




Descomposición en tópicos

06-10-23



¿Cómo describimos matemáticamente un texto?

≡ MENÚ

LA NACION

LA NACION | SOCIEDAD | DEBATE POR EL ABORTO

Con un nuevo proyecto de ley, mujeres marcharon al Congreso por el aborto legal



¿Palabras más frecuentes? No parece una buena idea...

Cantidad de veces que aparece la palabra "de" en el documento

Palabra	Frecuencia
de	57
la	40
el	37
...	...
las	9
un	9
proyecto	8

Primer palabra informativa...

¿Cómo describimos matemáticamente un texto?

MENÚ

LA NACION

LA NACION | SOCIEDAD | DEBATE POR EL ABORTO

Con un nuevo proyecto de ley, mujeres marcharon al Congreso por el aborto legal



Eliminando preposiciones y palabras muy comunes...

Palabra	Frecuencia
proyecto	8
campana	7
congreso	7
aborto	5
diputados	5

Mucho mejor...

¿Cómo describimos matemáticamente un texto?

LA NACION

LA NACION | SOCIEDAD | DEBATE POR EL ABORTO

Con un nuevo proyecto de ley, mujeres marcharon al Congreso por el aborto legal



Palabra	Frecuencia
proyecto	8
campana	7
congreso	7
aborto	5

LA NACION

LA NACION | POLÍTICA | DEBATE POR EL ABORTO

El relanzamiento de un partido contrario al aborto reúne a referentes celestes

Valores para mi País, de la exdiputada Hotton, hará campaña en defensa de la vida

29 de marzo de 2019

Comentar (14) Me gusta Compartir

Pañuelos y banderas celestes, señal clara de los símbolos provida que prevalecerán en la campaña, dominaron ayer en un clima festivo el lanzamiento de Valores para mi País, la agrupación política impulsada hace unos años por la exdiputada Cynthia Hotton y recreada ayer con un acto en la Federación Argentina de Box, con vistas a las elecciones de octubre.

El lanzamiento reunió a agrupaciones y dirigentes políticos provida, como el

Palabra	Frecuencia
valores	4
hotton	3
país	3
aborto	2

LA NACION

LA NACION | POLÍTICA | DEBATE POR EL ABORTO

Un obispo recomendó no votar a los candidatos que apoyan el aborto




Palabra	Frecuencia
obispo	9
aborto	5
elecciones	4
apoyan	3

¿Cómo describimos matemáticamente un texto?

LA NACION | SOCIEDAD | DEBATE POR EL ABORTO


Con un nuevo proyecto de ley, el Congreso por el aborto



Palabras muy informativas de la nota

Palabras poco informativas: aparecen en varios documentos (verbos, adjetivos, o palabras comunes en un corpus específico)

datos que



Palabra	Frecuencia
proyecto	8
campana	7
congreso	7
aborto	5

Palabra	Frecuencia
valores	4
hotton	3
país	3
aborto	2

Palabra	Frecuencia
obispo	9
aborto	5
elecciones	4
apoyan	3

¿Cómo describimos matemáticamente un texto?

Palabras poco informativas: aparecen en varios documentos (verbos, adjetivos, o palabras comunes en un corpus específico)

Depende del corpus con el que le adjudique la valorización (la palabra “Macri” tendrá diferente especificidad en un corpus de notas políticas que en un corpus de notas de diferentes secciones).

¿Qué hacemos con estas palabras? Las pesamos según su especificidad:

Inverse document frequency (para cada término)

$$idf_t = \log\left(\frac{N}{n_t}\right)$$

Cantidad de documentos en el corpus

Cantidad de documentos donde aparece el término t

¿Cómo describimos matemáticamente un texto?

MENÚ LA NACION

LA NACION | SOCIEDAD | DEBATE POR EL ABORTO

Con un nuevo proyecto de ley, mujeres marcharon al Congreso por el aborto legal



MENÚ LA NACION

LA NACION | POLÍTICA | DEBATE POR EL ABORTO

El relanzamiento de un partido contrario al aborto reúne a referentes celestes

Valores para mi País, de la exdiputada Hotton, hará campaña en defensa de la vida

29 de marzo de 2019

Comentar (14) Me gusta Compartir

Pañuelos y banderas celestes, señal clara de los símbolos provida que prevalecerán en la campaña, dominaron ayer en un clima festivo el lanzamiento de Valores para mi País, la agrupación política impulsada hace unos años por la exdiputada Cynthia Hotton y recreada ayer con un acto en la Federación Argentina de Box, con vistas a las elecciones de octubre.

