

# Análisis de Texto e Imágenes con Deep Learning | Tarea #1

Análisis Integral de Corpus en Español

---

*César M. Aguirre Calzadilla*



*Centro de Investigación en Matemáticas*

Maestría en Cómputo Estadístico

*Catedráticos:*

Dr. Miguel Álvarez Carmona

Dr. Francisco Javier Hernández López

7 de septiembre de 2025

# Tabla de contenidos

	Página
<b>Ejercicio #1   Descripción del corpus</b> . . . . .	<b>3</b>
0.1 Descripción del corpus . . . . .	3
0.2 Número de documentos, tokens y vocabulario . . . . .	4
0.3 Hapax legomena y su proporción . . . . .	6
0.4 Porcentaje de <i>stopwords</i> . . . . .	7
<b>Ejercicio #2   Ley de Zipf</b> . . . . .	<b>9</b>
0.5 Distribución de términos y Ley de Zipf . . . . .	9
0.6 Representación Zipf de nuestro corpus . . . . .	10
Discusión del modelo Zipfiano . . . . .	12
0.7 Conclusiones del modelo Zipfiano . . . . .	13
<b>Ejercicio #3   Palabras importantes por clase</b> . . . . .	<b>14</b>
0.8 Términos discriminativos . . . . .	14
0.9 Análisis para Type . . . . .	16
<b>Ejercicio #4   Patrones gramaticales (POS 4-gramas)</b> . . . . .	<b>23</b>
0.10 Part of Speech (POS) . . . . .	23
0.11 Análisis de Estructuras Sintácticas mediante N-gramas de POS . . . . .	24
0.12 Resultados . . . . .	26
<b>Ejercicio #5  </b> . . . . .	<b>31</b>
0.13 Representaciones BoW: TF y TF-IDF . . . . .	31
<b>Ejercicio #6   Bigramas</b> . . . . .	<b>46</b>
0.13.1 Resultados . . . . .	49
0.13.2 Aportación semántica de los bigramas . . . . .	57
<b>Ejercicio #7   Word2Vec y analogías</b> . . . . .	<b>58</b>
0.13.3 Distancia consenso . . . . .	58
0.13.4 Analogías Semánticas . . . . .	59
0.13.5 Resultados . . . . .	59
<b>Ejercicio #8   Embeddings de documento y clusterización</b> . . . . .	<b>62</b>
0.13.6 Word2Vec Promediado . . . . .	62
0.13.7 Resultados . . . . .	63

0.14 Análisis . . . . .	68
0.15 Análisis . . . . .	72
<b>Ejercicio #9   Clasificando con Partición 70/3 . . . . .</b>	<b>73</b>
0.15.1 Resultados . . . . .	73
<b>Ejercicio #10   LSA con 50 tópicos . . . . .</b>	<b>76</b>
0.16 Latent Semantic Analysis (LSA) . . . . .	76
0.17 Resultados . . . . .	77

## Ejercicio #1 | Descripción del corpus

Analiza el corpus y reporta:

- Número de documentos, tokens y vocabulario.
- Hapax legomena y su proporción.
- Porcentaje y su proporción.
- Estadísticas por clase (número de documentos, tokens y vocabulario).

### 0.1. Descripción del corpus

Antes de comenzar, hablemos un poco acerca de nuestro corpus: MeIA.csv. Este archivo contiene un conjunto de datos en español con reseñas de establecimientos turísticos y gastronómicos de lugares turísticos en México. Existen cinco columnas dentro del dataset:

- **Review:** es donde viene contenido el texto de la reseña dejada por el usuario.
- **Polarity:** una puntuación numérica (de 1.0 a 5.0) que indica la valoración dejada por el usuario.
- **Town:** ciudad o localidad donde se encuentra el establecimiento del que se habla en la reseña.
- **Region:** estado o región de México donde se encuentra el establecimiento.
- **Type:** tipo de establecimiento (restaurante, hotel o atractivo).

Cuenta con un total de 5,000 reseñas, todas en español aunque algunas de ellas tienen mezcla de acentos o formas de escribir. Entre los estados que se mencionan están Quintana Roo, Jalisco, Puebla, Chiapas, Estado de México o Yucatán. Entre las reseñas se pueden leer comentarios de la calidad de la comida, servicio, ambiente, limpieza, instalaciones, ubicación; o lugares como cenotes, ruinas, museos, playas o miradores.

En general, creo que se trata de un corpus turístico bastante bueno que además nos permitiría trabajar en tareas como el análisis de sentimiento basado en la polaridad, el estudio de tendencias turísticas por región o el tipo de establecimiento, así como la identificación de aspectos clave que permitan conocer mejor las tendencias y necesidades de los clientes para ofrecer un mejor servicio.

## 0.2. Número de documentos, tokens y vocabulario

El cuadro 1 nos muestra las métricas básicas de nuestro corpus, compuesto de un total de 5,000 filas, las cuales son por completo “no nulas”, i.e. todas cuentan con descripciones y ninguna está vacía. Esto se revisó con una métrica sencilla, en la cual solicitamos que se revisara que todos los documentos (o *reviews*) contuvieran al menos un token.

Cuadro 1: Métricas básicas del corpus de reseñas.

Métrica	Valor
Filas totales (CSV)	5,000
Filas con Review no nula	5,000
Documentos válidos ( $\geq 1$ token)	5,000
Número total de tokens (N)	348,838
Tamaño del vocabulario ( $ V $ )	20,486

Podemos encontrarnos en el cuadro 1 que se generaron un total de 348,838 tokens con un tamaño de vocabulario de 20,486 palabras. El tokenizador se personalizó por defecto desde spaCy, mediante las siguientes configuraciones:

- **Preservación de entidades sociales:** definimos algunas expresiones regulares para que, en caso de que se hubieran usado, se tomaran en cuenta hashtags, arrobas o URLs, de tal manera que se toman como tokens únicos.
- **Emojis:** añadimos un diccionario de casos particulares de emojis para que no fuesen segmentados.

Lo anterior se realizó pensando en que las reseñas pueden contener lenguaje informal o propio de redes sociales como Twitter, TikTok o Instagram. Un tokenizador estándar como el que se tomó de base segmentaría *#Recomendado* en *#* y *Recomendado*, perdiendo el significado contextual del hashtag. Este enfoque intenta preservar la intención del usuario al utilizar ese lenguaje informal y común de redes sociales.

También se realizó una elongación para reducir la variabilidad morfológica causada por la repetición excesiva de caracteres. Es decir, implementamos un componente dentro del pipeline de spaCy para aplicar expresiones regulares que detectaran secuencias de tres o más caracteres idénticos consecutivos, como podría ser *maravillooooooso*, y así reducirlo solo a *maravilloso*. Sin este tratamiento, podríamos estar generando ruido al tomar como dos palabras diferentes alguna expresión elongada, cuando en realidad es la misma. Este

tratamiento contrasta con la capacidad actual de los LLM de contextualizar la información y tokenizar, no por palabras, sino por subpalabras.

Esto significa que no ven **buenísimooooo** como una palabra desconocida. En su lugar, la descomponen en partes más pequeñas y reconocibles. Por ejemplo:

**buenísimooooo** → [buen] [ísim] [o] [o] [o] [o] [o]

Pero bueno, eso es tema para otra ocasión. El pipeline de este problema toma:

- **Tokenizador social** como se describió anteriormente.
- **Normalizador de elongaciones** como también se describió anteriormnete.

Además, para cada token en cada documento procesado, se aplicaron las siguientes reglas:

- **Filtrado**: se descartaron espacios, signos de puntuación, comillas, brackets, y números puros (e.g. 10, 3.1415, 07).
- **Normalización**: cada token válido se convirtió en minúsculas.

De esa manera, se intenta asegurar que el vocabulario esté compuesto de unidades léxicas como sustantivos, adjetivos o verbos. La normalización a minúsculas y la reducción de elongaciones consolidad dos distintas representaciones superficiales de una misma palabra en una única forma.

