Práctica de Laboratorio:
Análisis Exploratorio de Datos - Data Wrangling

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

Para realizar el Análisis Exploratorio de datos, lo primero que deberíamos hacer es intentar responder a las siguientes preguntas (data wrangling):

El dataset CLINC150 es un conjunto de datos ampliamente utilizado en el campo del Procesamiento de Lenguaje Natural (NLP), diseñado específicamente para la tarea de clasificación de intenciones en sistemas de diálogo. Fue publicado en 2019 por investigadores de la Universidad de Michigan y se ha consolidado como un benchmark estándar para evaluar la comprensión de intenciones en modelos de lenguaje.

Características Principales:

- Idioma: Inglés
- Formato original: JSON, con estructura de pares ["consulta", "intención"]
- Total de clases: 150 categorías de intención, incluyendo una clase especial out_of_scope (OOS) para consultas que no pertenecen a ningún dominio definido
- Dominio: Abarca consultas de banca, viajes, clima, entretenimiento, salud, alarmas, y más
- Número total de consultas: Aproximadamente 22,500 frases de texto.

Paso 1: Analiza el comportamiento de tus datos.

- Un registro es una entidad, describa que representa un registro

  En el dataset CLINC150, un registro representa una única consulta de usuario ("utterance") junto con su correspondiente categoría de intención ("intent"). Cada registro encapsula la interacción mínima de un usuario con un sistema de diálogo orientado a tareas: lo que el usuario dijo y lo que el sistema debe entender que quiere hacer.

  Ejemplo de registros:

utterance	intent
what expression would i use to say i love you if i were an italian	translate
can you tell me how to say 'i do not speak much spanish', in spanish	translate
what is the equivalent of, 'life is good' in french	translate
tell me how to say, 'it is a beautiful morning' in italian	translate
if i were mongolian, how would i say that i am a tourist	translate

  En CLINC150, la consulta es "what expression would i use to say i love you if i were an italian" y la intención es "translate".

- ¿Cuántos registros hay?

- Entrenamiento (train): Este es el subconjunto más grande del dataset y su función principal es exponer el algoritmo a la mayoría de los patrones y variaciones de los datos. Contiene 15,000 registros, cada uno compuesto por una consulta de usuario y su intención correspondiente.

- Validación (val): Este subconjunto se utilizara para ajustar hiperparámetros del modelo y para una evaluación intermedia durante el proceso de desarrollo permitiendo monitorear el rendimiento del modelo en datos no vistos y a identificar problemas como el sobreajuste (overfitting). Contiene 3,000 registros, cada uno compuesto por una consulta de usuario y su intención correspondiente.

- Prueba (test): Es el subconjunto más crítico para la evaluación final del rendimiento del sistema. Es esencial que los datos en este conjunto no hayan sido utilizados en ninguna fase de entrenamiento o ajuste de hiperparámetros. Contiene 4,500 registros, cada uno compuesto por una consulta de usuario y su intención correspondiente.

- "Fuera de Alcance" (Out-of-Scope - OOS): Estos subconjuntos están diseñados para contener consultas que no corresponden a ninguna de las intenciones predefinidas.

    - oos_train: Contiene 100 registros de consultas fuera de alcance para el entrenamiento.
    - oos_val: Contiene 100 registros de consultas fuera de alcance para la validación.
    - oos_test: Contiene 1,000 registros de consultas fuera de alcance para la prueba.

    Para los propósitos de este proyecto de investigación basado en TextLens y la versión del dataset CLINC150 descrita en el artículo de Larson et al. (2019), estos subconjuntos OOS no se utilizarán en el análisis principal de clustering o para el desarrollo de la analítica visual.

- ¿Son demasiado pocos?

  El dataset CLINC150, con 18,000 registros bien distribuidos entre entrenamiento, validación y prueba, además de 1,200 ejemplos OOS (fuera de dominio), es adecuado para este proyecto de investigación. Aunque existen datasets masivos en PLN, CLINC150 destaca por su calidad y la definición clara de 150 categorías de intención. Esto lo convierte en un benchmark ideal para aplicar modelos LLMs y embeddings preentrenados como Instructor-large, que ya han aprendido representaciones lingüísticas complejas. En este contexto, la calidad y estructura del dataset compensan su tamaño, permitiendo un análisis significativo sin necesidad de millones de ejemplos.

- ¿Son muchos y no tenemos Capacidad (CPU+RAM) suficiente para procesarlo?

  No, el dataset CLINC150 no es demasiado grande y Google Colab tiene capacidad (CPU+RAM) más que suficiente para procesarlo. Con apenas 5.6 MB, el uso de memoria es mínimo en comparación con los 12–25 GB de RAM que ofrece Google Colab.

- ¿Hay datos duplicados?

