Webinar Nuclio:
tratamiento de fotos con Python

Nico Popescul

NUCLIO
DIGITAL SCHOOL

01

INTRO DEL PONENTE

NUCLIO DIGITAL SCHOOL

Nico Popescul: Senior Data Scientist @ CaixaBank BI

Área de especialización clasificación binaria, clusterización y series temporales.

Formación profesional

- Economía en UPNA
- Postgrado de Data Science en UB
- Fan de Kaggle
- Mooc lover

Habilidades y conocimientos

- Σ 6 años de experiencia en Data Analytics
- Σ 4 años de experiencia en Python
- Σ 3 años de experiencia en SQL
- Σ 4 años de experiencia en VBA
- Σ 6 años de experiencia con Excel

Datos de contacto:

LinkedIn: https://www.linkedin.com/in/nico-popescul-a8a1a084/

Kaggle: https://www.kaggle.com/python10pm

Email: 1890np@gmail.com

Móvil: +34 660 365 323

KPMG (3 años)	Banco Sabadell (9 meses)	CaixaBank BI (2 años)
Senior Consultant Financial Services 1. Excel 2. VBA 3. Python 4. IDEA	**Data Analyst** 1. SQL 2. Excel 3. VBA 4. SAS	**Senior Data Scientist** 1. Python 2. SQL 3. Excel 4. SAS

02

ROADMAP DEL WEBINAR

NUCLIO DIGITAL SCHOOL

Roadmap del Webinar

El objetivo principal del Webinar es calcular una métrica de similitud entre fotos.
Dadas un conjunto de fotos, calcular un coeficiente de cúanto se parecen las imágenes entre sí.

Como input del proyecto vamos a tener una carpeta donde tendremos las fotos que queremos analizar.

Pasos a seguir	Comentarios
Crear la carpeta de output	Vamos a crear con Python una carpeta donde vamos a guardar las fotos “procesadas”. ¿Porque?
Procesamos las fotos	Debido a que muchos algoritmos de ML necesitan que los inputs que sean del mismo tamaño.
Convertimos las fotos procesadas en numpy arrays	Una foto es un conjunto de pixeles de tres colores RGB que toman un valor entre 0 – 255. Por tanto los podemos representar como una matriz con estos valores.
Calculamos la similitud de las fotos	Una vez que tengamos nuestras fotos en forma de matriz, podemos calcular una métrica de similitud usando la similitud de coseno.
Crearemos un resumen de 2 fotos más parecidas	Dada una foto usaremos matplotlib (librerías para hacer gráficas) para crear un resumen con la foto que mayor score tiene.

NUCLIO DIGITAL SCHOOL

Stack tecnológico del Webinar

Vamos a usar librerías open – source de Python (totalmente gratuitas para el público).

