PROYECTO 5

Similitud del Coseno

Descripción breve

Desarrollar un programa en R que permita calcular la similitud del coseno a partir del conjunto de datos: Abstract COVID Papers.csv

Jose Angel Eduardo Garcia De Arcos

Instrucciones

1.Analizar el conjunto de datos que contiene 3 atributos: “title”, “abstract” y “url”.

2.Revisar el corpus de datos y los metadatos directamente de la página: https://www.kaggle.com/datasets/anandhuh/covid-abstracts

3.Seleccionar 100 abstracts y aplicarles la similitud del coseno. Cada equipo debe seleccionar diferentes instancias para que no sean iguales. El dataset está compuesto por 10,000 instancias.

4.Posteriormente, rankear los abstracts por grado de similitud y visualizarlos por el título con una representación gráfica (opcional tipo de gráfico).

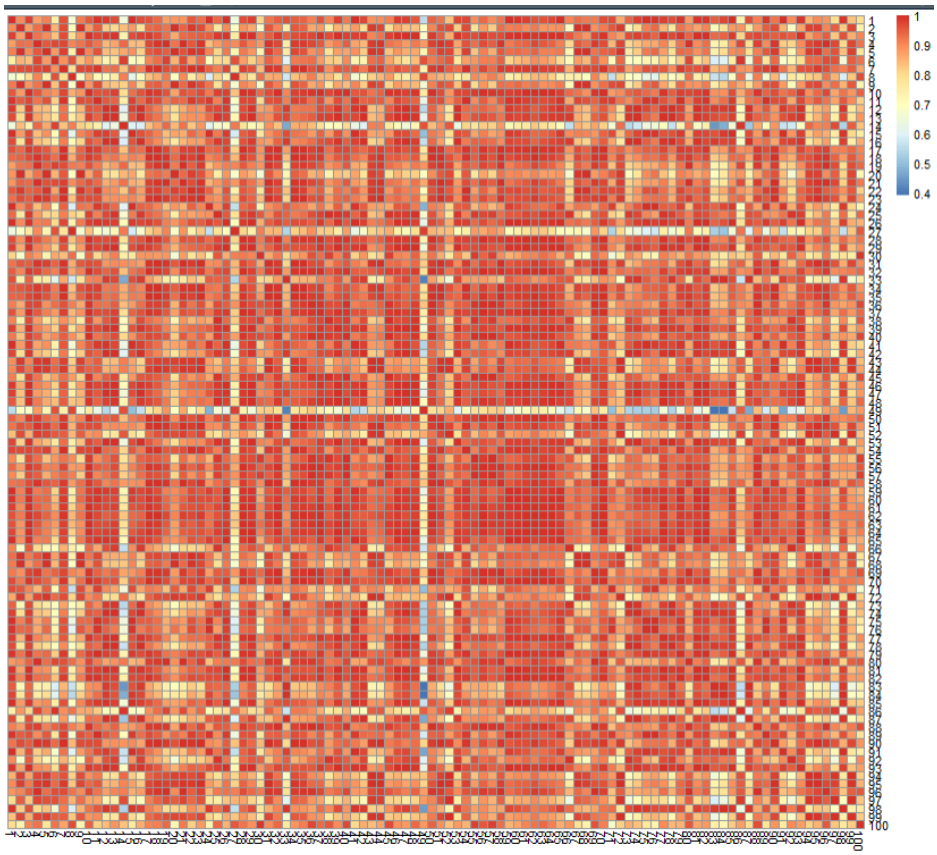
Resultados obtenidos

Diagrama obtenido de similitud

Diagrama de calor

◊ Entre más rojo sea el cuadrado mayor similitud

◊ Entre más azul menor será la similitud



Impresión del ranking de los 15 abstract con mayor similitud

1. "The dynamic change of SARS-CoV-2 variants in Sierra Leone"
2. "Invasive Fungal Sinusitis in Patients With Coronavirus Disease 2019 Seen in South India"
3. "A pilot study assessing the feasibility and acceptability of Project UPLIFT adapted for patients with psychogenic nonepileptic seizures"
4. "Determinants of Immune Response to Anti-SARS-CoV-2 mRNA Vaccines in Kidney Transplant Recipients A Prospective Cohort Study"
5. "DURATION OF ANTIGEN SHEDDING AND DEVELOPMENT OF ANTIBODY TITERS IN MALAYAN TIGERS PANTHERA TIGRIS JACKSONI NATURALLY INFECTED WITH SARS-CoV-2"
6. "Impact of lockdown during Covid-19 pandemic on physical activity and arrhythmia burden in heart failure patients"
7. "Comparison of high-frequency in-pipe SARS-CoV-2 wastewater-based surveillance to concurrent COVID-19 random clinical testing on a public U S university campus"
8. "Women men and COVID-19"
9. "HIV psychological resilience and substance misuse during the COVID-19 pandemic A multi-cohort study"
10. "The impact of COVID-19 on physical activity behaviour in Italian primary school children a comparison before and during pandemic considering gender differences"
11. "COVID-19 a trigger for severe thrombotic microangiopathy in a patient with complement gene variant"
12. "Triage and monitoring of COVID-19 patients in intensive care using unsupervised machine learning"
13. "Innovations in infectious disease testing Leveraging COVID-19 pandemic technologies for the future"
14. "Impact of lockdown during Covid-19 pandemic on physical activity and arrhythmia burden in heart failure patients"
15. "Clinical reasoning in dire times Analysis of cognitive biases in clinical cases during the COVID-19 pandemic"