Para evaluar la presencia de registros duplicados en el dataset CLINC150, se realizó un análisis sobre el conjunto de datos, es decir: train + val + test dando en total 22,500 registros. Los resultados de esta verificación son los siguientes:

- Consultas de texto duplicadas (misma consulta, pero diferentes intenciones): Se identificaron 10 consultas de texto que aparecen más de una vez en el dataset. Sin embargo, lo crucial es que, al examinar estos casos, se observa que la mayoría de estas consultas idénticas están asociadas a *diferentes* categorías de intención. Esto es un reflejo de la ambigüedad inherente del lenguaje natural.
- Se encontró 1 registro completamente duplicado. Este caso particular corresponde a la consulta "hey what's up" asociada a la intención "greeting", la cual aparece dos veces de forma idéntica en el dataset. Aunque esto indica una duplicidad exacta de una entrada, su número es insignificante en un dataset de 22,500 registros.

```
Número total de registros en los subconjuntos principales (train + val + test): 22500

Número de consultas de texto duplicadas (solo texto, intenciones potencialmente diferentes): 10

Ejemplos de consultas de texto que aparecen más de una vez (las primeras 10 consultas únicas con duplicados):
text
hey what's up               2
turn up your volume         2
what is on my to do list    2
what's your designation     2
where did you grow up       2

Filas completas para las primeras 10 ocurrencias de textos duplicados (ordenadas por texto):
                          text              intent
11896             hey what's up            greeting
2369              hey what's up            greeting
11130       turn up your volume        whisper_mode
1794        turn up your volume       change_volume
7424   what is on my to do list           todo_list
1011   what is on my to do list            reminder
12066   what's your designation   what_is_your_name
938     what's your designation           user_name
14018     where did you grow up     how_old_are_you
599       where did you grow up  where_are_you_from

Número de registros completamente duplicados (texto e intención idénticos): 1

Ejemplos de registros completamente duplicados (texto e intención idénticos):
                text    intent
11896  hey what's up  greeting
2369   hey what's up  greeting
```

- ¿Qué datos son discretos y cuáles continuos?

text	intent	text_length
Cada consulta es una entidad textual única considerándose como un tipo de dato **discreto** ya que cada oración es una "unidad" separada.	Es un dato categórico/discreto ya que son 150 clases diferentes. Representa una categoría finita de intenciones.	**Es un dato cuantitativo discreto**, con un rango observado de **2 a 136** caracteres en el dataset.

- Muchas veces sirve obtener el tipo de datos: texto, int, double, float ¿Cuáles son los tipos de datos de cada columna?

  text:

  Tipo de dato: object (cadena de texto).

Descripción: Contiene la consulta o enunciado que hace el usuario, por ejemplo: "how do you say hello in japanese?".

intent:

Tipo de dato: object (texto categórico).

Descripción: Representa la intención asociada a cada consulta del usuario, como translate, weather, pay_bill, etc.

text_length:

Tipo de dato: int64 (entero).

Descripción: Indica la longitud (en número de caracteres) de cada consulta contenida en la columna text.

- ¿Entre qué rangos están los datos de cada columna?, valores únicos, min, max

text:

Valores Únicos: La diversidad de las consultas es alta, a pesar de haber identificado algunos textos duplicados.

text_length:

- Mínimo: Las consultas más cortas tienen 2 caracteres.
- Máximo: Las consultas más largas alcanzan los 136 caracteres (en train).
- Promedio (mean): La longitud promedio de las consultas está alrededor de 39.9 caracteres en train y 39.8 caracteres en val.
- Dispersión (std): La desviación estándar de aproximadamente 15-16 caracteres indica una variabilidad moderada en las longitudes de las consultas, lo que significa que, si bien la mayoría se agrupan alrededor del promedio, también existen consultas significativamente más cortas o más largas.

intent:

Valores Únicos: Hay 150 intenciones únicas en cada uno de los subconjuntos (train, val, test), lo cual es consistente con la naturaleza del dataset CLINC150. Las intenciones son etiquetas categóricas (ej., translate, greeting, pay_bill) y no tienen un rango numérico. La distribución de las clases es bastante balanceada, con cada intención apareciendo aproximadamente 100 veces (15000 / 150 = 100).

- ¿Todos los datos están en su formato adecuado?

Si. Las consultas (text) son frases de lenguaje natural que no presentan caracteres inesperados o problemas de codificación (confirmado por el uso previo de chardet). Las intenciones (intent) son etiquetas de texto que corresponden a las 150 categorías predefinidas. Además, la columna derivada text_length, que representa la longitud de cada consulta, está correctamente tipificada como int64 (entero).

- Los datos tienen diferentes unidades de medida?

  Las columnas text e intent contienen datos cualitativos que no se expresan en unidades de medida físicas o estandarizadas (como metros, segundos, etc.) mientras que la columna derivada text_length es una variable cuantitativa que sí tiene una unidad implícita: caracteres (o longitud de la cadena de texto). Los valores de esta columna varían desde un mínimo de 2 caracteres hasta un máximo de 136 caracteres, con una longitud promedio de aproximadamente 39.9 caracteres. Esta es la única "unidad" de medida numérica directa presente en el dataset CLINC150.