El lanzamiento reunió a agrupaciones y dirigentes políticos provida, como el

MENÚ LA NACION

LA NACION | POLÍTICA | DEBATE POR EL ABORTO

Un obispo recomendó no votar a los candidatos que apoyan el aborto



Si mi corpus consiste en 3 notas:

$$\text{idf}(\text{"aborto"}) = \log(3/3) = 0$$

$$\text{idf}(\text{"hotton"}) = \log(3/1) = 1.09$$



En este corpus “aborto” es poco específica, mientras que “hotton” toma el máximo valor (suponiendo que solo aparece en un solo documento).

¿Cómo describimos matemáticamente un texto?



Descripción más específica de cada texto



Frecuencia (tf) multiplicada por el idf del término

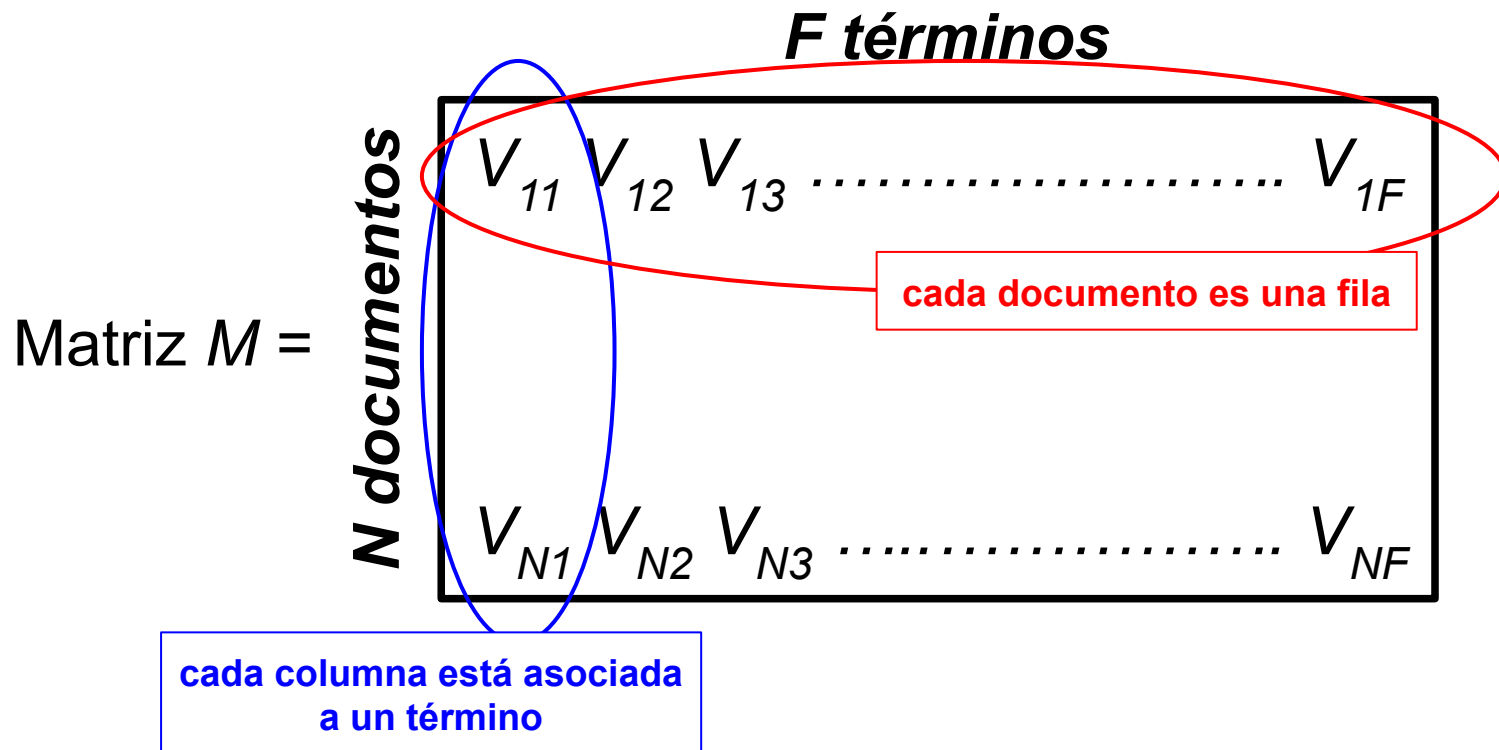


Palabra	Tf x idf
campaña	16.01
proyecto	16.00
congreso	15.45
diputados	12.88

Palabra	Tf x idf
valores	13.55
hotton	11.57
exdiputada	9.55
lanzamiento	8.73

Palabra	Tf x idf
obispo	31.67
cristo	14.32
ossola	14.32
diócesis	13.10

Representación de los textos



Representación de los textos

Palabras, bigramas,
trigramas, lemas, solo la
raíz de la palabra...

