Escuela de verano

Universidad Autónoma del Estado de Morelos

Centro de Investigación en Ciencias

Laboratorio de Semántica Computacional

Presentación

David Torres Moreno

Licenciatura en Ciencias área terminal Computación

Maestría en Ciencias Cognitivas

Estudiante del Doctorado en Ciencias

Contenido

1. Modelado del lenguaje

2. N-grams

3. Modelo vectorial – Bag of words

Modelado de lenguaje

Centro de Investigación en Ciencias

1. El niño corrió detrás de la _____.

2. El _____ ladró al cartero.

3. La _____ está podrida.

LABSEMCO Data Science

Modelado de lenguaje

Centro de Investigación en Ciencias

1. El niño corrió detrás de la **pelota**.

2. El **perro** ladró al cartero.

3. La **manzana** está podrida.

LABSEMCO Data Science

Modelado de lenguaje

¿Cómo modelamos este fenómeno en la computadora?

Modelado de lenguaje

Centro de Investigación en Ciencias

Enfoque estadístico

LABSEMCO Data Science

Modelado de lenguaje

Centro de Investigación en Ciencias

Hoy es martes. Mañana será _____.

Empezó a llover, me tendré que llevar mi _____.

LABSEMCO Data Science

Modelado del lenguaje

Centro de Investigación en Ciencias

Para calcular la probabilidad conjunta de un lanzamiento de dos monedas en el que podemos obtener águila en una (suceso x) y en la otra un sol (suceso y).

En este caso, la probabilidad del suceso x es del 50% y la probabilidad del suceso y también es del 50%. Entonces tenemos:

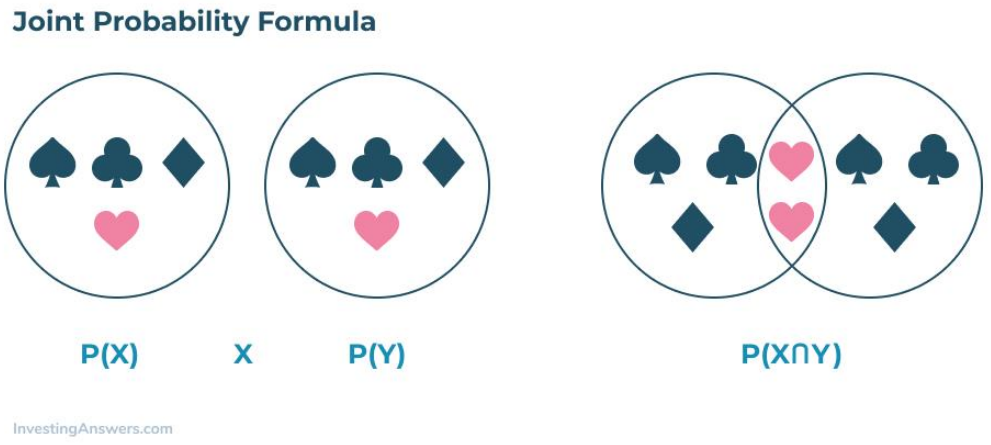
$$P(x=\text{“águila”}, y=\text{“sol”}) = P(x)\,P(y \mid x)$$
$$= P(y)\,P(x \mid y)$$

LABSEMCO Data Science

Modelado del lenguaje

Centro de Investigación en Ciencias

$$P(x,y) = P(x)\,P(y|x)$$



$$P(y|x)=P(y)$$

$$P(y|x)$$

LABSEMCO Data Science

Modelado del lenguaje

Centro de Investigación en Ciencias

Para medir la probabilidad conjunta, ambos sucesos deben ocurrir al mismo tiempo y deben ser independientes entre sí.