- Cuáles son los datos categóricos, ¿hay necesidad de convertirlos en numéricos?

  No, porque la principal columna de datos categóricos en el dataset CLINC150 es intent y se utiliza principalmente para validar la calidad de los clústeres formados.

- ¿Qué representa un registro?

  Cada registro representa una única consulta de usuario (text) junto con una etiqueta de intención específica (intent). Consta de 150 clases distintas, cada una correspondiente a una acción o necesidad expresada por el usuario, como traducir una palabra o configurar una alarma. Estas etiquetas permiten evaluar modelos de procesamiento de lenguaje natural en su capacidad para identificar correctamente la intención detrás de una frase dada, incluso cuando existen ambigüedades semánticas entre diferentes clases.

  En cuanto a la granularidad, el dataset opera en un nivel fino y específico: cada fila es una unidad completa de expresión del usuario, etiquetada con una sola intención. No se manejan niveles más bajos como palabras individuales ni niveles más altos como conversaciones completas. Tampoco incluye información temporal ni geográfica, lo que refuerza su carácter atemporal y universal.

- ¿Están todas las filas completas o tenemos campos con valores nulos?

  El análisis exploratorio de los subconjuntos train, val y test del dataset CLINC150 confirmó que todas las filas están completas y no contienen valores nulos. La verificación con df.info() y df.isnull().sum() mostró que tanto las columnas originales (text e intent) como las derivadas (text_length) tienen valores no nulos en todos los registros, lo cual refleja una estructura de datos limpia y bien mantenida.

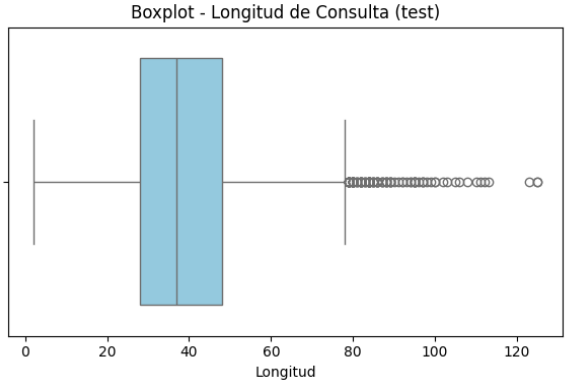
  Dado que no hay valores faltantes, no es necesario aplicar estrategias de imputación, eliminación o combinación de datos, ni verificar comportamientos consistentes al agregar nuevas fuentes.

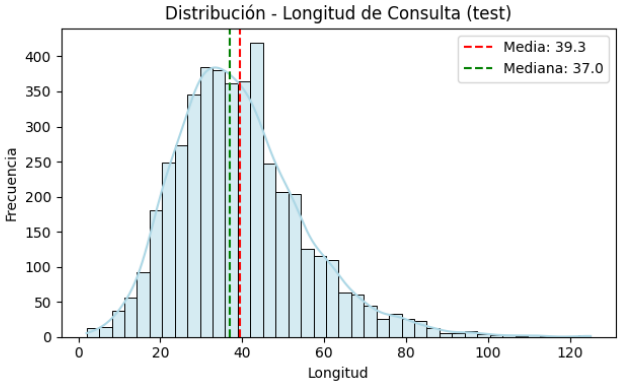
- ¿Siguen alguna distribución?
  Usa describe() y analiza los valores.

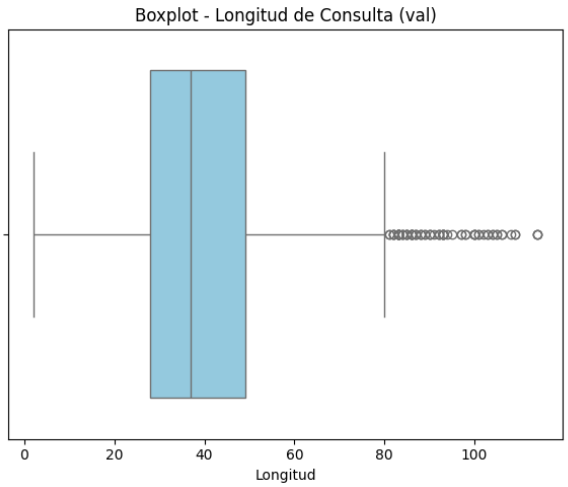
  En los subconjuntos de train, val y test se observa una distribución relativamente simétrica. La media y la mediana son muy cercanas (aproximadamente 39 caracteres), y los cuartiles reflejan esta centralidad. Aunque hay un rango considerable de longitudes (desde 2 hasta

