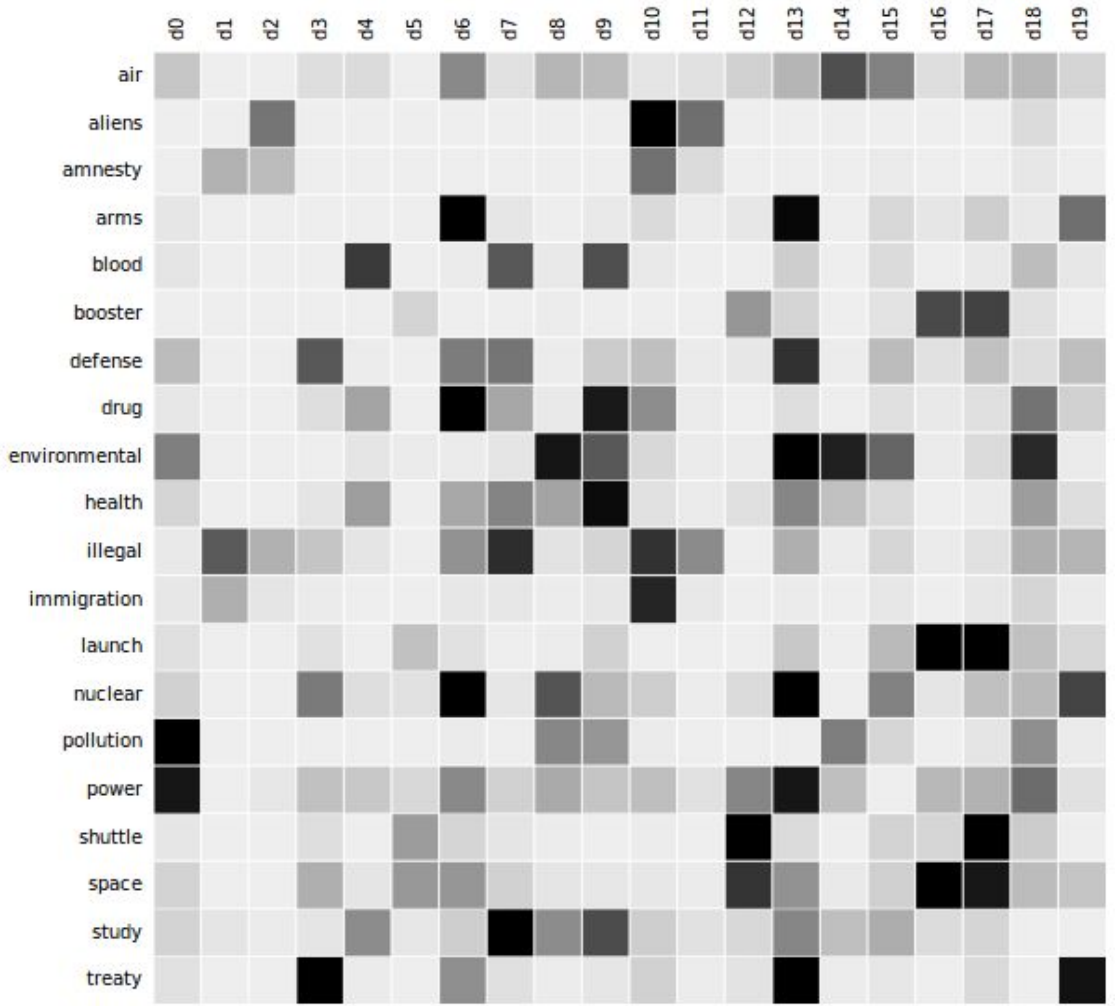
Temáticas (*topics*) y Reducción Dimensional

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**)

