

# **Reporte: Tarea 1**

Bórquez Guerrero Angel Fernando - 219208106  
Machado Félix Flor María - 223220679  
Universidad de Sonora  
Procesamiento de Lenguaje Natural

## **1. Corpus 1: RestMex2021-train**

El primer corpus está conformado por 4416 documentos, contiene un título y una opinión, acompañado de las siguientes categorías: place, gender, age, country, date, label.

Podemos clasificar los textos por medio del lugar al que se le ha realizado la opinión, siendo los lugares: Alhondiga (487), Basilica Colegiate (213), Callejón del beso (638), Casa de Diego Rivera (292), Jardín de la Unión (418), Mercado Hidalgo (239), Monumento Pípila (599), Museo de las Momias(722), Teatro Juarez (378) y Universidad de Guanajuato (431).

También podemos identificar las opiniones por el país de origen de la persona, siendo en total gente de 43 países diferentes, con mayor cantidad México (3434).

Sobre los textos, están escritos mayormente en español (con excepción de un par que estaban en inglés) y el tono, se puede identificar como no muy formal. Sobre el estado de los documentos, se observa que tanto el genero como el país se encuentra la respuesta N/I, y la edad existen múltiples ocasiones que aparece como -1. Sobre la fecha, no existe un formato único, ya que aveces aparece el día, mes y año juntos, y en otras únicamente el año

Como se observa en la siguiente figura, donde se muestra las palabras más usadas en los títulos de los documentos, se encuentran palabras como Guanajuato, ciudad, museo, arquitectura, se entiende que los documentos se relacionan a lugares turísticos dentro de Guanajuato. Por palabras como interesante, excelente, hermoso, se entiende que hablamos de reseñas a lugares turísticos dentro de Guanajuato.



Figura 1: Nube de palabras en referencia a la categoría "Título"

Las top palabras más usadas en los documentos son los siguientes:

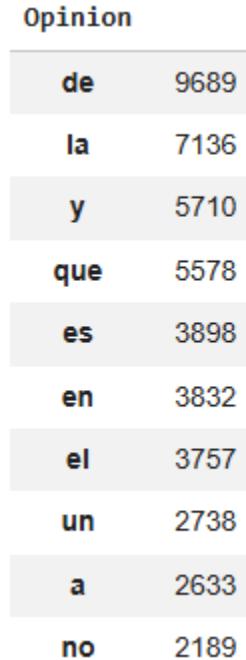


Figura 2: Palabras más usadas, y su cantidad de usos

Sin embargo, esto deja mucho que desear, entonces realizamos una filtración simple de los 25 conectores más usados en el texto:

– de, la, es, y, el, que, en, un, a, no, para, una, por, los, se, las, con, del, lo, te, al, si, su, esta, o –

Dejándonos el siguiente resultado:

Opinion	
<b>muy</b>	1953
<b>lugar</b>	1473
<b>pero</b>	1008
<b>hay</b>	945
<b>este</b>	765
<b>historia</b>	745
<b>como</b>	718
<b>más</b>	718
<b>guanajuato</b>	640
<b>museo</b>	620

Figura 3: Palabras más usadas, y su cantidad de usos

Aunque sea poco, podemos ver mejor sobre la relación del texto con la temática.

Sobre la longitud de los documentos: En total son 182228 palabras en el apartado de Opinion. Incluye un vocabulario (palabras únicas) de 17954 palabras. Eso nos da en la densidad de la léxica un total de 9.85.

## 2. Corpus 2: cnn

### 2.1. Descripción general del corpus

El corpus **cnn** está conformado por un conjunto amplio de noticias provenientes de la cadena informativa CNN. Cada documento corresponde a una noticia completa, y los documentos se encuentran separados mediante el marcador @delimiter.

- **Número total de documentos:** 92,464
- **Idioma del corpus:** Inglés
- **Categorías:** El corpus no presenta categorías explícitas (por ejemplo, política, economía, deporte). Todas las noticias se analizan como un solo conjunto.

### 2.2. Temas principales

El tema principal del corpus es **noticias**, cubriendo distintos eventos informativos reportados por CNN.

Aunque no existen etiquetas temáticas explícitas, el vocabulario dominante sugiere contenidos relacionados con:

- política
- salud
- conflictos internacionales
- sucesos relevantes y reportes oficiales

Esto convierte al corpus en un conjunto generalista de noticias periodísticas.

### 2.3. Tono predominante

El tono del corpus es formal.

Las noticias presentan:

- lenguaje informativo
- estilo objetivo
- uso frecuente de citas indirectas y directas
- estructura típica del periodismo profesional

### 2.4. Nivel de “limpieza” del corpus

El corpus puede considerarse relativamente limpio:

- No contiene etiquetas HTML
- No se observan URLs incrustadas
- No hay marcas de formato web

- El único elemento estructural artificial es el delimitador @delimiter, utilizado exclusivamente para separar documentos

No obstante, se detectan ciertos detalles típicos de texto crudo periodístico, como:

- contracciones separadas (s, como aparece en el top de palabras)
  - signos de puntuación adheridos a palabras antes del preprocesamiento

En general, el corpus es adecuado para análisis NLP sin limpieza intensiva.

## 2.5. Posibles tareas de NLP aplicables

Este corpus es especialmente adecuado para **Topic modeling** para identificar automáticamente los temas latentes en las noticias.

## 2.6. Análisis cuantitativo del corpus

### 2.6.1. Nube de palabras



Figura 4: Nube de palabras en referencia al corpus cnn

### 2.6.2. Histograma de longitudes de documentos

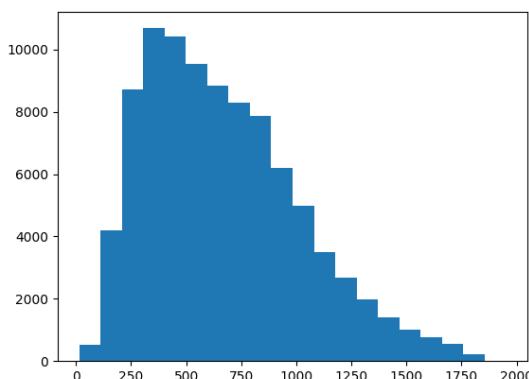


Figura 5: Distribución de longitudes de las noticias en número de palabras

### **2.6.3. Promedio de longitud por documento**

- Promedio de palabras por noticia: 670.63

### **2.6.4. Top 10 palabras más usadas**

Palabra	Frecuencia
the	3,624,734
to	1,693,174
of	1,521,217
and	1,477,344
a	1,467,066
in	1,305,697
s	842,617
that	714,287
for	562,612
is	535,370

Estas palabras corresponden principalmente a artículos, preposiciones y verbos auxiliares, lo cual es esperado en textos en inglés.

### **2.6.5. Longitud del vocabulario**

- Número total de palabras: 62,009,537
- Número de palabras únicas: 237,087

### **2.6.6. Densidad léxica**

La densidad léxica se calculó mediante la fórmula:

$$\text{Densidad léxica} = \frac{\text{palabras únicas}}{\text{palabras totales}} \times 100$$

- Densidad léxica: 0.38 %