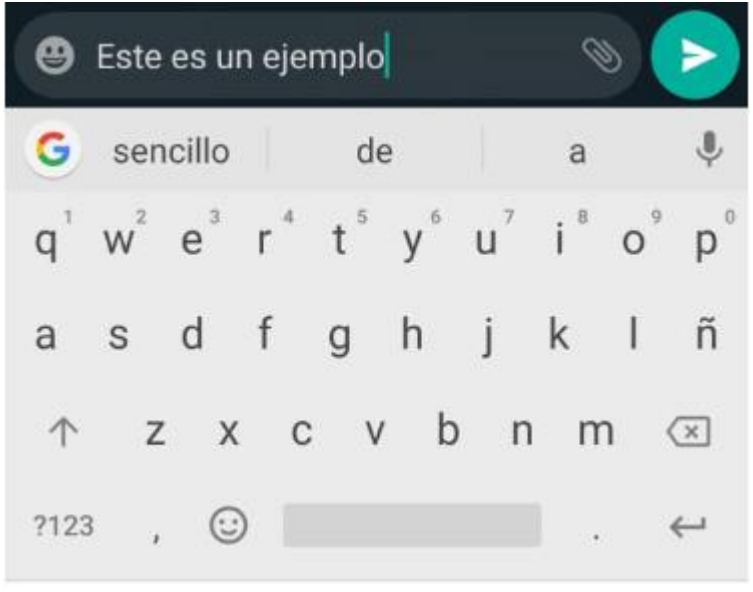
Esto significa que el resultado de un suceso no puede afectar al impacto del resultado del otro suceso.

El objetivo del modelado estadístico del lenguaje, es aprender la función de probabilidad conjunta de las secuencias de palabras de una lengua.

LABSEMCO Data Science

Modelado del lenguaje

El modelo de lenguaje es una parte fundamental de varias tareas de NLP, como reconocimiento de lenguaje hablado, reconocimiento de lenguaje escrito, traducción automática, corrección de ortografía, sistemas de predicción de escritura, etc.



Modelado del lenguaje

Centro de Investigación en Ciencias

Frase= “Today is wednesday”

x = “Today”

y = “is”

z = “Wednesday”

$P(x, y, z) = P(x)\ P(y)\ P(z)$?

LABSEMCO Data Science

Modelado del lenguaje

El modelo estadístico del lenguaje es representado por la probabilidad condicional de la siguiente palabra, dadas las anteriores (contexto).

Es decir, para una palabra W_t (que es la t-ava palabra) de la secuencia:

$$\hat{P}(w_1^T) = \prod_{t=1}^{T} \hat{P}(w_t | w_1^{t-1}),$$

Una de las ventajas al construir un modelo estadístico de lenguaje es el orden de las palabras, y que las palabras contiguas son más dependientes entre ellas.

Modelado del lenguaje

Centro de Investigación en Ciencias

Frase= “Today is wednesday”

x = “Today”

y = “is”

z = “Wednesday”

$$P(x, y, z) = P(x)\, P(y \mid x)\, P(z \mid x\, y)$$

LABSEMCO Data Science

Modelado del lenguaje

Centro de Investigación en Ciencias

Esto es intrínsecamente difícil debido al problema de la **dimensionalidad.**

Por ejemplo, si se quiere modelar la distribución conjunta de 10 palabras consecutivas en un lenguaje natural, con un vocabulario V de tamaño 100,000, hay potencialmente $100{,}000^{10} - 1 = 10^{50} - 1$ parámetros libres.

LABSEMCO Data Science

Modelado del lenguaje

Una secuencia de palabras con la que se va a probar el modelo, es probable que sea diferente de todas las secuencias de palabras vistas en un corpus.

La mayoría de las posibles secuencias de palabras no serán observadas en el corpus. Una solución es hacer la hipótesis de que la probabilidad sea posible.

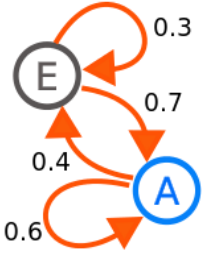
Los enfoques tradicionales y muy exitosos para un primer acercamiento del modelado del lenguaje, son los de n-gramas, estos obtienen una generalización del lenguaje concatenando secuencias cortas (n) y vistas en el corpus.

Corpus: un conjunto de textos.

N-grams

Centro de Investigación en Ciencias

Cadena de Márkov o modelo de Márkov a un tipo especial de proceso estocástico discreto en el que la probabilidad de que ocurra un evento depende solamente del evento inmediatamente anterior.



$$P(w_1, \ldots, w_m) = \prod_{i=1}^{m} P(w_i | w_1, \ldots, w_{i-1}) \approx \prod_{i=1}^{m} P(w_i | w_{i-(n-1)}, \ldots, w_{i-1})$$

