INSTITUTO POLITÉCNICO NACIONAL

ESCUELA SUPERIOR DE CÓMPUTO

Ingeniería en Sistemas Computacionales

# DOCUMENT SIMILARITY

Práctica 2

## NATURAL LANGUAGE PROCESSING

**Integrantes:**

Garcia Quiroz Gustavo Ivan

Hernández Medina Ulises

Reyes Nunez Sebastian

Saucedo Moreno César Enrique

**Profesor:**

Juarez Gambino Joel Omar.

**Grupo 7CV2**

**03 / Abril / 2025**

**2025 ~ 1**

# TABLAS DE EVIDENCIA

**Test documento 1 arxiv_cl_2.bib:** We propose a novel approach for generating complex outputs that significantly improves accuracy in text-to-SQL tasks. Our method leverages execution results to select the most semantically consistent query from multiple candidates, enabling smaller, cost-effective models to surpass computationally intensive reasoning methods such as o1, o3-mini, and DeepSeek R1 while reducing inference cost by as much as 30 times. It integrates effortlessly with existing models, offering a practical and scalable pathway to state-of-the-art SQL generation.

| **Documento del corpus** | **Representación vectorial** | **Características extraídas** | **Elemento de comparación** | **Valor de similitud** |
|---|---|---|---|---|
| 10.48550./arXiv.2503.10486 | TF-IDF | Unigramas | Abstract | 0.3127 |
| 10.48550./arXiv.2503.10460 | TF-IDF | Unigramas | Abstract | 0.1532 |
| 10.48550./arXiv.2503.11074 | TF-IDF | Unigramas | Abstract | 0.1532 |
| 10.1001/jamahealthforum.2024.5586 | TF-IDF | Unigramas | Abstract | 0.1507 |
| 10.1186/s13561-025-00611-0 | TF-IDF | Unigramas | Abstract | 0.1413 |
| 10.3390/jimaging10080192 | TF-IDF | Unigramas | Abstract | 0.1365 |
| 10.48550./arXiv.2503.10814 | TF-IDF | Unigramas | Abstract | 0.1320 |

| | | | | |
|---|---|---|---|---|
| 10.48550./arXiv.2503.10666 | TF-IDF | Unigramas | Abstract | 0.1242 |
| 10.3390/jimaging10090215 | TF-IDF | Unigramas | Abstract | 0.1173 |
| 10.1136/jitc-2024-011149 | TF-IDF | Unigramas | Abstract | 0.1116 |

**Test documento 2 arxiv_cv_2.ris:** SU-YOLO: Spiking Neural Network for Efficient Underwater Object Detection

| **Documento del corpus** | **Representación vectorial** | **Características extraídas** | **Elemento de comparación** | **Valor de similitud** |
|---|---|---|---|---|
| 10.48550./arXiv.2503.11005 | Binaria | Bigrama | Título | 0.2673 |
| 10.48550./arXiv.2503.11389 | Binaria | Bigrama | Título | 0.2673 |
| 10.3390/jimaging10080197 | Binaria | Bigrama | Título | 0.2500 |
| 10.48550./arXiv.2503. | Binaria | Bigrama | Título | 0.2357 |
| 10.48550./arXiv.2503. | Binaria | Bigrama | Título | 0.2236 |
| 10.48550./arXiv.2503. | Binaria | Bigrama | Título | 0.2041 |

**Test documento 3 AS_3.ris:** Employee task performance plays a critical role in driving organizational success, and understanding its interaction with employee psychological status is essential for unlocking a workforce's full potential. Psychological ownership has been shown to significantly influence performance outcomes, making it crucial to explore how these dynamics shape individual effectiveness. This study attempts to gain a deeper understanding of how employees' sense of ownership influences their intrapreneurial behavior and contributes to enhanced task performance outcomes within organizational settings. A sample of full-time employees based in the United States provided 523 responses on an online questionnaire. The hypotheses were tested using SmartPLS. The findings support that intrapreneurial behavior exhibits full mediation of task performance's relationship with psychological ownership. The outcomes indicate that when employees feel a sense of personal responsibility and attachment to their work, it significantly fosters their innovative actions and enhances their performance, thereby contributing to organizational success. This study contributes to the existing literature by arguing that employees who feel attached to the organization take more responsibility, improve performance, and proactively establish creative innovations to foster organizational success. Study limitations and recommendations are discussed.

