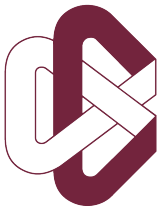


CIMAT

Centro de Investigación en Matemáticas
Unidad Monterrey

Análisis de Texto e Imágenes
Tarea 1

Gustavo Hernández Angeles

6 de septiembre de 2025

Índice

Ejecutar el código

Esta sección detalla los pasos para configurar el entorno y ejecutar el código de los problemas.

1. Instalación de Dependencias

Para evitar errores de compatibilidad, se recomienda crear un entorno virtual e instalar las dependencias especificadas. El proyecto utiliza `uv` para una gestión rápida del entorno. Desde la raíz del proyecto, ejecute el siguiente comando para crear el entorno y sincronizar las dependencias del archivo `pyproject.toml`:

```
uv sync
```

Una vez finalizado, active el entorno virtual. El comando varía según su sistema operativo:

- **Linux o macOS:**

```
source ./.venv/bin/activate
```

- **Windows (PowerShell):**

```
.\.venv\Scripts\Activate.ps1
```

- **Windows (Command Prompt):**

```
.\.venv\Scripts\activate
```

2. Configuración de la Ruta de Datos

Antes de ejecutar cualquier script, es necesario especificar la ubicación del archivo de datos. Abra el archivo `codigo/config.py` y modifique la variable correspondiente con la ruta correcta a su archivo.

3. Ejecución del Código

Con el entorno activado y la ruta configurada, puede ejecutar los scripts desde la raíz del proyecto.Para ejecutar todos los problemas en secuencia:

```
python -m codigo.main
```

Para ejecutar un problema específico, utilice el siguiente comando, reemplazando `XX` por el número del problema que desea ejecutar:

```
python -m codigo.problemas.problemaXX
```

Nota importante: Para los problemas 5 y 6, que están combinados en un solo script, debe usar `5_6` en lugar de un número. Por ejemplo: `python -m codigo.problemas.problema5_6`.

Descripción del Corpus

> **1**
>
> Analiza el corpus y reporta:
>
> a) Numero de documentos, tokens y vocabulario.
>
> b) Hapax legomena y su proporcion.
>
> c) Porcentaje de stopwords.
>
> d) Estadisticas por clase (numero de documentos, tokens y vocabulario).

Inciso a)

El corpus consiste en **5000 reseñas/documentos** clasificadas por su *Polaridad*, consistiendo en un entero del rango 1-5, donde 5 representa una opinión completamente favorable para el local y 1 una opinión pesimista del mismo (el negocio puede variar por reseña, incluso en el giro: Restaurante, Atractivo, Hotel). En esta práctica utilizaremos la tokenización por palabras en su mayoría, haciendo que coincida con el vocabulario. Obtuvimos que:

- El número de palabras/tokens que hay en el corpus es de 350,871.
- El vocabulario consiste en 22,319 palabras.

Inciso b)

El concepto de *Hapax Legomena* refiere a palabras cuya frecuencia de aparición en un corpus es única, es decir, aparece solo una vez en todo el corpus. En este caso obtuvimos que **existen 11,671 hapax's** en todo el corpus, consituyendo un **52.3 %** de todo el vocabulario. Una causa de esta alta proporción es el tamaño limitado del corpus.

Inciso c)

Encontré un total de 176k apariciones de stopwords en todo el corpus, constituyendo un 50 % del corpus. Sin embargo, también obtuve que fueron 229 stopwords las que constituyeron todas las apariciones. ¡Cerca de la mitad del corpus no nos da información alguna sobre el contenido de los textos!

Inciso d)

Las estadísticas por clase se pueden resumir en la tabla 2.1. Aquí podemos observar que la clase de reviews con una polaridad de 5 son las de mayor frecuencia en el corpus. Sin embargo, al calcular el número de palabras promedio por review por polaridad, obtenemos que las reviews más *descriptivas* suelen ser las de baja polaridad. Esto quiere decir que, a medida que la polaridad disminuye, la longitud de las reviews es más alta. No se encuentra una diferencia signicativa entre el tamaño de vocabulario por cada polaridad en las reviews.

Polaridad	**1**	**2**	**3**	**4**	**5**
# Documentos	800	900	1000	1100	1200
# Palabras	62249	75380	69868	68992	74382
# Palbrs. / # Docs.	77.81	83.76	69.87	62.72	61.98
Tamaño de vocabulario	8609	9323	8444	8204	9334

Cuadro 2.1: Polarity Statistics

Ley de Zipf

2

Calcula la frecuencia absoluta $f(w)$ de cada palabra w en el corpus y ordénalas de mayor a menor. A cada palabra así ordenada se le asigna un rango r, donde $r = 1$ corresponde a la palabra más frecuente, $r = 2$ a la segunda, y así sucesivamente.