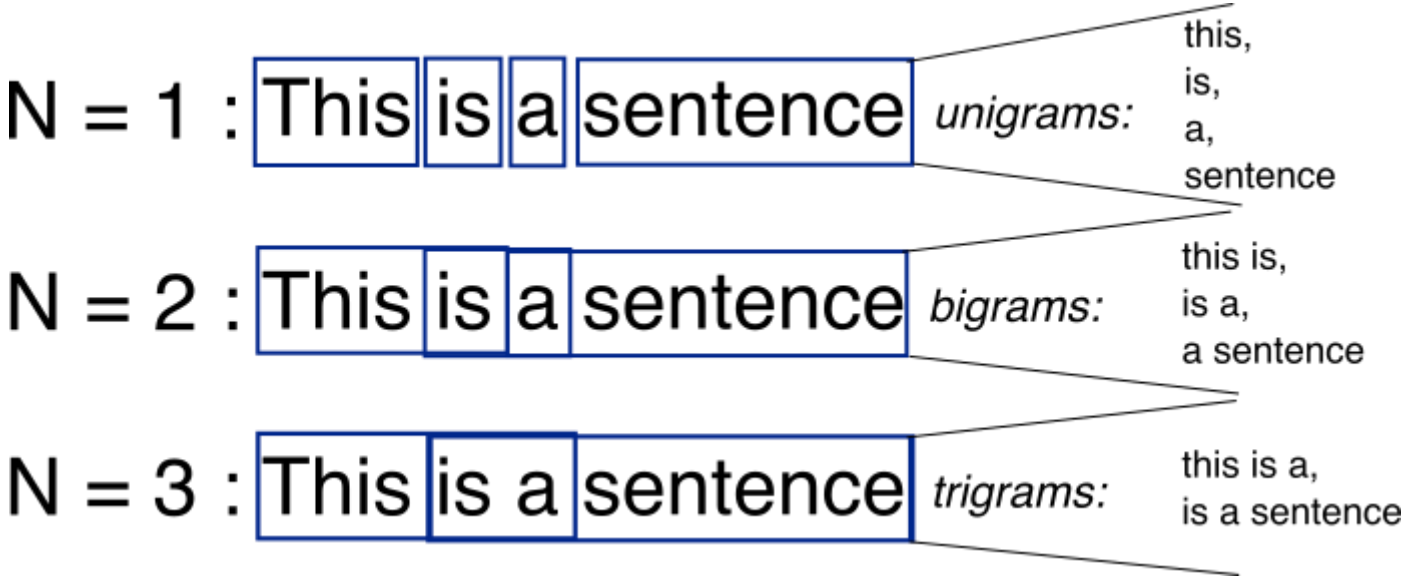
$$\hat{P}(w_t | w_1^{t-1}) \approx \hat{P}(w_t | w_{t-n+1}^{t-1}).$$

LABSEMCO Data Science

N-grams

Centro de Investigación en Ciencias

Un n-grama es una subsecuencia de n elementos de una secuencia dada.



LABSEMCO Data Science

N-grams

Centro de Investigación en Ciencias

Frase: *I saw the red house*

Bigrama

$$P(I, saw, the, red, house) \approx P(I|<s>)P(saw|I)P(the|saw)P(red|the)P(house|red)P(</s>|house)$$

Trigrama

$$P(I, saw, the, red, house) \approx P(I|<s>,<s>)P(saw|<s>,I)P(the|I, saw)P(red|saw, the)P(house|the, red)P(</s>|red, house)$$

LABSEMCO Data Science

N-grams

Centro de Investigación en Ciencias

Estos modelos usan un corpus y se puede obtener una probabilidad condicional para una frase de longitud *n*.

¿Cómo calculamos la probabilidad de que ocurra la siguiente palabra w_i?

LABSEMCO Data Science

N-grams

Centro de Investigación en Ciencias

¿Cómo calculamos la probabilidad de que ocurra la siguiente palabra w_i?

$$\text{Unigram LM} : p(w_1^N) = \prod_{n=1}^{N} p(w_n)$$

$$\text{Bigram LM} : p(w_1^N) = \prod_{n=1}^{N} p(w_n|w_{n-1})$$

$$\text{Trigram LM} : p(w_1^N) = \prod_{n=1}^{N} p(w_n|w_{n-2}, w_{n-1})$$

LABSEMCO Data Science

N-grams

Centro de Investigación en Ciencias

¿Cómo calculamos la probabilidad de que ocurra la siguiente palabra w_i?

$$P(w_i|w_{i-(n-1)},\ldots,w_{i-1}) = \frac{count(w_{i-(n-1)},\ldots,w_{i-1},w_i)}{count(w_{i-(n-1)},\ldots,w_{i-1})}$$

LABSEMCO Data Science

N-grams

Centro de Investigación en Ciencias

Pero que pasa para las frases compuestas que no se encuentran o no se ven en el corpus.