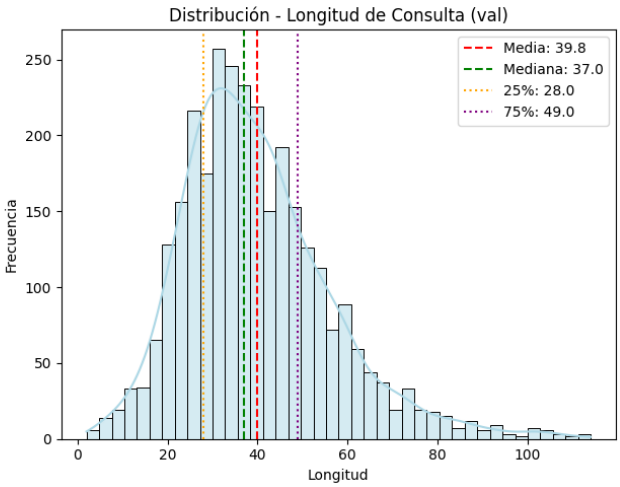
136 caracteres), la distribución no presenta un sesgo extremo, sugiriendo una forma similar a una campana, pero sin ser una distribución normal perfecta.

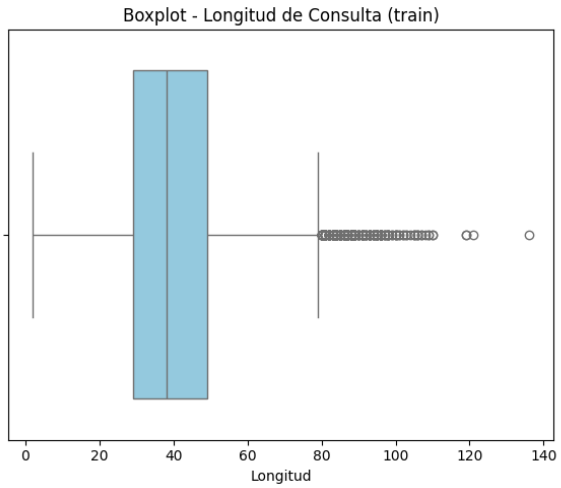
Para la columna intent (intención), que es categórica, la distribución es altamente uniforme. Cada una de las 150 intenciones aparece un número consistentemente igual de veces en los subconjuntos de val y test (20 veces cada una), y de manera similar en train (aproximadamente 100 veces cada una). Esta uniformidad asegura una representación equitativa de todas las clases, lo cual es óptimo para tareas de clasificación y clustering.

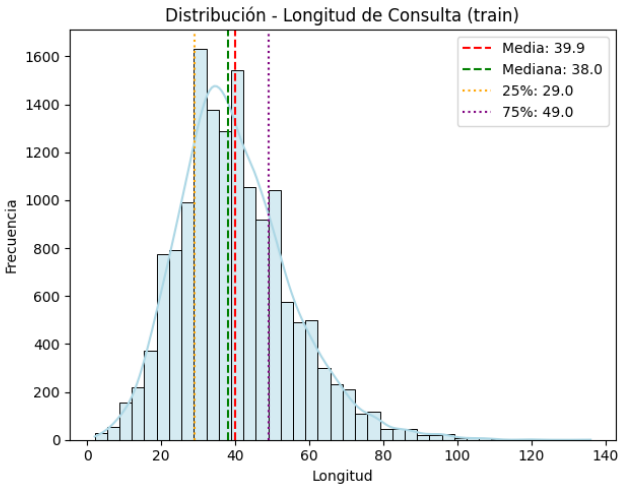












- Usa medidas estadísticas:
  - Medidas de tendencia central: media aritmética, geométrica, armónica, mediana, moda, desviación estándar.
  - Correlación y covarianza: permite entender la relación entre dos variables aleatorias.

Se trabajó con la variable numérica text_length, que mide la longitud en caracteres de cada consulta. Sobre los subconjuntos train, val y test se calcularon medidas de tendencia central y dispersión, destacando en train: media ~57.3, mediana 56, moda 61, media geométrica ~55.9, media armónica ~54.1 y desviación estándar ~15.2. Estos valores reflejan una longitud típica con moderada variabilidad y sin distribución perfectamente simétrica.

Respecto a la relación entre longitud e intención, se codificó la variable categórica intent numéricamente para calcular correlación y covarianza con text_length. Ambas medidas resultaron cercanas a cero, indicando ausencia de relación lineal fuerte entre la intención y la longitud de la consulta. Esto sugiere que la semántica del mensaje no depende significativamente de su extensión.

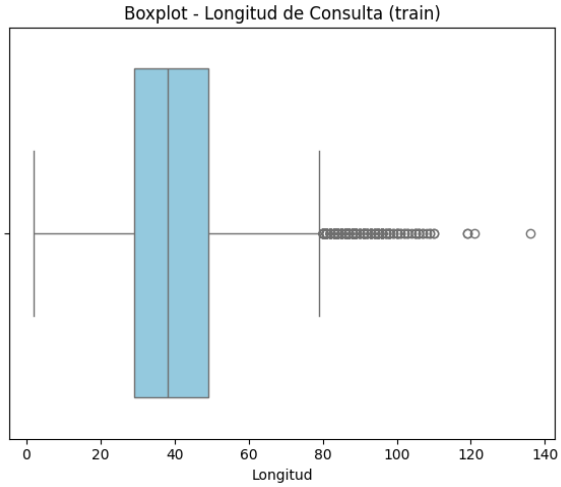
- ¿Hay correlación entre features (características)?