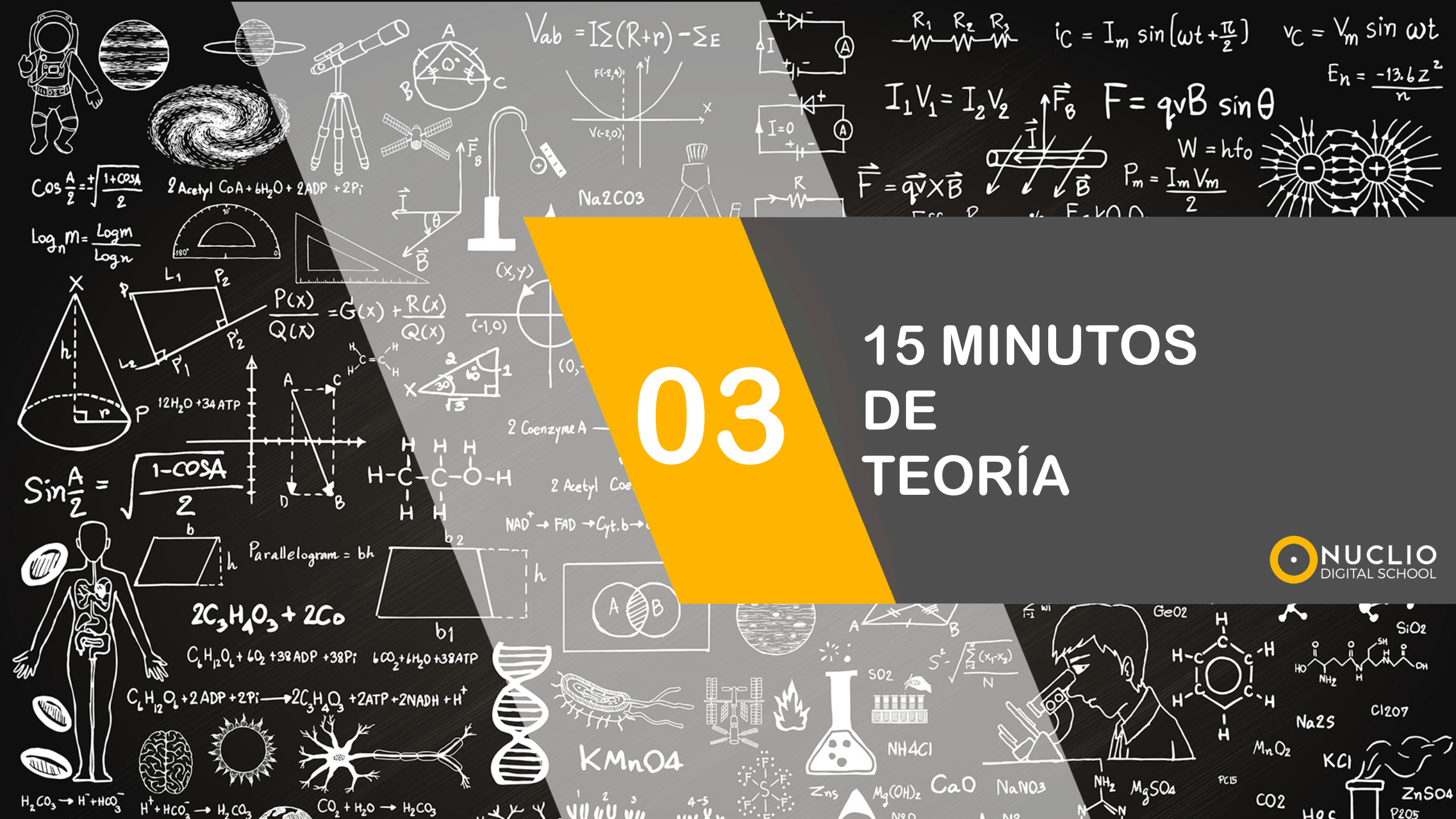
Tech Stack	Comentarios
import os import sys	Son dos librerías que nos permiten interactuar con nuestro sistema operativo.
import numpy as np import pandas as pd	Son las principales librerías para el análisis de datos del ecosistema Python (junto con scipy)
import sklearn	Usaremos de la librería de sklearn funciones y clases para el procesamiento de datos y cálculo de la similitud
import matplotlib.pyplot as plt	Matplotlib nos ayudará a visualizar las fotos
import PIL	PIL es una librería muy famosa de Python que permite trabajar con imágenes.
scripting de Python	En vez de usar Jupyter Notebook (que es el que se usa en el máster de DS de Nuclio), vamos a ejecutar todo el código como un script de Python (vía terminal).
OOP	Usaremos Programación Orientada a Objetos para construir una clase que nos ayude en toda la manipulación.

NUCLIO DIGITAL SCHOOL

Comandos básicos del terminal

Dado que vamos a ejecutar nuestro programa vía terminal, vamos a tener que aprender unos comandos básicos.

Comando MacOS	Comando Windows	Comentarios
cd nombre_carpeta	cd nombre_carpeta	Para navegar por las carpetas vía terminal
cd ..	cd..	Moverse a la carpeta anterior
ls	dir	Mostrar el contenido de una carpeta
python nombre_script.py	python nombre_script.py	Usaremos este comando para ejecutar un programa Python vía terminal
clear	cls	Limpiar el contenido del terminal