No se le puede asignar una probabilidad nula puesto que se pueden obtener en otros corpus.

¿Cómo tratar este problema?

LABSEMCO Data Science

N-grams

Palabras desconocidas

Podemos entrenar nuestro modelo con una palabra adicional que denota palabras desconocidas. Una estrategia es tomar las palabras con frecuencia baja y sustituirlas por <ukn> y establecer su probabilidad.

N-grams

Centro de Investigación en Ciencias

Laplace:

Este suavizado cambia la forma en que se calculan las probabilidades de los N-gramas, agregando un conteo a todos los N-gramas, incluso si no están en el corpus de entrenamiento, de esta forma evita que las probabilidades se hagan 0 para N-gramas que no estén en el corpus.

$$p(w_n|w_1^{n-1}) \approx \frac{C(w_n|w_1^{n-1}) + 1}{\sum_{w \in V} C(w|w_1^{n-1}) + |V|}$$

LABSEMCO Data Science

N-grams

Centro de Investigación en Ciencias

Backoff:

Aproxima la probabilidad de un N-grama no encontrado como la probabilidad de un (N-1)-grama similar, por ejemplo si se busca la probabilidad del N-grama:

$$P(corre|el, perro) \approx P(corre|perro)$$

$$P(w_n|w_1^{n-1}) \approx P(w_n|w_2^{n-1}) \approx \ ... \approx P(w_n|w_i^{n-1}) \approx \ ... \approx P(w_n)$$

LABSEMCO Data Science

Evaluación del modelo

La evaluación de modelos se hace calculando la perplejidad en una muestra de textos de prueba, que no fueron utilizados para entrenar los modelos.

La perplejidad es la medida más frecuentemente reportada en modelos de lenguaje. Mejores modelos tienen valores más bajos de perplejidad.

$$LP(W) = -\frac{1}{N}\sum_{i=1}^{N}\log P(w_i|w_{i-1})$$

N-grams

Centro de Investigación en Ciencias

¿Qué tipos de fenómenos lingüísticos se capturan en estas estadísticas?

LABSEMCO Data Science

Centro de Investigación en Ciencias

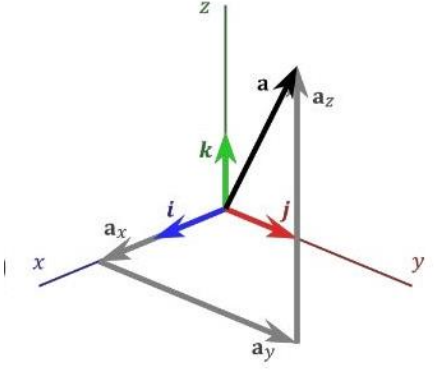
Modelo vectorial – Bag of words - TF-IDF

LABSEMCO Data Science

Modelo vectorial – Bag of words

Un espacio vectorial es una estructura matemática creada a partir de un conjunto no vacío, con una operación suma interna al conjunto y una operación producto externa entre dicho conjunto y un cuerpo, cumpliendo una serie de propiedades o requisitos iniciales.

A los elementos de un espacio vectorial se les llamará vectores y a los elementos del cuerpo se les llamará escalares.

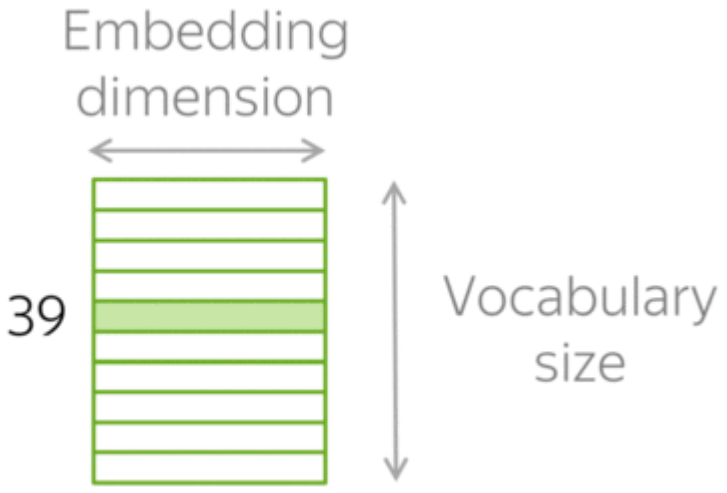


Modelo vectorial – Bag of words

Centro de Investigación en Ciencias

Queremos asignar un vector numérico a cada palabra del vocabulario de un texto.

