Universidad Nacional de Río Negro
Ingeniería en Computación
Ciencia de datos Aplicada

Amazon Books Reviews - Redes Neuronales - Text Mining con Orange

Docentes:
Profesora Dra. Paola Britos
Profesora Giuliana Fois
Profesor Pablo Argañaras

Estudiante :
Laura Velazquez
Fecha 19 de Noviembre de 2024

Índice

1. Introducción

En la era digital, las reseñas en línea se han convertido en una fuente esencial de información para los consumidores, especialmente en plataformas de comercio electrónico como Amazon. Las opiniones de los usuarios no solo influyen en las decisiones de compra, sino que también reflejan tendencias y percepciones colectivas sobre productos y servicios. En este contexto, el análisis de las reseñas de libros en Amazon ofrece una oportunidad valiosa para comprender las preferencias de los lectores, evaluar la calidad percibida de las obras y explorar patrones en las opiniones de los usuarios.

Este estudio se centra en la base de datos *Amazon Books Reviews*[1], que compila una amplia colección de reseñas de libros disponibles en Amazon. A través de técnicas de análisis de datos y minería de texto, se busca abordar los siguientes objetivos:

1. **Análisis de Sentimientos de las Reseñas**:
   - **Objetivo**: Determinar la polaridad (positiva, negativa o neutral) de las reseñas de los libros.
   - **Justificación**: Comprender cómo perciben los lectores los libros y cómo estas percepciones se reflejan en las calificaciones.
2. **Identificación de Temas Recurrentes**:
   - **Objetivo**: Extraer y analizar los temas más comunes presentes en las reseñas.
   - **Justificación**: Identificar patrones en las opiniones de los lectores y comprender las características que valoran o critican en los libros.
3. **Clustering Jerárquico y Visualización de Reseñas mediante MDS:**
   - **Objetivo:** Analizar las similitudes y diferencias entre las reseñas para identificar patrones y agrupaciones naturales basados en el contenido textual.
   - **Justificación:** Este análisis permite explorar cómo se agrupan las reseñas según sus similitudes y cómo estas agrupaciones pueden reflejar aspectos comunes como estilo, temática o percepciones compartidas de los libros.
4. Predicción de Calificaciones mediante Aprendizaje Supervisado:

   Objetivo: Entrenar un modelo de red neuronal para predecir las calificaciones numéricas (1 a 5 estrellas) de los libros basándose en las reseñas textuales de los usuarios.

   Justificación: Este modelo busca explorar la relación entre las palabras utilizadas en las reseñas y las calificaciones asignadas por los lectores. Además, pretende evaluar el desempeño de las redes neuronales como una herramienta para tareas de clasificación en el análisis de datos textuales.

Al abordar estos objetivos, se pretende ofrecer una visión integral de las dinámicas presentes en las reseñas de libros en Amazon, proporcionando *insights* que puedan ser útiles tanto para autores y editores como para futuros lectores en la toma de decisiones informadas.

Repositorio de Trabajo

Los archivos utilizados y generados durante este trabajo, incluyendo el archivo de Orange con el flujo completo y la base de datos depurada, se encuentran disponibles en el siguiente repositorio de GitHub:

`https://github.com/lauravelazquez25/Ciencia-de-Datos-con-Orange`

[1] `https://www.kaggle.com/datasets/mohamedbakhet/amazon-books-reviews`

Este repositorio tiene como objetivo facilitar la revisión, la reproducibilidad del análisis y servir como recurso para futuros trabajos relacionados con minería de texto y aprendizaje supervisado.

2. Análisis de Sentimientos: Visualización y Resultados

2.1. Flujo de Trabajo para el Análisis de Sentimientos

El flujo de trabajo para el análisis de sentimientos se diseñó utilizando la herramienta Orange. Este flujo consta de múltiples pasos que procesan los datos desde su entrada inicial hasta la visualización final de los resultados. A continuación, se describe cada componente del flujo:

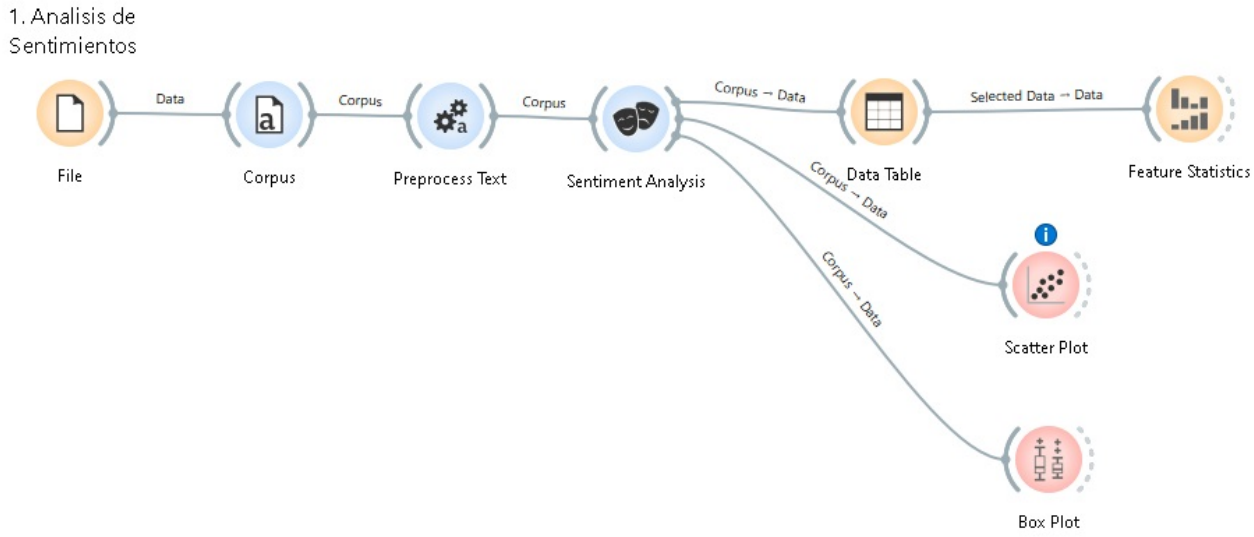


Figura 1: Flujo de trabajo para el análisis de sentimientos en Orange.