| **Documento del corpus** | **Representación vectorial** | **Características extraídas** | **Elemento de comparación** | **Valor de similitud** |
|---|---|---|---|---|
| 10.48550./arXiv.2503.09896 | Frecuencia | Unigrama | Abstract | 0.2889 |
| 10.48550./arXiv.2503.10542 | Frecuencia | Unigrama | Abstract | 0.2601 |
| 10.48550./arXiv.2503.10671 | Frecuencia | Unigrama | Abstract | 0.2586 |
| 10.48550./arXiv.2503.10679 | Frecuencia | Unigrama | Abstract | 0.2582 |
| 10.48550./arXiv.2503.10659 | Frecuencia | Unigrama | Abstract | 0.2574 |

|  |  |  |  |  |
|---|---|---|---|---|
| 10.1001/jamanetworkopen.2025.0728 | Frecuencia | Unigrama | Abstract | 0.2237 |
| 10.1093/bjs/znaf005 | Frecuencia | Unigrama | Abstract | 0.2125 |
| 10.1007/s40265-025-02162-4 | Frecuencia | Unigrama | Abstract | 0.2120 |
| Non | Frecuencia | Unigrama | Abstract | 0.2116 |
| 10.7150/thno.104858 | Frecuencia | Unigrama | Abstract | 0.2113 |

**Test documento 5 LG_3.bib:** How a single fertilized cell gives rise to a complex array of specialized cell types in development is a central question in biology. The cells grow, divide, and acquire differentiated characteristics through poorly understood molecular processes. A key approach to studying developmental processes is to infer the tree graph of cell lineage division and differentiation histories, providing an analytical framework for dissecting individual cells' molecular decisions during replication and differentiation. Although genetically engineered lineage-tracing methods have advanced the field, they are either infeasible or ethically constrained in many organisms. In contrast, modern single-cell technologies can measure high-content molecular profiles (e.g., transcriptomes) in a wide range of biological systems. Here, we introduce CellTreeQM, a novel deep learning method based on transformer architectures that learns an embedding space with geometric properties optimized for tree-graph inference. By formulating lineage reconstruction as a tree-metric learning problem, we have systematically explored supervised, weakly supervised, and unsupervised training settings and present a Cell Lineage Reconstruction Benchmark to facilitate comprehensive evaluation of our learning method. We benchmarked the method on (1) synthetic data modeled via Brownian motion with independent noise and spurious signals and (2) lineage-resolved single-cell RNA sequencing datasets. Experimental results show that CellTreeQM recovers lineage structures with minimal supervision and limited data, offering a scalable framework for uncovering cell lineage relationships in challenging animal models. To our knowledge, this is the first method to cast cell lineage inference explicitly as a metric learning task, paving the way for future computational models aimed at uncovering the molecular dynamics of cell lineage.

| **Documento del corpus** | **Representación vectorial** | **Características extraídas** | **Elemento de comparación** | **Valor de similitud** |
|---|---|---|---|---|
| 10.1073/pnas.2420466122 | Bigrama | Binaria | Abstract | 0,0845 |
| 10.48550JarXiv.2503.11044 | Bigrama | Binaria | Abstract | 00801 |
| 10.1172/JC1185217 | Bigrama | Binaria | Abstract | 0.0620 |
| 10.48550JarXiv.2503.10620 | Bigrama | Binaria | Abstract | 00612 |
| 10.4855WarXiv.2503.10661 | Bigrama | Binaria | Abstract | 0.0546 |
| 10.1007/500018-025-05642-8 | Bigrama | Binaria | Abstract | 0,0540 |
| 10.4855WarXiv.2503.11101 | Bigrama | Binaria | Abstract | 0,0529 |
| 10.1038/543018-025-00933-2 | Bigrama | Binaria | Abstract | 0.0525 |
| 10.1038/s43018-025-00928-z | Bigrama | Binaria | Abstract | 0.0510 |
| 10.48550./arXiv.2503.11164 | Bigrama | Binaria | Abstract | 00481 |

**Test documento 6 pubmed_natcard_2.bib:** Post-injury remodeling is a complex process involving temporal specific cellular interactions in the injured tissue where the resident fibroblasts play multiple roles. Here, we performed single-cell and spatial transcriptome analysis in human and mouse infarcted hearts to dissect the molecular basis of these interactions. We identified a unique fibroblast subset with high CD248 expression, strongly associated with extracellular matrix remodeling.