F términos

Matriz $M =$

V_{11}	V_{12}	V_{13}	V_{1F}
V_{N1}	V_{N2}	V_{N3}	V_{NF}

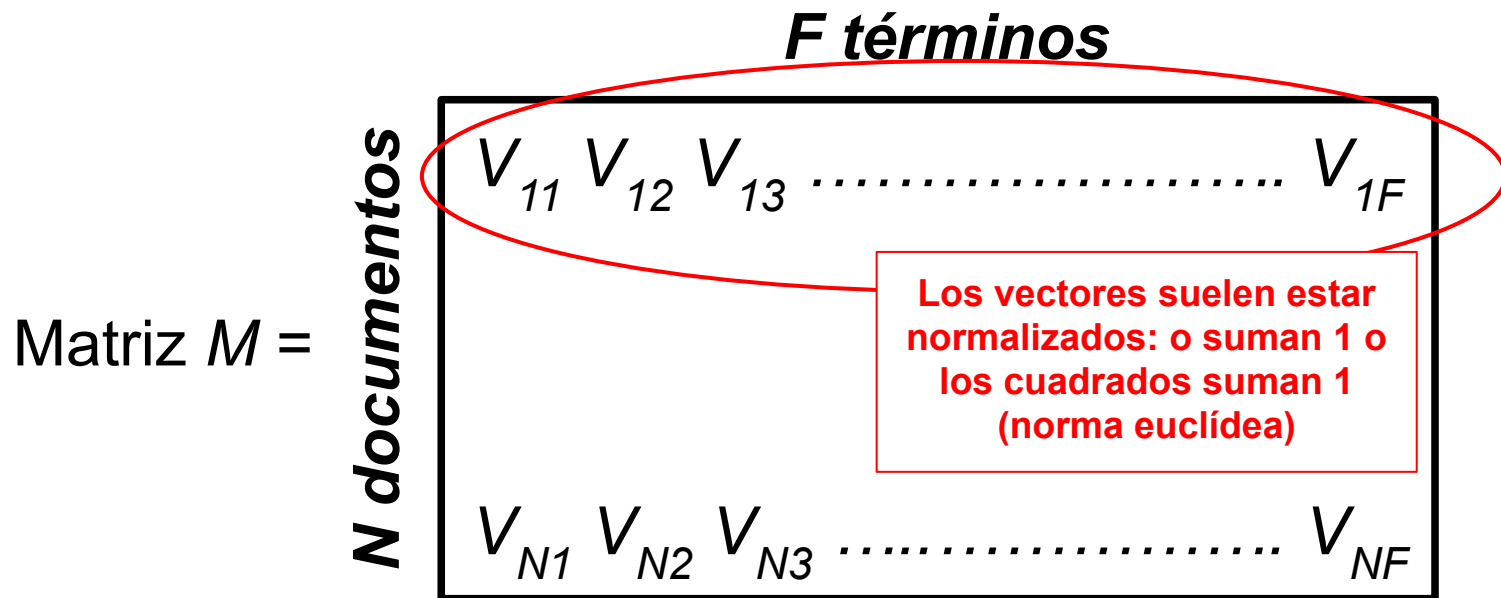
Frecuencia del término (tf)
o frecuencia x
especificidad (idf)

