Procesamiento de Lenguaje Natural

Olivia Gutú y Julio Waissman

Maestría en Ciencia de Datos
Semana 3: Modelos de espacios vectoriales

Introducción

Luis estudia solo piano
Luis estudia solo una materia

↑

diferente significado

no me gusta practicar
odio hacer ejercicios

↑

mismo significado

Contexto (*word by word*)

Ventana de distancia k de un token: tokens que se encuentran a una distancia menor o igual a k.

los cronopios según Cortázar cronopios y famas de Cortázar

e. g.
Frecuencia w-w en una ventana de tamaño $k = 2$:

	los	cronopios	según	y	famas	de
Cortázar	0	2	1	1	1	1

Cortázar tendría un representación en $\mathbb{R}^{|V|}$, V el vocabulario de tipos.

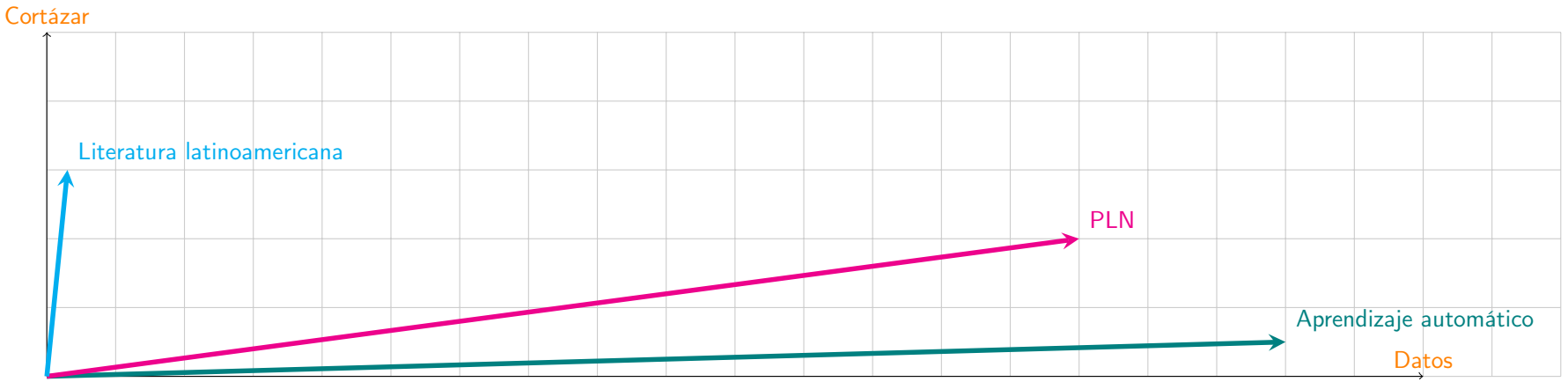
Diseño (*word by document*)

Frecuencia de un token aparece en alguna categoría:

	Aprendizaje automático	PLN	Literatura latinoamericana
Datos	345445	233341	102
Cortázar	207	2345	34456

Podemos ver al vocabulario en términos de las categorías, o bien, las categorías en términos del vocabulario.