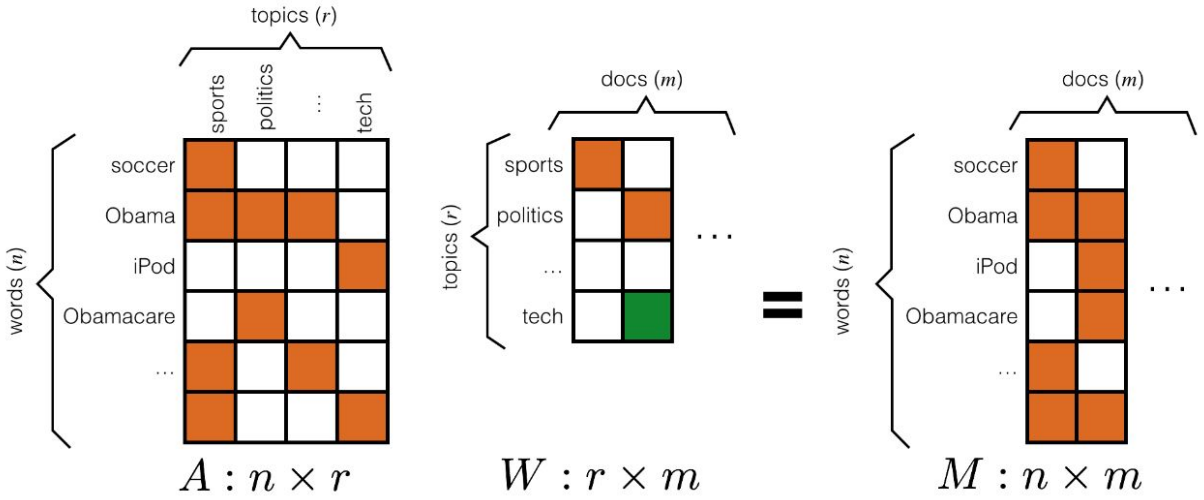
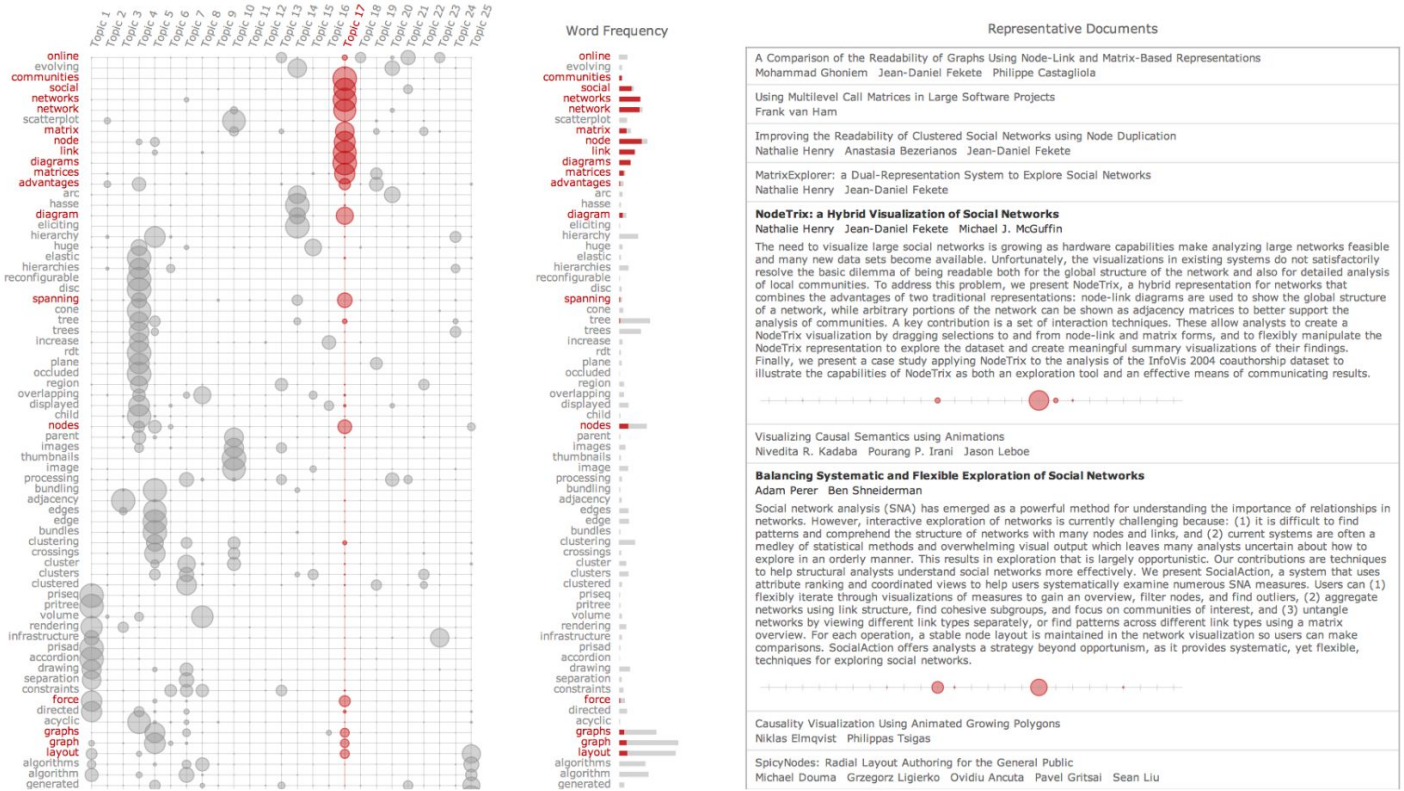


