Análisis de datos

Brayan Camilo Rodríguez Díaz
Juliana Alejandra Nieto Cárdenas
David Camilo Cortes Salazar
Juan Manuel De La Torre Sánchez
Jose Leonardo Guavita Hernandez

Ciencia de datos

La **ciencia de datos** es un campo interdisciplinario que involucra métodos científicos, procesos y sistemas para extraer conocimiento o un mejor entendimiento de datos

Análisis de datos

Consiste en inspeccionar y transformar datos con el objetivo de resaltar información útil, para sugerir conclusiones y apoyo en la toma de decisiones

Minería de datos

Es el análisis automático o semi-automático de grandes cantidades de datos para extraer patrones

Permite organizar información por relevancia y disipar ruido en los datos. Es comúnmente utilizado para machine learning e inteligencia artificial

AlphaStar DeepMind

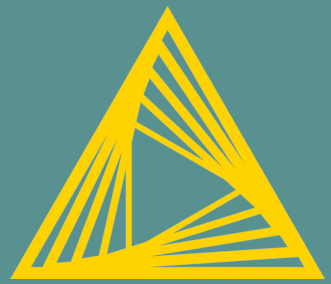
AlphaStar analizó hasta 200 años de repeticiones de la comunidad (aprox. 2’000.000)

Por lo tanto su estilo de juego recuerda al de un humano

Herramientas de análisis de datos

splunk>

Knime



Es una plataforma de código abierto de minería de datos que permite el desarrollo de modelos en un entorno visual

Big Data

Big Data es un conjunto de tecnologías que permiten la recopilación, almacenamiento, gestión, análisis y visualización, potencialmente en condiciones de tiempo real, de grandes conjuntos de datos con características heterogéneas.
El concepto de Big Data se explica mediante 5 características principales conocidas como las ‘5 vs’ (velocidad, volumen, variedad, veracidad y valor).

Características del Big Data

VOLUMEN

• Se refiere al gran volumen de información que es generada cada segundo, minuto , dias en nuestro entorno, es la característica más asociada al big data.

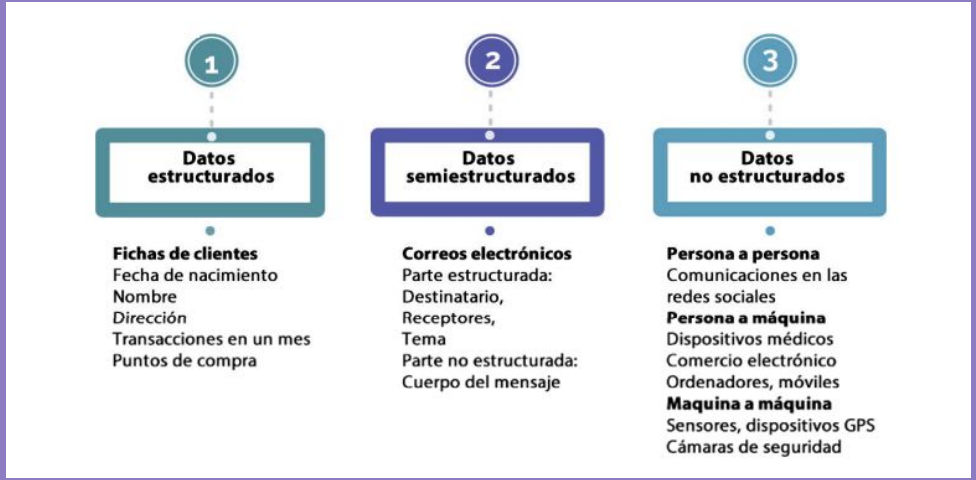
VELOCIDAD

•La velocidad se refiere a los datos en movimiento por los constantes interconexiones que realizamos, es decir, a la rapidez en la que son creados, almacenados y procesados en tiempo real.

Características del Big Data

VARIEDAD

• Necesidad de agregar información procedente de una amplia variedad de fuentes de información independientes: redes sociales, sensores, máquinas o personas individuales.