Distancia entre palabras



Distancia entre palabras: producto punto

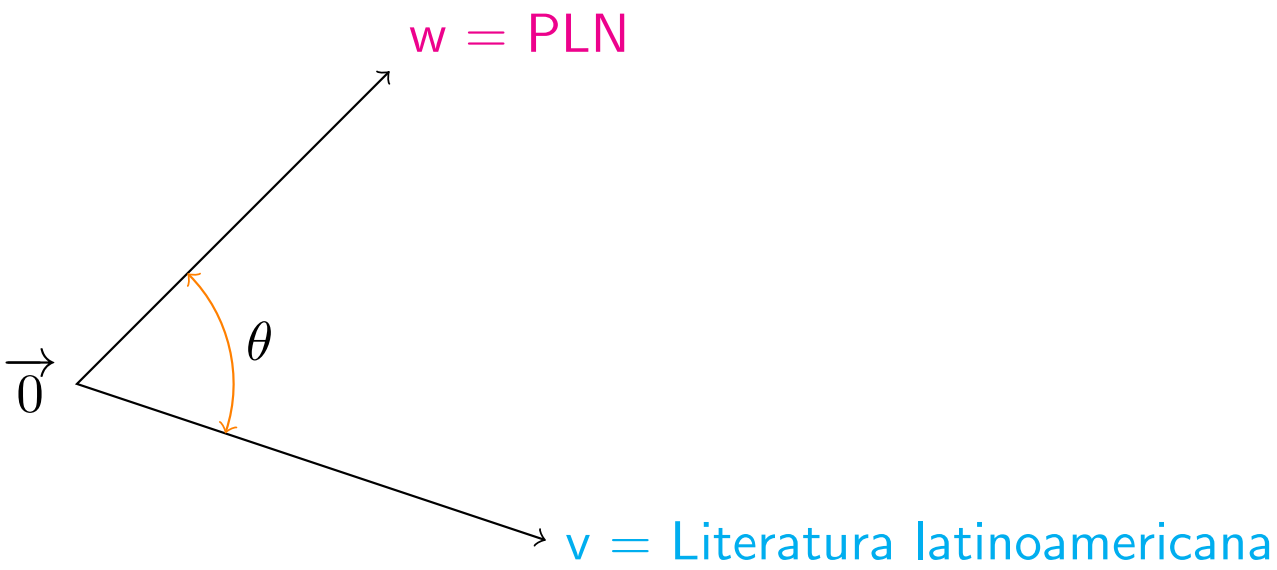
$v = (102, 34456)$
$w = (233341, 2345)$

$$v \text{ dot } w = (102 \times 233341) + (34456 \times 2345)$$

$$\begin{aligned} v \text{ dot } w &\leftarrow \text{producto punto de } v \text{ y } w \\ \|v\| = \sqrt{v \text{ dot } v} &\leftarrow \text{norma euclidiana de } v \\ \|v - w\| &\leftarrow \text{distancia euclidiana entre } v \text{ y } w \end{aligned}$$

Distancia entre palabras: similaridad coseno

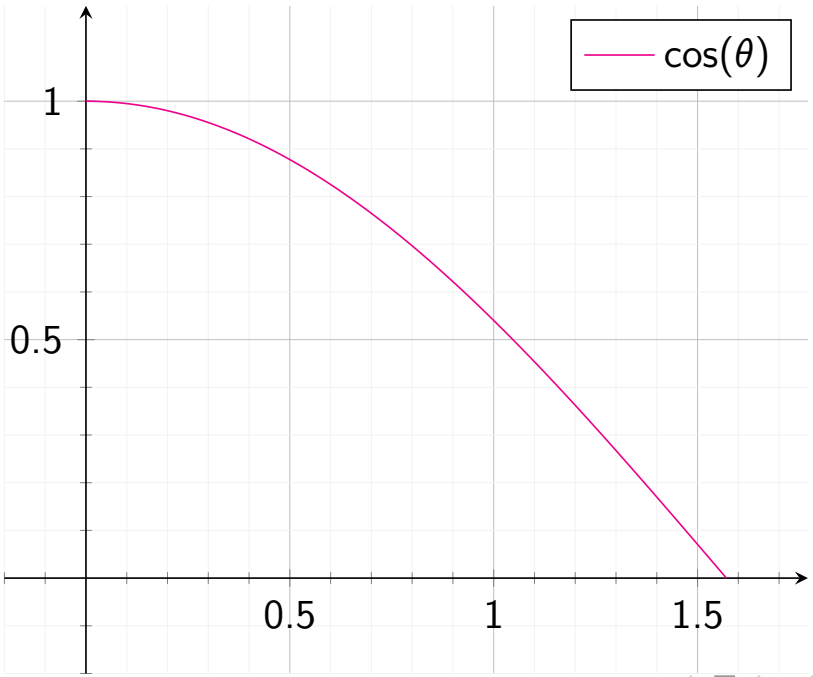
$$\text{sim}(v, w) = \frac{v \text{ dot } w}{\|v\|\|w\|}$$