Para el cálculo de vocabulario únicamente aprovechamos la estructura de datos tipo set (conjunto) para de esa maenra recolectar las formas canónicas de los tokens válidos, aprovechando que un conjunto no permite duplicados. El uso de un set garantiza que cada palabra única se cuente una sola vez, calculando eficientemente el tamaño del vocabulario. Estas métricas son fundamentales para entender la riqueza léxica del corpus y para tareas posteriores de modelado de lenguaje.

### 0.3. Hapax legomena y su proporción

Cuando hablamos de *Hapax legomena* nos referimos a palabras que aparecen exactamente solo una vez en el corpus. Es un término de lingüística que se usa para identificar las palabras más raras o únicas en un texto, en la obra completa de un autor, o incluso en todo un idioma documentado. Estos sirven bastante para perfilado de autor, por ejemplo. Los hapax legomena son indicadores importantes de dos cosas en particular:

- La riqueza léxica del corpus.
- La presencia de términos especializados o poco comunes.

La estructura de datos utilizada fue el objeto Counter de la biblioteca collections de Python para mantener un registro de la frecuencia de cada palabra en el corpus. Durante el procesamiento de cada documento (reseña), se actualizaba el contador con cada token válido en su forma canónica. Además, el contador almacenó pares de palabra-frecuencia para todo el corpus. Las métricas calculadas se presentan en el cuadro 2.

Cuadro 2: Estadísticas de Hapax Legomena.

Métrica	Valor
Número de hapax (#hapax)	10,783
Tamaño del vocabulario ( $ V $ )	20,486
Número total de tokens ( $N$ )	348,838
Proporción tipo (#hapax/ $ V $ )	0.526,4
Proporción token (#hapax/ $N$ )	0.030,9

Entre las métricas podemos encontrarnos con 10,783 hapax legomena, valor que nos sirve para calcular la Proporción tipo y la *Proporción token*. La primera de las proporciones indica qué tan rico es el vocabulario de un texto, una alta proporción tipo nos habla de uno muy diverso mientras que una baja nos diría que la mayoría de las palabras aparecen múltiples veces. Asimismo, la proporción token nos indica la distribución general de las frecuencias.

En nuestro caso, tenemos una proporción tipo de 0.5264, lo cual nos dice que más de la mitad de todas las palabras únicas en el corpus aparecen exactamente una sola vez. Este valor me preocupó un poco al inicio, pero se supone que en el caso particular de las reseñas (sea de productos, visitas, restaurantes) es común detectar un vocabulario muy diverso. Lo anterior sucede porque los usuarios emplean descripciones variadas y creativas, así como la presencia de términos muy particulares dejados por cada persona.

Ahora bien, la proporción token de 0.0309 revela que, pese a la alta diversidad léxica, los hapax legomena representan solo una pequeña fracción de las apariciones totales de palabras. Es decir, un pequeño número de palabras muy frecuentes domina la mayoría de las apariciones, existe una gran cantidad de palabras de baja frecuencia y la distribución sigue el patrón característico de  $f \propto 1/r$  observado en lenguaje natural (algo que veremos más adelante como ley de Zipf).

#### 0.4. Porcentaje de *stopwords*

Para este caso, utilizamos la lista base de *stopwords* de spaCy en español como punto de partida. Se extrajeron todas las palabras de parada de cada módulo con `nlp.Defaults.stop_words`. Así, todas las palabras se convirtieron a minúsculas para garantizar la consistencia.

Implementamos la función `is_stopword_form()` para determinar si una forma canónica es *stopword*. Aplicamos normalización de tildes a la palabra canónica, se busca en el conjunto personalizado de *stopwords* y se retorna un booleano que indica si la palabra es o no *stopword*. Los resultados pueden verse en el cuadro 3.

Cuadro 3: Resumen de estadísticas del corpus: Stopwords.

Métrica	Valor
<b>Estadísticas de Stopwords</b>	
Número total de tokens ( $N$ )	348,838
Tokens que son stopwords	198,837
Porcentaje de stopwords (%)	57.00

En cuanto al porcentaje de *stopwords*, tenemos que el 57 % de nuestro texto en el corpus corresponde a palabras funcionales i.e. *stopwords*. Este porcentaje es bastante alto, pero creo que tiene bastante sentido si pensamos que nuestro corpus se compone de reseñas, las cuales no suelen ser tan largas. Es decir, son documentos cortos para el estandar de un documento de texto.

Conocer el contenido de palabras funcionales en nuestro corpus nos puede permitir realizar un mejor preprocesamiento de los datos, pues podríamos eliminarlas para reducir el tamaño del dataset en caso de que eso no afecte al análisis que queremos realizar. Computacionalmente hablando, esto podría ser bastante ahorrativo. Asimismo, la reducción de ruido permite un mejor análisis de frecuencias para palabras que de verdad aportan al texto.

Sin embargo, la eliminación indiscriminada de *stopwords* podría afectar la interpretación del texto. En otras palabras, podríamos perder contexto. Pensemos en negaciones

como “no”, “nunca”, o “tampoco”; así como intensificadores como “muy”, “bastante” o “chingos”; o en conectores lógicos como “pero”, “aunque” o “sin embargo”. Los ejemplos anteriores enriquecen el contexto del documento al indicarnos, valga la redundancia, contexto.

Por lo que, dependiendo de la tarea que se quiera hacer con un corpus, nos convendrá en mayor o menor medida, realizar ajustes al filtrado de *stopwords*. Así como no eliminar aquellas palabras funcionales que pueden enriquecer la semántica y contexto.

## Ejercicio #2 | Ley de Zipf

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

Ajusta una recta mediante regresión lineal sobre los puntos  $(\log r, \log f(w))$ , de la forma:

$$\log f(r) = \log C - s \cdot \log r$$

lo cual equivale al modelo Zipfiano  $f(r) \approx \frac{C}{r^s}$ .

En la anterior 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 de lenguajes naturales.

## 0.5. Distribución de términos y Ley de Zipf

En muchas ocasiones, dentro del procesamiento del lenguaje natural, nos interesa comprender cómo se distribuyen los términos dentro de un documento. Es decir, la frecuencia y posición de cada palabra no es aleatoria, su estructura revela el tono, significado o esencia de un documento. En ese sentido, la Ley de Zipf proporciona un modelo matemático para lograr ver cómo es la distribución de cierto texto, sirviendo como punto de partida para diversas tareas.

Analizar cómo se distribuyen las palabras nos deja una brecha para estudiar la “huella digital” de cada texto. Por ejemplo, las palabras que son inusualmente frecuentes en el

documento, al compararse con su uso en el lenguaje en general, suelen ser palabras clave que definen el tema del escrito. Asimismo, dos o más documentos pueden utilizar palabras similares, pero si su distribución y frecuencia es distinta, esto nos puede decir la naturaleza del mismo. De igual forma, para la detección de tópicos, es importante conocer la distribución de términos para poder identificar y clasificar el tópico al que se puede relacionar el trabajo.

En términos matemáticos, la Ley de Zipf establece que, si el primer término  $t_1$  ese 1 más común en el conjunto,  $t_2$  el segundo más común,  $t_3$  el tercero, y así sucesivamente, entonces la frecuencia del conjunto  $c f_i$  del  $i$ -ésimo término más común es proporcional a  $\frac{1}{i}$ :

$$c f_i \propto \frac{1}{i}$$

De manera más simple, esta ley describe que la frecuencia de una palabra es inversamente proporcional a su rango en una tabla de frecuencias. Es decir: una de las palabras más frecuentes dentro de un texto puede ser "el" y esta aparecerá aproximadamente el doble de veces que la segunda más frecuente, digamos "de"; y a su vez aparecerá el triple de veces que la tercera más frecuente "que".

Nuestra intuición nos indica algo bastante claro, la frecuencia disminuye de manera muy abrupta. De manera equivalente, podemos describir la Ley de Zipf como  $c f_i = c i^k$  o como  $\log c f_i = \log c + k \log i$ , donde  $k = -1$  y  $c$  es una constante que puede ser definida. Así, podemos hablar de una "ley de potencias" o *power law*, con exponente  $k = -1$ . Hay diversos fenómenos en la ciencia, economía y computación que siguen esta estructura, como la Distribución de Pareto, el espectro de un cuásar, la Ley de Kleiber o la de Stefan-Boltzmann.

