



Facultad de  
Ingeniería

coursera

MIAD



Maestría  
en Inteligencia  
Analítica de Datos

La Universidad de los Andes como institución que almacena, y recolecta datos personales, requiere obtener su autorización para que de manera libre, previa, expresa, voluntaria y debidamente informada, permita a todas las dependencias académicas y/o administrativas, recolectar, recaudar, almacenar, usar, circular, suprimir, procesar, compilar, intercambiar, dar tratamiento, actualizar y disponer de los datos que han sido suministrados y que se han incorporado en distintas bases o bancos de datos, o en repositorios electrónicos de todo tipo con que cuenta la Universidad. Esta información es, y será utilizada en el desarrollo de las funciones propias de la Universidad en su condición de institución de educación superior, de forma directa o a través de terceros, a no ser que usted le manifieste lo contrario de manera directa, expresa, inequívoca y por escrito a la cuenta de correo electrónico dispuesta para tal efecto: [habeasdata@uniandes.edu.co](mailto:habeasdata@uniandes.edu.co) – Conozca más aquí: [www.uniandes.edu.co/datospersonales](http://www.uniandes.edu.co/datospersonales)

# MIAD



Maestría  
en Inteligencia  
Analítica de Datos

## Machine Learning y Procesamiento de Lenguaje Natural

**Semana 5:** Mayo 02 - 09 de 2023

# Embeddings

ASIN	Description
Adidas running shoes	These men's adidas running-inspired shoes are durable to stand up to everyday usage, and they're lightweight to help you make your next time out the best one yet. A Cloudfom Super midsole provides elevated cushioning, while the textile upper is super soft and comfy. The rubber outsole grips the ground because you never know when you might go off road.
Nike running shoes	Give strength to your step with the Nike Air Zoom Pegasus 37 shoe. Ensuring the fit loved by the runners, the shoe is equipped with a brand new cushioning unit on the forefoot and foam for maximum responsiveness. The result is a durable and lightweight shoe designed for your everyday runs. Nike React foam in the midsole for lightness, elasticity and durability. More foam means more cushioning, no bulk. The Air Zoom unit in the forefoot is twice as large as previous versions, thus offering greater elasticity at each step. It is closer to the foot for better responsiveness.

# Embeddings

ASIN	Description
Adidas running shoes	These men's adidas running-inspired shoes are durable to stand up to <b>everyday</b> usage, and they're <b>lightweight</b> to help you make your next time out the best one yet. A <b>Cloudfoam</b> Super midsole provides elevated cushioning, while the textile upper is super soft and comfy. The rubber outsole grips the ground because you never know when you might go off road.
Nike running shoes	Give strength to your step with the Nike Air Zoom Pegasus 37 shoe. Ensuring the fit loved by the runners, the shoe is equipped with a brand new cushioning unit on the forefoot and foam for maximum responsiveness. The result is a durable and <b>lightweight</b> shoe designed for your <b>everyday</b> runs. Nike React foam in the midsole for lightness, elasticity and durability. More foam means more <b>cushioning</b> , no bulk. The <b>Air Zoom</b> unit in the forefoot is twice as large as previous versions, thus offering greater elasticity at each step. It is closer to the foot for better responsiveness.

- Palabras idénticas: **lightweight**, **cushioning**, **everyday**
- Palabras con mismo significado en contexto particular: **AirZoom – CloudFoam**
- Frases con significado similar: “**to help you make your next time out the best one**” – “**ensuring the fit loved by the runners**”

## CountVectorizer

El embedding más sencillo es el conteo de palabras en el texto:

1. Crear un “corpus” (vocabulary) con todas las posibles palabras en el contexto
2. Cada dimensión representa un token
3. Un token puede ser una palabra o un n-grama
4. **Cada palabra la vamos a representar en el espacio como un one-hot encoding**
5. Un texto es la suma de las representaciones de las palabras que aparecen en él: el conteo de palabras

