Uso de modelos de NLP para el estudio del lenguaje en el cerebro

Clase 1: Representaciones vectoriales

Dr. Bruno Bianchi
Laboratorio de Inteligencia Artificial Aplicada
Dpto Computación - FCEN - UBA
Instituto Cs Computación - CONICET - UBA

[ELBA]

Julio 2024

Sobre el curso

¿Podemos usar los modelos de Inteligencia Artificial para comprobar hipótesis o generar intuiciones sobre cómo funciona el cerebro?

Sobre el curso

¿Podemos usar los modelos de Inteligencia Artificial para comprobar hipótesis o generar intuiciones sobre cómo funciona el cerebro?

- Los modelos de IA funcionan cada vez mejor

Sobre el curso

¿Podemos usar los modelos de Inteligencia Artificial para comprobar hipótesis o generar intuiciones sobre cómo funciona el cerebro?

- Los modelos de IA funcionan cada vez mejor
- ¿Funcionan este modelos como el cerebro?
  - Nivel Comportamental
  - Nivel Neuronal

Sobre el curso

¿Podemos usar los modelos de Inteligencia Artificial para comprobar hipótesis o generar intuiciones sobre cómo funciona el cerebro?

- Los modelos de IA funcionan cada vez mejor
- ¿Funcionan este modelos como el cerebro?
  - Nivel Comportamental
  - Nivel Neuronal
- Cada vez más trabajos trabajando con esto
  - Veremos algunos

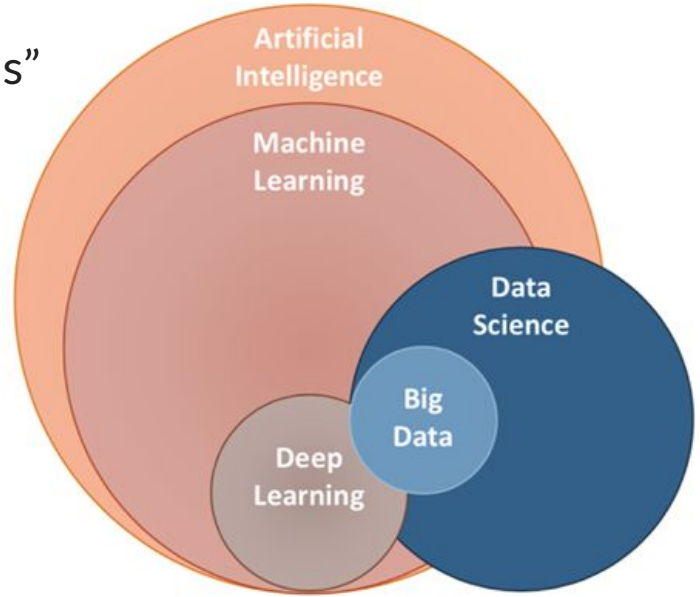
Objetivos de esta clase

- Presentar la necesidad de representar las palabras de forma numérica
- Presentar el concepto de representaciones vectoriales
- Intro de LSA y word2vec como formas de generar embedding
- Cerrar con RNN como modelo de lenguaje?

No es todo lo mismo

Inteligencia Artificial:

Que la computadora resuelva tareas “humanas”



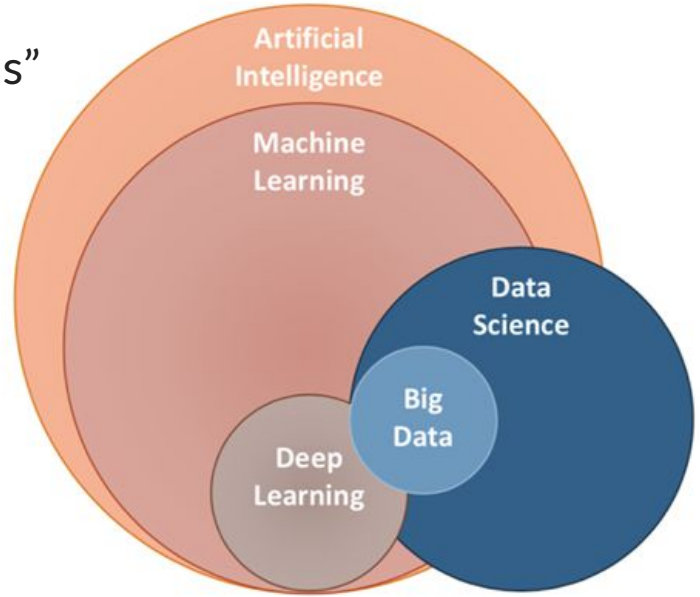
No es todo lo mismo

Inteligencia Artificial:

Que la computadora resuelva tareas “humanas”

Aprendizaje Automático:

Que lo haga aprendiendo de datos



No es todo lo mismo

Inteligencia Artificial:

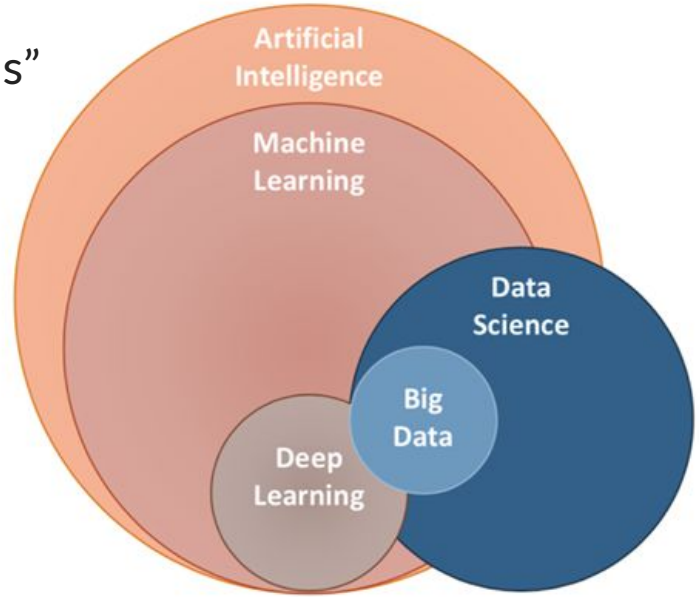
Que la computadora resuelva tareas “humanas”

Aprendizaje Automático:

Que lo haga aprendiendo de datos

Aprendizaje Profundo:

Que lo haga usando redes profundas



No es todo lo mismo

Inteligencia Artificial:

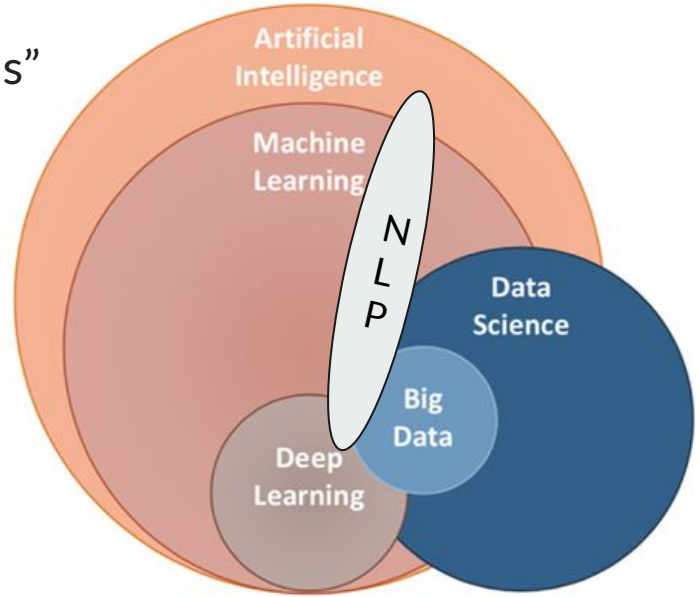
Que la computadora resuelva tareas “humanas”

Aprendizaje Automático:

Que lo haga aprendiendo de datos

Aprendizaje Profundo:

Que lo haga usando redes profundas



Procesamiento del Lenguaje Natural:

Área de IA que busca sistemas que entiendan lenguaje natural

El problema del lenguaje en la computadora

El problema del lenguaje en la computadora

¿25?