## 0.6. Representación Zipf de nuestro corpus

Para el caso de nuestro corpus, "MeIA.csv", el perfil de Zipf se muestra como podemos ver en la figura 1.

La gráfica muestra claramente la relación entre el rango de una palabra y su frecuencia. Tenemos unos cuantos puntos muy cargados a la izquierda del rango (donde se encuentran las palabras más frecuentes), y una gran cantidad de palabras muy raras cargadas a la derecha del rango (palabras menos frecuentes dentro del texto). Podemos ver más información acerca de nuestra modelación Zipf en el cuadro 4.

El ajuste lineal para este modelo zipfiano es de  $y = 5.219 - 1.248x$ , el cual es el modelo matemático que describe la tendencia de nuestro corpus MeIA.csv. La pendiente, con una carga de  $-1.248$  sigue la tendencia de la Ley de Zipf, una pendiente negativa cercana

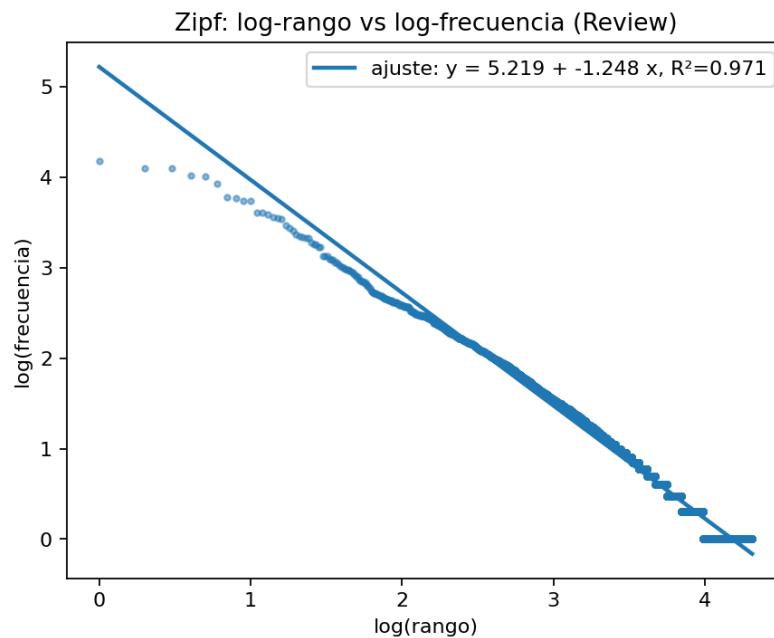


Figura 1: Perfil de Zipf dentro de nuestro corpus.

al  $-1$  que confirma de manera cuantitativa que la distribución de palabras en el corpus se apega sólidamente al modelo. Además, contamos con un coeficiente de determinación  $R^2 = 0.971$ , indicando que nuestros datos se ajustan bastante bien a la línea predicha por  $y = 5.219 - 1.248x$ . Es decir, el 97.1 % de la variabilidad en la frecuencia de las palabras puede ser explicada por su rango, i.e. indica que el modelo lineal es un excelente descriptor de los datos del corpus.

## Discusión del modelo Zipfiano

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 nuestro caso, tenemos un  $s \approx 1.25 > 1$  que supera el decaimiento más rápido del Zipf canónico  $s \approx 1$ . Es decir, la masa de probabilidad se concentra más en la cabeza del ranking y las palabras raras aparecen menos de lo esperado bajo un Zipf puro. Sin embargo, estos resultados son consistentes con lo que tenemos en el corpus, pues el 57 % de los tokens son *stopwords*, lo que infla las primeras posiciones y provoca una caída más abrupta en la pendiente. Además, hay que recordar que el peso de los *hapax legomena* en tokens es pequeño, apenas el 3.1 % (10,783 *hapax*/348,838 tokens), de modo que la cola aporta poca masa y no logra aplanar más la recta.

Cuadro 4: Parámetros del ajuste lineal para la Ley de Zipf.

Métrica	Valor
<b>Parámetros del ajuste: <math>y = a + bx</math></b>	
Ecuación de la recta	$y = 5.219 - 1.248x$
Ordenada en el origen ( $a$ )	5.219
Pendiente ( $b$ )	-1.248
Coeficiente de determinación ( $R^2$ )	0.971
<b>Parámetros derivados de Zipf</b>	
Exponente de la ley ( $s = -b$ )	1.248
Constante de normalización ( $C \approx 10^a$ )	$1.657 \times 10^5$
Frecuencia real del 1er rango ( $f(1)$ )	15,247
Relación $C/f(1)$	$\approx 9.87$

Si lo pensamos un poco, en las reseñas de todo tipo, suelen escribirse documentos cortos y, en muchas ocasiones, repetitivos. Debe haber muchas que contengan solo frases como "muy buen lugar", "excelente lugar", "grandioso lugar", etc. Esto refuerza que un pequeño conjunto de términos sea muy frecuente, provocando una cabeza pesada y haciendo que  $s \uparrow$ . De igual modo, nombres de lugares, adjetivos o clichés concentran frecuencia en pocas palabras.

## 0.7. Conclusiones del modelo Zipfiano

La presencia de la Ley de Zipf en el lenguaje natural revela una interesante estructura de economía en la comunicación. Nuestro corpus se ajusta de manera excelente al modelo Zipfiano, como lo demuestra el coeficiente de determinación  $R^2 = 0.971$ . Esto confirma que, pese a la diversidad de opiniones y estilos de escritura en las reseñas, la distribución de las palabras sigue siendo un patrón predecible y universal, dirigido por una ley de potencias.

Asimismo, el exponente  $s \approx 1.25$  con un valor mayor a 1, indica un decaimiento de frecuencia más rápido que el modelo tradicional zipfiano. De ese modo, nos encontramos con una alta concentración de frecuencia en un vocabulario muy reducido y común, en la cabeza de la distribución. A su vez, hay una menor diversidad de palabras raras de lo que se esperaría en un texto más general, pues la cola de la distribución es ligera.

Finalmente, podemos asegurar que la Ley de Zipf es una interesante herramienta analítica que permite obtener información útil de los corpus utilizados en el Procesamiento del Lenguaje Natural. Para MeIA.csv, pudimos visualizar que se trata de un corpus con lenguaje pragmático, repetitivo y dominado por un pequeño conjunto de términos.

## Ejercicio #3 | Palabras importantes por clase

- Elimina palabras vacías y normaliza el texto.
- Identifica las palabras más frecuentes en cada clase.
- Reflexiona si las palabras más repetidas son realmente discriminativas.

La solución para este problema sigue un pipeline que lleva a cabo los siguientes pasos:

- **Carga:** utilizamos nuestro corpus MeIA.csv.
- **Tokenizador:** la tokenización se realiza con un enfoque de redes sociales y lematización.
- **Filtrado:** se eliminan signos de puntuación, números puros y URLs. Se incorporan *stopwords*.
- **Conteo:** se realiza la cuenta de términos por clase (Polarity, Town, Type).
- **Rankings:** produce dos rankings por clase:
  - Top por frecuencia.
  - Top por "discriminatividad" usando log-odds con prior informativo (Monroe et al. 2008).
- **Guardado:** almacenamos en un CSV e imágenes PNG las gráficas y datos del análisis. Se generan gráficos de barras (frecuencia y log-odds), heatmap (z de log-odds) y dispersión (frecuencia vs. z)

### 0.8. Términos discriminativos

Cuando queremos encontrar términos o palabras que sean realmente característicos de cierta clase (como Hotel o Parque), la frecuencia cruda de palabras no es suficiente. Hay palabras que, pese a ser comunes, se repiten a lo largo de todas las clases, son *stopwords* o hablan de algo de manera general. Por ello, es necesario de un criterio que responga a la pregunta:

¿Este término aparece desproporcionadamente más en mi clase que en las demás, más allá del ruido por tamaños de muestra y conteos pequeños?

El método de log-odds con prior informativo (Monroe et al. 2008) se encarga justo de ese problema. Se trata de un método robusto para conteos pequeños, pues evita la “explosión” de términos muy raros que sólo aparecen un par de veces. Además, entrega un z-score que nos indica términos más distintivos cuando su valor es más alto, así como One-vs-rest: para cada clase  $k$ , se comparara contra el resto  $r$ . En términos prácticos, lo que busca es darnos un ranking donde los términos en las primeras posiciones son aquellos distintivos de la clase, no solo que sean los más frecuentes.

