Reporte Integral de Análisis Exploratorio de Datos (EDA)

Equipo 4
Joan Antonio Lazaro Silva
Caro Perez Horacio
Joaquín Alfredo Castro Córdova

20 de enero de 2026

Resumen

Este documento presenta el análisis exploratorio realizado sobre tres corpus distintos: **CNN Articles** (Periodismo), **ARvs** (Reseñas de Amazon) y **GTBG** (Project Gutenberg). Se detallan las características volumétricas, estilísticas y de calidad de cada conjunto de datos, así como su potencial aplicación en tareas de Procesamiento de Lenguaje Natural (NLP).

1 Análisis del Corpus: CNN Articles

1.1 Caracterización General

El corpus está compuesto por noticias globales, predominando la política internacional y nacional de EE.UU.

- **Número total de documentos:** 2,065,892.
- **Total de palabras:** 60,490,380.
- **Longitud del vocabulario:** 809,367 palabras únicas.

- **Idioma:** Inglés (Confirmado por vocabulario).
- **Tono:** Formal / Periodístico.

1.2 Interpretación de Métricas

- **Densidad Léxica (1.34 %):** Es extremadamente baja. Esto es típico de textos periodísticos estandarizados y masivos donde predominan las palabras funcionales y se repiten constantemente los mismos términos noticiosos.
- **Extensión (29.28 palabras/doc):** El promedio indica que el corpus no contiene artículos completos, sino fragmentos breves, titulares o "leads"(noticias de último minuto).

1.3 Categorización Temática (Muestreo)

Al no contar con categorías explícitas, se identificaron mediante muestreo:

- **Política:** ∼251,786 documentos.
- **Deportes:** ∼140,036 documentos.
- **Salud/Ciencia:** ∼99,676 documentos.

1.4 Calidad y Limpieza

Se considera **Moderadamente Sucio**.

- Contiene marcas técnicas como `@delimiter` (aparece casi 100,000 veces funcionando como separador).
- Presencia de errores de formato como guiones dobles (−−) al inicio de líneas.
- Requiere limpieza de *stopwords* para ser útil.

1.5 Tareas de NLP Recomendadas

1. **Clasificación de Texto:** Etiquetar el 76 % de noticias "sin tema.$^{\text{en}}$ categorías específicas.

2. **Generación de Titulares:** Ideal debido a la brevedad de los textos.
3. **Reconocimiento de Entidades (NER):** Extracción de nombres de políticos y lugares.

2 Análisis del Corpus: ARvs (Amazon Reviews)

2.1 Caracterización General

Este corpus consiste en opiniones de usuarios sobre productos electrónicos, específicamente tabletas **"Kindle"** y **"Fire"**.

- **Volumen de Datos:** 1,597 documentos.
- **Idioma:** Inglés predominante.
- **Temas Recurrentes:** Hardware, pantalla (*screen*), batería (*battery*) y precio.

2.2 Análisis de Tono y Estilo

El tono es **Informal y Positivo**.

- Uso de lenguaje coloquial directo.
- Alta frecuencia de adjetivos positivos como *"great"*, *"love"* y *"good"* (Ver Figura 2).

2.3 Estadísticas Clave

- **Promedio de longitud:** 166.38 palabras. Son textos de extensión media (párrafos), no frases cortas.
- **Densidad Léxica (4.50 %):** Valor bajo, indicando repetición de palabras clave, lo cual facilita el hallazgo de patrones.

2.4 Visualizaciones

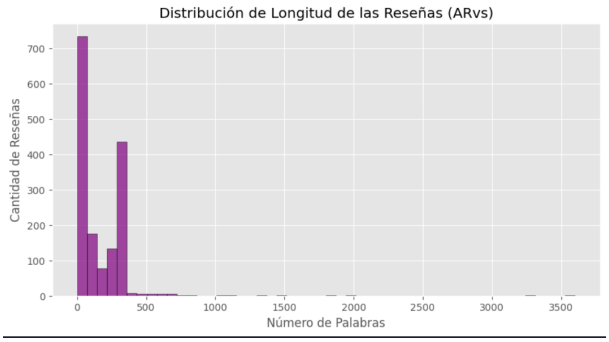

Figura 1: Histograma de longitud. Concentración en textos breves.

Figura 2: WordCloud. Términos destacados: Kindle, Fire, Tablet.

2.5 Calidad y Anomalías

- **Limpieza:** Bastante limpio (0 etiquetas HTML), aunque contiene 23 URLs dispersas.
- **Anomalías (Outliers):** Como se observa en la Figura 1, existe una distribución de çola larga". Aunque la mayoría son breves (<200 palabras), hay reseñas extremas de más de 3,000 palabras que requieren truncamiento.

2.6 Tareas de NLP Recomendadas

1. **Análisis de Sentimiento:** Determinar polaridad (positiva/negativa).
2. **Resumen Automático:** Para las reseñas atípicas de gran extensión.

3 Análisis del Corpus: GTBG (Project Gutenberg)

3.1 Caracterización General

Subconjunto de obras literarias de dominio público. A diferencia de los otros corpus, aquí cada "documento.es un libro completo.

- **Total de documentos:** 1,500.
- **Categoría Dominante:** Literatura Infantil (*Children's*) con 213 documentos.
- **Otras Categorías:** Ciencia, Crimen, Ficción.
- **Idioma:** Inglés (100 %).
- **Tono:** Formal, Literario y Clásico (Siglos XVIII-XIX).

3.2 Estadísticas Masivas

- **Promedio de longitud:** 91,601 palabras por documento (Libros enteros).
- **Vocabulario único:** 1,343,371 palabras.
- **Densidad Léxica (0.98 %):** Valor extremadamente bajo, esperado dado que en libros completos se reutiliza todo el vocabulario disponible del idioma.
- **Top Palabras:** Predominio absoluto de stopwords (*the, of, and*), con "the.ªpareciendo más de 8 millones de veces.

3.3 Calidad y Limpieza

Nivel de suciedad: **Moderado**.

- **Metadatos:** Contiene licencias de Project Gutenberg e índices que no son parte de la narrativa.
- **Etiquetas de Transcripción:** Marcas como `[Illustration:...]` o `[Note:...]`.
- **Formato:** Saltos de línea irregulares típicos de archivos .txt antiguos.

3.4 Hallazgos Curiosos

- **Categoría ÇIA":** Se encontraron 21 documentos etiquetados bajo ÇIA". Esto sugiere la inclusión de reportes de inteligencia de dominio público (como el *World Factbook*) dentro de la colección literaria, lo cual es atípico.

3.5 Tareas de NLP Recomendadas

1. **Clasificación de Género Literario:** Predecir si un libro es Çhildren's.º Çrime"basado en vocabulario.
2. **Estilometría:** Comparar la complejidad gramatical entre autores o épocas.
3. **Resumen Automático (Summarization):** Generar sinopsis para evitar la lectura de textos masivos.