$$idf_t = \log\left(\frac{N}{n_t}\right)$$

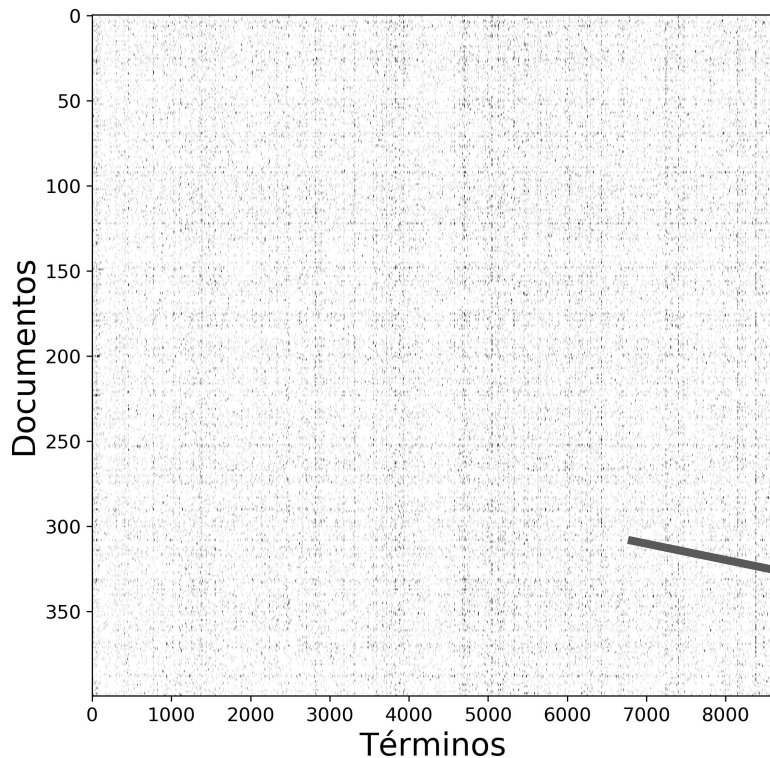
Cantidad de documentos en el
corpus

Cantidad de documentos donde
aparece el término t

Representación de los textos



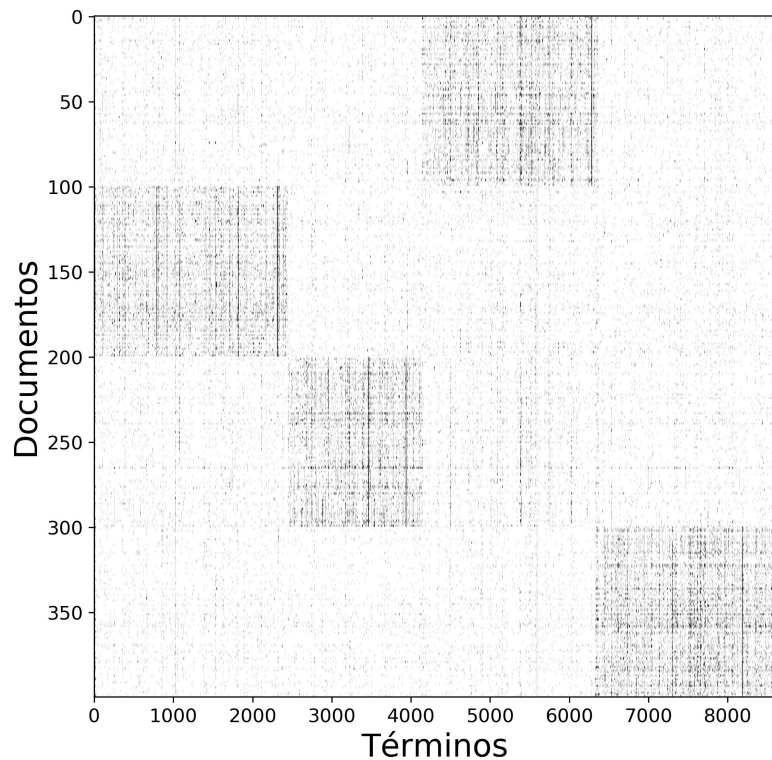
Tópicos



¿Cómo se ve una matriz
de documentos por
términos real?