- Representa gráficamente la relación entre **log-rango** y **log-frecuencia**. Es decir, para cada palabra graficar el punto $(\log r, \log f(w))$. La Ley de Zipf predice que los puntos deberían aproximarse a una línea recta decreciente.
- Ajusta una recta mediante regresión lineal sobre los puntos $(\log r, \log f(w))$, de la forma:
$$\log f(r) = \log C - s \cdot \log r, \tag{3.1}$$
lo cual equivale al modelo Zipfiano $f(r) \approx \frac{C}{r^s}$. En esta formulación:
  - C es una constante de normalización que se aproxima a la frecuencia de la palabra más común ($f(1) \approx C$).
  - s es el exponente de Zipf, que controla la rapidez con que decrecen las frecuencias conforme aumenta el rango. Valores cercanos a $s \approx 1$ son típicos en lenguajes naturales.
- Interpreta el valor del exponente s: si $s > 1$, la frecuencia cae más rápido de lo esperado; si $s < 1$, las palabras raras aparecen relativamente más seguido.
- Discute posibles desviaciones: por ejemplo, la presencia de *stopwords* muy frecuentes, el tamaño limitado del corpus, o palabras raras (*hapax legomena*) que afectan la cola de la distribución.

En la figura 3.1, se presenta la relación lineal y empírica entre log-rango y log-frecuencia que predice la ley de Zipf. También se añade el ajuste de una regresión lineal para esta misma relación, obteniendo una puntuación $R^2 = 0.971$.

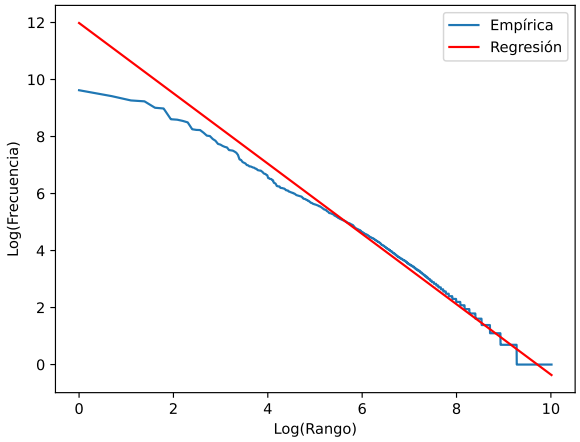


Figura 3.1: Ley de Zipf en el corpus de reseñas. Gráfica empírica (color azul) y gráfica de la regresión lineal ajustada (color rojo).

Siguiendo la Ley de Zipf (eq. 3.1), el ajuste de la regresión lineal resulta en un valor para $\log(C) \approx 11.98$ y $-s \approx -1.23$, por tanto:

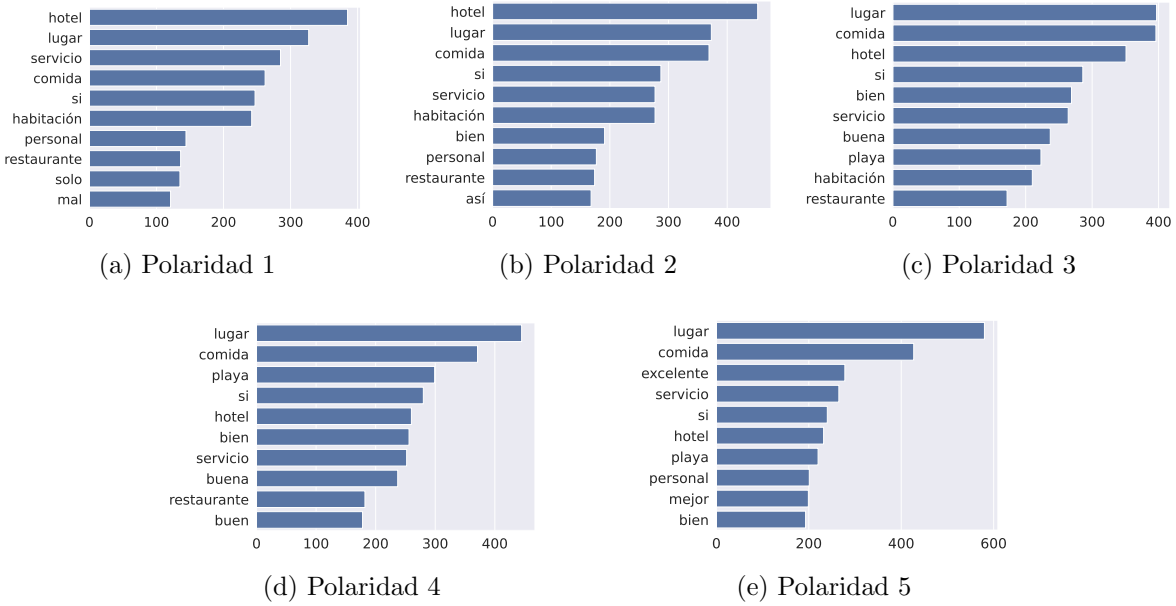
- $C \approx 159,988$. Mientras que $f(1) = 15,131$, haciendo que ambas variables difieran en un orden de magnitud, este efecto puede apreciarse en la gráfica 3.1.
- $s = 1.233$. Esto quiere decir quel a frecuencia decae más rápido de lo esperado, lo cual es esperable debido al tamaño reducido del corpus. Recordemos que obtuvimos que un $52\,\%$ de las palabras del vocabulario eran Hapax.