1. **File:** Se carga el conjunto de datos en formato CSV, que contiene información sobre las reseñas de libros.
2. **Corpus:** Los datos textuales de las reseñas se convierten en un corpus para ser procesados como texto.
3. **Preprocess Text:** Se aplican técnicas de preprocesamiento como tokenización, eliminación de palabras vacías (*stopwords*), y conversión a minúsculas, asegurando la limpieza y uniformidad de los datos.
4. **Sentiment Analysis:** Se utiliza este widget para realizar un análisis de sentimientos en el texto de las reseñas, clasificando cada una en términos de polaridad positiva, negativa y neutral.
5. **Data Table:** Se visualizan los datos procesados y las columnas generadas por el análisis de sentimientos.
6. **Scatter Plot:** Se generan gráficos de dispersión para explorar la relación entre las calificaciones de las reseñas (*reviewer rating*) y las polaridades positiva, negativa y neutral.
7. **Box Plot:** Se emplea este widget para analizar la distribución de las polaridades de sentimientos, identificando patrones y posibles outliers en los datos.
8. **Feature Statistics:** Se incluyen estadísticas descriptivas para examinar las características derivadas del análisis de sentimientos y validar los resultados obtenidos.

Este flujo permite procesar las reseñas de libros de manera estructurada , proporcionando visualizaciones claras y análisis cuantitativos sobre las percepciones de los usuarios.

2.2. Visualización del Scatter Plot

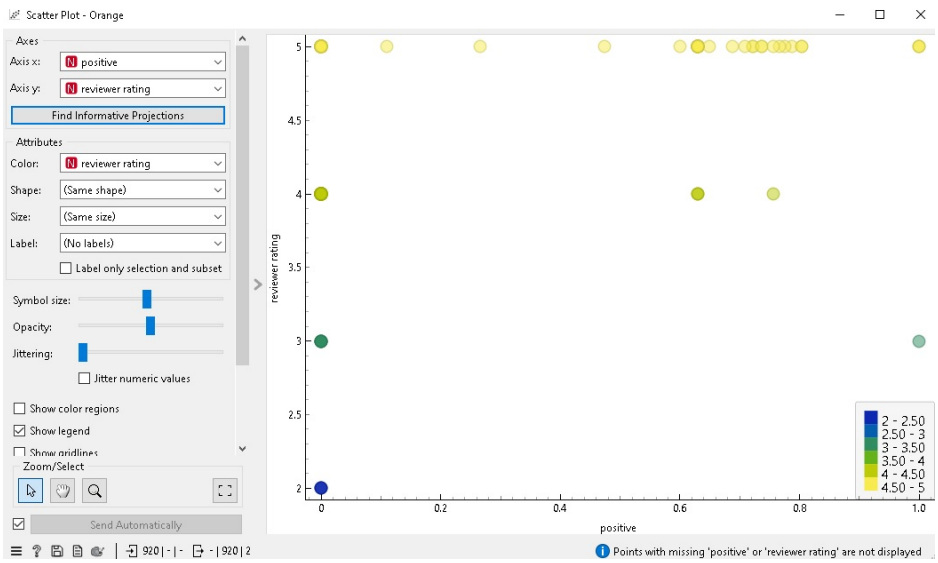
En este apartado se presentan tres gráficos de dispersión (*scatter plots*) que representan la relación entre la calificación de las reseñas (`reviewer rating`) y la polaridad de las mismas (positiva, negativa y neutral). Cada punto del gráfico corresponde a una reseña individual, con el color indicando la calificación otorgada.

- **Gráfico Positivo:** El eje y representa el puntaje de polaridad positiva. Se observa una tendencia en la que las calificaciones más altas (e.g., valores de `reviewer rating` cercanos a 5) tienden a asociarse con mayores puntajes positivos, lo cual es consistente con las expectativas.
- **Gráfico Negativo:** Aquí, el eje y muestra el puntaje de polaridad negativa. Existe una menor cantidad de puntos con valores negativos significativos, lo que puede indicar una inclinación general hacia opiniones más positivas.
- **Gráfico Neutral:** El eje y muestra el puntaje de neutralidad. Aunque hay reseñas con calificaciones altas que también tienen puntajes neutrales, esto podría sugerir que algunas reseñas muy positivas no reflejan completamente una polaridad positiva.