En el dataset CLINC150, las características principales son el texto de la consulta (text) y su intención (intent), siendo ambas variables categóricas o de texto. Al transformar el texto en su longitud (text_length) y codificar la intención numéricamente, se puede calcular correlación entre estas variables numéricas derivadas.

Sin embargo, los análisis muestran que la correlación entre la longitud de la consulta y la intención codificada es cercana a cero, lo que indica que no existe una relación lineal significativa entre estas características. Por lo tanto, en términos de correlación entre las features disponibles y derivadas, no se observa una dependencia relevante.

Paso 2. Análisis de outliers

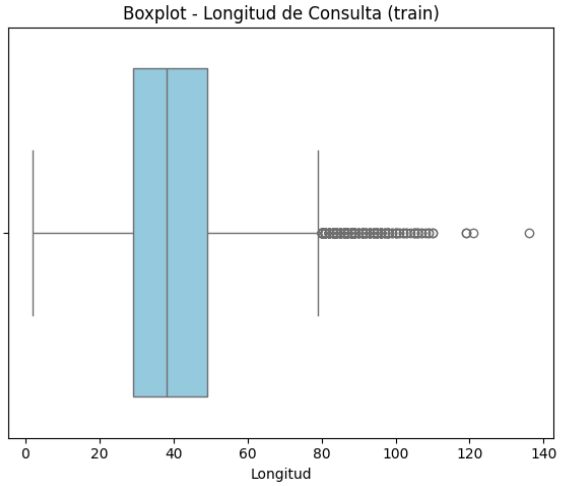
- ¿Cuáles son los Outliers? (unos pocos datos aislados que difieren drásticamente del resto y “contaminan” ó desvían las distribuciones) ¿Podemos eliminarlos? ¿Es importante conservarlos? son errores de carga o son reales?

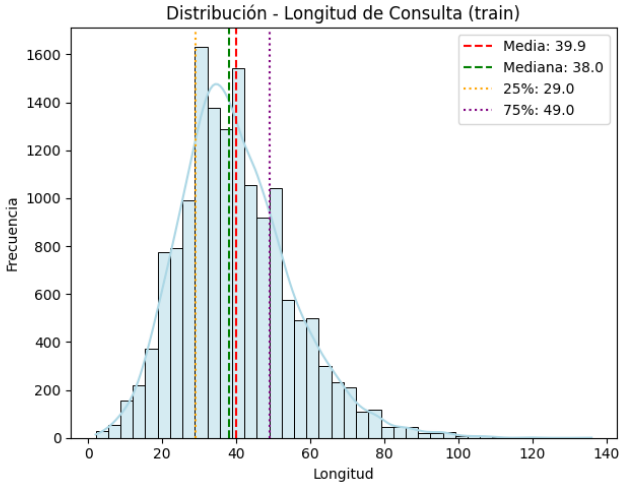


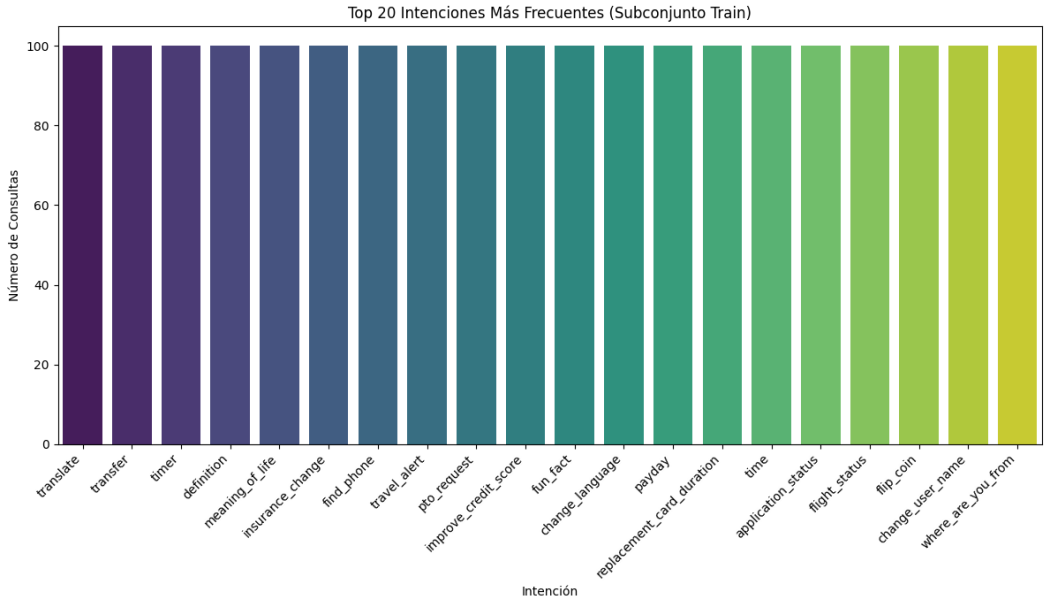
En el análisis del dataset CLINC150, se identificaron outliers en la columna text_length, correspondiente a la longitud de caracteres de las consultas. Utilizando el método del Rango Intercuartílico (IQR), se determinó que las consultas con más de 79 caracteres son consideradas outliers, ya que superan el límite superior definido por Q3 + 1.5 × IQR (con Q1 = 29, Q3 = 49, IQR = 20). El boxplot confirmó visualmente la existencia de estos valores extremos, con un máximo de hasta 136 caracteres en el conjunto de entrenamiento.