Palabras importantes por clase

3

Elimina palabras vacias y normaliza el texto, despues:

- Identifica las palabras mas frecuentes en cada clase.
- Reflexiona si las palabras mas repetidas son realmente discriminativas.



(a) Polaridad 1 (b) Polaridad 2 (c) Polaridad 3

(d) Polaridad 4 (e) Polaridad 5

Figura 4.1: Palabras más comunes por cada clase. (a-d) Muestra la gráfica de barras con las palabras y su frecuencia, según la polaridad de las reseñas.

El conjunto de figuras 4.1 muestra las palabras más frecuentes en cada tipo de review (según su polaridad). Podemos observar que, aunque las palabras nos pueden dar una pista del tema que se trata en las opiniones, no dan información crítica para la discriminación de las clases.

Por ejemplo, las palabras “lugar”, “comida”, “hotel”, “servicio”, “restaurante”. Estas palabras son comunes en casi todas las clases. Su alta frecuencia indica que son centrales al contenido, pero dado que aparecen tanto en clases positivas como negativas, no pueden usarse para determinar el sentimiento de una nueva reseña.

Sin embargo, también podemos encontrar algunas palabras discriminativas. Estas son aquellas que tienen un claro sentimiento positivo o negativo. Por ejemplo, “mal” es muy específica de la polaridad 1. Por el contrario, palabras como “buena”, “buen”, “excelente” y “mejor” están asociadas a clases de mayor polaridad (3, 4 y 5) y son fuertes indicadores de una reseña positiva.

Patrones gramaticales (POS 4-gramas)

4

Etiqueta con POS cada documento, y luego:

- Extrae las secuencias gramaticales más frecuentes de longitud 4 en cada clase.
- Discute si estas estructuras difieren entre clases y explica por qué.

En este ejercicio determiné las 5 secuencias gramáticales más comúnes por cada clase, estas son casi idénticas en todas las clases, con variaciones menores en el orden y las frecuencias exactas.

Las secuencias más comunes son:

- `NOUN ADP DET NOUN`: Presente en las 5 clases y siempre en primer lugar.
- `ADP DET NOUN PUNCT`: Presente en las 5 clases y siempre en segundo lugar.
- `DET NOUN ADP NOUN`: Presente en las 5 clases y siempre en tercer lugar.
- `ADP DET NOUN ADP` y `DET NOUN ADP DET`: Varían entre el cuarto y quinto lugar.
- `VERB ADP DET NOUN`: La única que parece importante, aparece únicamente en las reseñas con polaridad 1 en el top 5. Sin embargo, no es discriminante porque sigue siendo una estructura gramatical muy común.

La similitud de las estructuras gramaticales se debe a que las reseñas, independientemente de su polaridad, tienden a seguir patrones de lenguaje similares. Las personas usan una gramática consistente para describir sus experiencias. La principal diferencia no está en la estructura de la oración.

Por ejemplo, el patrón `NOUN ADP DET NOUN` podría ser çomida del hotel". Esta estructura se mantiene ya sea que la reseña sea positiva ("la comida del hotel fue excelente") o negativa ("la comida del hotel fue horrible"). La información clave para diferenciar entre una reseña positiva y una negativa no reside en la sintaxis, sino en la semántica, es decir, en las palabras con carga emocional que expresan el sentimiento del autor.

Representaciones BoW

5

Construye representaciones BoW con TF y con TF-IDF del corpus dado.

- Aplica alguna medida estadística (chi-cuadrado, información mutua o information gain), y obtén el top 20 de características más importantes en cada representación.
- Analiza diferencias entre ambas representaciones.

En este ejercicio utilizamos la librería *Scikit-learn* para crear las representaciones de BoW y TF-IDF de las reviews, cada uno con las funciones `CountVectorizer` y `TfidfVectorizer` con argumentos predeterminados, respectivamente. En la misma librería utilizamos el modulo `feature_selection` para obtener el top 20 características más importantes.

El análisis de las características más importantes para la clasificación de polaridad se realizó utilizando el método de Chi-Cuadrado (χ^2), examinando las dos representaciones de datos propuestas: TF y TF-iDF .