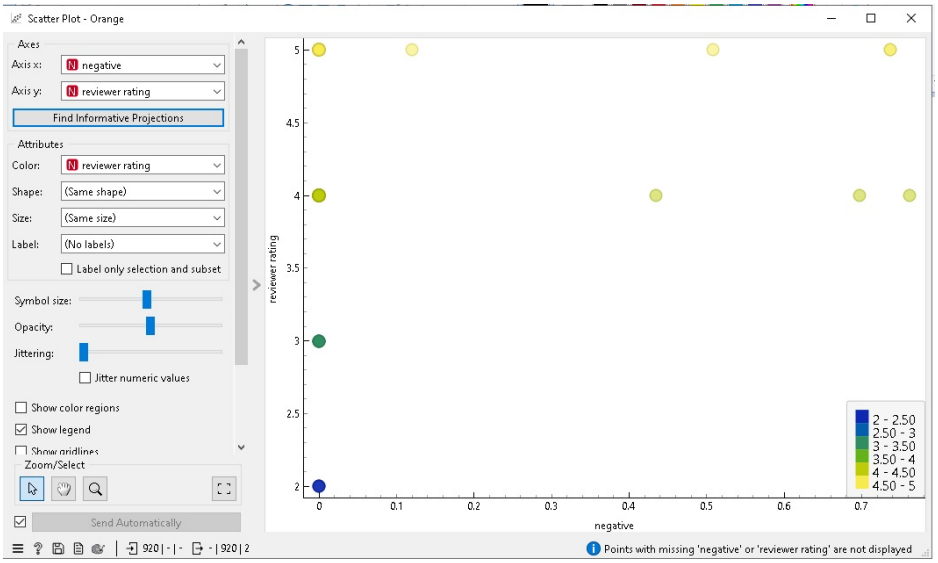
2.3. Análisis de las gráficas

De la observación de los gráficos, surgen los siguientes puntos críticos:

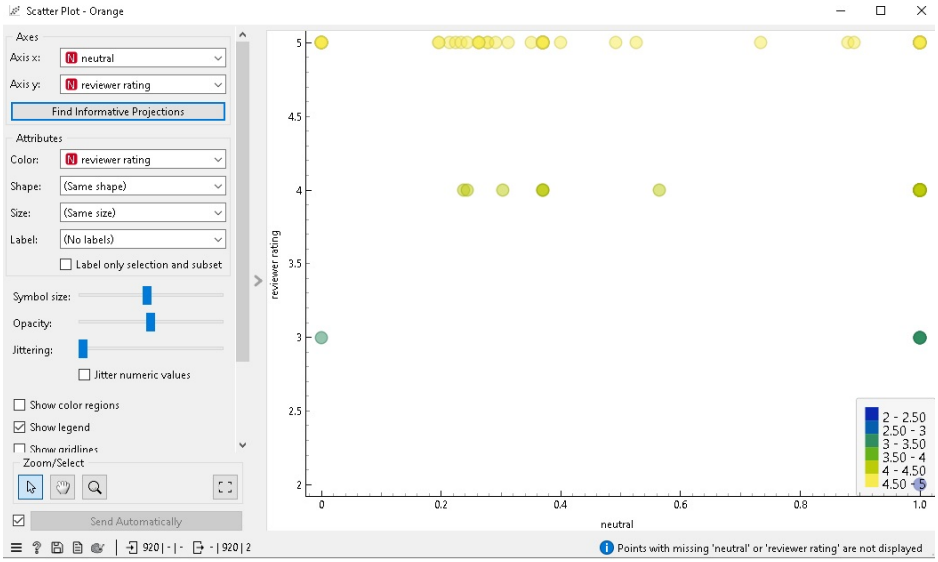
1. **Correlación esperada:** Existe una correlación clara entre reseñas altas y polaridades positivas, lo que valida la interpretación automática de sentimientos.
2. **Polaridades negativas:** La baja frecuencia de reseñas con alta negatividad puede indicar que la mayoría de los lectores tienen experiencias satisfactorias.
3. **Polaridad neutral:** El comportamiento del puntaje neutral muestra que no todas las reseñas extremas en términos de calificación numérica están alineadas con polaridades emocionales. Esto puede requerir una revisión del modelo de análisis de sentimientos o la incorporación de un mayor contexto.



(a) Polaridad positiva



(b) Polaridad negativa



(c) Polaridad neutral

Figura 2: Gráficos de dispersión: Relación entre calificaciones y polaridades.

Esta información puede ser utilizada para comprender mejor las percepciones de los lectores sobre los libros y mejorar los modelos de análisis de sentimientos.

2.4. Análisis de las Distribuciones con Box Plot

Las siguientes figuras representan el análisis de las distribuciones de las variables de polaridad positiva, negativa y neutral en las reseñas de los libros, utilizando diagramas de caja (Box Plot). Este análisis permite identificar tendencias y dispersiones en los datos de cada categoría de sentimiento.

2.5. Distribución de Sentimientos Positivos

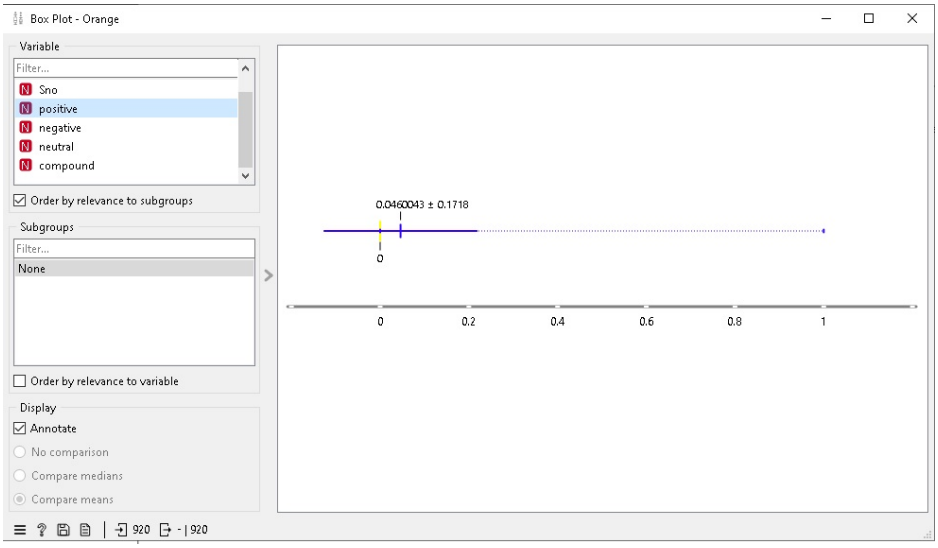
En la Figura 3a, se observa que la mayoría de las reseñas presentan valores bajos de polaridad positiva, con una media cercana a 0,046 y una desviación estándar reducida ($\pm$0,1718). Esto indica que, aunque hay sentimientos positivos, estos tienden a ser de baja intensidad.