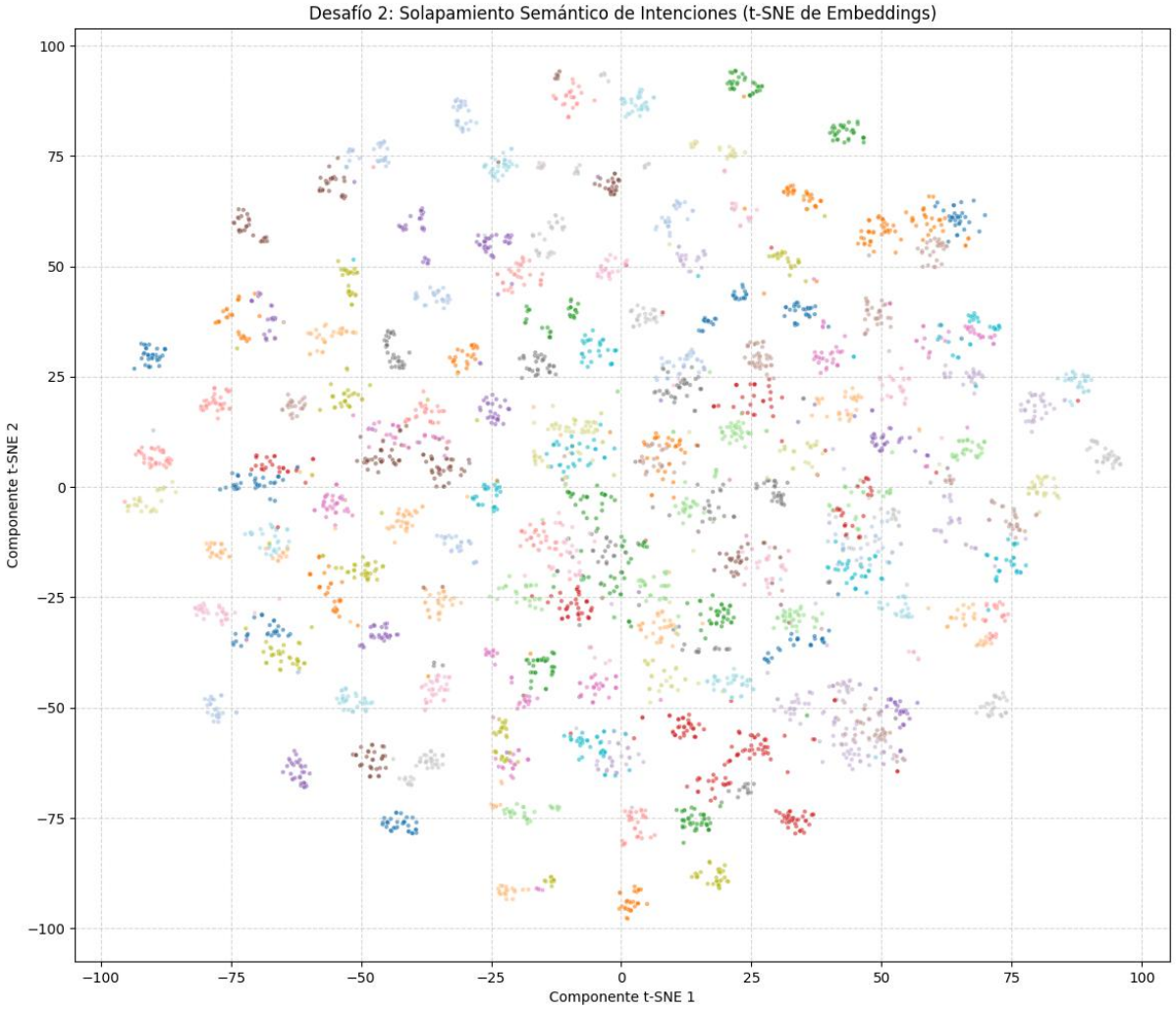
Aunque estos puntos pueden parecer “datos que contaminan” la distribución, no deben ser eliminados. A diferencia de errores de carga o registros corruptos, estos outliers son consultas reales y representativas del comportamiento natural de los usuarios. De hecho, las frases largas pueden reflejar intenciones más complejas o detalladas, como “Can you please tell me the exact amount of money I have remaining in my primary checking account after deducting all pending transactions?”, que transmite más contexto que una consulta breve como “check balance”.