Las 20 características más relevantes por *TF* son “cuando”, “dijeron”, “excelente”, “habitación”, “había”, “hotel”, “increíble”, “la”, “mal”, “mala”, “me”, “nada”, “ni”, “no”, “nos”, “peor”, “pero”, “por”, “pésimo” y “que”.

Las características por *TF-IDF* son “buen”, “deliciosa”, “dijeron”, “excelente”, “gracias”, “habitación”, “horrible”, “increíble”, “mal”, “mala”, “nada”, “ni”, “no”, “peor”, “pero”, “pésima”, “pésimo”, “que”, “regular” y “terrible”.

Es notable que ambas representaciones terminan con stopwords dentro de sus características más importantes; para la representación por TF existen 11 stopwords dentro de sus 20 palabras más importantes, mientras que para la representación TF-iDF esta cantidad se reduce a 5 stopwords. Este resultado es a causa de la penalización del pesado TF-IDF que da a las palabras que aparecen en muchos documentos.

Para ambas representaciones también podemos encontrar palabras que expresan la polaridad de las opiniones (“excelente”, “mal”, “horrible”, “buena”...), lo que nos indica la importancia semántica de las palabras para la tarea de clasificación. Ambas representaciones también varían por palabras, aunque todas son en base a un adjetivo sobre la polaridad.

Bigramas

6

Repite el ejercicio anterior pero utilizando bigramas de palabras.

- Compara resultados y discute si los bigramas aportan mayor discriminación semántica.

Realizando un procedimiento similar al anterior, obtenemos las representaciones TF y TF-IDF tomando como cada término un bigrama, es decir, un término equivale a dos palabras concurrentes. En ambos casos se añade el argumento `ngram_range=(2,2)` al inicializar cada vectorizador.

Nuevamente, los bigramas más importantes se obtienen utilizando el método χ^2 para ambas representaciones.

Así, obtenemos para la representación TF los bigramas: “dijeron que”, “dijo que”, “este hotel”, “excelente servicio”, “la habitación”, “las ruinas”, “me dijeron”, “muy mal”, “muy mala”, “ni siquiera”, “no es”, “no había”, “no hay”, “no lo”, “nos dijeron”, “pero no”, “pésimo servicio”, “que no”, “sin embargo” y “un poco”.

Para el caso de TF-iDF los bigramas más importantes son: “comida deliciosa”, “comida no”, “dijeron que”, “dijo que”, “el peor”, “es buena”, “excelente servicio”, “la habitación”, “mal servicio”, “mala experiencia”, “muy mal”, “muy mala”, “ni siquiera”, “no es”, “no había”, “no lo”, “nos dijeron”, “pésimo servicio”, “sin embargo” y “un poco”.

La representación en bigramas nos permite explicar cómo se utilizan las palabras calificativas que encontramos utilizando solo unigramas. Por ejemplo, la palabra “deliciosa” que encontramos en el problema anterior reaparece en la lista de bigramas TF-IDF como “comida deliciosa”, mostrando que se refiere a la calidad de los alimentos en algún restaurante, sucede también con la stopword “ni” donde ahora obtenemos “ni siquiera”, sugiriendo una queja. También hallamos sentido a unigramas anteriores como “dijeron”: ¿Qué nos podría decir esta palabra por si sola sobre la polaridad de una opinión? Sin embargo, en los bigramas encontramos que esta palabra aparece como “dijeron que” o “dijo que”, lo cual podemos atribuir a una queja (responsabilizar a otra persona).

Esto quiere decir que podemos hallar sentido a el uso de las palabras vacías cuando vienen acompañadas de contexto. Esto también funciona para reforzar los adjetivos calificativos más considerados para la polarización de la opinión, por ejemplo podemos transformar “excelente” a “excelente servicio” que puede o no aplicarse a un segmento específico de negocios.

Word2Vec y analogías

7

Entrena un modelo Word2Vec sobre el corpus, después:

- Realiza al menos 5 analogías interesantes y discute los resultados.

La capacidad de Word2Vec para resolver analogías se basa en la aritmética de vectores, donde las relaciones entre palabras se capturan mediante la suma y resta de sus vectores. La fórmula general para una analogía de la forma .A es a B como C es a D"se puede expresar como:

$$vector(A) - vector(B) + vector(C) \approx vector(D)$$

En el contexto de la función `most_similar` de `gensim`, las palabras que se restan se colocan en la lista de 'negative' y las que se suman en la lista de 'positive'. A continuación, se discuten los resultados de las analogías proporcionadas.