2.6. Distribución de Sentimientos Negativos

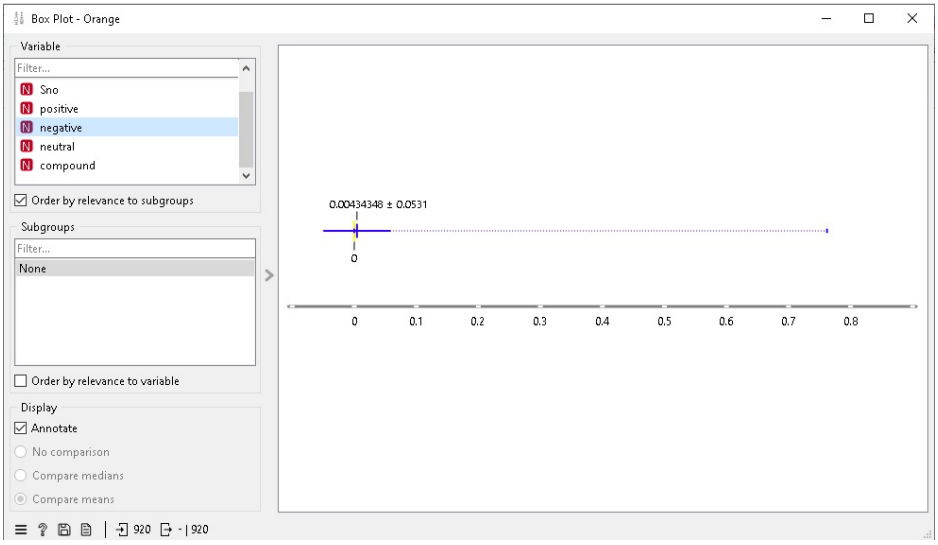
En la Figura 3b, los valores de polaridad negativa tienen una media significativamente baja ($0{,}0043 \pm 0{,}0531$), lo que sugiere que las reseñas negativas son muy escasas o de baja intensidad.

2.7. Distribución de Sentimientos Neutrales

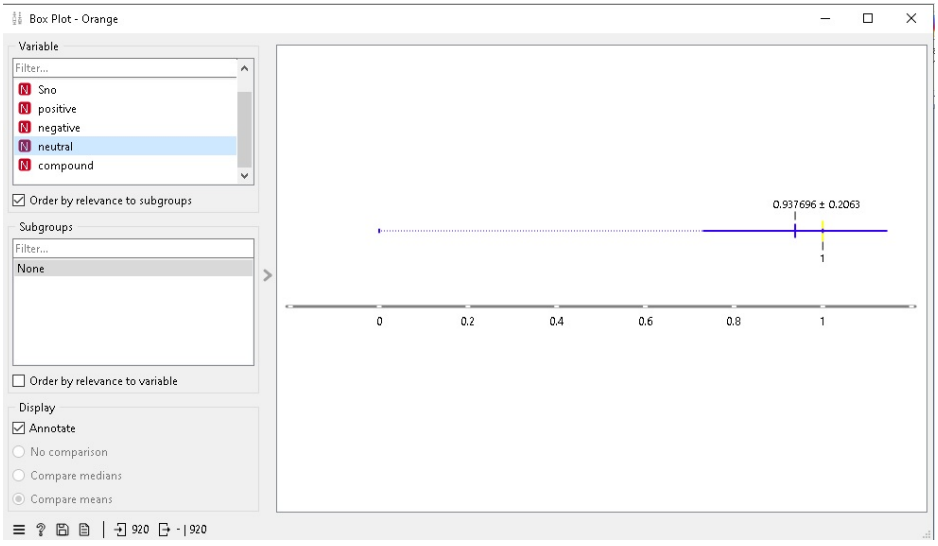
En la Figura 3c, se muestra que los valores neutrales tienen una alta concentración en $0{,}9376 \pm 0{,}2063$, lo que indica que la mayoría de las reseñas son percibidas como neutras en su sentimiento. Esto podría deberse a una tendencia general de los reseñadores a proporcionar comentarios objetivos o carentes de fuertes emociones.



(a) Polaridad Positiva



(b) Polaridad Negativa



(c) Polaridad Neutral

Figura 3: Análisis de Distribución de Polaridades de Sentimientos en las Reseñas.

3. Identificación de Temas Recurrentes

Objetivo: Extraer y analizar los temas más comunes presentes en las reseñas de libros para identificar patrones en las opiniones de los lectores.

Flujo del modelo implementado: Se utilizó el modelo de flujo mostrado en la Figura 4. Este flujo incluye los siguientes pasos:

- **Preprocesamiento del texto:** Se aplicaron técnicas de normalización, tokenización y eliminación de palabras vacías para preparar los datos.
- **Modelado de temas:** Con el widget `Topic Modeling`, se identificaron los temas más representativos en las reseñas.
- **Visualización:** Los resultados fueron explorados mediante nubes de palabras, tablas de datos y diagramas de dispersión.