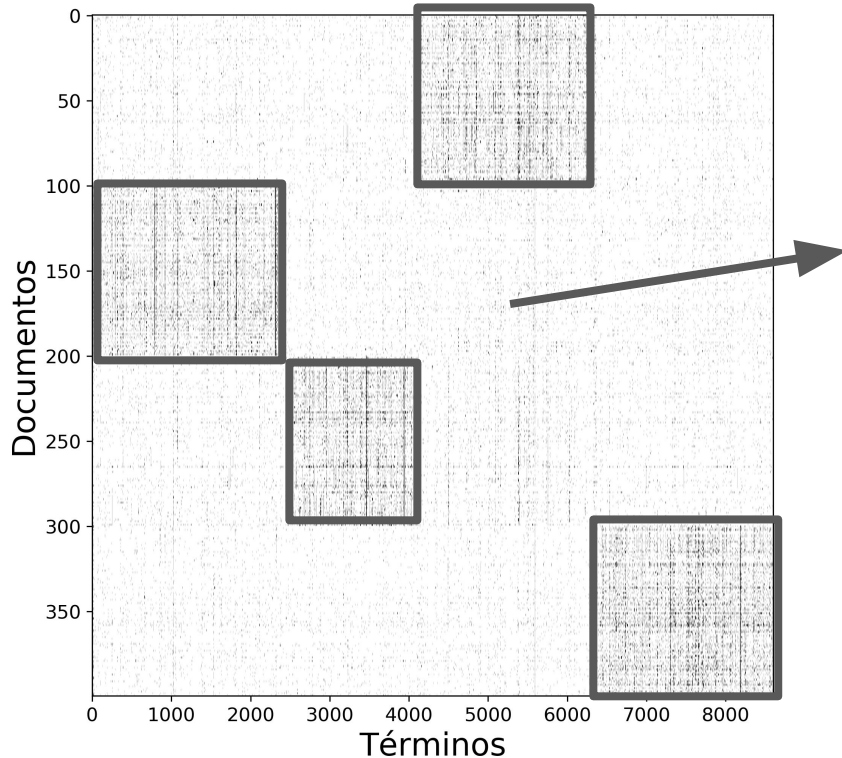
En blanco las componentes
igual a cero; en negro las
componentes distintas de cero.

Tópicos



Ordenando la matriz,
tanto en filas como en
columnas...

Tópicos



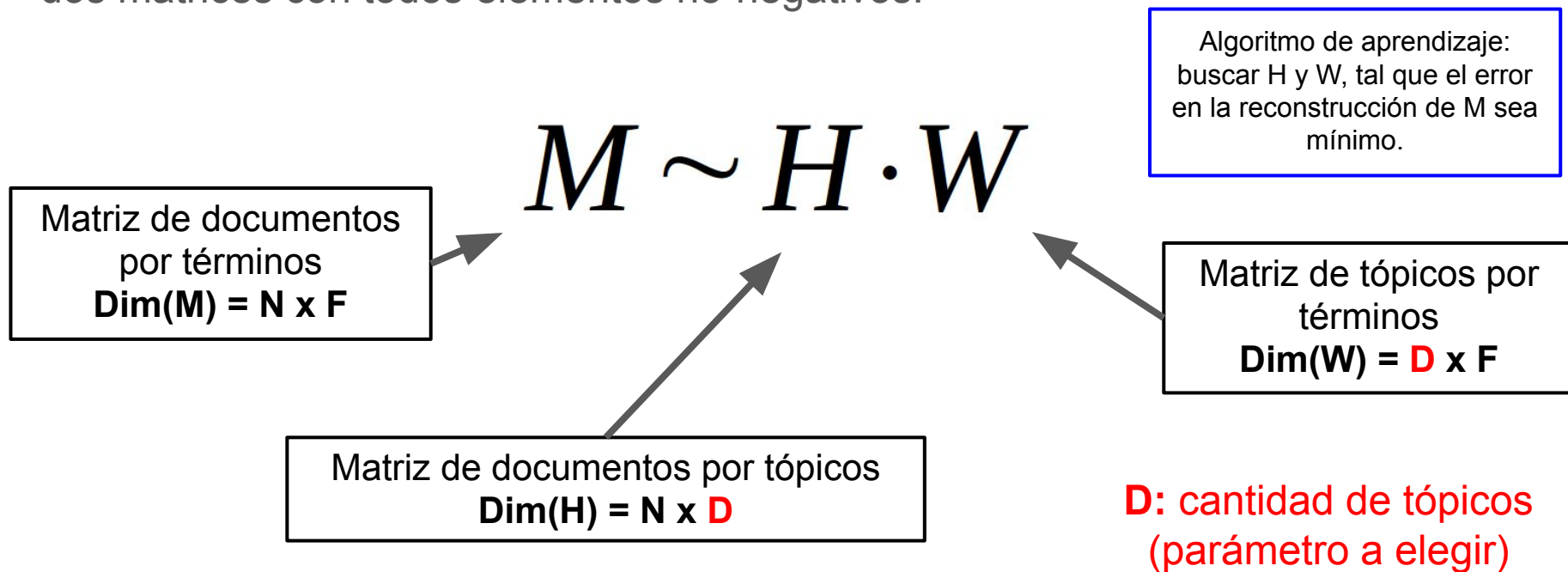
Emergencia de bloques: **Conjunto de documentos que usan términos similares.** Estos bloques **emergen naturalmente** del “ordenamiento” de la matriz de documentos por términos.