Vocabulario: {tomate, queso, jamón, lentejas, lechuga, pizza, hamburguesa, pasta}

tomate:	[ 1 , 0 , 0 , 0 , 0 , 0 , 0 , 0 , 0 ]
queso:	[ 0 , 1 , 0 , 0 , 0 , 0 , 0 , 0 , 0 ]
jamón:	[ 0 , 0 , 1 , 0 , 0 , 0 , 0 , 0 , 0 ]
lentejas:	[ 0 , 0 , 0 , 1 , 0 , 0 , 0 , 0 , 0 ]
lechuga:	[ 0 , 0 , 0 , 0 , 1 , 0 , 0 , 0 , 0 ]
pizza:	[ 0 , 0 , 0 , 0 , 0 , 0 , 1 , 0 , 0 ]

# CountVectorizer

El embedding más sencillo es el conteo de palabras en el texto:

1. Crear un “corpus” (vocabulary) con todas las posibles palabras en el contexto
2. Cada dimensión representa un token
3. Un token puede ser una palabra o un n-grama
4. **Cada palabra la vamos a representar en el espacio como un one-hot encoding**
5. Un texto es la suma de las representaciones de las palabras que aparecen en él: el conteo de palabras

Tres platos:

	tomate	queso	jamón	lentejas	lechuga	pizza	hamburg.	pasta
1. Pizza con tomate y queso:	[ 1 , 1 , 0 , 0 , 0 , 1 , 0 , 0 ]							
2. Pasta con jamón, tomate y queso:		[ 1 , 1 , 1 , 0 , 0 , 0 , 0 , 1 ]						
3. Hamburguesa de lentejas con lechuga:			[ 0 , 0 , 0 , 1 , 1 , 0 , 1 , 0 ]					
4. Queso campesino y queso mozzarella:				[ 0 , 2 , 0 , 0 , 0 , 0 , 0 , 0 ]				

## CountVectorizer

El embedding más sencillo es el conteo de palabras en el texto:

1. Crear un “corpus” (vocabulary) con todas las posibles palabras en el contexto
2. Cada dimensión representa un token
3. Un token puede ser una palabra o un n-grama
4. **Cada palabra la vamos a representar en el espacio como un one-hot encoding**
5. Un texto es la suma de las representaciones de las palabras que aparecen en él: el conteo de palabras

Tres platos:

- |   | tomate | queso | jamón | lentejas | lechuga | pizza | hamburg. | pasta |
|---|--------|-------|-------|----------|---------|-------|----------|-------|
| 1. Pizza con tomate y queso:            | [ 1 ,  | 1 ,   | 0 ,   | 0 ,      | 0 ,     | 1 ,   | 0 ,      | 0 ]   |
| 2. Pasta con jamón, tomate y queso:     | [ 1 ,  | 1 ,   | 1 ,   | 0 ,      | 0 ,     | 0 ,   | 0 ,      | 1 ]   |
| 3. Hamburguesa de lentejas con lechuga: | [ 0 ,  | 0 ,   | 0 ,   | 1 ,      | 1 ,     | 0 ,   | 1 ,      | 0 ]   |

Principal problema: da peso a palabras que en ciertos contextos pueden ser irrelevantes o contener poca información.

## TF-IDF (Term frequency – Inverse document frequency)

- Hay ciertas palabras que son más importantes que otras, dependiendo del contexto
- En restaurantes mexicanos, un plato que contiene “fríjoles” no es muy informativo – la mayoría de platos contienen fríjoles
- Un plato que contiene “sashimi” o “pad thai” es muy informativo – son palabras muy “escasas”
- TFIDF es el producto de:
  - Term frequency: qué tanto aparece una palabra en un texto

$$tf(t, d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}},$$

- Inverse document frequency: qué % de los textos contienen la palabra

$$idf(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

- Pizza con tomate y queso: [ 1/3\*ln(3/2) , 1/3\*Ln(3/2) , ... ]

# CONTÁCTANOS

PÁGINAS WEB - Ingeniería Industrial:

<https://industrial.uniandes.edu.co/es/programas-academicos/maestrias>

CORREO Maestría: [solicitudes-miad@uniandes.edu.co](mailto:solicitudes-miad@uniandes.edu.co)