Imagen por Catalin Voss, Stanford.

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.



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/~Ian.Wood/termite/50-run1-LIWCvocab/public_html/

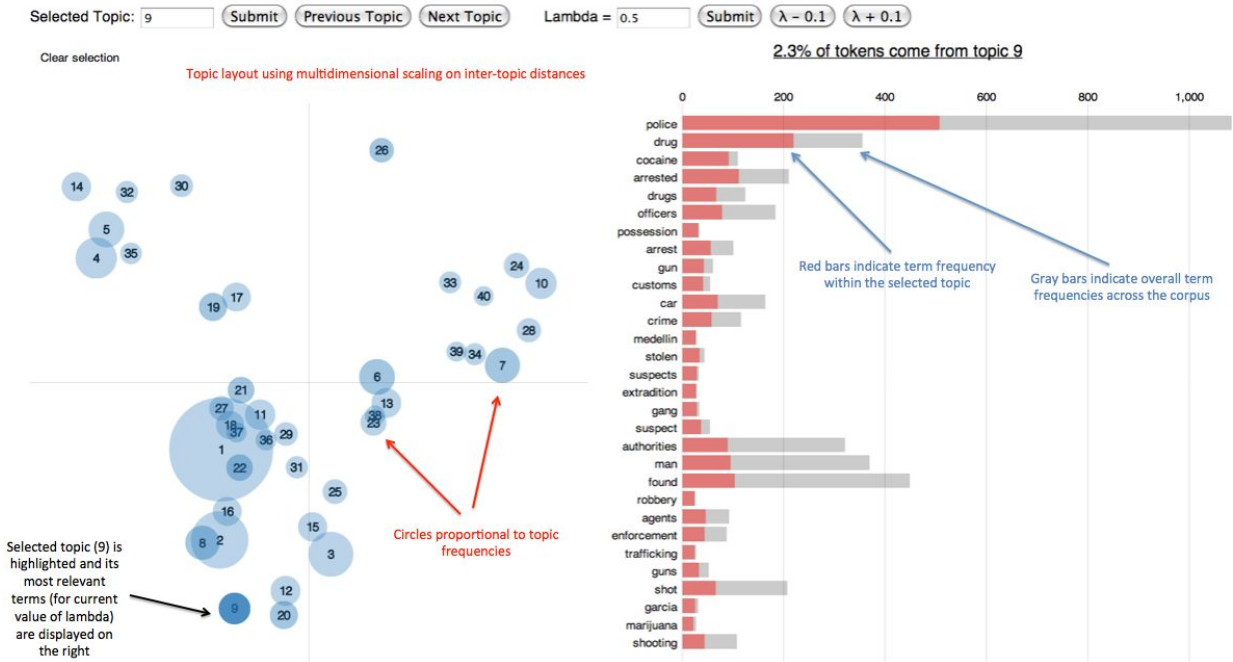
LDA Vis

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).