Analogía 1: Habitación - Hotel + Restaurante

- **Positivo**: `['habitacion', 'restaurante']`
- **Negativo**: `['hotel']`
- **Resultado**: `('bebida', 0.929), ('ropa', 0.926), ('cocina', 0.920)`

La analogía busca encontrar el equivalente en un restaurante de una habitación en un hotel. Es decir, "una habitación en un hotel es a un restaurante como una habitación es a...". Las respuestas 'bebida' y 'cocina' son lógicas, ya que son componentes clave de un restaurante, al igual que la 'habitación' lo es de un hotel. La aparición de 'ropa' es menos obvia, pero puede estar ligada a la indumentaria del personal (camarero, chef) o a la ropa de cama que se encuentra en una habitación.

Analogía 2: Mesa - Restaurante + Hotel

- **Positivo**: `['mesa', 'hotel']`
- **Negativo**: `['restaurante']`
- **Resultado**: `('puerta', 0.902), ('llegada', 0.884), ('tercera', 0.879)`

En esta analogía, se plantea "una mesa en un restaurante es a un hotel como una mesa es a..". Se espera una entidad física o un concepto relacionado con la entrada o el registro en un hotel, como 'recepción' o 'cama'. El resultado 'puerta' es semánticamente coherente, ya que tanto una 'mesa' como una 'puerta' son elementos físicos y funcionales dentro de sus respectivos establecimientos.

Analogía 3: Deliciosa + Platillo

- **Positivo**: `['deliciosa', 'platillo']`
- **Negativo**: `None`

- **Resultado**: `('sencilla', 0.979), ('tradicional', 0.978), ('magnífica', 0.976), ('exquisita', 0.975)`

Cuando se suman los vectores de 'deliciosa' y 'platillo', el modelo busca palabras que sean semánticamente cercanas a la combinación de ambos conceptos. Los resultados 'sencilla', 'tradicional', 'magnífica' y 'exquisita' son adjetivos que comúnmente se usan para describir un 'platillo'. La palabra 'exquisita' es un sinónimo directo de 'deliciosa', lo que muestra una fuerte relación de similitud.

Analogía 4: Comida - Atención + Servicio

- **Positivo**: `['servicio', 'comida']`
- **Negativo**: `['atención']`
- **Resultado**: `('desayuno', 0.814), ('sabor', 0.767), ('menú', 0.762), ('sazón', 0.746)`

La analogía se puede interpretar como: "la comida es a el servicio como ... es a la atención". O más precisamente, "la comida es a la atención lo que el servicio es a...". La fórmula es $\vec{\text{servicio}} - \vec{\text{atención}} + \vec{\text{comida}}$. En este caso, la *atención* es un aspecto o una característica del *servicio*. La analogía busca una característica de la *comida* que sea equivalente a cómo la atención se relaciona con el servicio. Los resultados 'sabor' y 'sazón' encajan perfectamente, ya que son atributos que definen la calidad de la comida, al igual que la atención define la calidad del servicio. Los resultados 'desayuno' y 'menú' son también coherentes, ya que son tipos de 'comida' o elementos relacionados con ella. El modelo ha logrado capturar una relación de característica-a-entidad.

Analogía 5: Comida - Platillo + Servicio

- **Positivo**: `['comida', 'servicio']`
- **Negativo**: `['platillo']`
- **Resultado**: `('atención', 0.730), ('calidad', 0.697), ('presentación', 0.678)`

Esta analogía ("la comida es a un platillo como el servicio es a...") busca un concepto que sea parte o un aspecto del 'servicio', de manera análoga a cómo un 'platillo' es parte de la 'comida'. La respuesta principal, 'atención', es muy precisa, ya que la calidad del servicio se mide a menudo por la 'atención' que se recibe. Las palabras 'calidad' y 'presentación' también son aspectos que definen tanto a la 'comida' como al 'servicio' en general.

Embeddings de documento y clusterización

8

Calcula embeddings de documentos como el promedio de Word2Vec. A continuación:

- Aplica K-means con $k = 5$.
- Reporta los 5 textos más cercanos al centroide de cada clúster.
- Discute si los clústeres se alinean con las etiquetas originales.

Analisis de centroides

A continuación, se presentan los documentos más cercanos al centroide de cada uno de los 5 clústeres, junto con su polaridad original.

Centroide 1

- **Polaridad**: 2.0
- **Documento**: "Este B&B nos lo habían recomendado por algunos buenos amigos que viajan mucho, así que esperábamos una estancia muy agradable. Sin embargo, el hotel no cumplió con nuestras expectativas. Los jardines y la zona principal son bastante encantador y relajante, pero la habitación deja mucho que desear. La cama era terriblemente incómoda, viejo y pequeño, y no había ningún otro tipo de muebles en la habitación para poner nuestras pertenencias. El cuarto de baño estaba bastante sucia y un estado general de deterioro. Es una relación calidad-precio decente pero si buscas comodidad recomiendo que se alojen en otro lugar."

El documento más cercano a este centroide tiene una polaridad de 2.0, lo que indica una reseña negativa. El texto describe una experiencia insatisfactoria, señalando aspectos negativos como la incomodidad de la cama y la falta de limpieza.