## OTROS ENLACES DE IMPORTANCIA

Matrículas: <https://matriculas.uniandes.edu.co/>

Registro: <http://registro.uniandes.edu.co/>

Bloque Neón: <http://bloqueneon.uniandes.edu.co/>



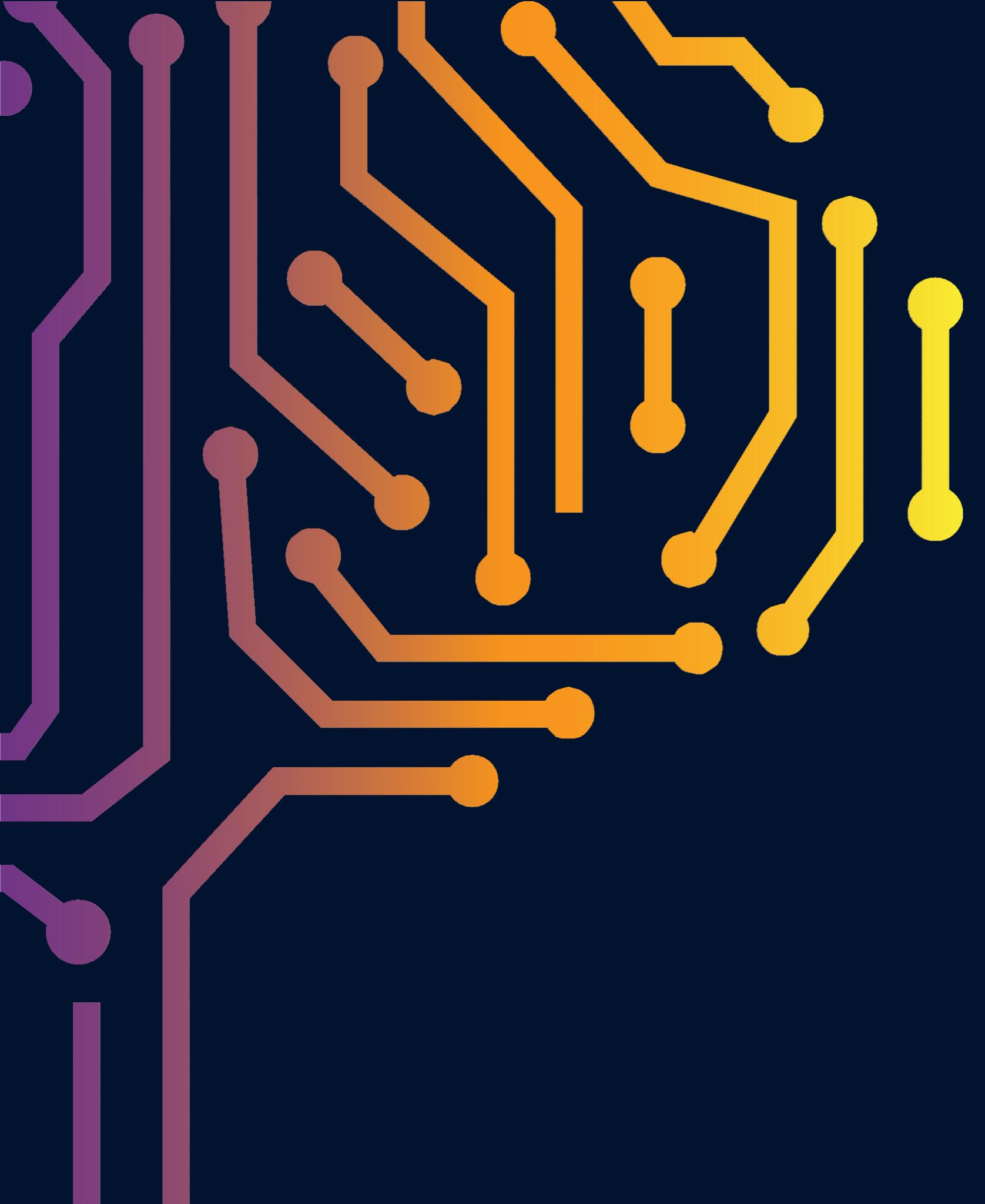
AnalyticsUA



AnalyticsUA



AnalyticsUA



MIAD