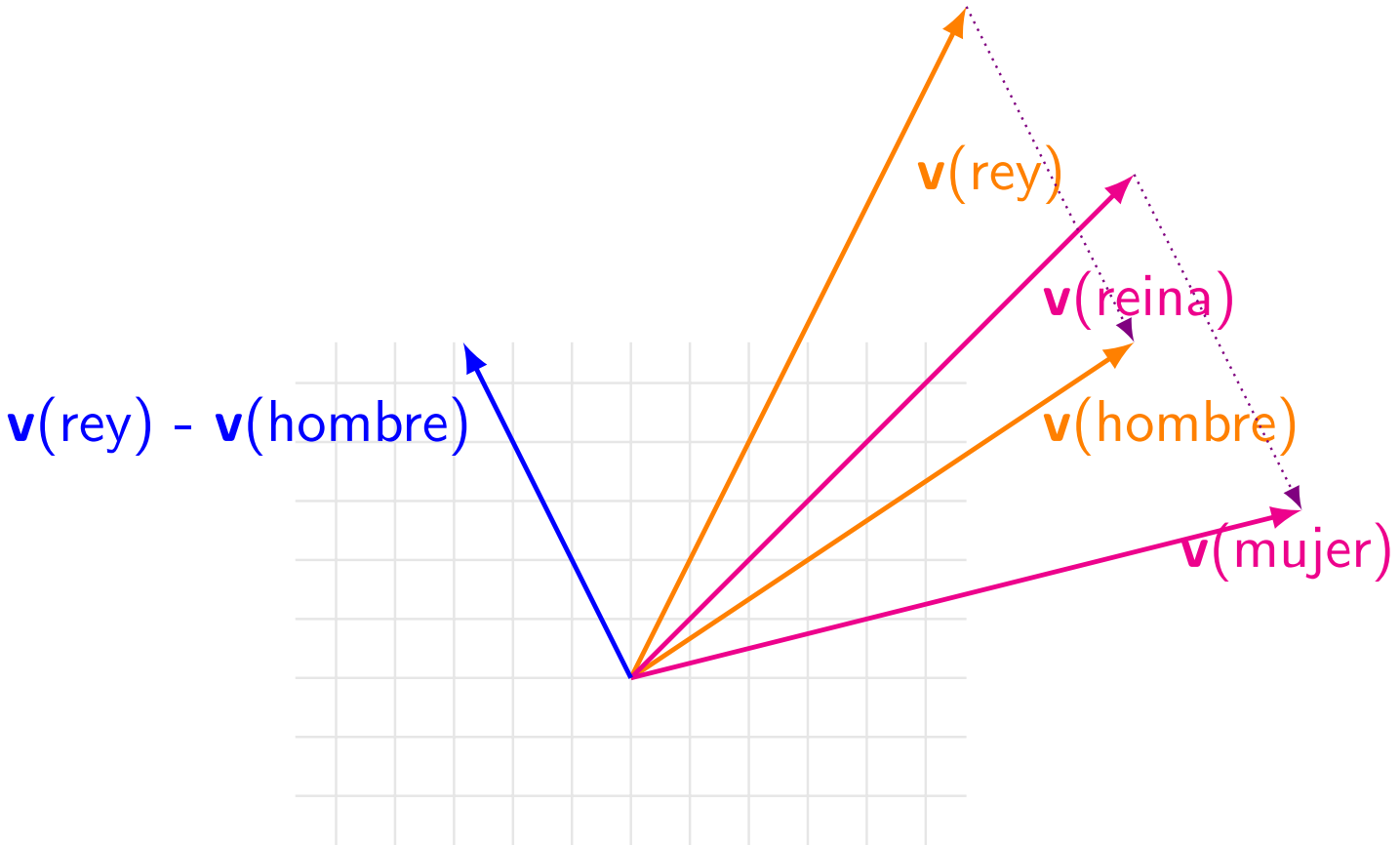
¡justamente es el coseno del ángulo entre los dos vectores!

Distancia entre palabras: similaridad coseno

- La distancia euclidiana no es adecuada si los corpus son de tamaños diferentes.
- $\text{sim}(v, w) = 0$ implica que v y w son ortogonales: son lo menos parecido posible.
- $\text{sim}(v, w) = 1$ implica que v y w son linealmente dependientes: son lo más similar posible.



