Informe Dashbord
DataSet CLINC150

Docente: Ana Maria CuadrosValdivia
Alumna: Cecilia del Pilar Vilca Alvites

Este informe detalla la estructura y funcionalidad del Dashboard Interactivo de Análisis de Intenciones, diseñado para explorar visualmente los problemas de granularidad y ambigüedad en agrupamientos de intenciones, utilizando el conjunto de datos CLINC150. El dashboard está construido con Dash de Plotly y permite una interacción dinámica con los datos, lo que facilita la comprensión de las complejidades semánticas de las intenciones de usuario.

1. Espacios de Visualización del Dashboard

El dashboard está organizado en **dos pestañas principales**, cada una dedicada a un aspecto específico del análisis:

1.1. Pestaña: "Visualización de Intenciones y Clusters"

Esta es la pestaña principal para la exploración visual de los embeddings de las consultas de usuario y los resultados del clustering.

- **1.1.1. Gráfico Principal de Dispersión (Main Scatter Plot)**
  - **Ubicación:** Ocupa la mayor parte del espacio izquierdo de esta pestaña.
  - **Descripción:** Un **gráfico de dispersión 2D** que visualiza los puntos de datos (consultas de usuario) en el espacio de baja dimensionalidad generado por **t-SNE**. Cada punto representa una consulta, y su posición indica su similitud semántica con otras consultas.

- **Interactividad:**
  - **Colorear Puntos:** Mediante un selector de radio (dcc.RadioItems), el usuario puede elegir si los puntos se colorean por su **intención real** (el "ground truth" del dataset) o por el **ID del cluster** asignado por el algoritmo K-Means.
  - **Número de Clusters (KMeans):** Un campo de entrada numérico (dcc.Input) permite al usuario ajustar dinámicamente el valor de K (el número de clusters) para el algoritmo K-Means. Al cambiar este valor, el clustering se recalcula y el gráfico se actualiza, permitiendo observar el impacto de K en la formación de clusters.
  - **Hover Data:** Al pasar el ratón sobre cualquier punto, se muestra información detallada en un *tooltip*, incluyendo el texto original de la consulta, su intención real y el ID de cluster asignado.
  - **Selección de Puntos:** El usuario puede hacer clic en puntos individuales o arrastrar una caja para seleccionar múltiples puntos. Esta selección alimenta el panel de detalles adyacente.
- Es fundamental para identificar visualmente:
  - **Solapamiento:** Regiones donde múltiples colores (intenciones reales) se mezclan, indicando que diferentes intenciones son semánticamente cercanas.
  - **Granularidad:** Cómo una única intención real puede dispersarse en varias áreas del espacio, o cómo los clusters de K-Means agrupan (o dividen) estas intenciones.
  - **Patrones de Agrupamiento:** Observar la forma y densidad de los clusters generados por K-Means en relación con la distribución de las intenciones reales.

Grafico por Id Cluster:

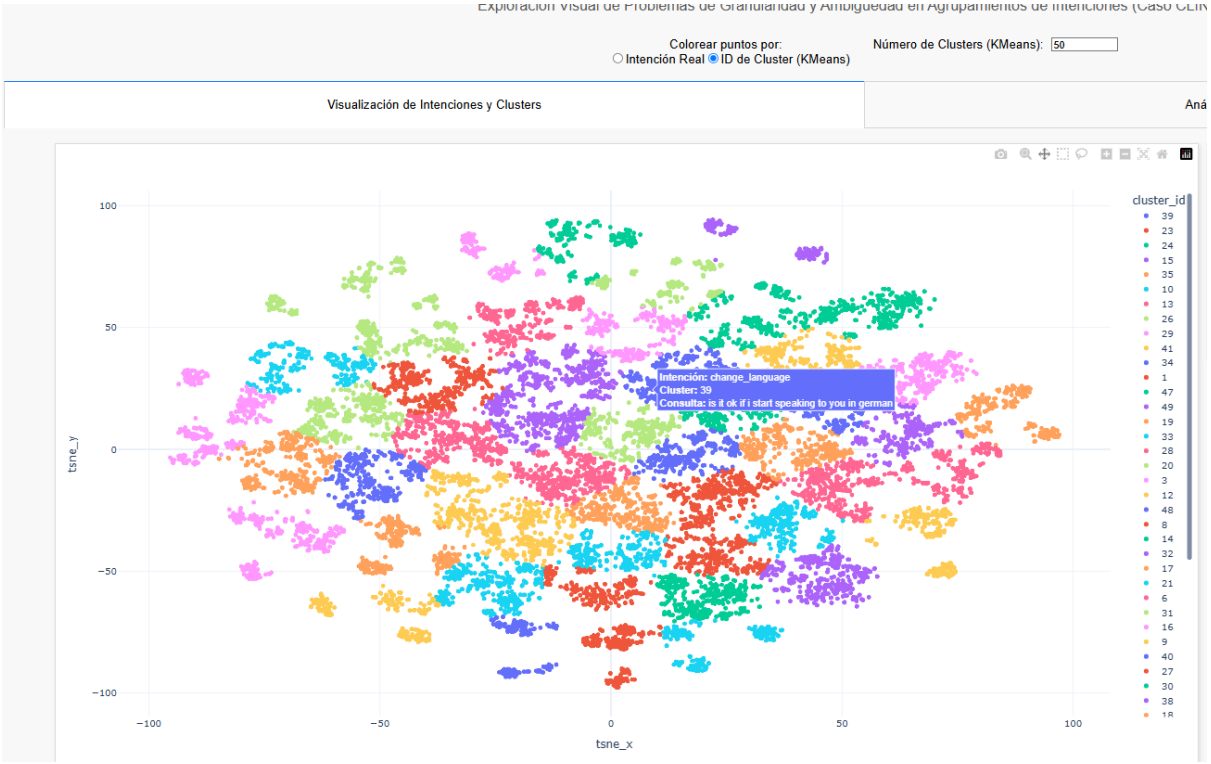
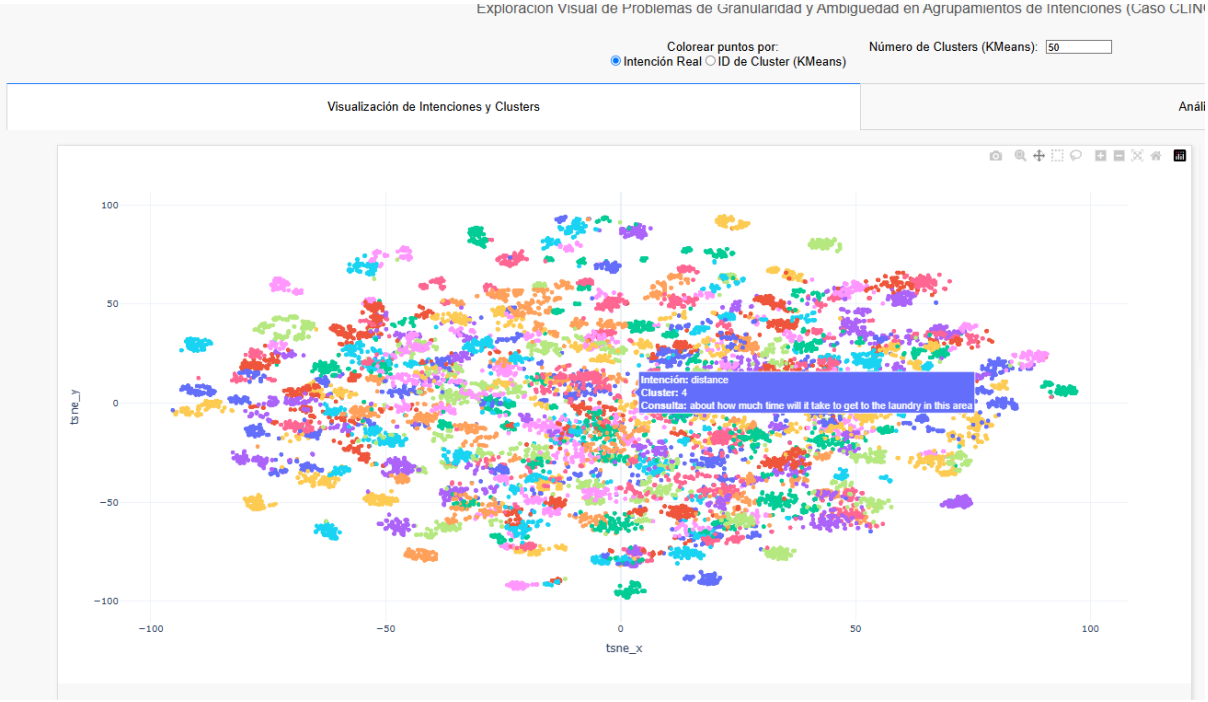
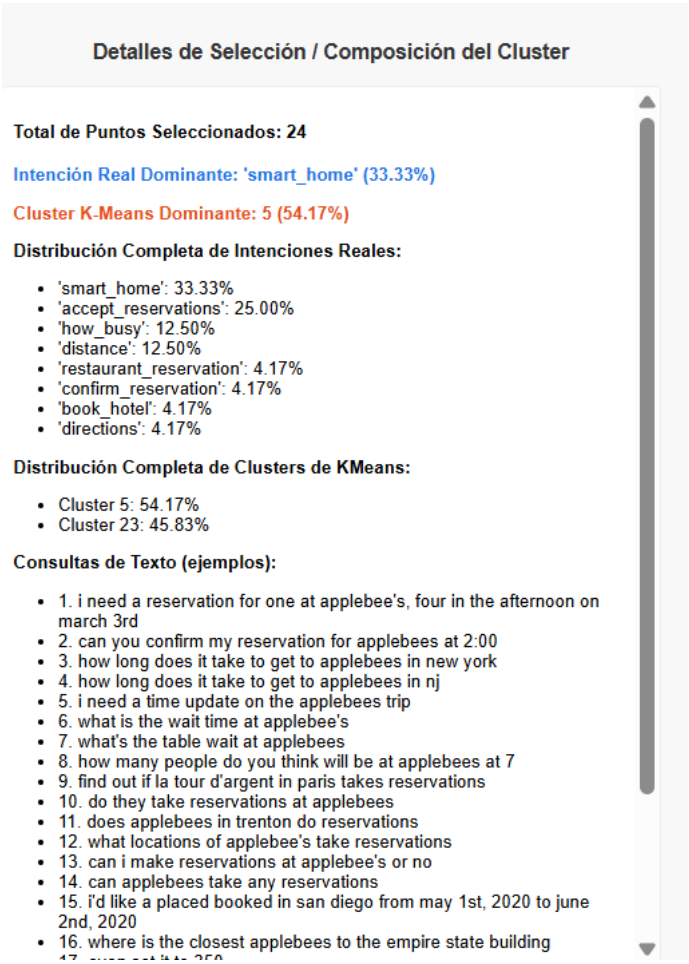