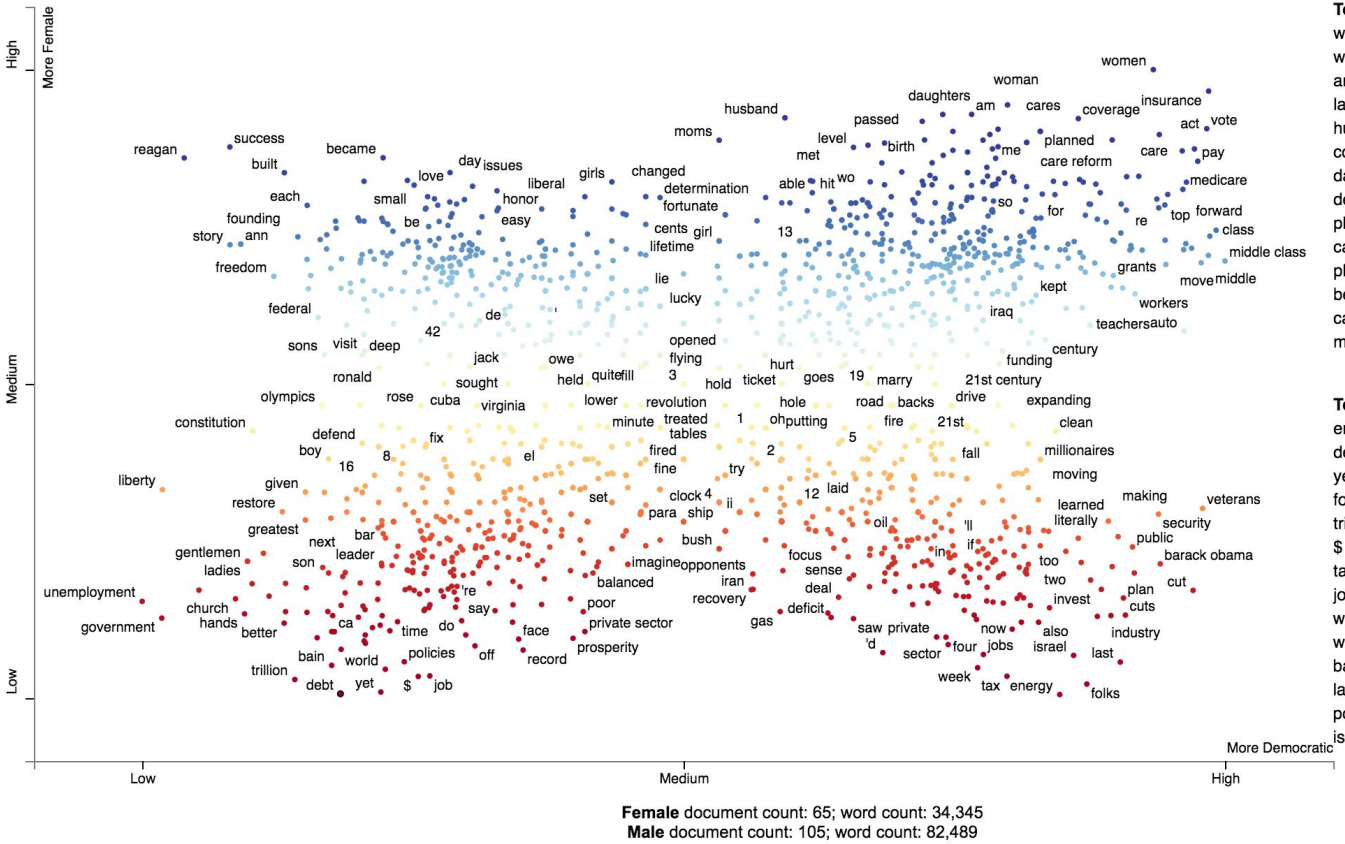


Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces* (pp. 63-70).

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.