El problema del lenguaje en la computadora

¿felicidad?

amor — comida

Metiendo palabras en la compu

Si vamos a querer operar con palabras en una computadora, tenemos que expresar las palabras como a la computadora le gusta

Metiendo palabras en la compu

Si vamos a querer operar con palabras en una computadora, tenemos que expresar las palabras como a la computadora le gusta

Representación numérica (o vectorial) de las palabras:

- Darle a cada palabra un índice:
  - Alfabético?
  - Algún ordenamiento con criterio semántico?

Metiendo palabras en la compu

Si vamos a querer operar con palabras en una computadora, tenemos que expresar las palabras como a la computadora le gusta

Representación numérica (o vectorial) de las palabras:

- Darle a cada palabra un índice:
  - Alfabético?
  - Algún ordenamiento con criterio semántico?
- Vectorial (más dimensiones):
  - Dimensiones manuales
  - Dimensiones arbitrarias

Semántica distribucional

“You shall know a word by the company it keeps” (J. R. Firth 1957:11)

Semántica distribucional: entender el significado de las palabras de acuerdo a su contexto. Muy usado en NLP! (es la base de word2vec hasta GPT)

Usamos un largo conjunto de instancias donde aparezca una palabra para intentar ver con qué palabras se relaciona más

Las fintech y	**bancos**	digitales del país implementan cada vez…
Salvo ciertos	**bancos**	del ámbito público, todas las otras entidades financieras…
La decisión del BCRA de que los	**bancos**	no puedan realizar operaciones con criptomonedas

Inventando embeddings

	Valence	Arousal	Dominance
courageous	8.05	5.5	7.38
music	7.67	5.57	6.5
heartbreak	2.45	5.65	3.58
cub	6.71	3.95	4.24

Osgood et al. (1957) noticed that in using these 3 numbers to represent the meaning of a word, the model was representing each word as a point in a three-dimensional space, a vector whose three dimensions corresponded to the word's rating on the three scales. This revolutionary idea that word meaning could be represented as a point in space (e.g., that part of the meaning of *heartbreak* can be represented as the point $[2.45, 5.65, 3.58]$) was the first expression of the vector semantics models that we introduce next.

Inventando embeddings

¿Qué hacemos con estos vectores?

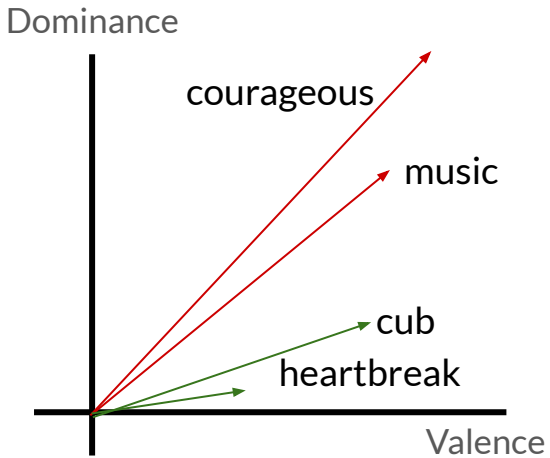
- Podemos operar matemáticamente
- La *similaridad coseno* nos da una métrica muy útil
  - Coseno de ángulo 0 = 1
  - Coseno de ángulo 90 grados = 0
  - Coseno de ángulo 180 grados = -1

Cosine similarity

$$cossim(\bar{v}_1, \bar{v}_2) = \cos(angle) = \frac{\bar{v}_1 . \bar{v}_2}{|\bar{v}_1| . |\bar{v}_2|}$$

$$cossim(\bar{v}_1, \bar{v}_2) \in [-1, 1]$$

	Valence	Arousal	Dominance
courageous	8.05	5.5	7.38
music	7.67	5.57	6.5
heartbreak	2.45	5.65	3.58
cub	6.71	3.95	4.24



Aprendiendo embeddings

Matriz términos-documentos:

- Tenemos un corpus con muchos documentos
- Cada documento tiene varias palabras
- Medimos la cantidad de las palabras en cada documentos (Bolsa de Pals - BoW)
- vemos cuando dos palabras tienen un *comportamiento similar* en el corpus

	Doc1	Doc2	Doc3	...	DocN-1	DocN
perro	4	0	2		0	0
la	23	32	23		21	22
de	13	21	12		21	17
gato	4	0	8		0	0
manzana	0	3	0		0	2
banana	0	7	0		0	5