Características del Big Data

• **VALOR**
• El valor de los datos está en que sean accionables, es decir, que los responsables de las organizaciones puedan tomar una decisión con base a esos datos

VERACIDAD
• Es probable que debido al gran volumen de datos que recibimos, algunos lleguen incompleto. Y es que todo lo que recibimos de Internet y sobre todo de las redes sociales no es fiable.
Por esa razón, es clave hacer un filtrado con la tecnología del *Big Data* de lo que puede ser falso o no

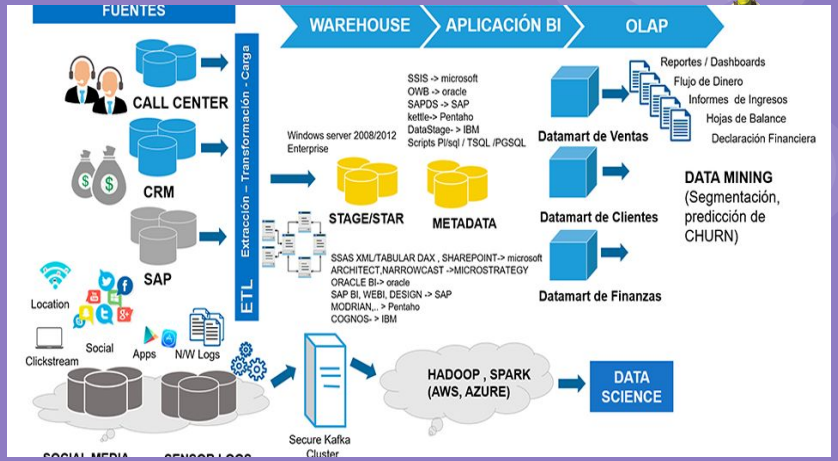
Conceptos básicos del Big Data

Big data y Business Intelligence

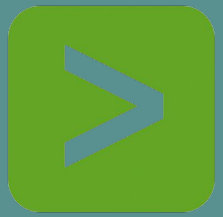
Big data y data warehouse

Big data y minería de datos

Big data y computación en la nube



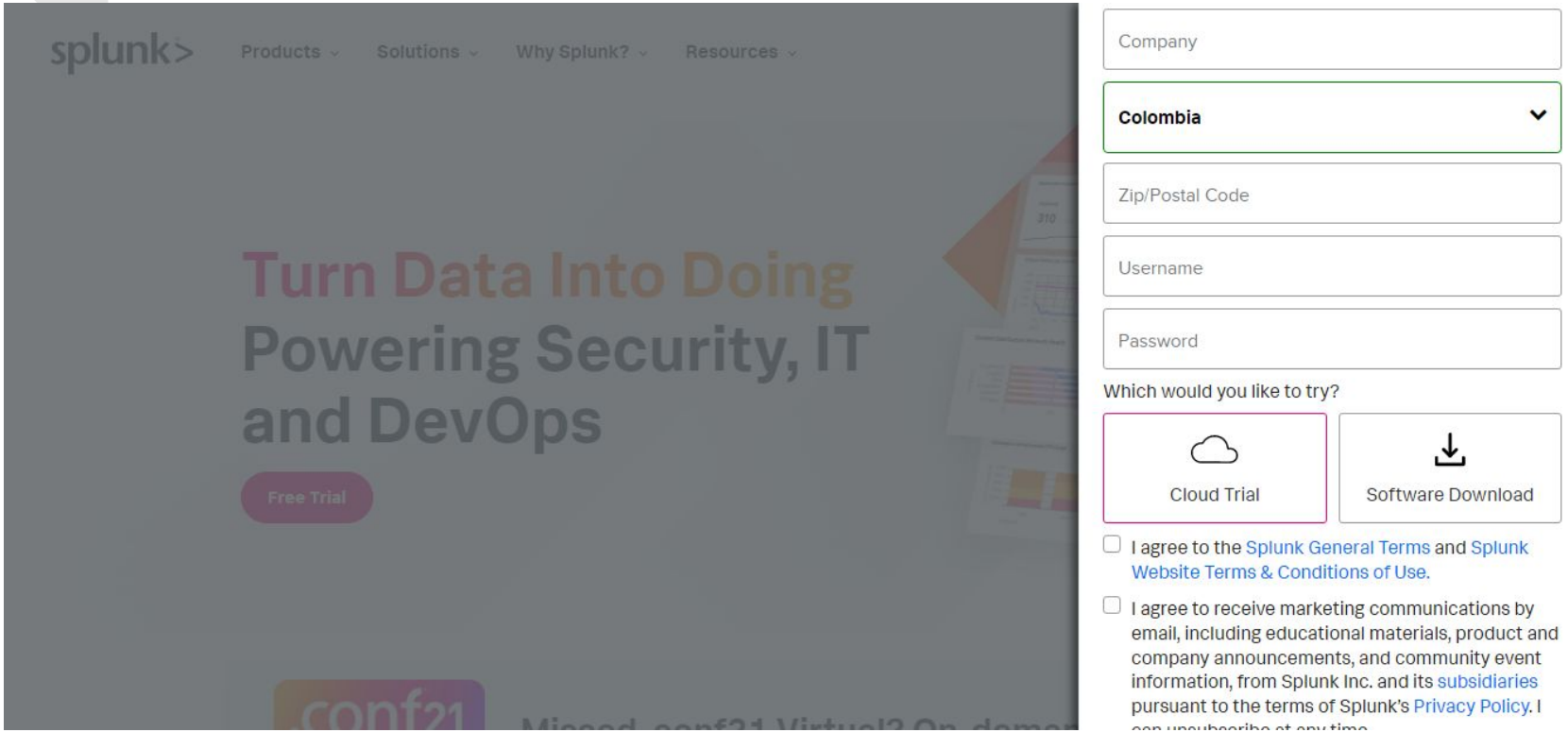
¿Qué es Splunk?