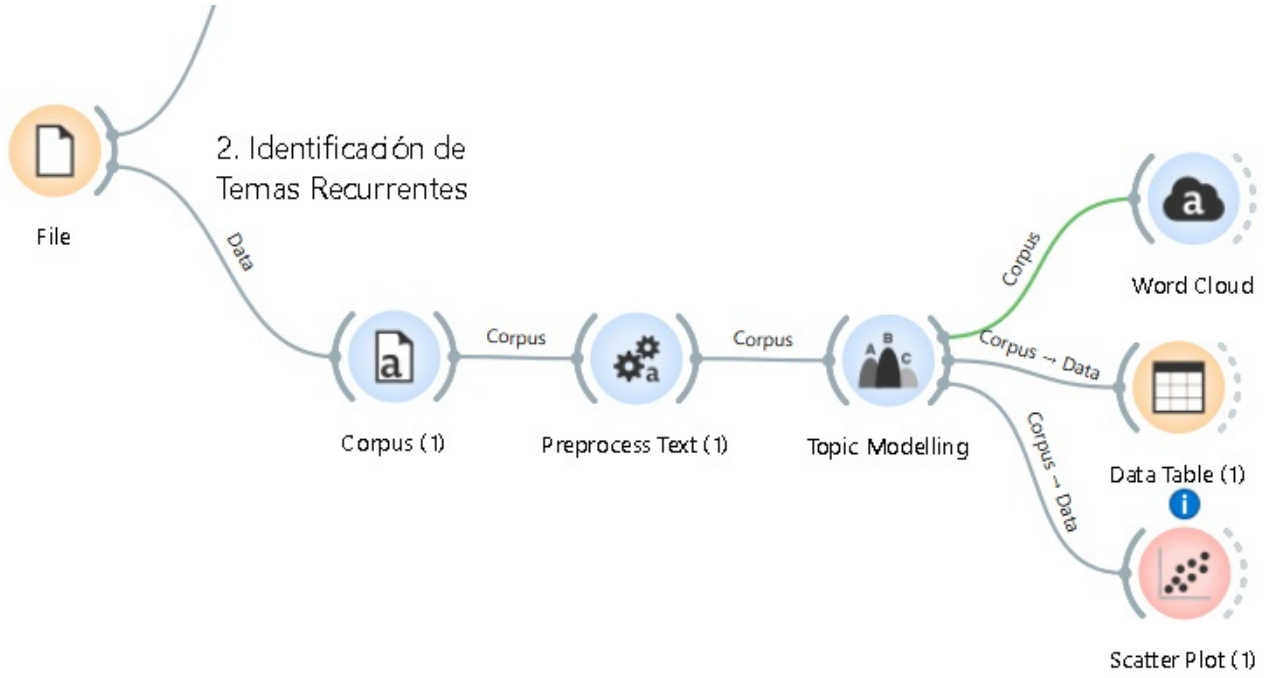


Figura 4: Flujo del modelo para la identificación de temas recurrentes.

3.1. Visualización de la Nube de Palabras

Objetivo: Visualizar los términos más representativos presentes en las reseñas de los libros para identificar patrones recurrentes y temáticas principales.

Análisis: La Figura 5 muestra la nube de palabras generada a partir de las reseñas. Entre los términos más destacados se encuentran *"book"*, *"board"*, *"kids"*, *çourt"*, y *"thorn"*. Esto sugiere que los lectores mencionan de manera recurrente conceptos relacionados con libros infantiles, novelas, y características específicas de ciertos géneros o títulos.

Palabras como *"game"*, *"science"*, *"shadow"*, y *Çhristmas"* reflejan temas diversos que abarcan desde elementos narrativos hasta libros temáticos. Este análisis preliminar indica que las opiniones de los usuarios

están influenciadas tanto por los géneros literarios como por temáticas específicas, permitiendo futuras exploraciones sobre su relación con las calificaciones y los sentimientos expresados.

Conclusión: La nube de palabras proporciona una representación visual intuitiva de los temas predominantes en las reseñas, facilitando la identificación rápida de patrones relevantes.

Figura 5: Nube de palabras generada a partir de las reseñas de los libros.

3.2. Análisis del Box Plot: Relación entre el Tema 1 y las Calificaciones de los Usuarios

Objetivo: Analizar la relación entre el Tema 1 identificado en el modelado de temas y las calificaciones otorgadas por los usuarios.

Análisis: La Figura 6 muestra un gráfico de dispersión que relaciona las calificaciones de los usuarios (*reviewer rating*) con el peso del Tema 1 (*Topic 1*) en las reseñas. En este análisis se observa que:

- Las reseñas con altas calificaciones (valores cercanos a 5) tienden a tener mayores pesos asociados al Tema 1.
- Existe una menor presencia de valores altos en el Tema 1 cuando las calificaciones son bajas (valores cercanos a 2), lo que podría sugerir que este tema está relacionado con aspectos positivos o atributos apreciados por los usuarios.
- Los valores intermedios del Tema 1 están distribuidos a lo largo de las calificaciones, indicando que este tema puede estar moderadamente presente en reseñas con diversas puntuaciones.

Conclusión: El análisis indica que el Tema 1 podría ser un factor asociado a percepciones positivas, aunque se requieren análisis adicionales para confirmar su impacto directo en las calificaciones.