Podemos pensar en el método log-odds como un modelo Dirichlet-Multinomial por clase:

- Para cada clase  $k$ , las palabras se generan de una Multinomial con probabilidades  $\theta_k$ .
- Antes de ver los datos, pones un prior Dirichlet Informativo centrado en la frecuencia global  $p_w$ , i.e. proporción del término  $w$  en todo el corpus con masa total  $\alpha_0$ :  $\alpha_w = \alpha_0 p_w$

Para una clase  $k$  vs el resto  $r$ :

- Conteos observados:  $c_k(w)$  y  $c_r(w)$ . Totales:  $n_k, n_r$ .
- Proporciones suavizadas con prior para  $w$ :

$$\hat{p}_k(w) = \frac{c_k(w) + \alpha_w}{n_k + \alpha_0} \quad \& \quad \hat{p}_r(w) = \frac{c_r(w) + \alpha_w}{n_r + \alpha_0}$$

- log-odds (diferencia de logits suavizados):

$$\delta_w = \log \frac{c_k(w) + \alpha_w}{(n_k - c_k(w)) + (\alpha_0 - \alpha_w)} - \log \frac{c_r(w) + \alpha_w}{(n_r - c_r(w)) + (\alpha_0 - \alpha_w)}$$

- Varianza aproximada (key para estandarizar):

$$\text{Var}(\delta_w) \approx \frac{1}{c_k(w) + \alpha_w} + \frac{1}{c_r(w) + \alpha_w}$$

- Z-score:

$$z_w = \frac{\delta_w}{\sqrt{\text{Var}(\delta_w)}}$$

Así:

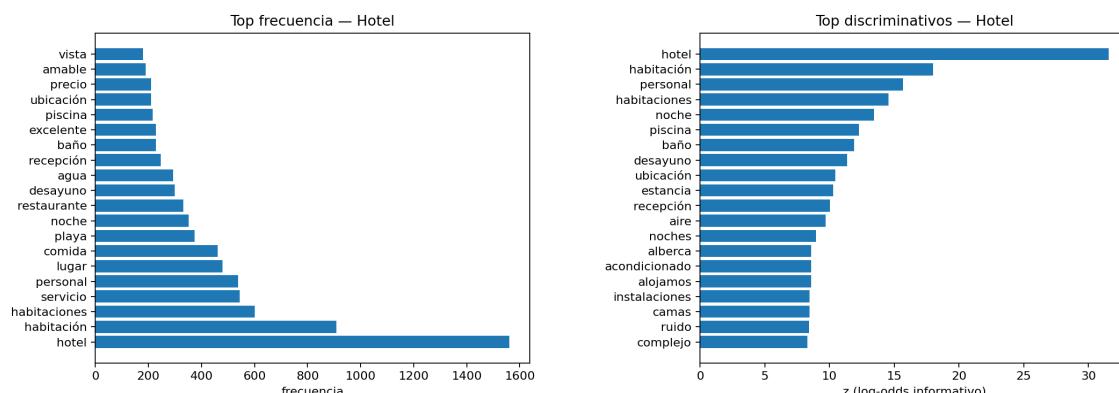
- $z_w \gg 0$ :  $w$  es mucho más probable en  $k$  que en el resto.
- $z_w \approx 0$ : parecido en todas partes.
- $z_w \ll 0$ :  $w$  es mucho más probable en el resto que en  $k$ .

El prior informativo ( $\alpha_w = \alpha_0 p_w$ ) actúa como “peso muerto”, estabilizando para términos raros. Si  $\alpha_0$  es grande, entonces hay mayor suavizado.

Intentaremos que, con el siguiente análisis, se presenten dos perspectivas complementarias para identificar los términos más relevantes en un corpus de reseñas. Para ello, utilizamos la frecuencia absoluta y el discriminativo, medido con el z-score de log-odds informativo. Mientras la frecuencia identifica términos comunes, el discriminativo aisla aquellas palabras que son más características o “de peso” en cada categoría (en nuestro caso: hotel, restaurante y atractivo).

## 0.9. Análisis para Type

### Hotel



(a) Top 20 de las palabras más frecuentes para la clase Hotel.

(b) Top 20 de las palabras más discriminativas para la clase Hotel.

En la figura 2a y en la 2b podemos encontrar el top 20 de las palabras más frecuentes y discriminativas, respectivamente, en la clase Hotel.

Un primer hallazgo importante es la convergencia en las dos primeras posiciones de los rankings, a los que podemos llamar como términos ancla. Estos son: “hotel” y “habitación”. Ellos no son solo los más frecuentes, sino también los más discriminativos. Este resultado es consistente con lo que se podría esperar de una reseña, pues son sustantivos

que actúan como puntero, i.e. definen el dominio semántico del documento. Su alta frecuencia es esperada ya que son palabras centrales en las reseñas de un hotel, y su alto peso discriminativo se debe a que su uso es desproporcionadamente más probable en la misma reseña de tipo Hotel, que en cualquiera de las otras dos categorías. A fin de cuentas, al escribir la reseña hablamos de lo vivido dentro del mismo Hotel, o habitación.

Ahora bien, a partir del tercer puesto podemos observar divergencia que muestra las diferencias entre lo común y lo característico de la clase. En el top de frecuencias encontramos términos como "servicio", "lugar", "comida", "excelente" y "amable", los cuales no figuran en el top de discriminativos. En esencia, estos términos constituyen un vocabulario genérico de la evaluación. Es decir, se trata de palabras que se pueden encontrar en las otras dos clases del corpus, lo que les resta especificidad. Aunque son muy frecuentes, su presencia no es señal significativa para definir que sean términos de peso para la clase Hotel. Pienso que podríamos decir que se trata de *stop words* si lo que queremos es identificar entre tipos de servicio.

Por el contrario, el ranking discriminativo destaca términos como "noche", "piscina", "desayuno", "recepción", "estancia", "aire" y "ruido". Estos términos sí capturan la esencia de la experiencia dentro de un hotel. No son necesariamente los más mencionados dentro de cada reseña, pero su presencia es de peso estadístico como para que se puedan identificar como términos ancla sobre una reseña de hotel. Se trata de palabras que describen atributos, servicios o eventos específicos de los alojamientos.

Una gráfica que también puede otorgarnos bastante información es la de dispersión, mostrada en la figura 3, que proporciona una visualización simultánea de la frecuencia logarítmica (eje X) y el discriminativo (eje Y), para cada término en la categoría Hotel. Este análisis permite no solo identificar los términos importantes, sino también comprender la naturaleza de su relevancia y estructura en la categoría.

Una primera conclusión a la que podemos llegar es que hay una fuerte correlación positiva entre la frecuencia de un término y su poder discriminativo. Es decir, la tendencia parece decírnos que cuanto más se utiliza una palabra dentro de la reseña (de hoteles), más probable es que sea una palabra característica de la categoría. Los conceptos centrales de un dominio se repiten con frecuencia, lo que a su vez, define dicho dominio.

Esta gráfica muestra una estratificación clara, de tres zonas. En el cuadrante superior derecho podemos encontrar los términos de mayor peso, separados del resto. Aparecen los dos términos ancla: "hotel" y "habitación". Una observación clara es la notable separación entre "hotel" y "habitación", reflejando el peso que tiene justo el término autoreferencial. Justo estos términos más cargados a la esquina superior derecha conforman los pilares semánticos del corpus. Se trata de términos de máxima frecuencia y máximo poder informativo.