Gráfico por Intención real:



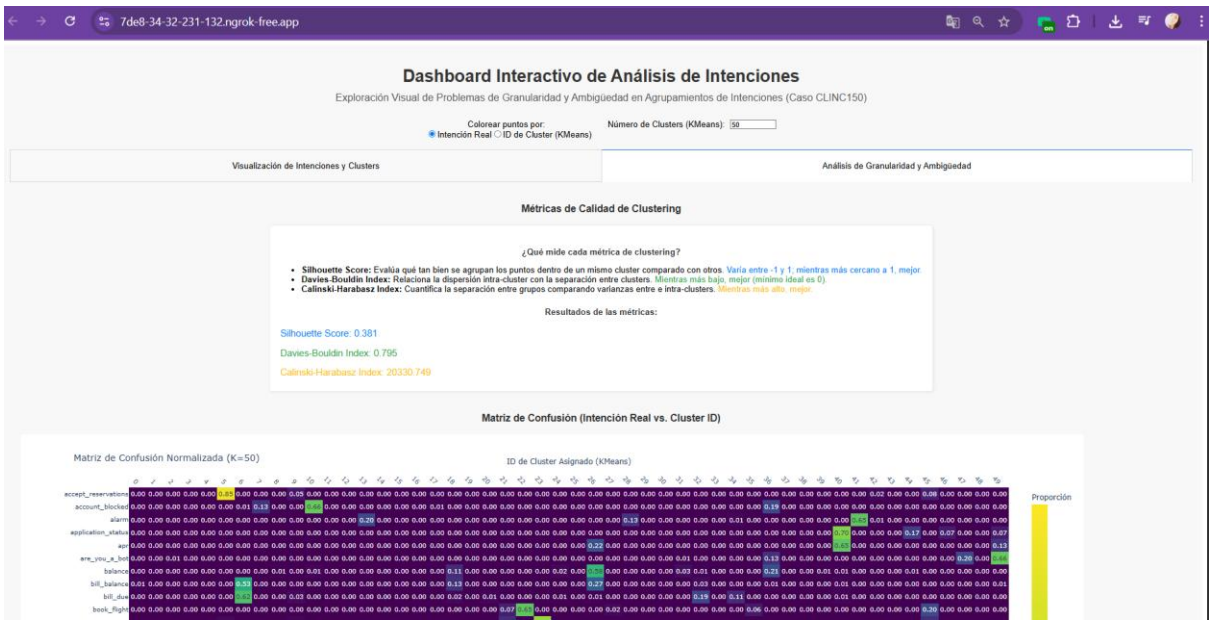
- **1.1.2. Panel de Detalles de Selección / Composición del Cluster**