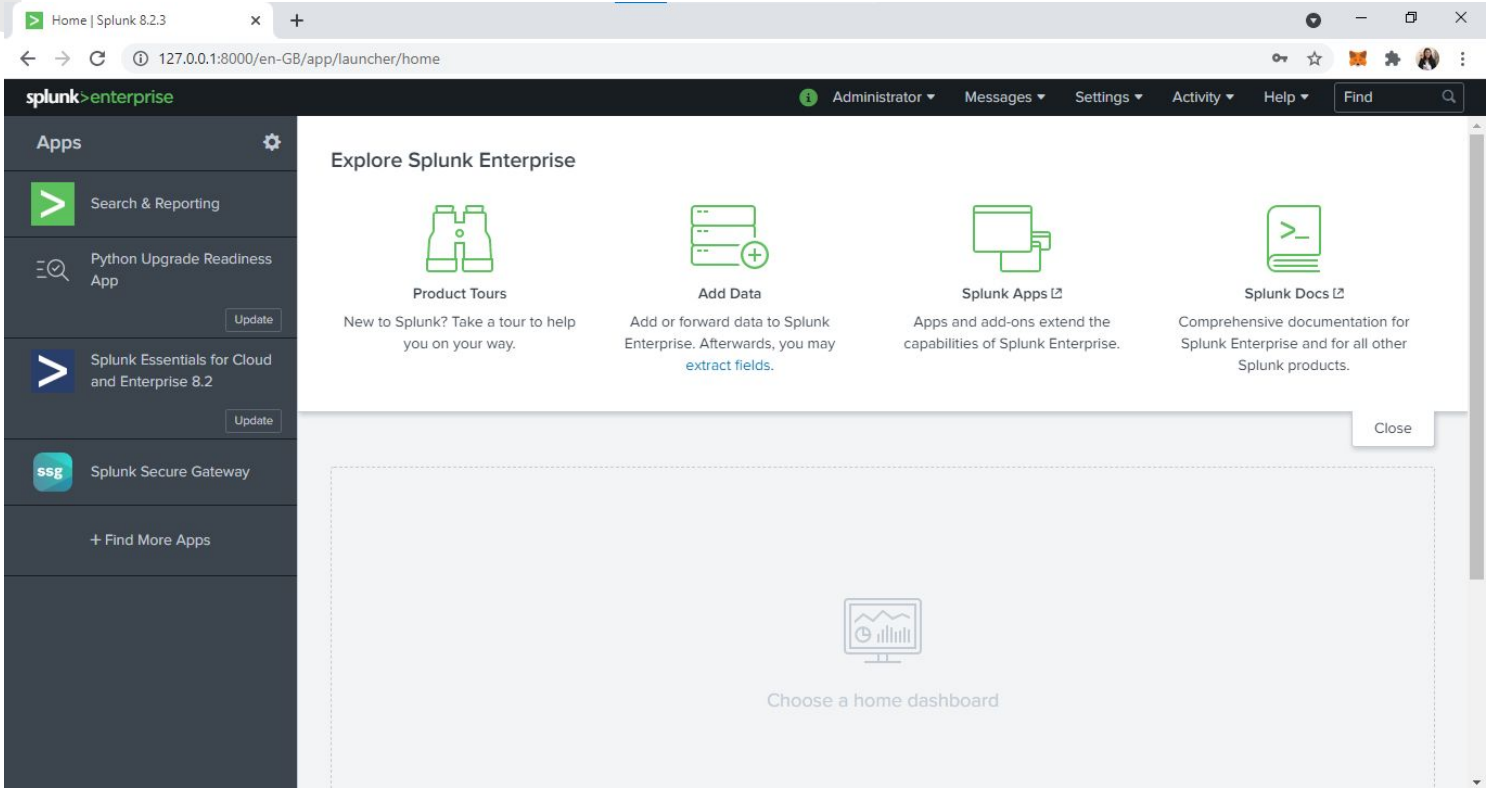
demo

Instalación

Instalación



Instalación



Usando Splunk Enterprise

- Datos
- Búsquedas
- Reportes y Alertas
- Visualizar

Datos

Datos

Buttercup Games

https://docs.splunk.com/Documentation/Splunk/8.2.3/SearchTutorial/Systemrequirements#Download_the_tutorial_data_files

Datos

tutorialdata.csv

Archivos de tipo log

prices.csv

Tabla con información sobre los productos: id, nombre, precio, etc

Subiendo los datos

- Segment in path: `\\(.*)\/`

Búsqueda

Búsqueda

Búsqueda Inicial

- Tabla de Eventos: índice, tiempo, evento
- Selected Fields
- Interesting Fields

```
buttercupgames
```

