# Graph-Based Structural Analysis of Terms in Text

Created @January 19, 2026 3:02 PM

## Graph-Based Structural Analysis of Terms in Text

### 1. Problema

Dado un texto y un conjunto de términos relevantes extraídos del mismo, el objetivo es analizar su estructura relacional, identificando:

- El peso total de la estructura que conecta todos los términos
- El término más central dentro de dicha estructura
- Los cinco términos más centrales según una métrica de distancia global

El reto es ir más allá de la frecuencia individual y analizar cómo los términos se organizan estructuralmente entre sí dentro del texto.

---

### 2. Enfoque General

La solución sigue un enfoque basado en grafos, compuesto por las siguientes etapas:

1. Convertir cada término detectado en un vector numérico (embedding)
2. Medir similitud entre términos a partir de dichos vectores
3. Construir un grafo ponderado que refleje esas similitudes
4. Extraer una estructura mínima que conecte todos los términos (MST)
5. Calcular medidas de centralidad y jerarquía sobre esa estructura

**texto → vectores → grafo → árbol → análisis estructural**

# 3. Detección y Normalización de Términos

El texto se normaliza (minúsculas, eliminación de acentos y puntuación) y se detectan términos relevantes como nombres propios y siglas mediante expresiones regulares.

Cada término detectado se modela como un nodo del grafo.

# 4. Representación Vectorial (Embedding)

Cada término se representa mediante un embedding TF-IDF basado en trigramas de caracteres, normalizado a norma 1.

Este tipo de representación captura similitud léxica y morfológica entre términos, y fue elegido porque:

- Funciona bien con nombres propios y siglas
- No depende de vocabularios o modelos preentrenados
- Es determinista, reproducible y adecuado para textos especializados

Este embedding no modela semántica profunda, sino similitud superficial basada en la forma de los términos. Sin embargo, permite construir una geometría consistente sobre la cual aplicar el análisis estructural.

El embedding define la geometría del espacio vectorial y, por tanto, determina las distancias entre términos.

# 5. Cálculo de Similitud y Construcción del Grafo

Se calcula la similitud coseno entre todos los pares de términos.

- Complejidad temporal: $\mathbf{O(n^2 \cdot d)}$
- Complejidad espacial: $\mathbf{O(n^2)}$

Este paso constituye el principal cuello de botella del pipeline y se asume explícitamente en el diseño.

Para limitar la densidad del grafo, cada término se conecta únicamente con sus **k vecinos más cercanos**, dando lugar a un grafo disperso pero representativo.

---

# 6. Árbol de Expansión Mínima (MST)

A partir del grafo disperso se calcula un Árbol de Expansión Mínima (MST) que:

- Conecta todos los términos
- Minimiza el peso total
- Elimina conexiones redundantes

El MST representa la estructura mínima de relaciones entre los términos bajo la métrica definida.

---

# 7. Jerarquía y Centralidad mediante Tree DP

Sobre el MST se define la siguiente medida de centralidad estructural:

$$S(u) = \Sigma_v \, dist(u, v)$$

donde `dist(u, v)` es la distancia entre los nodos `u` y `v` dentro del árbol.

## Solución naive

Ejecutar un algoritmo de caminos mínimos desde cada nodo, con complejidad **$O(n^2)$**.

## Optimización propuesta

Dado que el MST es un árbol, se aplica programación dinámica sobre árboles (Tree DP) mediante dos recorridos:

1. **Post-order**: cálculo de tamaños de subárbol y distancias parciales
2. **Pre-order**: propagación de resultados usando la relación:

$$S(hijo) = S(padre) + peso \times (n - 2 \times subtree\_size)$$

Esto reduce el cálculo completo de la jerarquía a **O(n)**.

## 8. Resultados

El sistema produce:

- El peso total del MST
- El término más central (menor S(u))
- El top 5 de términos más centrales

Estas métricas reflejan importancia estructural dentro del grafo, no únicamente frecuencia de aparición.

## 9. Análisis de Complejidad

| Etapa | Tiempo | Espacio |
|---|---|---|
| Normalización | O(L) | O(L) |
| Detección de términos | O(L · n) | O(n) |
| Embeddings | O(n · d) | O(n · d) |
| Similitud | $O(n^2 \cdot d)$ | $O(n^2)$ |
| MST | O(m · log n) | O(n) |
| Jerarquía (Tree DP) | O(n) | O(n) |

El término dominante es **$O(n^2 \cdot d)$**, correspondiente al cálculo de similitudes.

## 10. Conclusión

- El método analiza relaciones estructurales entre términos a partir de similitud léxica
- El embedding define la geometría del espacio, pero no semántica profunda
- El MST extrae una estructura mínima representativa
- El uso de Tree DP permite calcular centralidad global de forma eficiente

La solución es correcta, eficiente y adecuada para el análisis estructural de términos en texto.