- **Ubicación:** Situado a la derecha del gráfico principal de dispersión.
- **Descripción:** Este panel muestra un resumen analítico de los puntos de datos seleccionados en el gráfico de dispersión.
- **Interactividad:** Se actualiza automáticamente cada vez que el usuario selecciona o deselecciona puntos en el "Main Scatter Plot".
- **Contenido:** Proporciona:
  - El número total de puntos seleccionados.
  - La **intención real dominante** y su porcentaje entre los puntos seleccionados.
  - El **ID de cluster K-Means dominante** y su porcentaje.
  - Una **distribución completa** (en porcentaje) de todas las intenciones reales presentes en la selección.
  - Una **distribución completa** (en porcentaje) de todos los IDs de cluster de K-Means presentes en la selección.
  - Una lista de **ejemplos de consultas de texto** de los puntos seleccionados.
- Crucial para:
  - **Identificar Ambigüedad:** Al seleccionar un cluster que ha agrupado múltiples intenciones reales (vista desde la matriz de confusión o el gráfico), este panel permite revisar los textos de las consultas para encontrar frases idénticas o muy similares asignadas a distintas intenciones, confirmando la ambigüedad semántica.
  - **Análisis de Granularidad Específica:** Comprender la composición de un clúster específico o de un grupo de intenciones, detallando cómo se mezclan o si una intención se fragmenta.

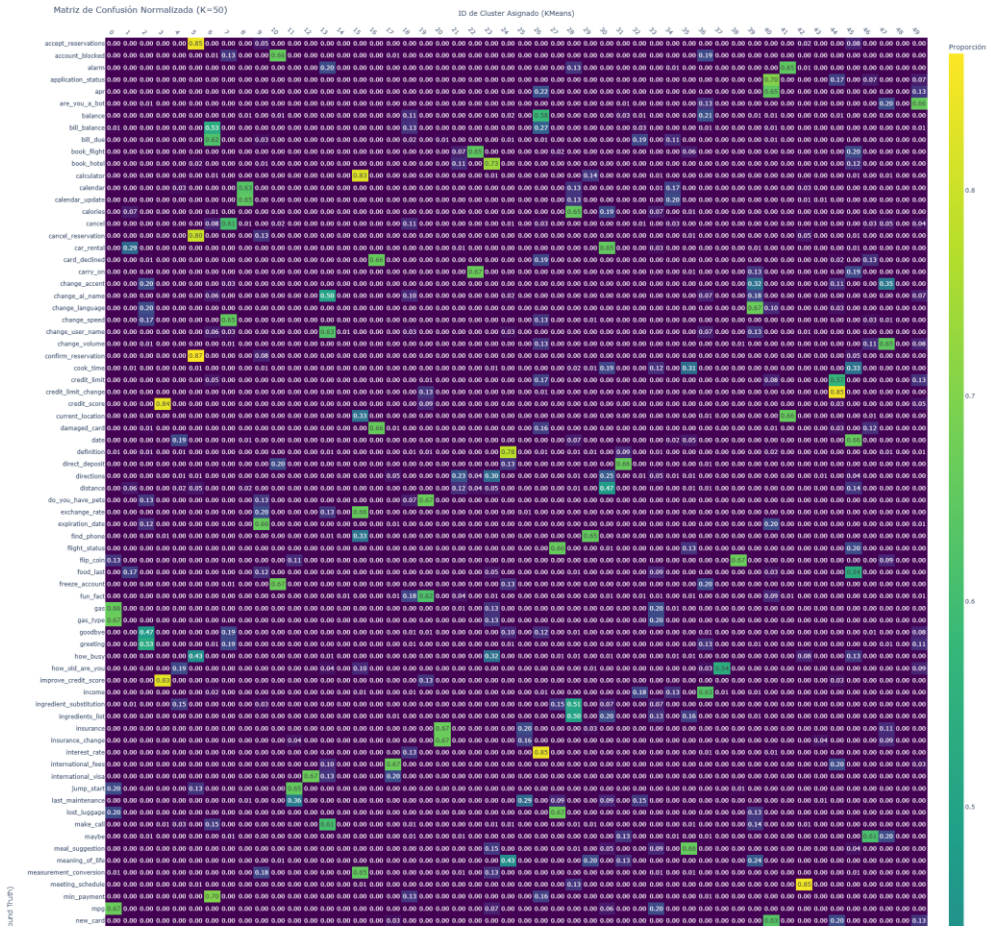
1.2. Pestaña: "Análisis de Granularidad y Ambigüedad"