El vector de características representa diferentes aspectos de la palabra: cada palabra se asocia a un punto en un espacio vectorial.



LABSEMCO Data Science

Modelo vectorial – Bag of words

Centro de Investigación en Ciencias

Codificaremos en un vector binario (vector de 0 y 1) de tamaño |V|, donde V es el vocabulario de palabras en un texto dado. Este vector tendrá ceros en todas partes excepto en el índice correspondiente al número que le asignamos a la palabra donde pondremos 1.

```
I think therefore I am.
```

```
I   think   therefore   am
1     2         3        4
```

```
             1  2  3  4
I           [1, 0, 0, 0]
think       [0, 1, 0, 0]
therefore   [0, 0, 1, 0]
am          [0, 0, 0, 1]
```

LABSEMCO Data Science

Modelo vectorial – Bag of words

- Un corpus es una colección de textos (o habla) del lenguaje que nos interesa.
- El vocabulario es una colección de palabras que ocurren en el corpus (o más general, en el lenguaje).
- La definición de palabra (token) depende de la tarea de NLP que nos interesa.

Modelo vectorial – Bag of words

Centro de Investigación en Ciencias

• Cada palabra está definida como una unidad separada por espacios o signos de puntuación.

•Los signos de puntuación pueden o no considerarse como palabras.

•Pueden considerarse palabras distintas las que tienen mayúscula y las que no.

•Pueden considerarse palabras distintas las que están escritas incorrectamente, o no.

LABSEMCO Data Science

Modelo vectorial – Bag of words

Centro de Investigación en Ciencias

•Pueden considerarse plurales como palabras distintas, formas en masculino/femenino, etc. (por ejemplo, en clasificación de textos quizá sólo nos importa saber la raíz en lugar de la forma completa de la palabra).

•Comienzos y terminación de oraciones pueden considerarse como “palabras” (por ejemplo, en reconocimiento de texto hablado).

LABSEMCO Data Science

Modelo vectorial – Bag of words

•Al proceso que encuentra todas las palabras en un texto se le llama **tokenización** o **normalización de texto**. Los **tokens** de un texto son las ocurrencias en el texto de las palabras en el vocabuario.

Modelo vectorial – Bag of words

Centro de Investigación en Ciencias

Representación de texto y vectorización de palabras en documentos

	the	book	is	good	I	like	it	books	a	table
The book is good	1	1	1	1	0	0	0	0	0	0
I like it	0	0	0	0	1	1	1	0	0	0
I like good books	0	0	0	1	1	1	0	1	0	0
Book a table	0	1	0	0	0	0	0	0	1	1

LABSEMCO Data Science

Modelo vectorial – Bag of words

Centro de Investigación en Ciencias

De la misma forma se puede realizar una representación que codifique fragmentos de texto en lugar de palabras individuales en vectores basados en sus palabras constituyentes.

```
Document 1: I think therefore I am
Document 2: I love dogs
Document 3: I love cats
```

Asignamos a cada palabra un número único, pero en lugar de representar palabras con estos números, usamos la frecuencia correspondiente para construir una representación útil para un documento dado.

```
I   think   therefore   am   love   dogs   cats
1     2         3       4     5      6      7
```

```
I  think  therefore  am  love  dogs  cats
Document 1: [ 2,   1,      1,      1,    0,    0,    0 ]
Document 2: [ 1,   0,      0,      0,    1,    1,    0 ]
Document 3: [ 1,   0,      0,      0,    1,    0,    1 ]
```

LABSEMCO Data Science

Modelo vectorial – Bag of words

El tamaño de nuestros nuevos vectores de documentos todavía está determinado por el tamaño del vocabulario y esto acarrea el problema de alta dimensionalidad.

Aunado a esto, existe el problema de palabras que no están en el vocabulario.

El orden en las oraciones puede cambiar sustancialmente el significado, por lo que tratar las oraciones como una mera colección de palabras sin orden ni contexto da como resultado que documentos como “El perro comió comida” y “La comida comió perro” se representen con el mismo vector.