A los bloques los identificamos como **tópicos o ejes temáticos.**

¿Cómo “ordenamos”? Con algoritmos de identificación de tópicos (**NMF**, LDA, etc...)

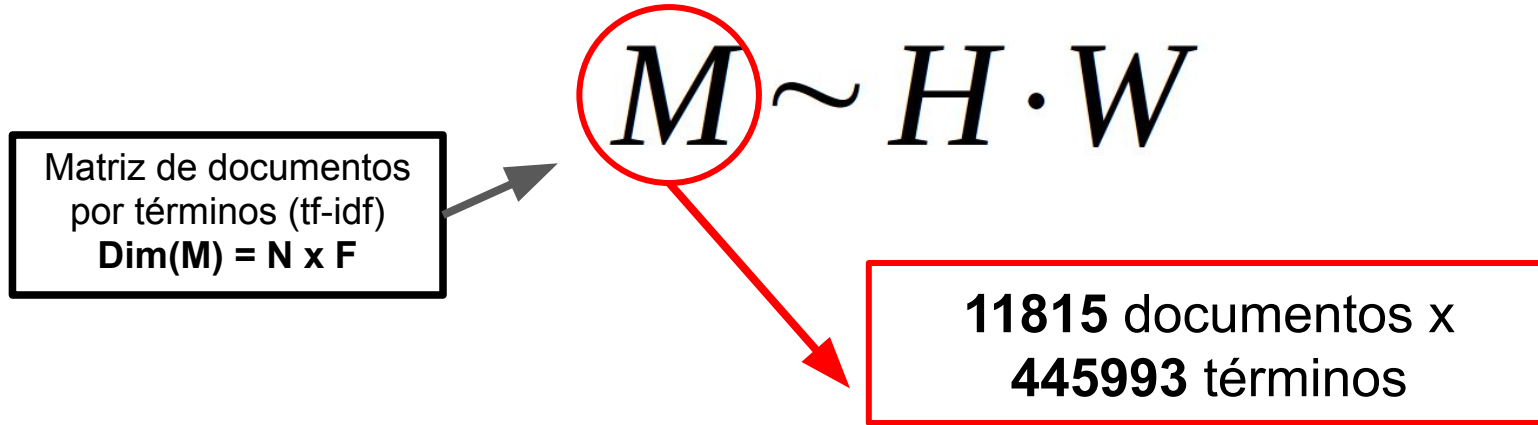
Non-negative factorization (NMF)

Algoritmo para detectar tópicos: describimos la matriz M como la multiplicación de dos matrices con todos elementos no-negativos.



Ejemplo

Notas de las secciones políticas de distintos medios, publicadas en el período del **31 de Julio al 5 de Noviembre del 2017:**



Interpretación de la matriz W

$$M \sim H \cdot W$$

Matriz de tópicos
por términos

$$\text{Dim}(W) = \mathbf{D} \times \mathbf{F}$$

Tópico 1

1.1 9.6 7.2 1.4 8.5

Tópico 2

9.4 0.4 1.2 8.3 1.3

Más todos los
tópicos que haya

Maldonado

Cristina

Kirchner

Santiago

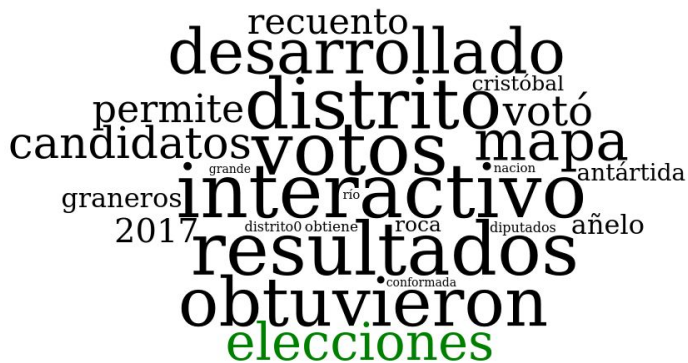
Cambiamos

Más todos los
términos que haya

Podemos **ordenar los términos de mayor a menor** peso para cada tópico e **interpretar**.