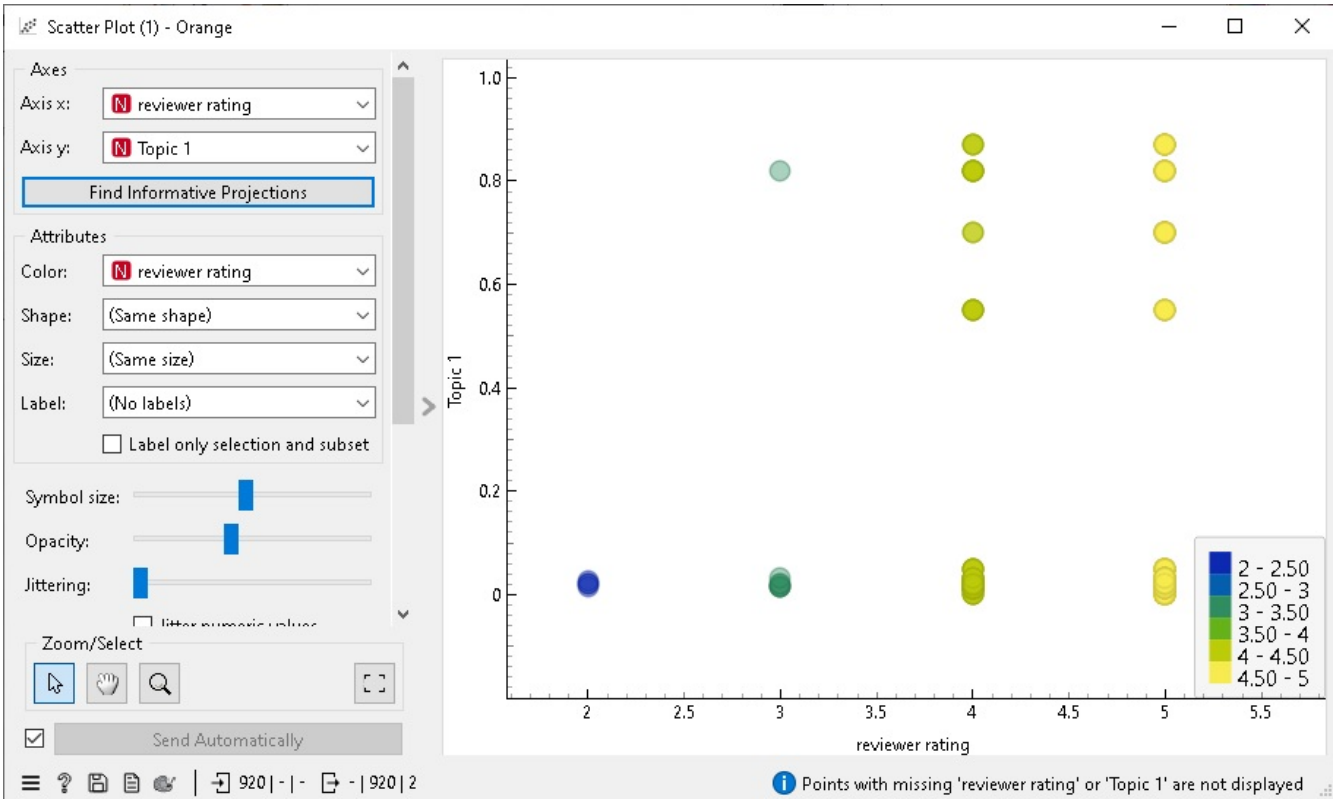


Figura 6: Scatter Plot mostrando la relación entre el Tema 1 y las calificaciones de los usuarios.

4. Clustering de Reseñas de Libros

4.1. Objetivo

El propósito de este modelo es agrupar las reseñas de libros de Amazon según su similitud textual, identificando patrones comunes en las opiniones de los usuarios. Esto permite observar cómo las reseñas se organizan en función de los temas o características dominantes y visualizar relaciones en un espacio de menor dimensión.

4.2. Descripción del Modelo

El flujo de trabajo para este análisis se ilustra en la Figura 7 y se compone de los siguientes pasos:

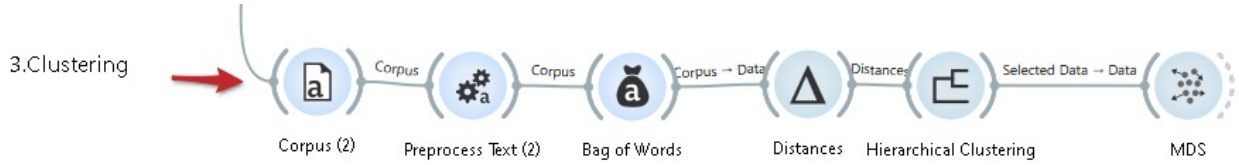


Figura 7: Flujo de trabajo del modelo de clustering implementado.

1. **Corpus:** Se carga el conjunto de datos de reseñas, que incluye texto no estructurado.

2. **Preprocess Text:** El texto es procesado para eliminar elementos no informativos, como puntuación, palabras vacías (stopwords) y convertir palabras a minúsculas.
3. **Bag of Words:** Se genera una representación vectorial del texto mediante el modelo "Bolsa de Palabras", en el que se cuenta la frecuencia de las palabras en cada reseña.
4. **Distances:** Se calcula una matriz de distancias entre las reseñas utilizando la distancia del coseno, para medir similitudes textuales.
5. **Hierarchical Clustering:** Se genera un dendrograma que representa cómo las reseñas se agrupan jerárquicamente según su similitud.
6. **MDS (Multidimensional Scaling):** Se proyectan las reseñas en un espacio bidimensional para visualizar los patrones de agrupamiento de manera intuitiva.

4.3. Resultados y Análisis

4.4. Descripción del Dendrograma

La visualización presentada corresponde a un dendrograma generado a partir del modelo de *Clustering Jerárquico* aplicado al conjunto de datos de reseñas de libros de Amazon. Cada nivel del dendrograma representa un paso en el proceso de agrupamiento, donde las reseñas se agrupan según su similitud textual.

- **Coloración por Reviewer Rating"**: Los colores de los nodos indican las calificaciones de los revisores, permitiendo observar cómo las calificaciones están distribuidas en los diferentes grupos.
- **Agrupaciones Principales**: Las principales agrupaciones se identifican en función del corte seleccionado (profundidad o número de clústeres). En este caso, los grupos como *"Just Because"* y *"Pookie's Thanksgiving"* destacan como clústeres dominantes debido a su alta similitud textual.
- **Escala Horizontal**: Representa la distancia de similitud, donde los valores más bajos (a la izquierda) corresponden a elementos más similares entre sí y los valores más altos (a la derecha) a elementos menos similares.