Hablando de la zona central, a lo largo del eje diagonal principal, se sitúan términos

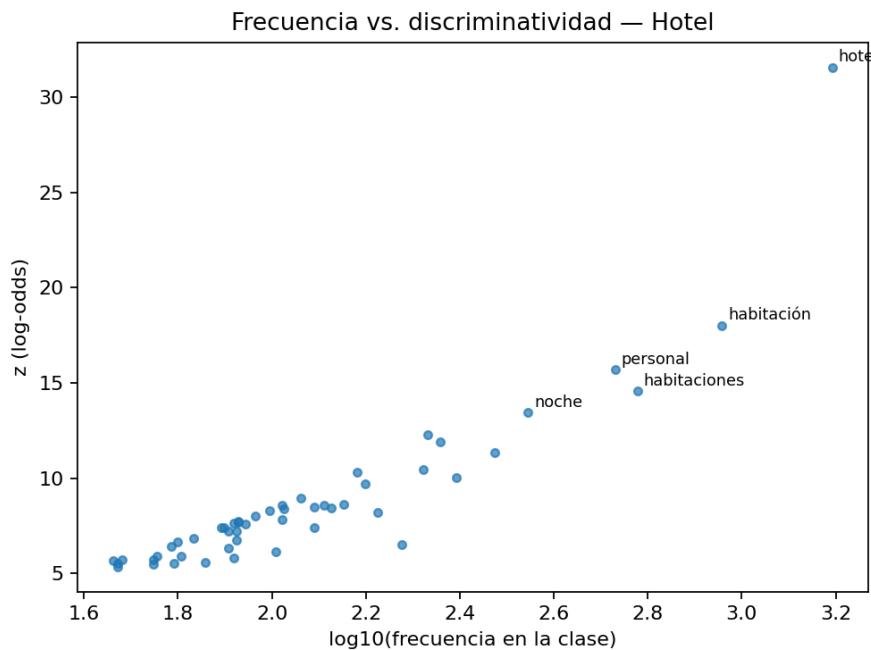
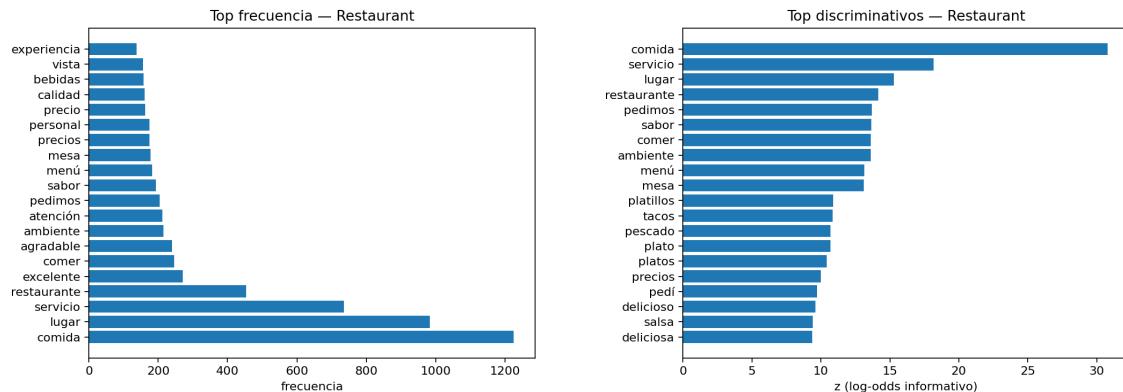


Figura 3: Relación entre Frecuencias y Discriminativo.

como “personal”, “noche” y “habitaciones”. Estos términos aparecen lo suficiente como para ser detectados como regulares, pero también lo suficientemente específicos como para describir con eficiencia su peso en las reseñas de hoteles. Como mencionamos antes, es vocabulario que describe servicios, atributos y eventos dentro del hotel.

Finalmente, en el cuadrante inferior izquierdo, la densa agrupación de puntos que se conforman en dicha zona corresponde a términos de baja frecuencia y bajo poder discriminativo. Para este caso en particular, se presenta una cola larga. Cada una de dichas palabras aporta poca información para clasificar el texto, pero colectivamente forman la reseña como tal.

## Restaurante



(a) Top 20 de las palabras más frecuentes para la clase Restaurante.

(b) Top 20 de las palabras más discriminativas para la clase Restaurante.

Para el caso de los restaurantes, podemos observar como la palabra anclaje de la clase Restaurante es: "comida". Este término domina de manera total tanto para frecuencia como para peso discriminativo. Además, podemos comprobar a través del gráfico de dispersión, ver figura 5, que "comida" es una palabra característica que se coloca como el núcleo semántico de la categoría Restaurante.

Ahora, la diferencia más clara entre nuestras listas es que el lod-odds logra filtrar adjetivos genéricos para identificar términos de poder discriminativo. De lado de la frecuencia se encuentran términos evaluativos como "excelente", "agradable" o "calidad". Por otra parte, log-odds los suprime y, en su lugar, le da mayor peso a verbos y sustantivos que describen el acto de comer con términos como: "pedimos", "menú", "mesa", "platillos", "sabor"; así como tipos de comida en específico como: "tacos" y "pescado". Esto revela como, si bien la gente frecuentemente califica la experiencia, las palabras que realmente distinguen una reseña de la clase Restaurante son aquellas que describen el proceso y los componentes de la experiencia de comer misma.

En cuanto al gráfico de dispersión, podemos ver cómo se ilustra la jerarquía del vocabulario. Comida se muestra directamente como el punto que más se aísla hacia la zona superior derecha. El grupo más cercano a la palabra de anclaje se conforma por "servicio", "lugar" y "restaurante".

A lo largo de la tendencia central, se encuentra la palabra "pedimos" reflejando su contexto de acción. Pese a que "comida", "lugar" y "servicio" sean palabras importantes para esta clase, palabras con un mayor valor analítico como "pedimos", "menú", "platillos" o

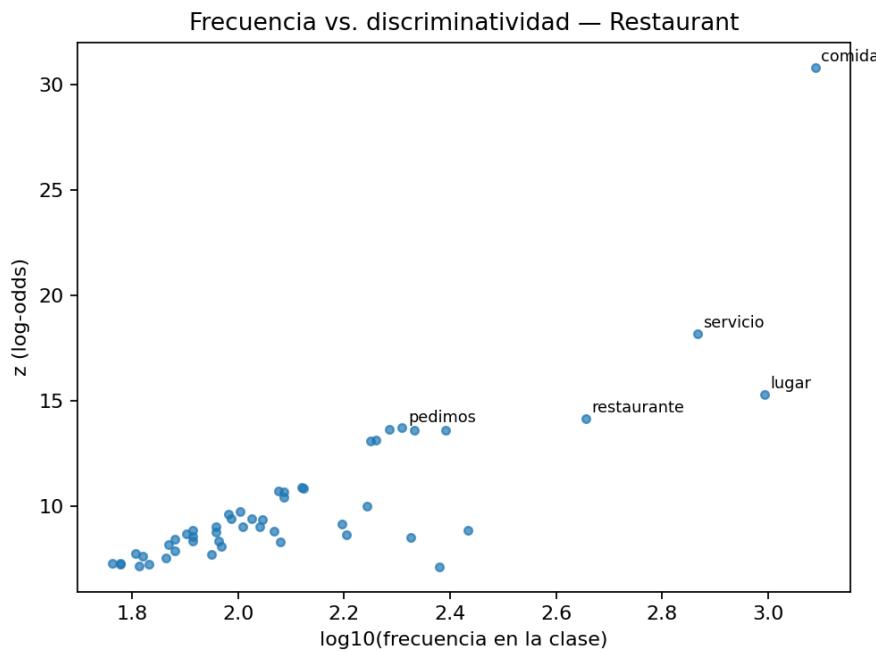
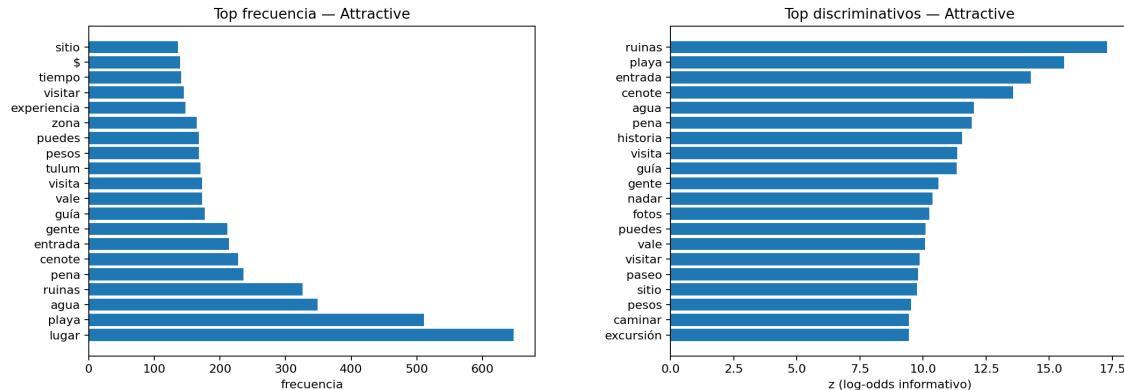


Figura 5: Relación entre Frecuencias y Discriminativo.