Esta pestaña ofrece una vista cuantitativa de la calidad del clustering y la relación entre intenciones reales y clusters asignados. Se actualiza automáticamente cada vez que se cambia el "Número de Clusters (KMeans)" en la pestaña principal, así como al cambiar a esta pestaña.



- **1.2.1. Métricas de Calidad de Clustering**

    - **Ubicación:** Parte superior de la pestaña.
    - **Descripción:** Presenta tres métricas comunes para evaluar la calidad de un clustering:
        - **Silhouette Score:** Mide qué tan similar es un objeto a su propio clúster en comparación con otros clústeres. (Rango: -1 a 1; **más cercano a 1 es mejor**).
        - **Davies-Bouldin Index:** Evalúa la relación entre la dispersión dentro del clúster y la distancia entre clústeres. (**más cercano a 0 es mejor**).
        - **Calinski-Harabasz Index:** Cuantifica la separación entre grupos comparando varianzas entre e intra-clusters. (**más alto es mejor**).
    - **Leyenda:** Incluye una explicación clara de lo que mide cada métrica y cómo interpretar sus valores ideales.
    - **Propósito en la Exploración:** Proporciona una **evaluación cuantitativa** del clustering, permitiendo comparar la calidad de las agrupaciones para diferentes valores de K y entender el grado de cohesión y separación que K-Means logra en este espacio semántico complejo. Los valores obtenidos ayudan a validar las observaciones visuales sobre el solapamiento y la dificultad de separación.
- **1.2.2. Matriz de Confusión (Intención Real vs. Cluster ID)**
    - **Ubicación:** Parte inferior de la pestaña.
    - **Descripción:** Un **mapa de calor** que visualiza la relación entre las **intenciones reales (filas)** y los **IDs de cluster asignados por K-Means (columnas)**. Cada celda muestra la **proporción** de consultas de una intención real específica que fueron asignadas a un cluster particular (normalizada por fila).
    - **Interactividad:** Se recalcula y actualiza cada vez que se ajusta el número de clusters en la interfaz.
    - Es la herramienta más potente para:
        - **Validar la Granularidad y Solapamiento:**
            - **Filas dispersas** (varios colores intensos en una fila): Indican que una sola intención real se está fragmentando en múltiples clusters de K-Means (alta granularidad).
            - **Columnas mezcladas** (varios colores intensos en una columna): Indican que un cluster de K-Means está agrupando consultas de múltiples intenciones reales diferentes (solapamiento semántico).
        - **Evaluar la separación de límites:** Una matriz de confusión con diagonales fuertes (y el resto de celdas cercanas a cero) indicaría una buena separación. La observación de que esto es difícil de lograr, incluso con ajustes de K, refuerza la hipótesis de la ausencia de límites claros.



En conjunto, estos espacios de visualización forman un sistema integral que permite al usuario navegar desde una vista macro del espacio semántico hasta el detalle de consultas individuales, y evaluar la calidad del agrupamiento mediante métricas y una matriz de confusión detallada. Esto facilita una profunda investigación de la granularidad, ambigüedad y la efectividad del clustering en el contexto de intenciones de usuario.