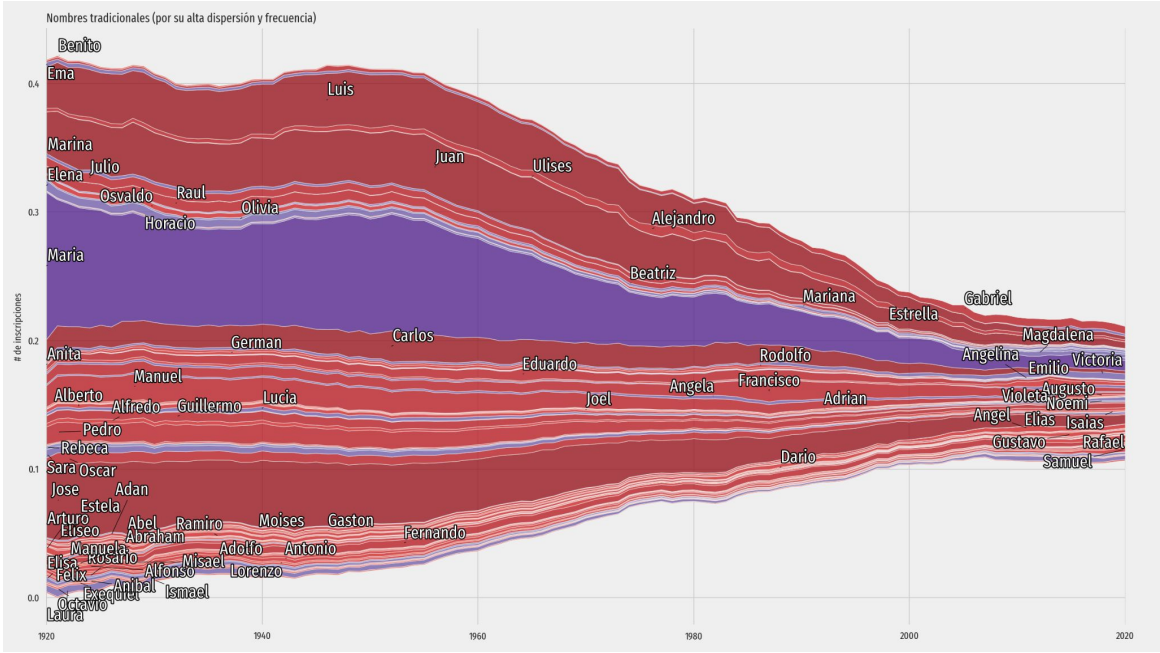
Kessler, J. (2017). Scattertext: a Browser-Based Tool for Visualizing how Corpora Differ. *Proceedings of ACL 2017, System Demonstrations*, 85-90.
Código: https://github.com/JasonKessler/scattertext

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)