“tacos”, que describen la acción de comer, son más informativos para hacer la discriminación de clases y la clasificación de un documento como de tipo Restaurante.

## Atractivo



(a) Top 20 de las palabras más frecuentes para la clase Atractivo.

(b) Top 20 de las palabras más discriminativas para la clase Atractivo.

El caso de Atracción, los resultados muestran que no está definida por un único concepto dominante, sino por un conjunto de términos que describen los tipos específicos de destinos y las actividades que se pueden asociar a ellos.

Contrario a lo que sucedía con la clase Restaurante y Hotel, en el caso de Atractivo tenemos un término en extremo frecuente: "lugar". Se trata de una palabra genérica que parece aportar poco valor semántico. En contraste, el ranking discriminativo relega dicho término y empuja términos más específicos: "ruinas", "playa", "entrada" y "cenote". El modelo log-odds filtra el ruido y extrae los tipos de atracciones que definen a la categoría.

De ese modo, podemos observar cómo esta clase no cuenta con un único término ancla como sucedía en las anteriores. En el gráfico de dispersión en la figura 7, nos encontramos que, en lugar de un único valor atípico aísle, tenemos un cluster de alto impacto en la esquina superior derecha.

Los términos: "ruinas", "playa", "entrada", "cenote" y "agua" pueden representar diversas subcategorías temáticas. En el fondo, todas ellas representan una experiencia distinta, como arqueología, naturaleza, tiempo de ocio acuático, etc. El análisis discriminativo, como tal, representa palabras descriptoras poderosas porque capturan lo que los usuarios hacen en la atracción, proporcionando contexto para la clase Atractivo y demostrando su variabilidad en cuanto a experiencias, descripciones y, por ende, reseñas.

Aquí podemos ver como la fuerte correlación entre frecuencia y discriminativo se rompe un poco, pues la dispersión se abrió hacia distintos lados, sin concentrarse todo en una recta ascendente como sucedía en, por ejemplo, Hotel. Esto revela como la generalidad de la clase parece estar apuntando a que, mientras más específica cierta categoría, mayor

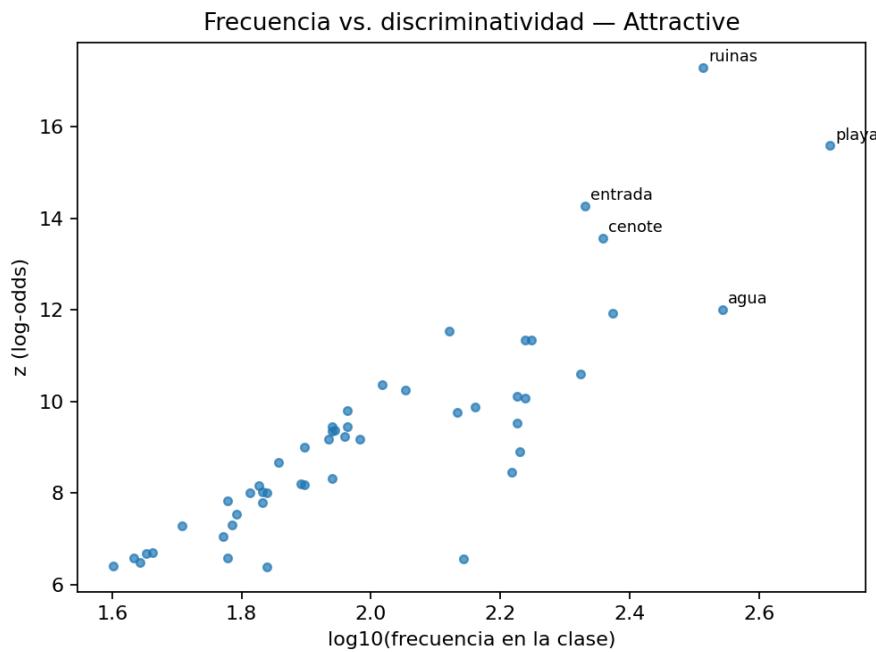


Figura 7: Relación entre Frecuencias y Discriminativo.

probabilidad de que haya correlación entre los dos rankings. Por otra parte, una categoría más general y variable, presenta una dispersión menos concentrada.

## Ejercicio #4 | Patrones gramaticales (POS 4-gramas)

- Etiqueta con POS cada documento.
- Extrae las secuencias gramaticales más frecuentes de longitud cuatro en cada clase.
- Discute si estas estructuras difieren entre clase y explica por qué.

### 0.10. Part of Speech (POS)

Para analizar el estilo y la estructura de un texto, no solo basta con observar las palabras que utiliza, también es fundamental entender cómo los organiza. Este enfoque se centra en el análisis de patrones gramaticales subyacentes, o la “huella digital sintáctica” de un texto. Para capturar estos patrones, se utilizan n-gramas de etiquetas de *Part of Speech*, o, en español, Parte de la Oración (POS).

Un n-grama de POS es una secuencia de etiquetas gramaticales. Por ejemplo, un 4-grama es una secuencia de cuatro etiquetas consecutivas como sería el caso de:

- **Determinante:** Precede al sustantivo para concretar su referencia (cantidad, posesión, proximidad). Ejemplos: **un** perro, **esta** manzana, **mi** casa.
- **Sustantivo:** Nombra personas, animales, cosas, lugares o ideas. Se clasifican en:
  - **Propios/Comunes:** *María* (único) vs. *niña* (clase).
  - **Concretos/Abstractos:** *mesa* (perceptible) vs. *amor* (idea).
  - **Individuales/Colectivos:** *oveja* (uno) vs. *rebaño* (conjunto).
- **Adjetivo:** Acompaña al sustantivo para expresar una cualidad o característica. Ejemplos: coche **rápido**, mujer **alta**.
- **Verbo:** Núcleo del predicado que expresa acción, estado o proceso. Varía en tiempo, modo, persona y número.
  - **Tipos principales:** De acción (*corre*), de estado (*es*, *está*) y transitivos/intransitivos según necesiten o no un complemento directo.

## 0.11. Análisis de Estructuras Sintácticas mediante N-gramas de POS

Formalmente dada una secuencia de palabras  $(w_1, \dots, w_T)$  y sus correspondientes etiquetas POS( $t_1, \dots, t_T$ ) donde cada  $t_i$  pertenece a un conjunto finito de etiquetas, como NOUN o VERB. Un 4-grama de POS se define como:

$$t_{i,i+3} = (t_i, t_{i+1}, t_{i+2}, t_{i+3})$$

Este modelo está construido para responder la pregunta: “¿Qué tan común es un patrón gramatical?”. Para ello, el primer paso es medir la “popularidad” de cada patrón gramatical dentro de una categoría específica, como en el caso de Type. Esto se logra con la Estimación de Máxima Verosimilitud (MLE), que de forma intuitiva no es más que calcular su frecuencia relativa.

En ese sentido, la probabilidad de un 4-grama  $t$  en una clase  $c$  se estima como:

$$\hat{p}_c(t) = \frac{N_c(t)}{\sum_{t'} N_c(t')}$$

Donde  $N_c(t)$  es simplemente el conteo de cuántas veces aparece nuestro patrón  $t$  en los documentos de la clase  $c$ , y el denominador es el número total de 4-gramas en esa misma clase. Un valor alto indica que el patrón es una estructura importante en el discurso.