Centroide 2

- **Polaridad**: 5.0
- **Documento**: "I viajado recientemente a Patzcuaro y me alojé en este B&B. increíble Yo era una cálida bienvenida y genuinamente mimado. Hay un grande y acogedora sala de desayuno en donde los huéspedes y la gerencia reunido cada mañana por la chimenea gigante de disfrutar realmente un gran desayuno, café y conversación. El Internet era gratis y siempre accesible. La gerencia, el propietario Victoria, Cynthia, Lon) ofrecía muchos extras no disponible en excursiones decoradas, incluyendo a otros B&B mexicano, con ropa de cama y toallas lujoso, enormes camas aparentemente grande, es acogedor calentadores de cama, las necesidades de champú, etc. , traducción, servicio de teléfono a los Estados Unidos a precios increíble, los servicios de lavandería, café disponible las 24 horas en un jardín"

El documento más cercano a este centroide tiene una polaridad de 5.0, una reseña altamente

positiva. El texto está lleno de adjetivos positivos como “increíble”, “cálida bienvenida” y “genuinamente mimado”, lo que refleja una experiencia excepcional.

Centroide 3

- **Polaridad**: 4.0
- **Documento**: “Fuimos en Semana Santa a comer al restaurante, el control sanitario es MUY BUENO (felicidades por eso) El servicio es muy bueno también, la comida no fue espectacular, bastante regular el sabor; tienen buenas bebidas de entrada y buenos postres El lugar es hermoso y...Más”

La polaridad de 4.0 indica una reseña mayoritariamente positiva. El documento destaca aspectos positivos como el “control sanitario” y el “buen servicio”, aunque la comida se describe como “regular”.

Centroide 4

- **Polaridad**: 4.0
- **Documento**: “Fuimos a pasar un fin de semana tranquilo a este lugar, es muy apacible y bonito hotel, limpio y las personas te atienden con amabilidad, tal vez por el costo podrian odrecer algunas cosas extras. Tiene buena ubicaciión esta muy cerca el centro, para ir caminando y la entrada de la carretera es accesible. Cabe destacar que fuimos en temporada baja y habia poca ocupación, sin embargo los ruidos de la alberca o del pasillo con facilidad de oyen hasta la habitación, pero sin ser molestos.”

La polaridad es 4.0. El documento describe el hotel como “apacible y bonito”, “limpio” y con “amabilidad” en el servicio. A pesar de mencionar algunos inconvenientes menores (como el ruido), el tono general es positivo.

Centroide 5

- **Polaridad**: 1.0
- **Documento**: “El propietario no ha registrado en este lugar en las últimas décadas. Hazte un favor y no te alojes aquí! (primero, permítanme decir que soy un hablante nativo española y otros críticos parecen pensar que la barrera del idioma es la raíz del problema. Confía en mí, no.) nos dijeron algunos locos precio a la llegada y cuando nos dijeron que la señora que un miembro de la familia había estado allí unos meses atrás y pagado mucho menos, dijo que era porque estábamos pagando con tarjeta de crédito. Nos ofrecieron efectivo y ella bajaron el precio $20 USD hasta los $65,00 USD. muy mediocre. Las habitaciones tienen una vista increíble, pero eso es todo. El lavabo (ver foto) en el cuarto de baño estaba hecho de plástico transparente con vida al mar Muerto dentro. ”

La polaridad de 1.0 refleja una reseña extremadamente negativa. El texto usa un lenguaje fuerte como “no te alojes aquí” y “locos precio”, además de describir detalles desagradables, lo que lo coloca firmemente en el extremo negativo del espectro de polaridad.

Conclusión

Los resultados demuestran que la clusterización de documentos basada en embeddings de Word2Vec promediados tiende a agrupar textos por su polaridad. Los clústeres 0 y 4 agrupan reseñas con polaridad baja (negativas), mientras que los clústeres 1, 2 y 3 agrupan reseñas con polaridades altas (positivas). Esto sugiere que el modelo de Word2Vec, al capturar las relaciones semánticas de las palabras, también permite que el promedio de sus vectores capture el tono o sentimiento general de un documento. La proximidad de los documentos a sus centroides indica una coherencia interna en los clústeres, lo que valida la efectividad de los embeddings de documento para tareas de agrupamiento no supervisado, incluso cuando la polaridad no se utiliza explícitamente durante el entrenamiento.

Clasificación con partición 70/30

9

Realiza cuatro experimentos acumulativos con un clasificador (SVM o regresión logística):

a) Sin preprocesamiento.

b) Con minúsculas.

c) Con minúsculas y stemming/lematización.

d) Con minúsculas, stemming y filtrando palabras con frecuencia mínima de 10.

Compara métricas (accuracy, F1 macro, matriz de confusión) y discute si el preprocesamiento es importante.