Distancia entre palabras: inferencia



$$\text{v(rey)} - \text{v(hombre)} + \text{v(mujer)} \rightarrow \text{v(reina)}$$

Distancia entre palabras: inferencia

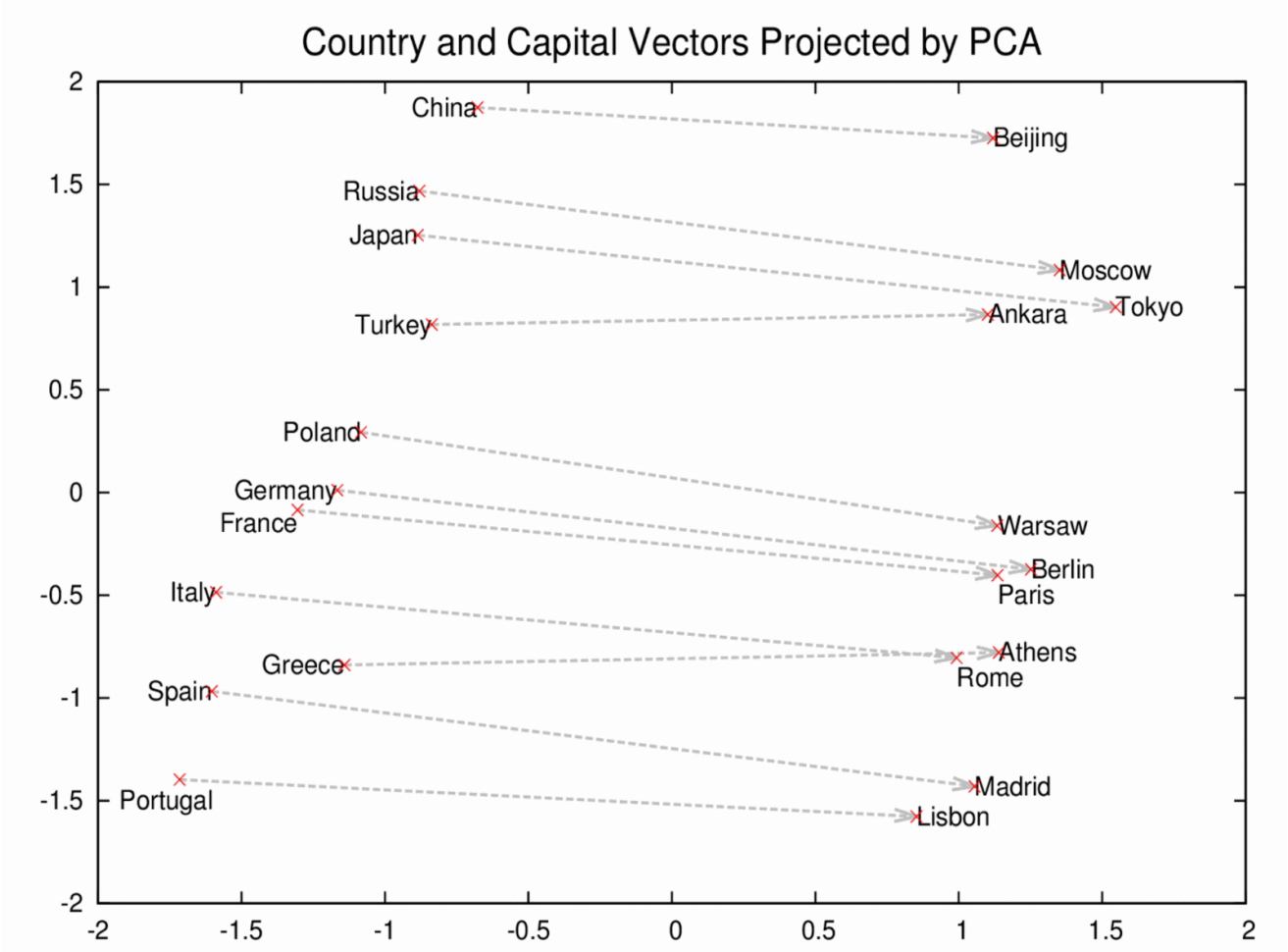


Figura tomada del artículo orgininal de Mikolov et. al (2014)

Visualización por PCA