Sin embargo, ser común no es lo mismo que tener peso significativo. Un patrón como [Determinante, Sustantivo, Verbo, Adverbio] puede ser frecuente en todas las categorías. Pero lo que de verdad no sinteresa es encontrar patrones que son mucho más comunes en una categoría en comparación con las demás. Se puede resolver ese problema utilizando log-odds nuevamente. Recordemos la formulación de prior Dirichlet ( $\alpha$ ) para estabilizar las estimaciones y manejar patrones poco frecuentes:

$$\delta_t = \log \frac{N_c(t) + \alpha_t}{N_c + \alpha_0 - (N_{\neg c}(t) + \alpha_t)} - \log \frac{N_{\neg c}(t) + \alpha_t}{N_{\neg c} + \alpha_0 - (N_c(t) + \alpha_t)}$$

Esta resta de logaritmos mide qué tan fuertemente está asociado el patrón  $t$  con la clase  $c$ . Un  $\delta_t$  positivo y grande significa que el patrón es característico de  $c$ . Para determinar si esta diferencia es estadísticamente significativa, y no producto del azar, se normaliza el puntaje y se recurre al z-score.

Con todo ello, podemos pasar de hacer análisis individuales de para medir la diferencia global entre los estilos de lenguaje entre categorías. En particular, se busca un valor que clasifique qué tan distinto es el lenguaje gramatical de las reseñas de, por ejemplo, Hoteles y Restaurantes.

Una manera de lograrlo es con la Divergencia de Jensen-Shannon (JSD). Esta es una medida simétrica que calcula la distancia entre dos distribuciones de probabilidad, para nuestro caso son las distribuciones de los 4-gramas de POS de cada clase:

$$JSD(P_c, P_d) = \frac{1}{2}KL\left(P_c \parallel \frac{P_c + P_d}{2}\right) + \frac{1}{2}KL\left(P_d \parallel \frac{P_c + P_d}{2}\right)$$

Donde  $P_c$  y  $P_d$  representan a las distribuciones de patrones para las clases  $c$  y  $d$ , respectivamente. Un JSD de 0 nos estaría diciendo que ambas clases usan los patrones gramaticales con exactamente la misma frecuencia, i.e. tienen estilos idénticos. Por otro lado, un valor mayor indicaría una gran divergencia estilística.

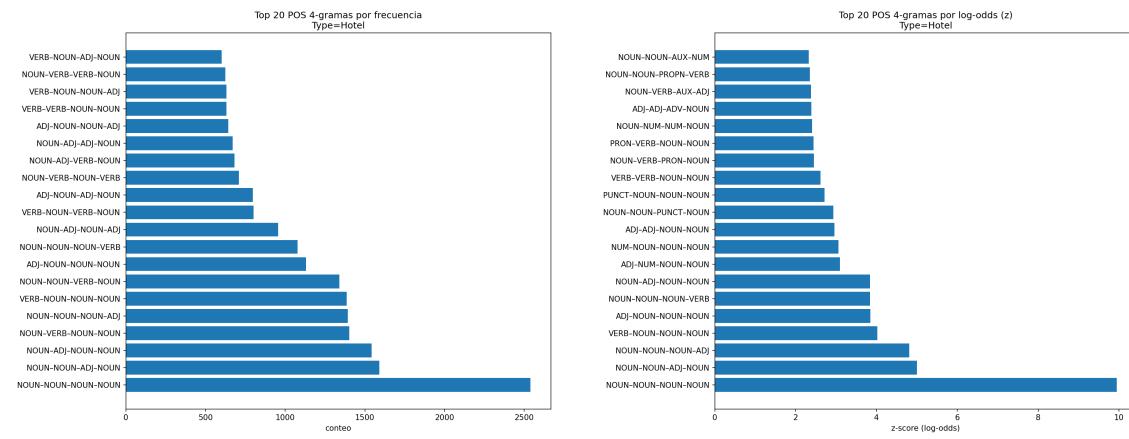
## Flujo del código

En ese sentido, nuestro código hace lo siguiente:

- **Definición formal de 4-gramas POS** → implementada mediante las funciones `sliding_ngrams` y `pos_sequence`, que generan las secuencias de etiquetas gramaticales.
- **Estimación de Máxima Verosimilitud (MLE)** → se calculan las frecuencias relativas de los 4-gramas dentro de cada clase; adicionalmente se generan tablas y, si se indica `-plots`, gráficas de barras con los patrones más frecuentes.
- **Log-odds con prior Dirichlet** → implementado en la función `dirichlet_log_odds`, siguiendo la formulación de Monroe et al. (2008), para identificar patrones discriminativos entre una clase y el resto; igualmente se guardan tablas y gráficos de los top patrones.
- **Divergencia de Jensen–Shannon (JSD)** → implementada en la función: `jensen_shannon_divergence`, utilizada para cuantificar la distancia estilística global entre distribuciones de 4-gramas de distintas clases.
- **Análisis explicativo** → mediante la función `jsd_per_token_contrib` se descompone la JSD en contribuciones por cada 4-grama, y con `sample_texts_with_pos4` se recuperan ejemplos textuales concretos de cada patrón, permitiendo una interpretación cualitativa de las diferencias entre clases.

## 0.12. Resultados

### Hotel



(a) Top 20 4-gramas POS para la clase Hotel.

(b) Top 20 4-gramas POS para la clase Hotel.

Para la clase Hotel, observamos que tanto en el conteo de 4-gramas por frecuencia como en los log-odds, el constructo más relevante es: NOUN-NOUN-NOUN-NOUN. Este patrón corresponde a bloques nominales en cadena, lo que refleja un discurso fuertemente descriptivo, típico cuando se habla de instalaciones, servicios o características de un lugar. Por ejemplo: “habitación suite lujo estándar”.

Además, entre los patrones más frecuentes encontramos:

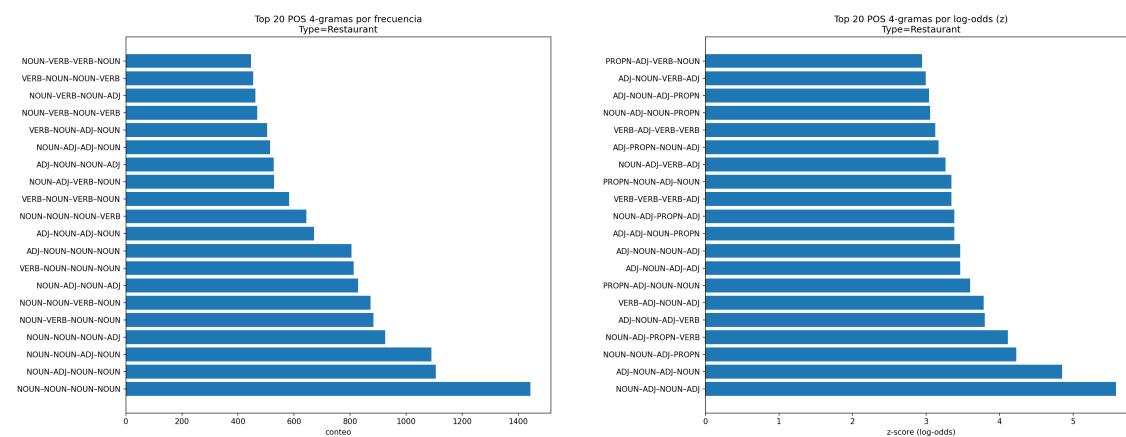
- NOUN-ADJ-NOUN-NOUN: describe un sustantivo acompañado de una cualidad que se relaciona con otros sustantivos, como en: “Habitación amplia, cama matrimonial” o “Hotel excelente, servicio habitación”.
- NOUN-VERB-NOUN-NOUN: un sustantivo funciona como sujeto de una acción que afecta a dos sustantivos relacionados, p. ej.: “Personal ofrece servicio alberca”, “Recepción tiene lista espera”.
- ADJ-NOUN-NOUN-NOUN: un adjetivo califica a un conjunto de tres sustantivos que forman una idea compuesta, como en: “Excelente ubicación, zona restaurantes”, “Pésimo servicio, atención cliente”.