Solución:

- Cada palabra es un vector
- Palabras parecidas, vectores parecidos

Problema:

- Muchas columnas
- Muchos ceros

Aprendiendo embeddings - LSA

A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge

Thomas K Landauer
University of Colorado at Boulder

Susan T. Dumais
Bellcore

How do people know as much as they do with as little information as they get? The problem takes many forms; learning vocabulary from text is an especially dramatic and convenient case for research. A new general theory of acquired similarity and knowledge representation, latent semantic analysis (LSA), is presented and used to successfully simulate such learning and several other psycholinguistic phenomena. By inducing global knowledge indirectly from local co-occurrence data in a large body of representative text, LSA acquired knowledge about the full vocabulary of English at a comparable rate to schoolchildren. LSA uses no prior linguistic or perceptual similarity knowledge; it is based solely on a general mathematical learning method that achieves powerful inductive effects by extracting the right number of dimensions (e.g., 300) to represent objects and contexts. Relations to other theories, phenomena, and problems are sketched.

Aprendiendo embeddings - LSA

Latent Semantic Analysis (1997):

- Bajamos las dimensiones de la matriz Términos-Documentos
- Se hace con la técnica SVD
- Históricamente se toman 300 dimensiones (columnas)
- Las dimensiones dejan de tener sentido

	Doc1	Doc2	Doc3	...	DocN-1	DocN
perro	4	0	2		0	0
la	23	32	23		21	22
de	13	21	12		21	17
gato	4	0	8		0	0
manzana	0	3	0		0	2
banana	0	7	0		0	5

SVD →

	d1	d2	...	d300
perro	15	1	...	24
la	9	50	...	31
de	7	2	...	5
gato	16	2	...	23
manzana	11	3	...	100
banana	10	7	...	98

Aprendiendo embeddings - LSA

Latent Semantic Analysis (1997):

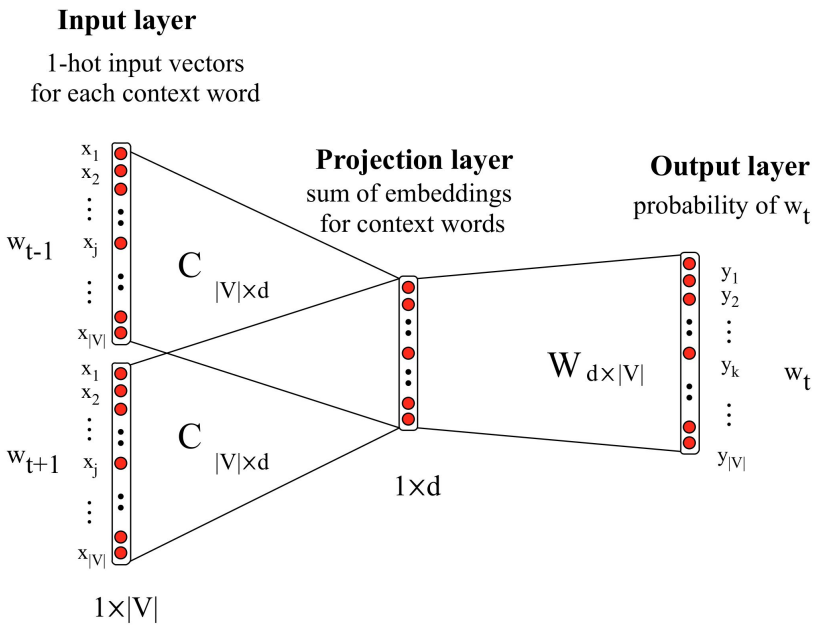
- Factorizar una matriz suele ser costoso (Si la matriz es de $\mathbb{R}^{m \times n}$ => costo O(m*n^2))
- Más o menos bueno para encontrar palabras similares
- No mejora la performance cuando los usamos en redes neuronales

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, *104*(2), 211.