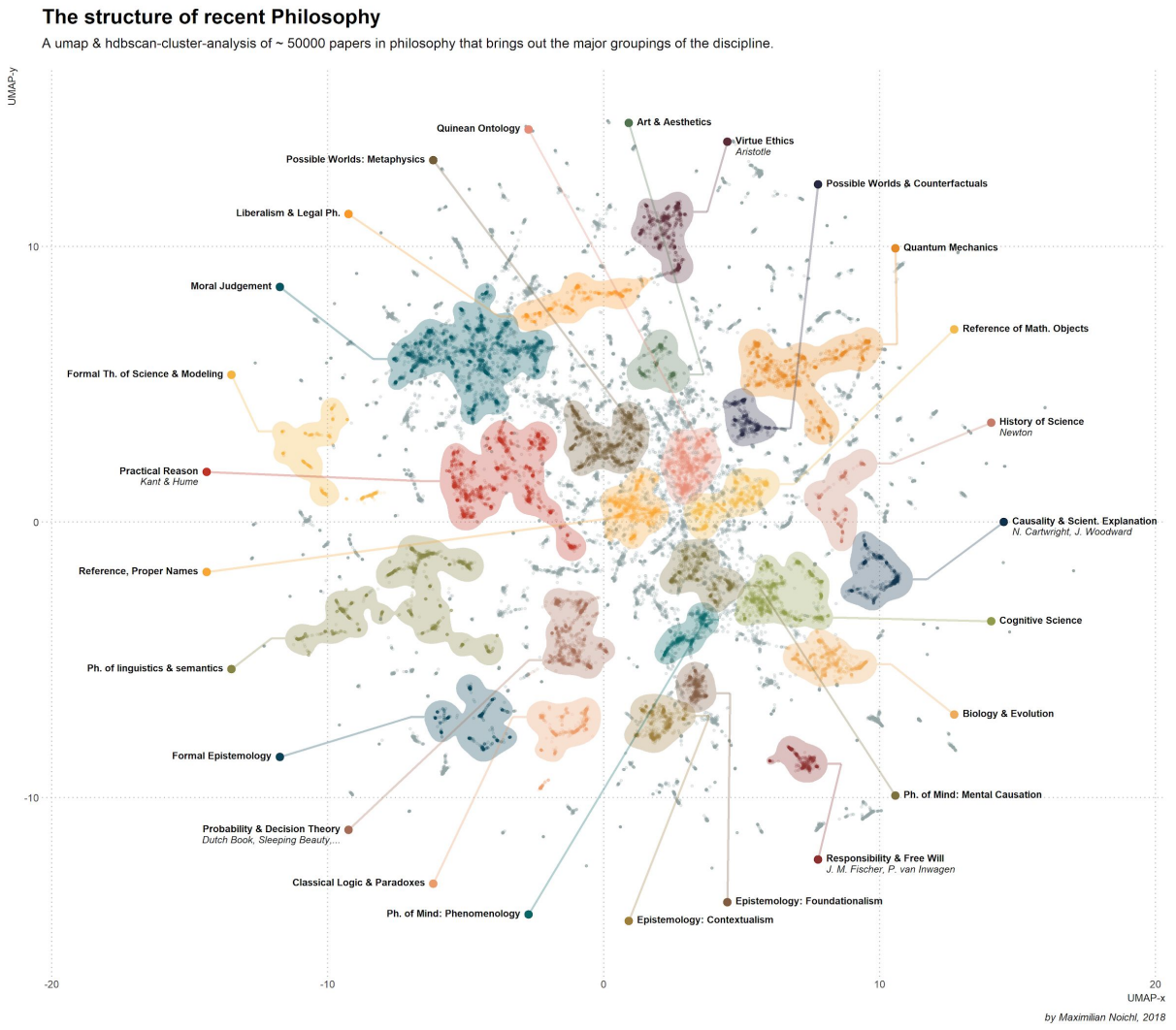


Reducción Dimensional con UMAP

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/posts/structure-of-recent-philosophy-ii/