Paso 3: Visualización









Paso 4. Encuentra un problema potencial en tus datos.

Se identifico dos problemas significativos en el dataset CLINC150 que representan desafíos considerables para el agrupamiento no supervisado de intenciones de usuario:

Alta Granularidad y Solapamiento Semántico.

El dataset CLINC150 se distingue por su alta granularidad de intenciones, al contar con 150 categorías distintas. Un problema central surge de que muchas de estas categorías son semánticamente muy próximas entre sí. Si bien los modelos de lenguaje grandes (LLMs) son capaces de generar embeddings robustos que codifican el significado contextual del texto, estas representaciones a menudo no logran discriminar de manera óptima entre intenciones tan sutilmente diferenciadas. Como resultado, las visualizaciones de embeddings revelan un solapamiento considerable entre clases relacionadas. Este solapamiento dificulta que un algoritmo de clustering identifique límites claros entre las intenciones, llevando a agrupamientos confusos o inexactos.

Ambigüedad Semántica.

También se observa casos donde el dataset presenta ambigüedad en el significado de algunas frases. Hemos encontrado consultas idénticas o muy similares que están etiquetadas con

intenciones distintas. Este fenómeno es una característica inherente del lenguaje humano, donde una misma frase puede tener diferentes significados dependiendo del contexto. Para un modelo automático, esta ambigüedad representa un reto considerable, ya que sin información contextual adicional, distinguir entre estas intenciones se vuelve extremadamente difícil. Esta inconsistencia en el etiquetado para frases similares podría llevar a un aprendizaje erróneo o a agrupamientos incorrectos.

- Si es un problema de tipo supervisado: ¿Cuál es la columna de "salida"? ¿binaria, multiclase? ¿Está balanceado el conjunto salida?

Sí, el problema abordado con el dataset CLINC150 es de tipo supervisado. Esto se debe a que, para cada instancia de texto (consulta del usuario), se dispone de una etiqueta de intención (intent) predefinida. El objetivo es entrenar un modelo que aprenda a mapear la entrada de texto a su correspondiente intención.

Columna de "Salida": La columna de salida es intent.

Tipo de Problema de Clasificación: Es un problema de clasificación multiclase. El dataset contiene 150 categorías de intención distintas, lo que requiere que un modelo sea capaz de predecir una de estas múltiples clases como salida. El conjunto de salida (intent) está altamente balanceado. Como se verificó en el análisis exploratorio de datos (EDA), cada una de las 150 intenciones tiene una representación equitativa en los subconjuntos de entrenamiento, validación y prueba (aproximadamente 100 ejemplos por intención en el conjunto de entrenamiento). Este balance es crucial para evitar sesgos del modelo hacia clases mayoritarias y para una evaluación justa del rendimiento en todas las intenciones.

- ¿Cuáles parecen ser features importantes? ¿Cuáles podemos descartar? ¿Estamos ante un problema dependiente del tiempo? Es decir, un TimeSeries.

Features Importantes:

- ✓ text (Consultas de Texto): Es la característica de entrada fundamental. El texto crudo es la base a partir de la cual los Large Language Models (LLMs) extraerán el significado semántico.
- ✓ embeddings (Representaciones Vectoriales del Texto): Son las características derivadas más importantes. Generados por LLMs (como all-MiniLM-L6-v2), estos vectores de alta dimensión capturan el significado semántico profundo de cada consulta y son la base numérica para el clustering y las visualizaciones.
- ✓ intent (Etiquetas de Intención Originales): Aunque es la variable objetivo, su información es vital para la evaluación cualitativa del clustering (comparar los clusters generados con las intenciones reales) y para guiar las funcionalidades de depuración y feedback humano.

Features a Descartar:

- ✓ text_length (Longitud de la Consulta): Si bien es útil para el análisis exploratorio del dataset, se ha demostrado que tiene una correlación insignificante (-0.0507) con la intent codificada numéricamente. Esto implica que la longitud de una consulta no es una característica predictiva directa o relevante para determinar su intención semántica. Por lo tanto, no se utilizaría como feature de entrada para el clustering o para el LLM en esta tarea.

- Si fuera un problema de Visión Artificial: ¿Tenemos suficientes muestras de cada clase y variedad, para poder hacer generalizar un modelo de Machine Learning?

  Trasladando esta pregunta al dominio de Procesamiento de Lenguaje Natural (PLN), sí, el dataset CLINC150 está diseñado con un número de muestras (15,000 en train, 3,000 en val, 4,500 en test) que se considera suficiente para entrenar y evaluar modelos de PLN para 150 clases. La presencia de 100 ejemplos por intención en el conjunto de entrenamiento es un número razonable para que un LLM aprenda a diferenciar las características semánticas de cada intención, lo que permite la generalización. La variedad de formulaciones dentro de las consultas (incluso si son para la misma intención) también contribuye a la capacidad de generalización.