Los tópicos pueden tener peso distinto de cero en todos los términos.

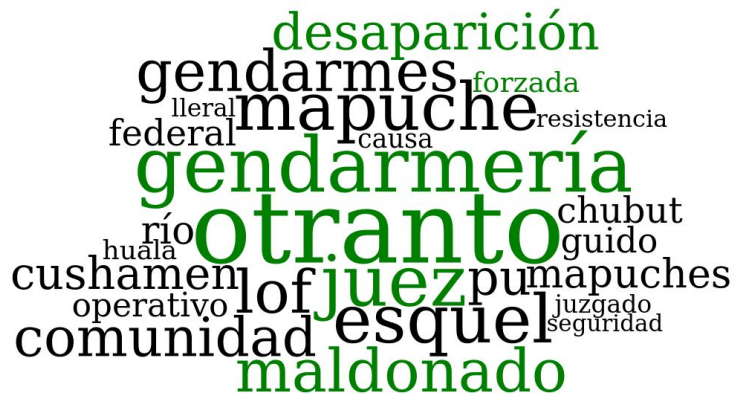
Tópicos en el corpus de noticias (D = 10)



Wordclouds
o nubes de
palabras

Elecciones

Tópicos en el corpus de noticias ($D = 10$)



Maldonado

Tópicos en el corpus de noticias (D = 10)

De Vido

A word cloud for the topic 'De Vido'. The most prominent words are 'desafuero' and 'vido'. Other visible words include 'detenido', 'corrupción', 'cámara', 'diputado', 'tragedia', 'causa', 'juicio', 'detención', 'federal', 'ministro', 'bonadio', 'juez', 'baratta', 'planificación', 'rusconi', 'diputados', 'pedido', 'minnicelli', 'rodríguez', 'once', and 'fueros'.

A word cloud for the topic 'Boudou'. The most prominent word is 'boudou'. Other visible words include 'vicepresidente', 'calcográfica', 'tribunal', 'juicio', 'enriquecimiento', 'lijo', 'núñez', 'vandenbroele', 'amado', 'oral', 'detenido', 'carmona', 'federal', 'ilícito', 'socio', 'juez', 'abogado', 'causa', 'impresión', 'ciccone', and 'detención'.

Boudou

A word cloud for the topic 'Milagro Sala'. The most prominent words are 'domiciliaria', 'sala', and 'milagro'. Other visible words include 'traslado', 'penal', 'llermanos', 'juez', 'comisión', 'dirigente', 'pullen', 'cidh', 'humanos', 'mercaderías', 'resolución', 'prisión', 'tupac', 'amaru', 'cautelar', 'preventiva', 'comedero', 'morales', 'interamericana', 'detención', 'juzuy', 'derechos', and 'bonadio'.

**Milagro
Sala**

Tópicos en el corpus de noticias (D = 10)