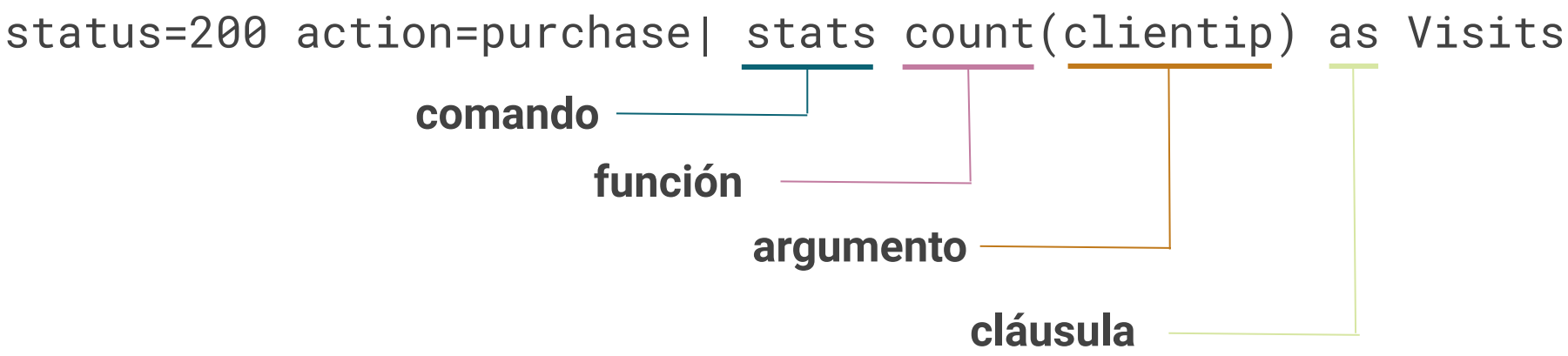
Búsqueda

Búsqueda usando Fields

```
sourcetype=access_* status=200 action=purchase
```

Búsqueda

Usando SPL (Splunk Search Language)



Búsqueda

Usando SPL (Splunk Search Language), **obteniendo el cliente más frecuente**

```
sourcetype=access_* status=200 action=purchase
| top limit=1 clientip
```

Búsqueda

Usando SPL, **subsearches**

```
sourcetype=access_* status=200 action=purchase
    [search sourcetype=access_* status=200 action=purchase
    | top limit=1 clientip
    | table clientip]
| stats count AS "Total Purchased",
distinct_count(productId) AS "Total Products",
values(productId) AS "Product IDs" by clientip
```

Reportes & Alertas

Reportes

Información sobre los productos comprados por el cliente más frecuente

```
sourcetype=access_* status=200 action=purchase [search sourcetype=access_* status=200
action=purchase | top limit=1 clientip | table clientip] | stats count AS "Total Purchased",
dc(productId) AS "Total Products", values(productName) AS "Product Names" BY clientip | rename
clientip AS "VIP Customer"
```

Alertas

Alertar cuando uno compra no pudo realizarse por problemas del servidor

```
sourcetype=access_* status=503 action=purchase
```

Visualizar

Gráficas + Dashboards

```
sourcetype=access_* status=200 | chart count AS views
count(eval(action="addtocart")) AS addtocart
count(eval(action="purchase")) AS purchases by productId
```

Playlist sobre Splunk

https://bit.ly/3GgLPzl

TRANSFORMERS Y MACHINE LEARNING

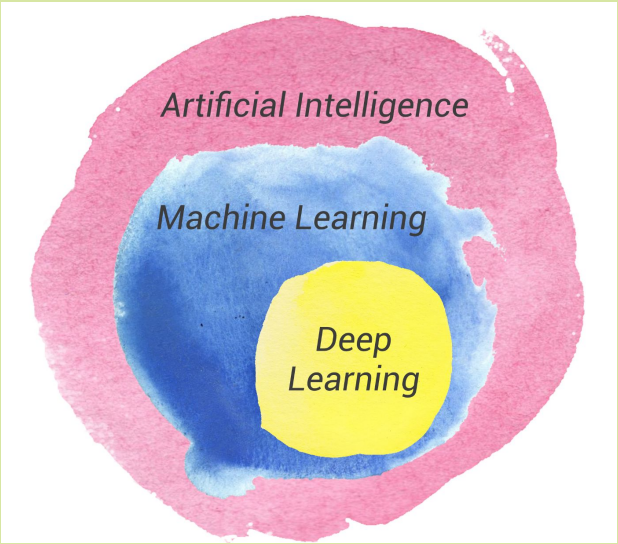
Antes de iniciar...

