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

4. Introducción a las representaciones incrustadas (word embeddings)

5. Métodos de reducción de dimensionalidad

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:

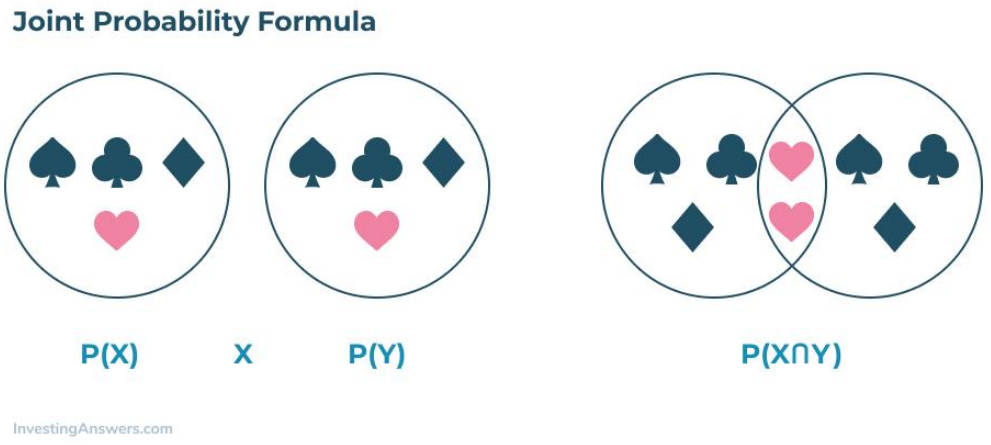
$$P(x=\text{“águila”}, y=\text{“sol”}) = P(x)\,P(y \mid x)$$
$$= P(y)\,P(x \mid y)$$

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

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

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.

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

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.

Un modelo estadístico del lenguaje es representado por la probabilidad condicional de la siguiente palabra dadas las anteriores.

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

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

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.

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.

LABSEMCO Data Science

Modelado del lenguaje

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

Así, los modelos de n-gramas construyen tablas de probabilidades condicionales para la siguiente palabra que queremos predecir, es decir combinaciones de las últimas n-1 palabras para cada contexto:

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

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

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

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*.

Pero que pasa para las frases compuestas que no se encuentran o no se ven en el entrenamiento. No se le puede asignar una probabilidad nula puesto que se pueden obtener en otros corpus.

LABSEMCO Data Science

N-grams

Centro de Investigación en Ciencias

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

LABSEMCO Data Science

N-grams

Centro de Investigación en Ciencias

Ejercicio en notebook

https://colab.research.google.com/drive/1SH3WSjzMVilwWZAwFXHf0eR2EfLrHmvH?usp=sharing

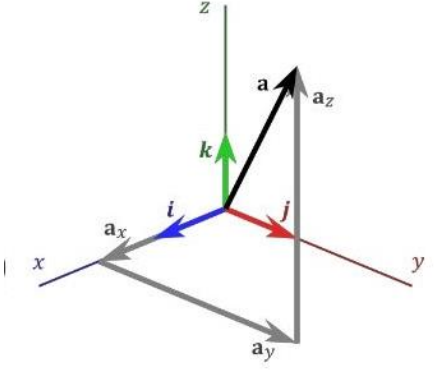
LABSEMCO Data Science

Modelo vectorial – Bag of words

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

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

Codificaremos el número 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

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, los usamos con su 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 ]
```

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 carecía de contexto era porque trataba 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

Si dividimos cada documento en trozos de 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

	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

Representación de texto y la vectorización de palabras.

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

Centro de Investigación en Ciencias

Representación de texto y la vectorización de palabras.

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.

```
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]
```

LABSEMCO Data Science

Modelo vectorial – Bag of words

Centro de Investigación en Ciencias

Palabras son semánticamente similares

Hombre – mujer **REY - Hombre + Mujer -> Reina**

Niño-hombre

El contexto juega un papel fundamental en la determinación de la similitud semántica.

Vectores resultantes serían mucho más representativos de sus palabras.

¿cómo se nos ocurren estos vectores?

LABSEMCO Data Science

Modelo vectorial – Bag of words

Centro de Investigación en Ciencias

1. ¿Es un ser humano?
2. Es masculino?
3. Es mujer?
4. ¿Es un gobernante estatal?
5. ¿Es inanimado?

```
word/context   1.    2.    3.    4.    5.
king         [ 0.9   1     0     1     0 ]
queen        [ 0.9   0     1     1     0 ]
man          [ 1     1     0     0.5   0 ]
woman        [ 1     0     1     0.5   0 ]
boy          [ 1     1     0     0.5   0 ]
girl         [ 1     0     1     0.5   0 ]
book         [ 0     0     0     0     1 ]
paper        [ 0     0     0     0     1 ]
```

$$king - man + woman = \begin{bmatrix} 0.9 \\ 1 \\ 0 \\ 1 \\ 0 \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0.5 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0.5 \\ 0 \end{bmatrix} = \begin{bmatrix} 0.9 \\ 0 \\ 1 \\ 1 \\ 0 \end{bmatrix} = queen$$

LABSEMCO Data Science

Word embeddings

Centro de Investigación en Ciencias

Proporcionan representaciones vectoriales de las palabras en las que estos vectores conservan la relación lingüística subyacente entre las palabras.

Estos vectores se calculan utilizando diferentes enfoques como las redes neuronales, la matriz de co-ocurrencia de palabras o las representaciones en términos del contexto en el que aparece la palabra. Algunas de las incrustaciones pre-entrenadas más utilizadas incluyen: word2vec, GloVe, fastText, BERT, etc.

LABSEMCO Data Science

Word embeddings

Centro de Investigación en Ciencias

Uno de los principales desafíos que enfrentan al despliegue los embeddings de palabras para medir la similitud es la deficiencia en la confluencia de significados.

Denota que los embeddings de palabras no atribuyen a los diferentes significados de una palabra, contaminando el espacio semántico con ruido al acercar las palabras irrelevantes entre sí.

Por ejemplo, las palabras ‘finance' y 'River' pueden aparecer en el mismo espacio semántico ya que la palabra ‘bank' tiene dos significados diferentes.

LABSEMCO Data Science

Word embeddings

Centro de Investigación en Ciencias

Es fundamental entender que los embeddings de palabras explotan la hipótesis distributiva para la construcción de vectores y dependen de grandes corpus, por lo tanto, se clasifican según los métodos de similitud semántica basados en Corpus.

Sin embargo, los métodos basados en la Deep-neural y la mayoría de los métodos de similitud semántica híbridos utilizan los embeddings de palabras para convertir los datos de texto a vectores de alta dimensión, y la eficiencia de estas integraciones desempeña un papel importante en el desempeño de los métodos de similitud semántica

LABSEMCO Data Science

N-grams

Centro de Investigación en Ciencias

Ejercicio en notebook

https://colab.research.google.com/drive/1SH3WSjzMVilwWZAwFXHf0eR2EfLrHmvH?usp=sharing

LABSEMCO Data Science

Word embeddings

Centro de Investigación en Ciencias

Gracias

LABSEMCO Data Science