Modelo vectorial – Bag of words

La razón por la que el enfoque de la bolsa de palabras carece de contexto es porque trata las palabras como unidades atómicas independientes.

El contexto, por otro lado, normalmente no se puede determinar a partir de una palabra, sino que surge en presencia de secuencias de palabras.

Otro enfoque consiste en n-gramas y, como antes, asignaremos un número único a cada elemento de vocabulario y representaremos los documentos con vectores que codifican las frecuencias de los elementos presentes en el documento.

Modelo vectorial – Bag of words

Centro de Investigación en Ciencias

```
Document 1: I think therefore I am
Document 2: I love dogs
Document 3: I love cats
```

Si dividimos cada documento en 2 palabras contiguas (es decir, bigrama), obtenemos el siguiente vocabulario:

```
I think:          1
think therefore:  2
therefore I:      3
I am:             4
I love:           5
love dogs:        6
love cats:        7
```

Tendremos la siguiente representación vectorial:

```
1  2  3  4  5  6  7
Document 1: [ 1, 1, 1, 1, 0, 0, 0 ]
Document 2: [ 0, 0, 0, 0, 1, 1, 0 ]
Document 3: [ 0, 0, 0, 0, 1, 0, 1 ]
```

LABSEMCO Data Science

Modelo vectorial – Bag of words

Centro de Investigación en Ciencias

En lugar de asignar números arbitrarios, queremos asociar cada palabra del documento con algún tipo de puntuación de importancia o relevancia y representar el documento con un vector de estas puntuaciones.

$$TF\text{-}IDF(t,d) = TF(t,d) \times IDF(t)$$

$$TF(t,d) = \frac{(Number\ of\ occurrences\ of\ term\ \mathbf{t}\ in\ document\ \mathbf{d})}{(Total\ number\ of\ terms\ in\ the\ document)}$$

$$IDF(t) = \log_e \frac{(Total\ number\ of\ documents)}{(Number\ of\ documents\ with\ term\ \mathbf{t}\ in\ them)}$$

LABSEMCO Data Science

Modelo vectorial – Bag of words


```
Document 1: I think therefore I am
Document 2: I love dogs
Document 3: I love cats
```

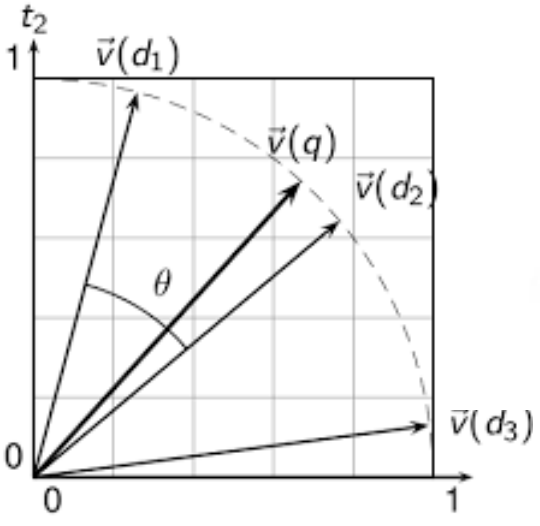
Representación de texto y la vectorización de palabras con TF-IDF.

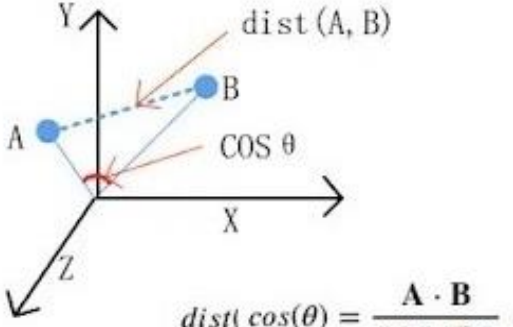
```
word    TF score      IDF score           TF-IDF score
I       1/3 = 0.33    log(3/3) = 0        0.33 * 0 = 0
love    1/3 = 0.33    log(3/2) = 0.18     0.33 * 0.18 = 0.059
dogs    1/3 = 0.33    log(3/1) = 0.48     0.33 * 0.48 = 0.158
I   think   therefore   am   love   dogs   cats
Document 2 vector: [0     0        0        0    0.059  0.158    0]
```

Information Retrieval System (IRS)

Centro de Investigación en Ciencias

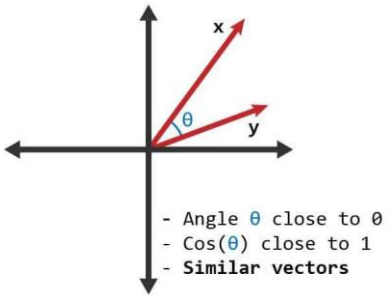
El modelo vectorial se basa en el grado de similaridad.