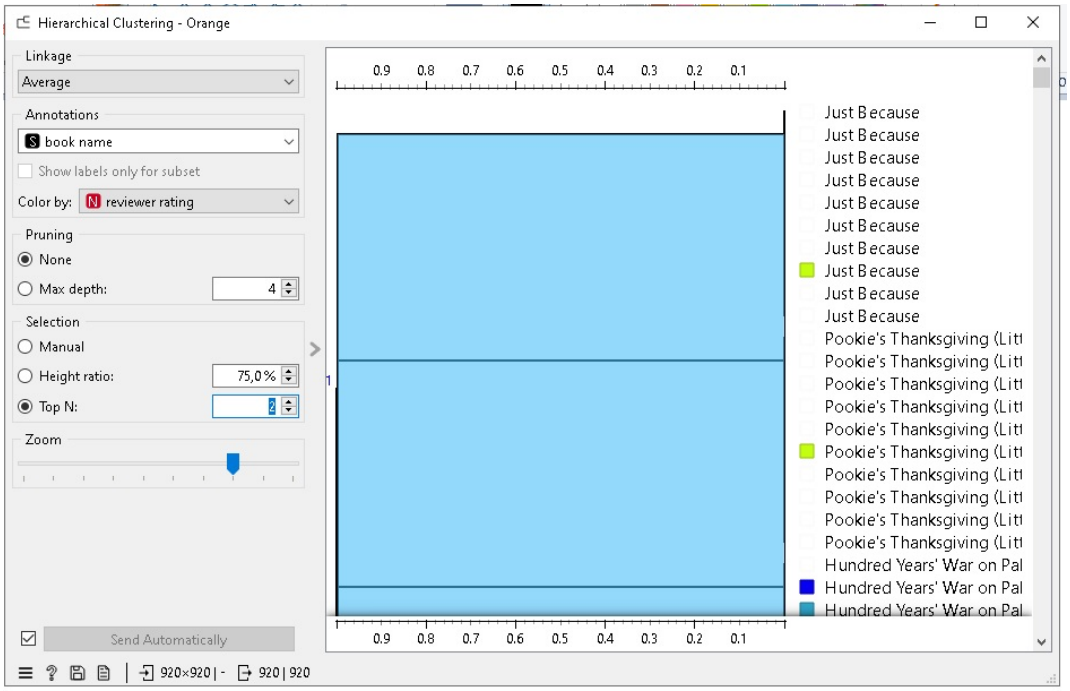


Figura 8: Dendrograma generado a partir del Clustering Jerárquico de las reseñas de libros de Amazon. Las agrupaciones reflejan la similitud textual entre las reseñas, destacando temas comunes o características compartidas.

Esta visualización es útil para analizar patrones en las opiniones y características compartidas en las reseñas, identificar temas recurrentes y explorar diferencias significativas entre los grupos.

4.5. Importancia del Análisis

Este modelo de clustering ayuda a identificar temas clave en las reseñas, lo que puede ser útil para los autores y editores de libros. Por ejemplo, se puede detectar si un libro tiene críticas frecuentes relacionadas con su precio o si un género literario específico tiende a generar más opiniones polarizadas.

4.6. Visualización de MDS para Clustering de Reseñas

Descripción: La visualización Multidimensional Scaling (MDS) muestra una representación bidimensional de la similitud entre las reseñas de libros de Amazon. Los puntos representan las reseñas, mientras que los colores reflejan la calificación asignada por los usuarios.

Análisis:

- **Agrupamiento:** Se observa que las reseñas con calificaciones similares tienden a agruparse en áreas cercanas, lo que indica una relación entre el contenido textual y las calificaciones otorgadas.
- **Distribución:** Los puntos más dispersos reflejan reseñas con opiniones más diversas, mientras que las áreas densamente pobladas representan opiniones más homogéneas.
- **Clusterización:** Los colores también resaltan cómo los clústeres identificados en el dendrograma previo se mantienen en esta representación visual.

Interpretación: Esta visualización proporciona una perspectiva intuitiva de cómo las calificaciones de las reseñas están relacionadas con el contenido textual, ayudando a identificar patrones y variaciones entre las opiniones de los usuarios.