Nisman

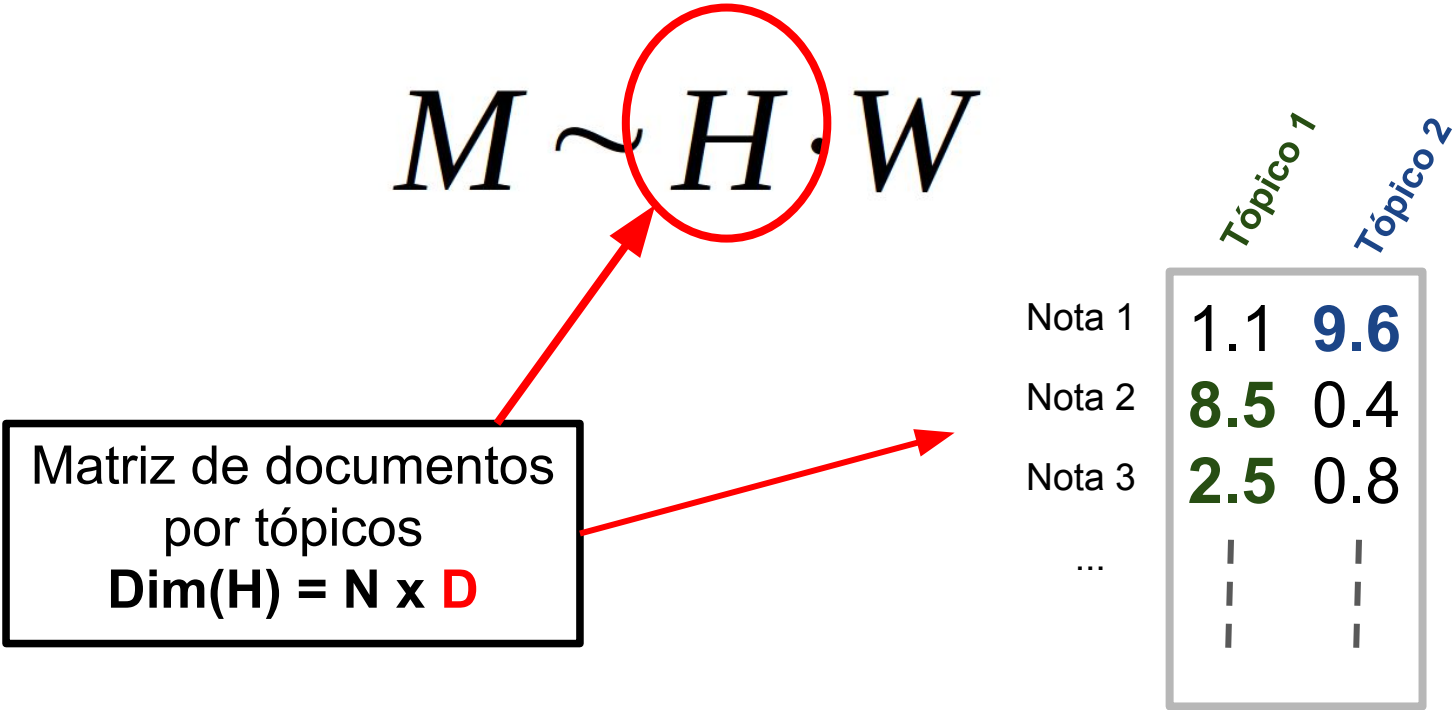
encubrimiento
pollicita atentado juez justicia
kirchner irán fiscal carbó
cristina amia causa
alberto gils
nisman
denuncia timerman iraníes
muerte lagomarsino bonadio federal
memorándum indagatoria

Macri

schmid trabajadores
presidente reformas
rosada mauricio
gobierno carbó
mauricio reunión
triaca ley
gils trabajo
laboral reforma
ministro gobernadores medina
jefe gabinete política argentina

Notar que nuestra
descripción es finalmente
con **7 tópicos**

Interpretación de la matriz H

$$M \sim H \cdot W$$


Matriz de documentos
por tópicos

$\text{Dim}(H) = N \times D$

Nota 1

1.1 9.6

Nota 2

8.5 0.4

Nota 3

2.5 0.8

...

⋮

⋮

Tópico 1

Tópico 2

Interpretación de la matriz H

$$M \approx H \cdot W$$

	Tópico 1	Tópico 2
Nota 1	1.1	9.6
Nota 2	8.5	0.4
Nota 3	2.5	0.8
...	⋮	⋮

Para cada nota hay un
tópico dominante, pero
**puede haber varios
tópicos que la
describan.**

**Lo pensamos como una
distribución.**

	Tópico 1	Tópico 2
Nota 1	0.11	0.89
Nota 2	0.95	0.05
Nota 3	0.76	0.24
...	⋮	⋮

Interpretación de la matriz H

$$M \sim H \cdot W$$

	Tópico 1	Tópico 2
Nota 1	0.11	0.89
Nota 2	0.95	0.05
Nota 3	0.76	0.24
...



Recomendaciones para NMF

- Suele andar mejor con la matriz de documentos por términos pesada por **idf** (*inverse document frequency*), es decir, la matriz **tf-idf**.
- Los vectores documentos se suelen entrar normalizados (por ejemplo, a norma euclídea igual a 1).
- Eliminar stopwords es siempre útil: si no lo hiciéramos es muy probable la emergencia de un tópico compuesto por sólo éstas palabras.

Características de NMF

- Los tópicos resultantes son no-ortogonales: esto suele dar una interpretación más natural de los tópicos, dado que hay temas que suelen tener overlap.
- Al normalizar los vectores documentos en el espacio de tópicos podemos interpretar dichos vectores como distribuciones (en el espacio de tópicos).
- Desventaja: la cantidad de tópicos D fija además un nivel de resolución. El algoritmo suele encontrar D tópicos de tamaño similar (los tópicos chicos suelen ser absorbidos por los grandes, o bien, uno grande suele partirse en varios).