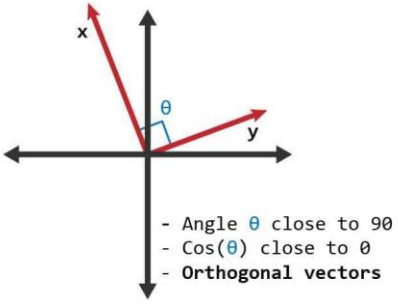


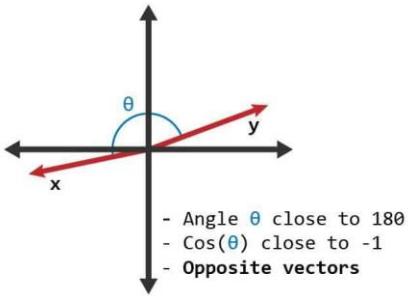


$$dist(\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \, \|\mathbf{B}\|}$$

$$Sim(q, d_i) = \frac{\sum_{j=1}^{m} t_{qj} \cdot t_{ij}}{\sqrt{\sum_{j=1}^{m} t_{qj}^2 \cdot \sum_{j=1}^{m} t_{ij}^2}}$$

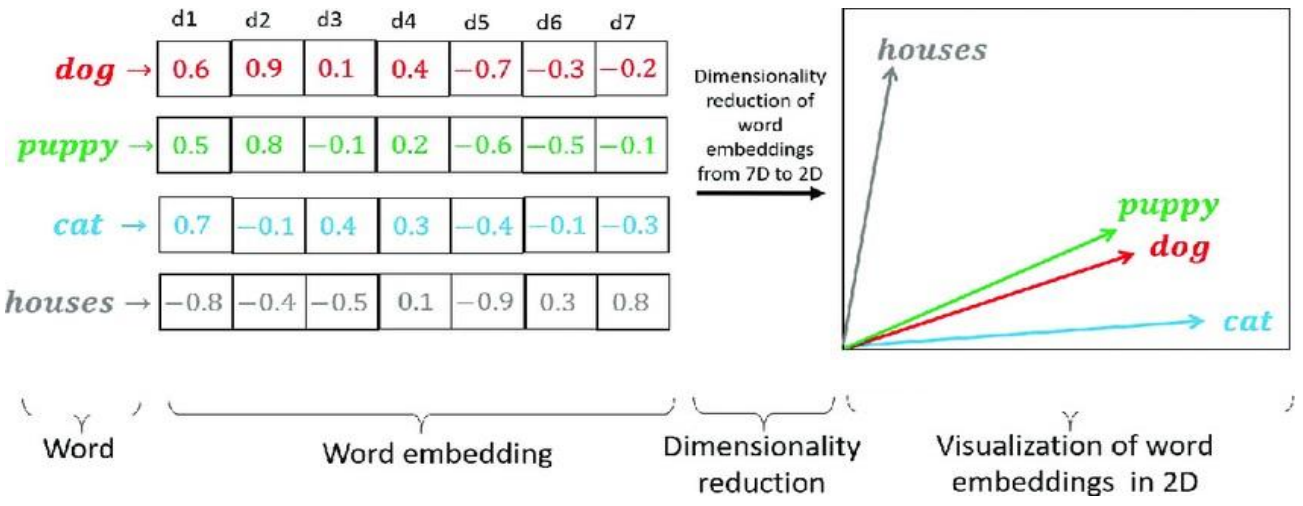






LABSEMCO Data Science

Reducción de dimensionalidad



¿Qué tipos de fenómenos lingüísticos se capturan estas representaciones?

Word embeddings

Centro de Investigación en Ciencias

Gracias

LABSEMCO Data Science

Referencias

- Jurafsky, D. (2018). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.
- Manning, C. (2000). *Foundations of Statistical Natural Language Processing*. London, England: The MIT Press.