Redes Neuronales

Natural Language Processing

Embedding

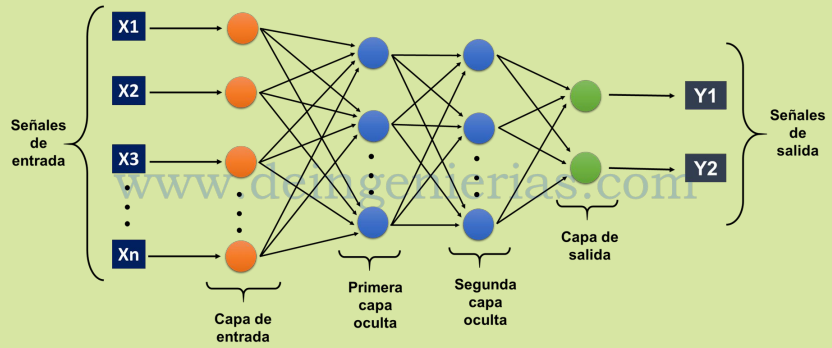
Mecanismos de Atencion



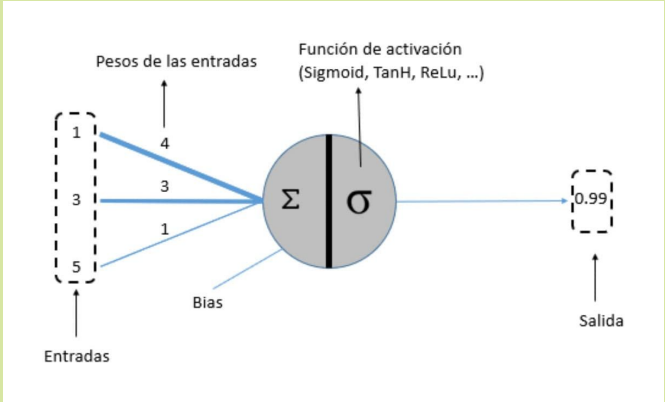
REDES NEURONALES

La red aprende examinando los registros individuales

Gráfico arquitectónico de un perceptrón multicapa con dos capas ocultas



Los círculos de colores representan a los nodos o neuronas



La red aprende sola con el backpropagation

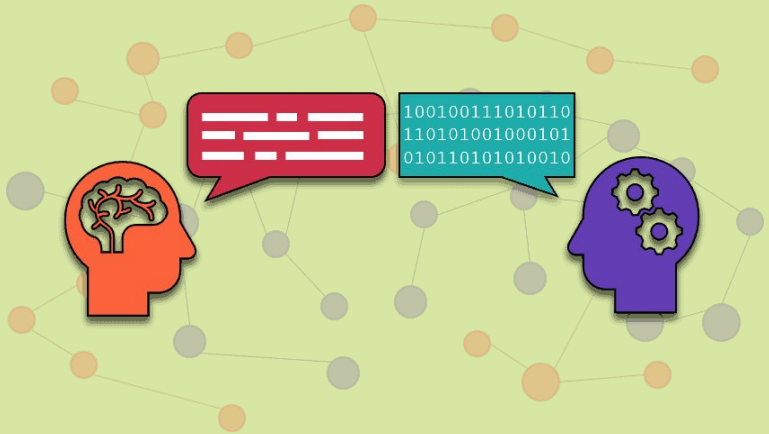
NATURAL LANGUAGE PROCESSING(NLP)

Las frases o textos se pueden subdividir de varias formas: caracteres, palabras o subpalabras , mejor conocidas como tokens

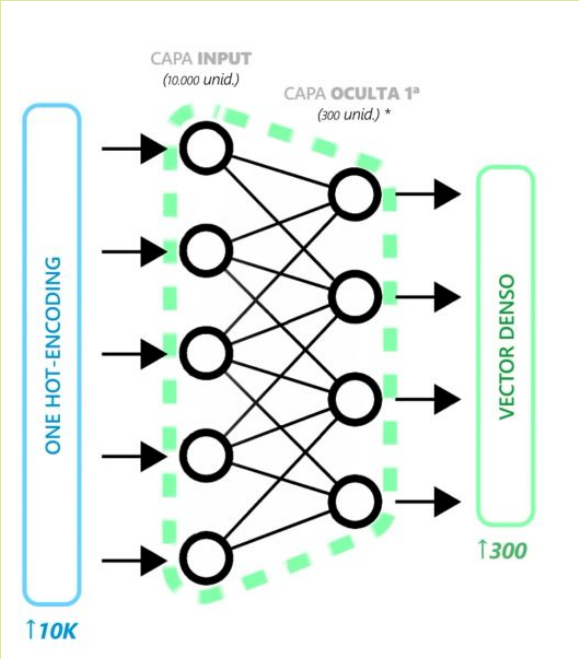
Método más sencillo: Hacer etiquetas de cada palabra y asignarle un numero