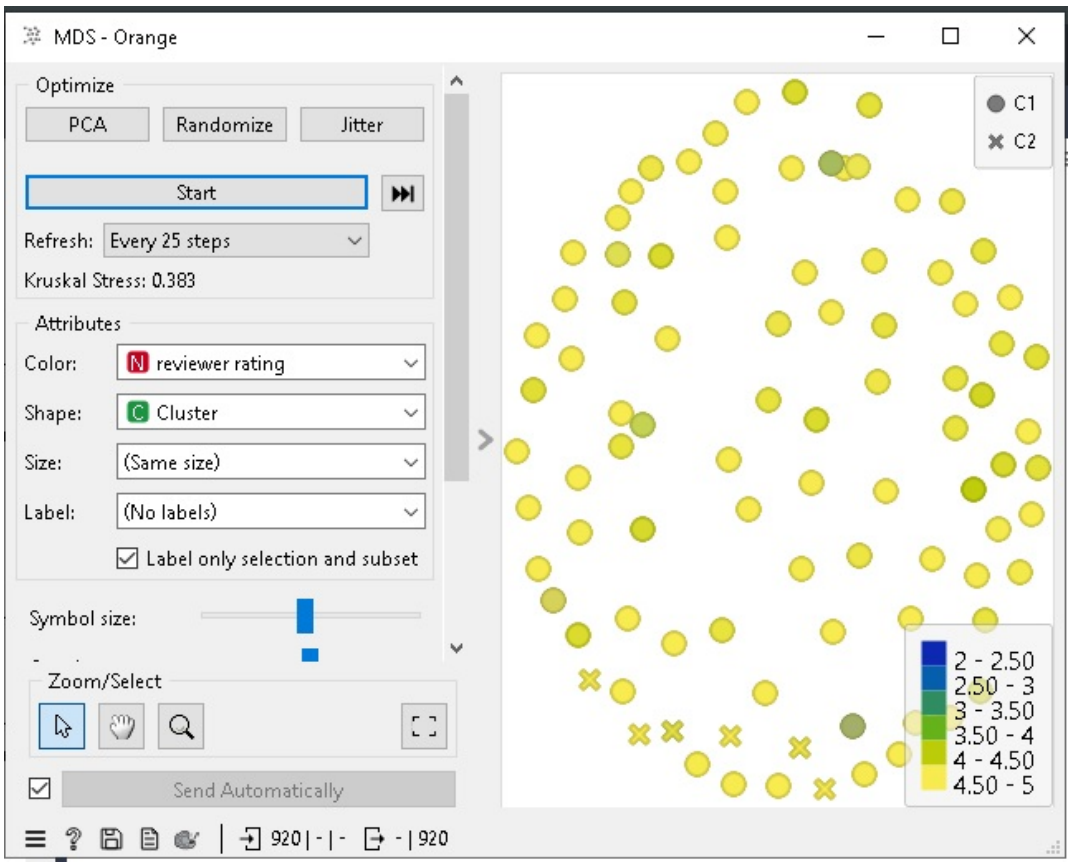


Figura 9: Visualización MDS de las reseñas de libros de Amazon, mostrando la relación entre la similitud textual y las calificaciones asignadas.

5. Modelo Supervisado con Redes Neuronales

El objetivo principal de esta sección fue implementar un modelo de aprendizaje supervisado utilizando redes neuronales para predecir la calificación de las reseñas (reviewer rating) en función de las características textuales del contenido. El flujo diseñado, mostrado en la Figura 10, incluye pasos esenciales para el procesamiento de texto, transformación a una representación numérica mediante *Bag of Words*, y posterior entrenamiento del modelo.

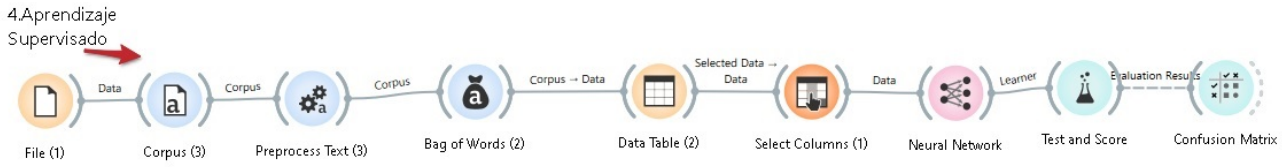


Figura 10: Flujo del modelo supervisado con redes neuronales en Orange.

5.1. Flujo de Trabajo y Configuración

El flujo de trabajo consistió en los siguientes pasos:

- **Preprocesamiento del texto:** A través de normalización, eliminación de palabras vacías (*stopwords*), y generación de tokens.

- **Vectorización:** Representación de las palabras mediante el modelo de *Bag of Words*.
- **Selección de Columnas:** Configuración del campo *reviewer rating* como objetivo (*target*) del modelo.
- **Entrenamiento:** Uso del modelo de red neuronal (*Neural Network*) con activación *tanh* y optimización mediante *Adam*, ajustando los hiperparámetros para minimizar el error.
- **Evaluación:** Se analizaron métricas de desempeño en el widget *Test and Score* y se utilizaron herramientas como *Confusion Matrix* para identificar áreas de mejora.

5.2. Dificultades y Observaciones

Durante la implementación, enfrenté desafíos relacionados con el preprocesamiento y configuración de variables, en particular:

- La alta dimensionalidad de las características textuales generadas por el modelo *Bag of Words* complicó la selección eficiente de atributos relevantes.
- La interpretación y representación de los resultados requerían ajustes adicionales en los parámetros del modelo.
- A pesar de los esfuerzos por optimizar la configuración de la red neuronal, persistieron discrepancias entre los valores predichos y las calificaciones reales, lo que resalta la necesidad de incluir más contexto o explorar arquitecturas alternativas.

5.3. Observaciones

Aunque no se logró completar las pruebas debido a la falta de generación de datos en el flujo, el diseño planteado buscaba evaluar la capacidad de las redes neuronales para predecir calificaciones basadas en texto. Este enfoque habría permitido analizar la relación entre las características textuales de las reseñas y las valoraciones numéricas, aprovechando la flexibilidad de las redes neuronales para modelar patrones complejos en datos no estructurados.

Se esperaba obtener un modelo que no solo identificara tendencias generales en las calificaciones, sino que también detectara discrepancias importantes, lo que sería valioso para aplicaciones como la mejora de productos o el análisis de la satisfacción del cliente.

6. Conclusión Global

Este trabajo representó mi primer acercamiento a la Ciencia de Datos, un campo que hasta ahora solo había conocido vagamente de forma teórica. A través de los cuatro modelos desarrollados, intenté mejorar mi comprensión del proceso de análisis de datos y aprender a utilizar herramientas como Orange para realizar tareas prácticas como el preprocesamiento, la clasificación y la exploración de datos.

Aunque enfrenté varios desafíos y no todos los resultados fueron los esperados, esta experiencia me permitió dar un paso inicial en un área de estudio y aplicaciones concreto y actual, y , según entiendo, con gran demanda profesional. Este trabajo despertó mi interés por seguir aprendiendo y explorando más sobre Ciencia de Datos en el futuro, con la intención de mejorar y aplicar estos conocimientos en proyectos más complejos.