Genetic Cd248 deletion in fibroblasts mitigated cardiac fibrosis and dysfunction following ischemia/reperfusion. Mechanistically, CD248 stabilizes type I transforming growth factor beta receptor and thus upregulates fibroblast ACKR3 expression, leading to enhanced T cell retention. This CD248-mediated fibroblast-T cell interaction is required to sustain fibroblast activation and scar expansion. Disrupting this interaction using monoclonal antibody or chimeric antigen receptor T cell reduces T cell infiltration and consequently ameliorates cardiac fibrosis and dysfunction. Our findings reveal a CD248+ fibroblast subpopulation as a key regulator of immune-fibroblast cross-talk and a potential therapy to treat tissue fibrosis.

| **Documento del corpus** | **Representación vectorial** | **Características extraídas** | **Elemento de comparación** | **Valor de similitud** |
|---|---|---|---|---|
| 10.1182/blood.2024025440 | Unigrama | Frecuencia | Abstract | 0.3808 |
| 10.1016/j.cce112025.02.009 | Unigrama | Frecuencia | Abstract | 0.3745 |
| 10.1038/543018-025-00927-0 | Unigrama | Frecuencia | Abstract | 0.3617 |
| 10.1172/JC1185217 | Unigrama | Frecuencia | Abstract | 0.3426 |
| 10.1093/brain/awaf096 | Unigrama | Frecuencia | Abstract | 0.3344 |
| 10.4855WarXiv.2503.11241 | Unigrama | Frecuencia | Abstract | 0.2800 |
| 10.4855WarXiv.2503.11439 | Unigrama | Frecuencia | Abstract | 0.2068 |
| 10.4855WarXiv.2503.10875 | Unigrama | Frecuencia | Abstract | 0.2016 |
| 10.4855WarXiv.2503.11465 | Unigrama | Frecuencia | Abstract | 0.2002 |
| 10.48550JarXiv.2503.11495 | Unigrama | Frecuencia | Abstract | 0.1999 |

**Test documento 7 pubmed_natcom_2.bib:** A genome-wide cross-trait analysis characterizes the shared genetic architecture between lung and gastrointestinal diseases**.**

| **Documento del corpus** | **Representación vectorial** | **Características extraídas** | **Elemento de comparación** | **Valor de similitud** |
|---|---|---|---|---|
| 10.1038/s41588-025-02136-y | TF-IDF | Unigramas | Título | 0.3347 |
| 10.1038/s41467-025-57452-y | TF-IDF | Unigramas | Título | 0.2386 |
| 10.48550./arXiv.2503.10655 | TF-IDF | Unigramas | Título | 0.1637 |
| 10.1186/512889-025-21910-5 | TF-IDF | Unigramas | Título | 0.1436 |
| 10.4855WarXiv.2503.10713 | TF-IDF | Unigramas | Título | 0.1376 |
| 10.1016/j.autrev.2025.103804 | TF-IDF | Unigramas | Título | 0.1371 |
| 10.1016/j.autrev.2025.103804 | TF-IDF | Unigramas | Título | 0.1371 |
| 10.4855WarXiv.2503.10740 | TF-IDF | Unigramas | Título | 0.1339 |
| 10.4855WarXiv.2503.09743 | TF-IDF | Unigramas | Título | 0.1325 |
| 10.48550./arXiv.2503.10354 | TF-IDF | Unigramas | Título | 0.1270 |

**Test documento 8 test.bib:“Bring Your Rear Cameras for Egocentric 3D Human Pose Estimation” [Simulación de documento con 100% de similitud]**

| Documento del corpus | Representación vectorial | Características extraídas | Elemento de comparación | Valor de similitud |
|---|---|---|---|---|
| 10.4855WarXiv.2503.11652 | TF-IDF | Unigramas | Título | 1.00 |
| 10.48550JarXiv.2503.11194 | TF-IDF | Unigramas | Título | 0.3863 |
| 10.1186/s13024-025-00819-y | TF-IDF | Unigramas | Título | 0.3501 |
| 10.4855WarXiv.2503.11143 | TF-IDF | Unigramas | Título | 0.2088 |
| 10.1038/541586-025-08873-8 | TF-IDF | Unigramas | Título | 0.2033 |
| 10.1126/science.adu6445 | TF-IDF | Unigramas | Título | 0.1928 |
| 10.4855WarXiv.2503.11345 | TF-IDF | Unigramas | Título | 0.1872 |
| 10.4855WarXiv.2503.11371 | TF-IDF | Unigramas | Título | 0.1702 |
| 10.1146/annurev-virology-092818-015907 | TF-IDF | Unigramas | Título | 0.1613 |
| 10.1016/j.cell.2025.02.009 | TF-IDF | Unigramas | Título | 0.1600 |