Metodo efectivo: Vectorización del Lenguaje

One Hot- Encoding

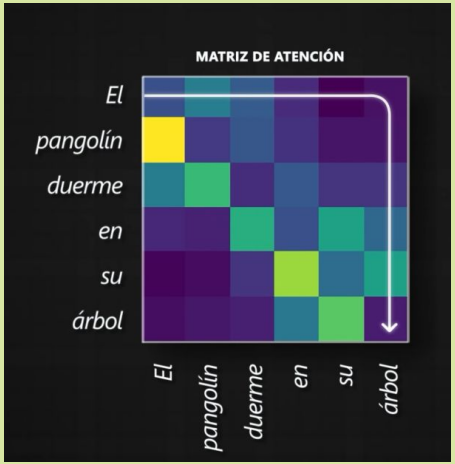


Embedding



Reduciremos la dimensionalidad de nuestros vectores, gracias a nuestra red neuronal

Para poder realizar los embeddings lo ideal es hacerlo con redes neuronales ya entrenadas con propósitos más básicos y menos específicos



Mecanismos de Atención

De una frase salen varios vectores, cada uno de una red neuronal diferente

Vector Query,Vector Key ,Vector valor

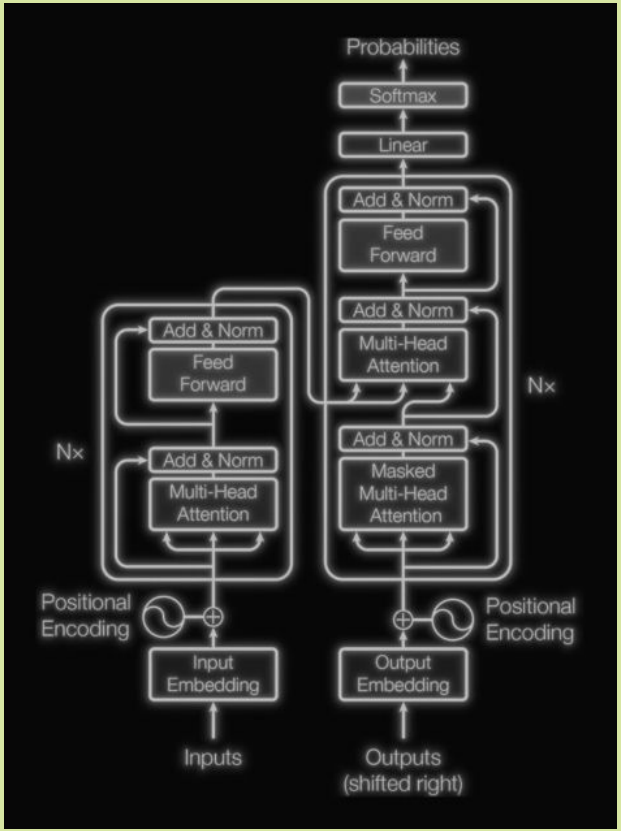
Producto punto entre los dos vectores, para conocer que tan similares son