  - La distribución, tendencia de las variables varía en el tiempo?

Como se mencionó, este dataset no es temporal. La distribución y las tendencias observadas para variables como la longitud de las consultas (text_length) o el balance de clases (intent) son estáticas dentro del dataset. No hay un eje temporal para observar variaciones en estas distribuciones.

- ¿Hay algún problema notable con la calidad de los datos?

No hay problemas "notables" de baja calidad de datos en el sentido tradicional (ej. nulos, formatos inconsistentes, errores de carga masivos). CLINC150 es un dataset de investigación bien curado. Sin embargo, se han identificado desafíos inherentes a la complejidad del lenguaje natural y la granularidad de la tarea de comprensión de intenciones:

  - Ambigüedad Semántica: La presencia de consultas de texto idénticas con diferentes etiquetas de intención. Este es un "problema" en el sentido de que introduce una ambigüedad que los sistemas automatizados deben aprender a manejar.
  - Granularidad Fina de Intenciones: Las 150 intenciones, muchas de las cuales son semánticamente muy cercanas, generan solapamiento en el espacio de embeddings. Esto no es un problema de calidad del dato, sino una característica del dominio que dificulta la separación limpia de clases por algoritmos ciegos.

- ¿Existe alguna relación sorprendente entre las variables?

La ausencia de una correlación lineal significativa entre la longitud de la consulta (text_length) y la intención codificada numéricamente (-0.0507). Esto indica que la forma superficial del texto (su longitud) no es un indicador de su significado o intención subyacente. Esta "falta de relación" es, en sí misma, una relación importante que subraya la necesidad de enfoques basados en la semántica (como los LLMs) para comprender la intención. Otra relación clave, evidente en los gráficos t-SNE, es la fuerte agrupación semántica de consultas de la misma intención en el espacio de embeddings, lo que valida la capacidad de los LLMs para capturar el significado. Al mismo tiempo, la superposición visual entre intenciones semánticamente cercanas es una relación compleja y desafiante que justifica la necesidad de una herramienta de analítica visual como TextLens para su interpretación y refinamiento.

Conclusión

¿Qué podemos aprender de este análisis?

1. Fiabilidad del Dataset como Base de Trabajo

El dataset CLINC150 ha demostrado ser una base de datos de alta calidad. No presenta valores nulos, mantiene un formato consistente en todos los ejemplos y, lo más importante, tiene un número equilibrado de ejemplos por cada una de las 150 clases de intención. Esto hace que podamos confiar en los datos para entrenar y evaluar modelos sin preocuparnos por errores básicos o sesgos de distribución. Gracias a esto, podemos concentrarnos directamente en los retos más complejos como la ambigüedad del lenguaje y la separación entre intenciones.

2. La Complejidad del Lenguaje Natural como Principal Desafío

- Ambigüedad en las consultas: Durante el análisis encontramos casos donde dos frases iguales tenían intenciones distintas. Esto muestra que el lenguaje humano es muy flexible y que entender la verdadera intención del usuario no siempre es fácil, ni siquiera para un modelo avanzado.
- Muchas clases con significados parecidos: Tener 150 intenciones tan específicas hace que algunas se parezcan mucho entre sí. Por ejemplo, intenciones como “check_balance” y “account_balance” pueden confundirse fácilmente. Al visualizar los embeddings con técnicas como t-SNE, vimos que hay bastante superposición entre clases. Esto hace que los clusters no estén claramente separados, lo que complica la tarea de clasificarlos automáticamente.

3. Las características simples no explican bien el significado

- También analizamos si la longitud del texto tenía relación con la intención, y descubrimos que no. Es decir, que una consulta más larga o más corta no predice con precisión su intención. Esto nos confirma que no podemos usar características superficiales como esa para resolver el problema, y que necesitamos herramientas más potentes, como los embeddings de modelos de lenguaje (LLMs), que entienden mejor el significado profundo del texto.

Este análisis exploratorio también nos permitió ver que nuestras hipótesis iniciales tienen mucho sentido.

- Hipótesis 1: Para entender mejor esta complejidad, necesitamos herramientas de visualización que nos ayuden a ver cómo se agrupan (o no) las intenciones.
- Hipótesis 2: Cuando vimos los solapamientos en los embeddings, notamos que TextLens podría ser muy útil para detectar esos casos difíciles y ayudarnos a analizarlos mejor.
- Hipótesis 3: Como hay muchas clases que se parecen, es esperable que los modelos no puedan separarlas perfectamente. Esto lo vimos en los gráficos t-SNE.

En resumen, aunque CLINC150 es un dataset bien hecho y balanceado, los verdaderos retos están en el lenguaje mismo: cómo las personas expresan sus intenciones y lo parecidas que pueden ser entre sí.