En este problema optamos por utilizar un SVM lineal como el clasificador para las pruebas. Además, se consideró el ajuste del hiper-parámetro de regularización C, utilizando la optimización por malla de Scikit-Learn `GridSearchCV` la cual también realiza *K-fold Cross Validation* con los datos de entrenamiento, obteniendo métricas más fiables para cada experimento. La optimización se realizó alrededor de la métrica F1 macro.

Métrica	**a)**	**b)**	**c)**	**d)**
Accuracy	0.396	0.380	0.417	**0.418**
F1 Macro	0.389	0.373	**0.408**	0.407

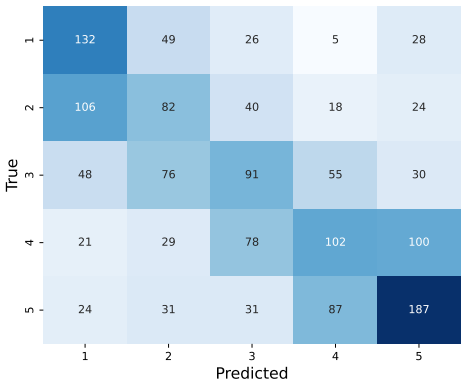
Cuadro 10.1: Comparación de Accuracy y F1 Macro en los distintos experimentos de preprocesamiento.

En la Tabla 10.1 se presentan las métricas de desempeño para cada uno de los cuatro experimentos. Se puede observar que la aplicación de sucesivos pasos de preprocesamiento no siempre resulta en una mejora monótona del rendimiento. Sorprendentemente, el segundo experimento (b), que introduce la conversión a minúsculas y la eliminación de *stopwords* (SS), muestra una ligera degradación en las métricas de *Accuracy* y F1 Macro en comparación con el modelo base sin preprocesamiento (a).

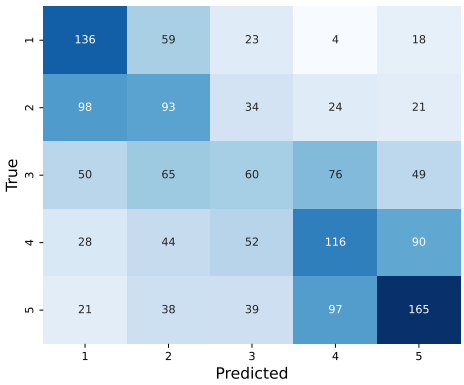
La mejora más sustancial se produce en el experimento (c) al incorporar el *stemming*, que logra el valor más alto de F1 Macro (0.408). Finalmente, el experimento (d), que añade un filtrado de palabras por frecuencia mínima (MF), alcanza el *Accuracy* más alto (0.418), aunque con un F1 Macro marginalmente inferior al de (c). Esto sugiere que los últimos dos pasos de preprocesamiento (stemming y filtrado por frecuencia) son los más determinantes para mejorar la capacidad de generalización del modelo.

Las matrices de confusión en la Figura 10.1 confirman visualmente estos resultados. En las matrices (a) y (b), se aprecia una considerable dispersión de las predicciones fuera de la diagonal principal, lo que indica un alto grado de confusión entre las clases. En cambio, en las matrices (c) y (d), la diagonal principal está mucho más marcada, lo que refleja un mayor número de aciertos. Específicamente, se observa que en todos los casos el clasificador tiende a confundir clases adyacentes, como la 4 con la 5. El último paso de preprocesamiento en

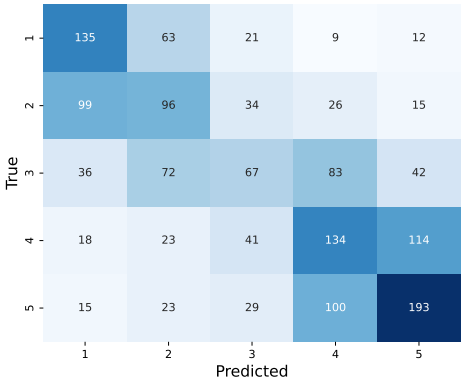
(d) parece beneficiar especialmente la clasificación de la clase 5, que alcanza 205 aciertos, el valor más alto de todos los experimentos para una sola clase. En conjunto, los resultados demuestran que el preprocesamiento es una etapa fundamental que impacta directamente en el rendimiento del clasificador.



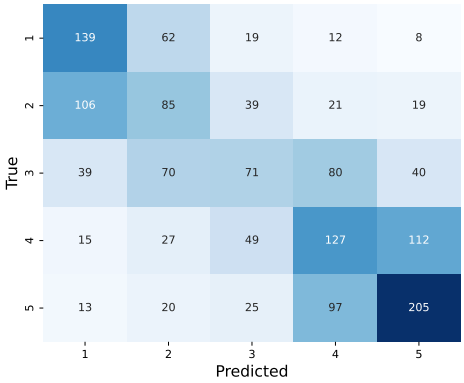
(a) Sin preprocesamiento



(b) Minúsculas, SS



(c) Minúsculas, SS, y stemming



(d) Minúsculas, SS, stemming y MF

Figura 10.1: Matrices de confusión para cada experimento realizado. SS refiere a la remoción de palabras vacías y MF al filtrado de palabras con frecuencia mínima de 10.