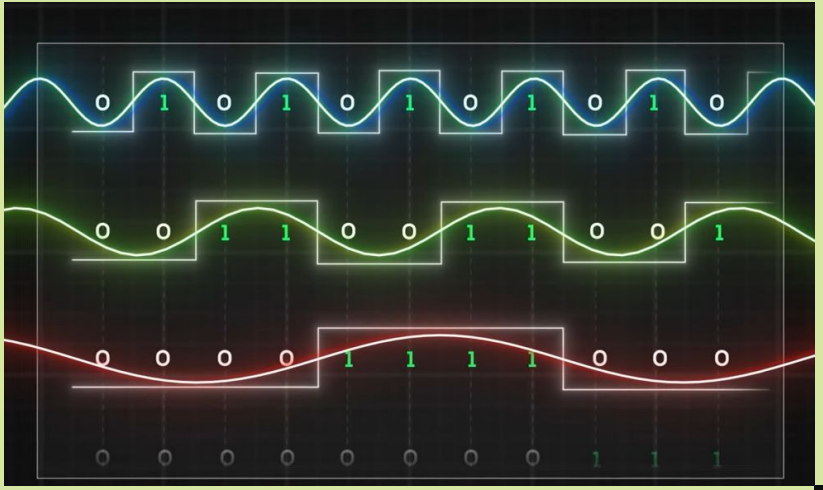
cuando le doy un input de una palabra, el me da el producto punto con cada vector key

El output de nuestro modelo seria un vector resultante en sumar los productos de la antencion con el vector valor de cada una de las demás palabras

Transformer

Se hace su respectivo embedding, pero es necesario saber en que parte de la frase se encuentra, así que se usa otro metodo para codificar su posicion

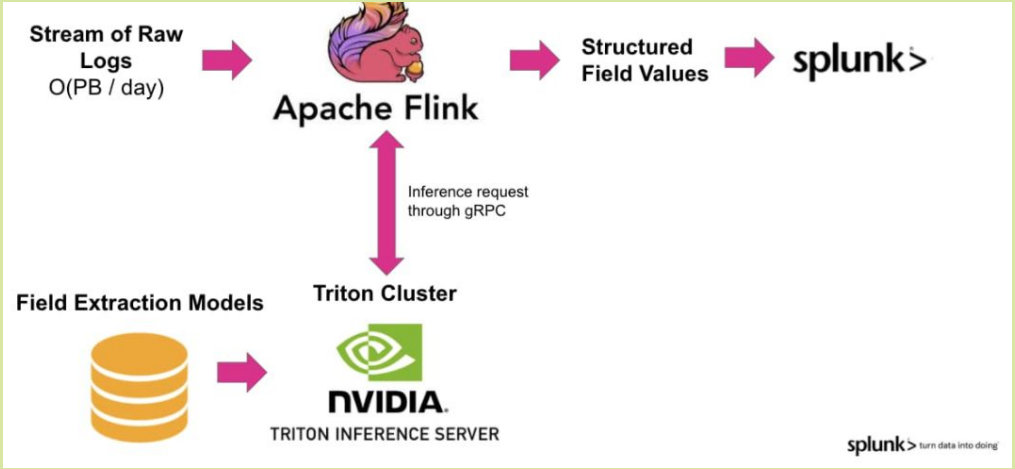




01	El	1 1 1 1 … 1 1 1 1	0 0 0 0 … 0 0 0 1
02	pangolín	2 2 2 2 … 2 2 2 2	0 0 0 0 … 0 0 1 0
03	es	3 3 3 3 … 3 3 3 3	0 0 0 0 … 0 0 1 1
04	tu	4 4 4 4 … 4 4 4 4	0 0 0 0 … 0 1 0 0
05	nuevo	5 5 5 5 … 5 5 5 5	0 0 0 0 … 0 1 0 1
06	dios	6 6 6 6 … 6 6 6 6	0 0 0 0 … 0 1 1 0

Como usa Splunk los transformer

Splunk usa una herramienta creada por Nvidia para poder hacer uso de los modelos de machine learningw



Links Relacionados

https://projector.tensorflow.org vsialuzador de wordvenc funcionando

https://colab.research.google.com/drive/1go6YwMFe5MX6XM9tv-cnQiSTU50N9EeT
Generador de imagenes con lenguaje natural

https://openai.com/blog/dall-e/ Dall-E

https://beta.openai.com/playground GPT-3 playground