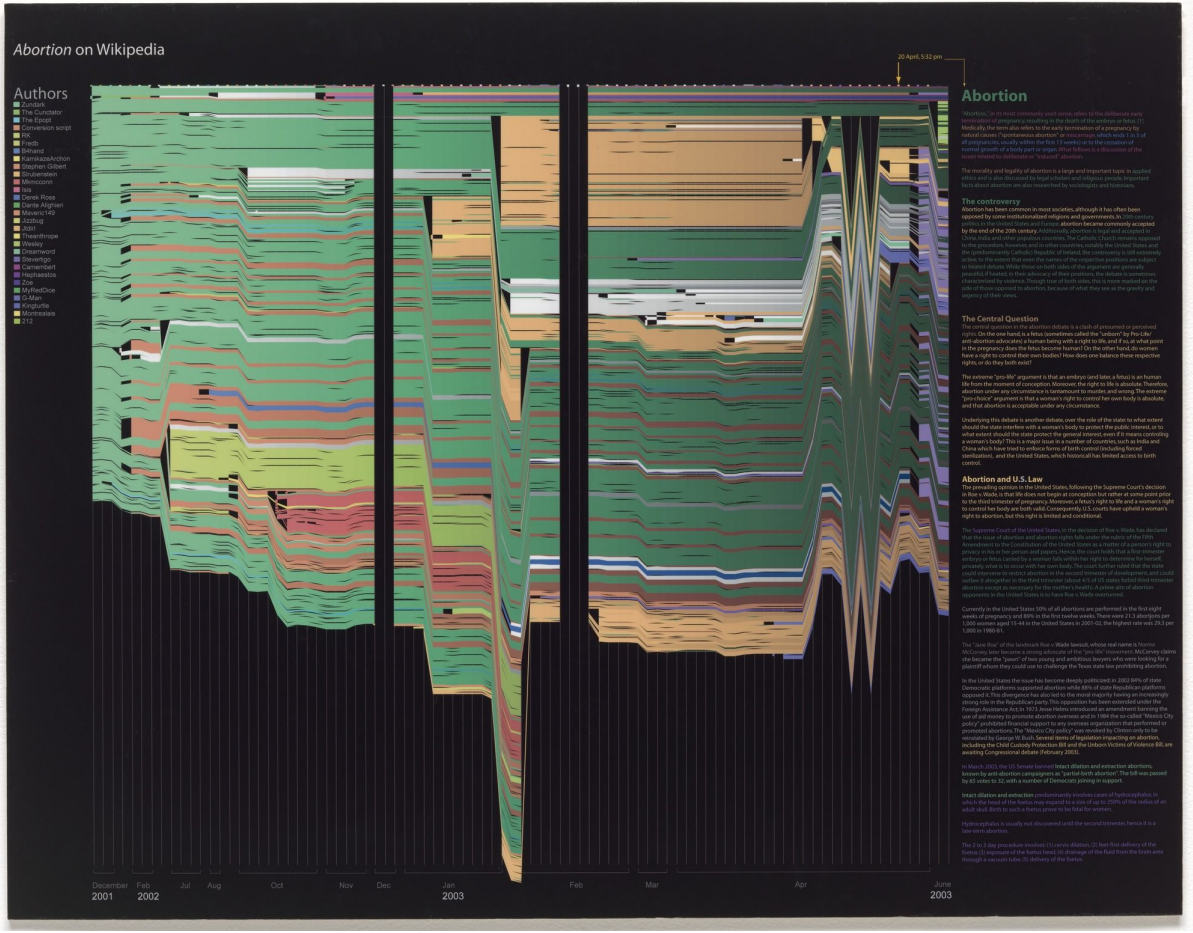


Evolución y Cambios

History Flow

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).



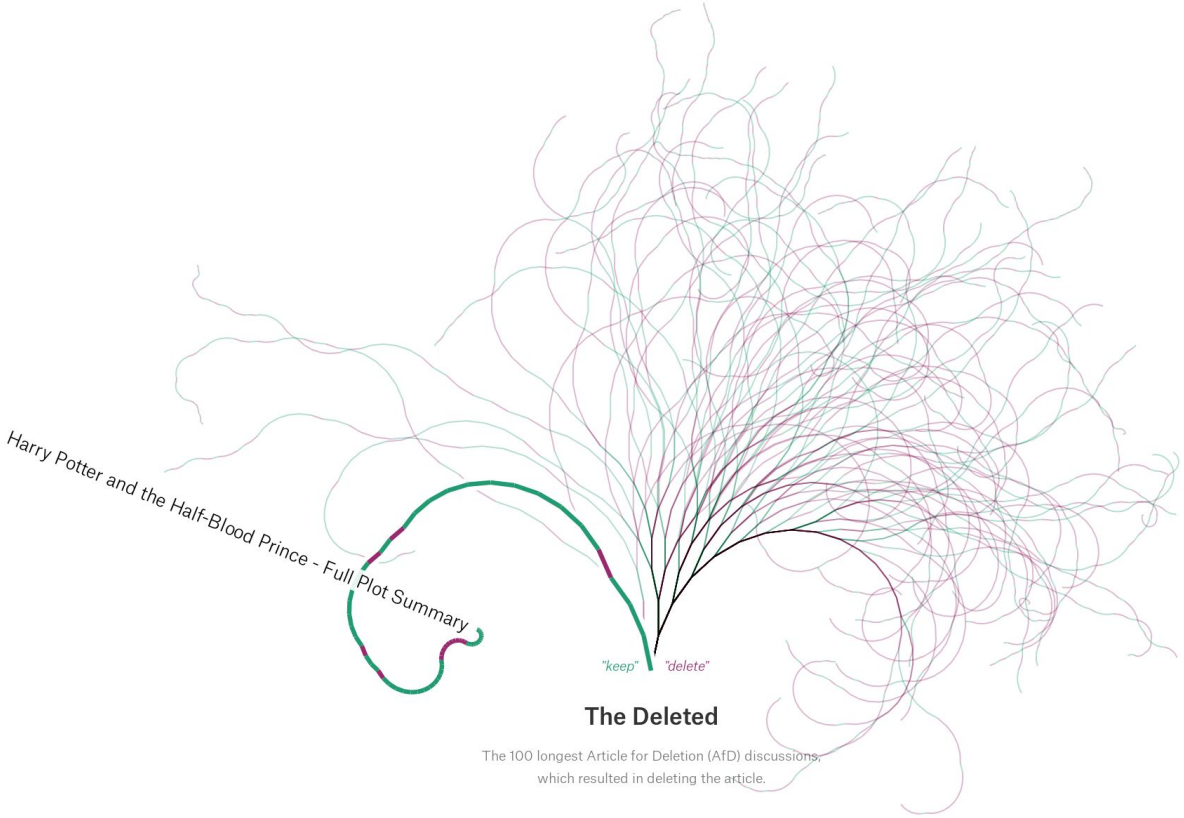
Viégas, F. B., Wattenberg, M., & Dave, K. (2004, April). Studying cooperation and conflict between authors with history flow visualizations. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 575-582). ACM.

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?