LSA con 50 tópicos

10

Aplica Latent Semantic Analysis (SVD truncado) con 50 tópicos.

- Muestra los términos más relevantes por tópico.
- Identifica qué tópicos son más informativos según una métrica estadística y analiza su coherencia.

Para resolver este problema, se aplicó LSA utilizando SVD truncado sobre una matriz de co-ocurrencia de palabras (con una ventana de 2 palabras), extrayendo un total de 50 tópicos. Los datos sobre los cuales se realizó la tarea fueron normalizados (convirtiendo a minúsculas, quitando palabras vacías). A continuación se presentan y analizan los resultados clave.

Términos Relevantes por Tópico

Cada uno de los 50 tópicos extraídos representa un concepto semántico, definido por un conjunto de palabras clave. A continuación, se muestran como ejemplo los términos más representativos para algunos de los tópicos más claros e interpretables:

- **Tópico 1 (Evaluación General):** comida, servicio, lugar, buena, hotel, bien, excelente, si, restaurante, personal.
- **Tópico 4 (Relación Valor-Experiencia):** pena, vale, lugar, merece, precio, comida, visitar, buena, visita, viaje.
- **Tópico 9 (Ambiente y Amabilidad):** lugar, agradable, bonito, personal, amable, hermoso, ambiente, servicial, increíble, atento.
- **Tópico 12 (Turismo en la Región):** playa, ruinas, tulum, amable, personal, norte, carmen, hermosa, zona.

Análisis Estadístico de Tópicos

Para identificar los tópicos más importantes, se utilizaron dos métricas complementarias: la varianza explicada, que mide la relevancia estadística, y la coherencia, que mide la interpretabilidad semántica.

Tópicos más Informativos

La **varianza explicada** indica la proporción de la información total del corpus que cada tópico es capaz de capturar. Un valor alto significa que el tópico es estadísticamente significativo. En la Tabla 11.1 se muestran los 5 tópicos más informativos. Destaca el Tópico 1, que por sí solo explica más del 32 % de la varianza, representando el tema más dominante en el conjunto de datos. Un resultado esperado al utilizar SVD truncado.

Tópico	Varianza Explicada
Tópico 1	0.3245
Tópico 2	0.0542
Tópico 3	0.0462
Tópico 4	0.0281
Tópico 5	0.0243

Cuadro 11.1: Tópicos con mayor varianza explicada.

Análisis de Coherencia

La **coherencia** mide qué tan relacionadas semánticamente están las palabras principales de un tópico. Un puntaje de coherencia alto generalmente indica que un tópico es más fácil de interpretar para un humano. La Tabla 11.2 presenta los 5 tópicos con mayor coherencia.

Se observa una fuerte correlación entre los tópicos más informativos y los más coherentes. Los tópicos 1, 2, 3 y 5 aparecen en ambas listas, lo que sugiere que los temas principales identificados por el modelo no solo son estadísticamente relevantes, sino también semánticamente consistentes y fáciles de interpretar.

Tópico	Puntaje de Coherencia
Tópico 2	0.5902
Tópico 1	0.5675
Tópico 3	0.5665
Tópico 5	0.5543
Tópico 8	0.5187

Cuadro 11.2: Tópicos con mayor coherencia semántica.

El análisis de los tópicos con mayor coherencia revela los temas más consistentes y semánticamente definidos dentro del corpus. El Tópico 2, el más coherente, se enfoca claramente en la calidad de la comida y el servicio, con términos como “calidad”, “deliciosa” y “atención”. De manera similar, el Tópico 1 describe la experiencia general en hotelería y restaurantes, abarcando “comida”, “servicio” y “personal”. El Tópico 3 también trata sobre la experiencia de servicio, pero introduce un contraste de sentimientos con palabras como “pésimo” junto a “bueno”, sugiriendo que agrupa opiniones fuertes. Por su parte, el Tópico 5 se centra en la evaluación subjetiva y el valor de la experiencia, dominado por adjetivos calificativos como “excelente”, “agradable” y la idea de que “vale la pena”. Finalmente, el Tópico 8 describe la atmósfera y el ambiente del lugar, combinando aspectos estéticos (“bonito”, “hermoso”, “vista”) con la calidad del servicio. En conjunto, estos tópicos demuestran la capacidad del modelo para agrupar las discusiones en temas bien definidos sobre la calidad, el valor y el ambiente de los servicios turísticos.