Aprendiendo embeddings - w2v

word2vec (2013):

- Uso de redes neuronales
- Mayor complejidad
- También es bolsa de palabras

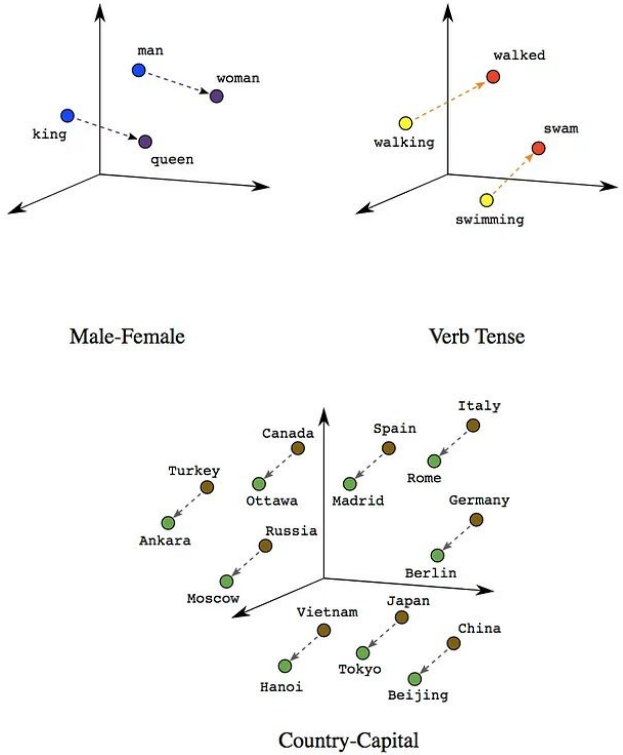


Mikolov et al. (2013) Distributed Representations of Words and Phrases and their Compositionality Jurafsky & Martin (2015) Speech and Language. Processing. 2nd edition

Aprendiendo embeddings - w2v

word2vec (2013):

- Uso de redes neuronales
- Mayor complejidad
- También es bolsa de palabras
- Aritmética de embeddings



Mikolov et al. (2013) Distributed Representations of Words and Phrases and their Compositionality Jurafsky & Martin (2015) Speech and Language. Processing. 2nd edition

Aprendiendo embeddings - w2v

word2vec (2013):

- Uso de redes neuronales
- Mayor complejidad
- También es bolsa de palabras
- Aritmética de embeddings
- Sus embeddings sirven para inicializar redes neuronales más complejas

Mikolov et al. (2013) Distributed Representations of Words and Phrases and their Compositionality Jurafsky & Martin (2015) Speech and Language. Processing. 2nd edition

Resumen

Hoy vimos:

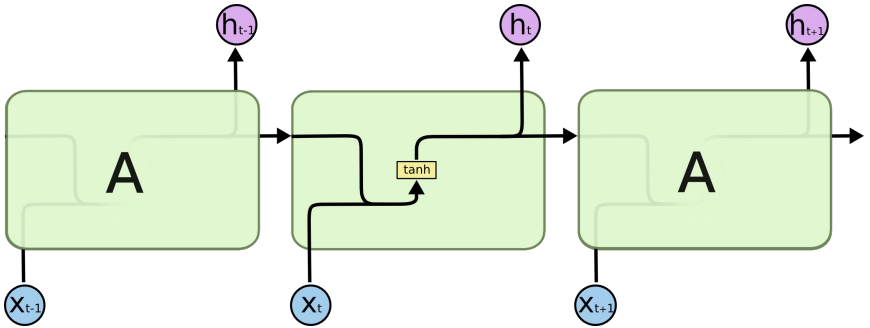
- Intruducción a IA-ML-DL-NLP
- Concepto de representación vectorial de palabras (embeddings)
- Generación manual de embeddings
- Generación automática de embeddings
    - LSA: conteo de palabras
    - Word2Vec: Primeras Redes neuronales

Hasta mañana!

Modelos de Lenguaje - RNN

Redes Neuronales Recurrentes (2016):

- Mucho más complejas que w2v
- Toman el texto como secuencia
- Mucho más que embeddings
- Permiten generar texto
- Vainilla, LSTM, GRU



Goodfellow, Bengio, and Courville (2016) *Deep learning*.

Semántica distribucional

La semántica de una palabra puede deducirse de su contexto

- Sarabaraban aparece de noche
- Cuando hay un peligro aparece sarabaraban
- Sarabaraban puede ayudarte
- Sarabaraban es un científico

“You shall know a word by the company it keep” J. R. Firth 1957