03

15 MINUTOS DE TEORÍA

NUCLIO DIGITAL SCHOOL

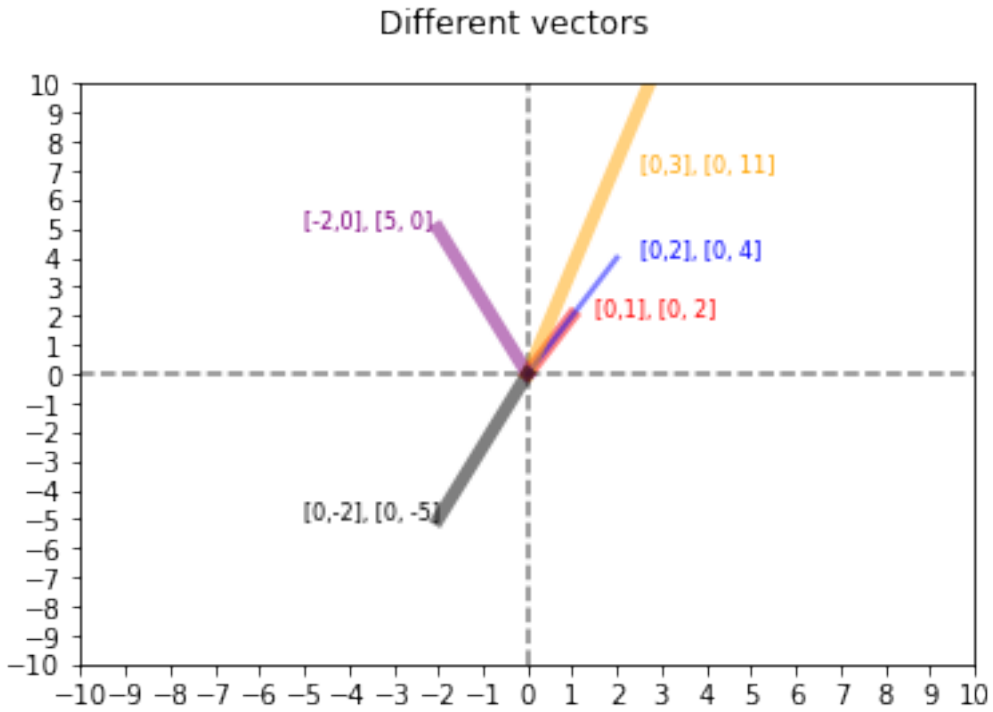
NUCLIO DIGITAL SCHOOL

Similitud de coseno

Dados dos vectores cualquiera (siempre y cuando tenga las mismas dimensiones), puedo calcular los ángulos que forman estos vectores (llamado similitud de coseno) y usar este ángulo como proxy de similitud.

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

-1 < similarity < 1
-1: vectores opuestos
 1: vectores iguales



NUCLIO DIGITAL SCHOOL

Similitud de coseno: ejemplo manual

Supongamos que tenemos el siguiente conjunto de vectores. Podemos calcular la similitud entre 2, usando la fórmula anterior en Python de la siguiente manera:

```
v1 = [1, 2]
v4 = [-2, -5]

((1 * -2) + (2 * -5))\
/(np.sqrt(np.power(1, 2) + np.power(2, 2)) * np.sqrt(np.power(-2, 2) + np.power(-5, 2)))
```

```
-0.9965457582448796
```

Conjunto de vectores

	Valor1	Valor2
Vector1	1	2
Vector2	1	2
Vector3	100	200
Vector4	-2	-5
Vector5	3	-3

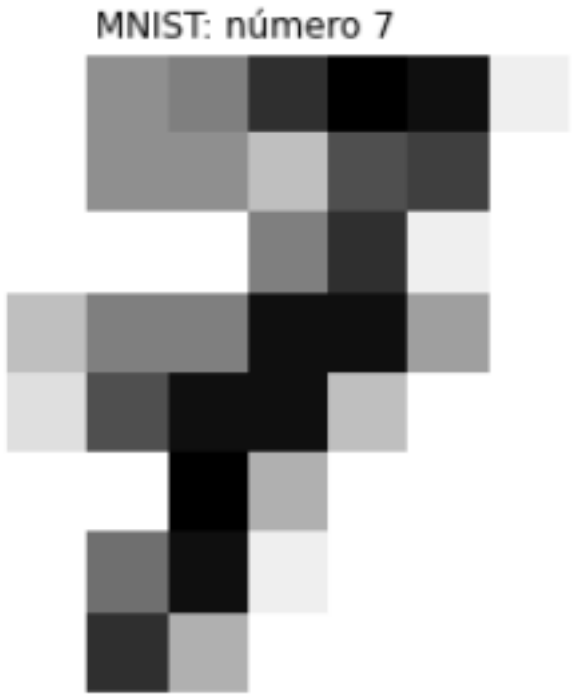
Matriz con las similitudes entre vectores

	Vector1	Vector2	Vector3	Vector4	Vector5
Vector1	1.000000	1.000000	1.000000	-0.996546	-0.316228
Vector2	1.000000	1.000000	1.000000	-0.996546	-0.316228
Vector3	1.000000	1.000000	1.000000	-0.996546	-0.316228
Vector4	-0.996546	-0.996546	-0.996546	1.000000	0.393919
Vector5	-0.316228	-0.316228	-0.316228	0.393919	1.000000

NUCLIO DIGITAL SCHOOL

¿Que es una foto?

A continuación tenemos una imagen y una matriz con números.
¿Representan la misma cosa estos valores?




```
array([[ 0.,  0.,  7.,  8., 13., 16., 15.,  1.],
       [ 0.,  0.,  7.,  7.,  4., 11., 12.,  0.],
       [ 0.,  0.,  0.,  0.,  8., 13.,  1.,  0.],
       [ 0.,  4.,  8.,  8., 15., 15.,  6.,  0.],
       [ 0.,  2., 11., 15., 15.,  4.,  0.,  0.],
       [ 0.,  0.,  0., 16.,  5.,  0.,  0.,  0.],
       [ 0.,  0.,  9., 15.,  1.,  0.,  0.,  0.],
       [ 0.,  0., 13.,  5.,  0.,  0.,  0.,  0.]])
```

Conclusiones

Las principales conclusiones extraídas del Webinar son:

1. Python es una excelente lenguaje de programación con una gran cantidad de librerías open source.

2. En desarrollo en Python es rápido y ágil. Con poco código se consiguen muchas cosas.

3. La similitud de coseno es una excelente métrica para calcular similitudes. Es fácil de interpretar para vectores pero se vuelve “contraintuitiva” cuando trabajamos con fotos **(dos fotos diferentes tendrán un coeficiente de similitud).**

Gracias y nos vemos en Nuclio ☺

NUCLIO DIGITAL SCHOOL