Al pasar al análisis discriminativo con log-odds, vuelve a destacar la secuencia NOUN-NOUN-NOUN-NOUN, confirmando que aparece significativamente más en Hotel que en las otras dos clases. También sobresalen otros constructos:

- NUM-NOUN-NOUN-NOUN y NOUN-NUM-NUM-NOUN: reflejan listados con números, propios de especificaciones como “2 camas queen almohada” o “gimnasio 24 horas servicio”.
- PUNCT-NOUN-NOUN-NOUN y NOUN-NOUN-PUNCT-NOUN: aparecen en enumeraciones de amenidades, como “alberca gimnasio , comedor”.

Finalmente, también encontramos secuencias de NOUN-NOUN-PROPN-VERB, que sugieren la presencia de nombres propios de hoteles o cadenas acompañados de acciones: “servicio cuarto Hilton decepciona”, “zona alberca One requiere”.

En conclusión, para la clase Hotel el patrón NOUN-NOUN-NOUN-NOUN no solo es el más común, sino que además constituye un rasgo distintivo de la categoría, marcando un estilo discursivo descriptivo y de catálogo.



(a) Top 20 4-gramas POS para la clase Restaurante.

(b) Top 20 4-gramas POS para la clase Restaurante.

En el caso de Restaurante, la cadena más frecuente sigue siendo la de NOUN-NOUN-NOUN-NOUN, igual que Hotel. Esto parecería indicar una alta presencia de bloques nominales largos. Le siguen las variantes:

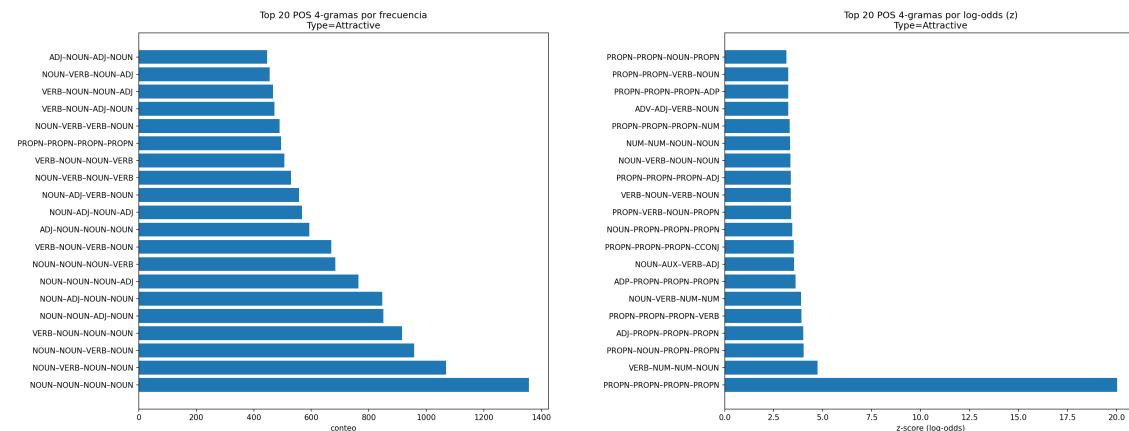
- NOUN-ADJ-NOUN-NOUN & ADJ-NOUN-NOUN-NOUN que muestran un uso intensivo de adjetivos calificativos para describir la experiencia, como: “excelente servicio atención cliente”.
- VERB-NOUN-NOUN-NOUN & NOUN-VERB-NOUN-NOUN: que apuntan a acciones relacionadas con la comida o el servicio, como: “trae comida mesa rápido” o “chef prepara platillo especial”.

Hablando solo de frecuencias, para la clase Restaurante, hay bastante correlación entre lo que se obtiene comparado con Hotel. Sin embargo, Restaurante incorpora constructos más verbales, que siguen sosteniendo la hipótesis del problema anterior en el que argumentamos que para esta clase se utilizan verbos que reflejan la “acción de comer”.

Ahora bien, si nos concentramos en la gráfica 9b referente a la discriminante log-odds, podemos ver cómo cambia el ranking, con la presencia de la cadena: “NOUN-ADJ-NOUN-ADJ” como dominante:

- ADJ-NOUN-ADJ-NOUN & NOUN-ADJ-VERB-ADJ: presencia de un intenso uso de los adjetivos, que refuerza la teoría de que estas reseñas describen el “acto de comer”.
- PROPN-ADJ-VERB-NOUN & NOUN-ADJ-NOUN-PROPN: uso de nombres propios que reflejan menciones de restaurantes, marcas o incluso chefs o meseras.
- VERB-VERB-VERB-ADJ & VERB-ADJ-VERB-VERB: la fuerte presencia de verbos consecutivos puede sugerir una reseña con toques narrativos “llegó muy rápido corriendo”.

Se podría decir que para la categoría Restaurante se refleja un estilo mucho más valorativo y cargado en la acción de comer. Probablemente abundan juicios, opiniones subjetivas y narrativa de la experiencia.



(a) Top 20 4-gramas POS para la clase Atractivo.

(b) Top 20 4-gramas POS para la clase Atractivo.

Para la clase Atracción, el patrón más frecuente es, nuevamente, NOUN-NOUN-NOUN-NOUN, reflejando la presencia de bloques nominales largos. Este tipo de construcciones son comunes en descripciones de lugares, espacios y características físicas de un atractivo turístico. Ejemplo: “zona arqueológica templo mayor”.

Otros patrones frecuentes son:

- NOUN-VERB-NOUN-NOUN VERB-NOUN-NOUN-NOUN: muestran que, además de la descripción, aparecen acciones vinculadas con los lugares, como “turistas visitan ruinas mayas” o “recorrer centro histórico ciudad”.
- NOUN-NOUN-VERB-NOUN: indica la interacción entre sustantivos y acciones, por ejemplo: “plaza alberga museo nacional”.
- NOUN-NOUN-ADJ-NOUN: introduce adjetivos calificativos, como “arquitectura colonial hermosa iglesia”.

Al pasar al análisis discriminativo con log-odds, se observa un cambio muy marcado respecto a las otras clases: los constructos dominados por PROPN (nombres propios) son los más característicos. El ranking lo encabezan secuencias como:

- PROPN-PROPN-PROPN-PROPN: cadenas largas de nombres propios, p. ej.: “San Juan Teotihuacán México Patrimonio”.
- ADJ-PROPN-PROPN-PROPN NOUN-PROPN-PROPN-PROPN: topónimos acompañados de adjetivos o sustantivos, como “hermoso Parque Nacional Nevado” o “zona Centro Histórico Ciudad”.
- PROPN-PROPN-PROPN-VERB: secuencias de nombres propios con verbos, p. ej.: “Cerro de la Silla impresiona”.

Estos resultados muestran que la categoría Attractive se distingue por su alta densidad de topónimos y nombres propios, lo cual es consistente con reseñas que nombran sitios, monumentos, barrios, zonas arqueológicas o reservas naturales.

## Divergencia Jensen-Shannon

Usando base 2, la JSD toma valores en  $[0, 1]$ . Obtuvimos lo mostrado en el cuadro 5, mostrado a continuación.

Cuadro 5: Matriz de divergencia Jensen–Shannon (JSD) entre clases.

	<b>Attractive</b>	<b>Hotel</b>	<b>Restaurant</b>
<b>Attractive</b>	0	0.0352	0.0438
<b>Hotel</b>	0.0352	0	0.0322
<b>Restaurant</b>	0.0438	0.0322	0

Estos valores indican diferencias pequeñas pero consistentes. El mayor valor entre Atractivo y Restaurante es coherente con que la primera presenta alta densidad de topónimos (secuencias con PROPN), mientras que Restaurante incorpora adjetivación valorativa y verbos. La menor distancia entre Hotel y Restaurante sugiere estilos más próximos por la abundancia de cadenas nominales; no obstante, Hotel se distingue por listados con NUM/PUNCT.

En conjunto, los resultados muestran que las estructuras sintácticas sí difieren entre clases, aunque con sus matices. La clase Hotel se caracteriza por bloques nominales largos y por el uso recurrente de números y signos de puntuación, lo cual refleja mucha carga descriptiva orientada a enumerar servicios y especificaciones. Por su parte, la clase Restaurante comparte la base nominal, pero se distingue por un mayor peso de verbos y adjetivos, en línea con un estilo experiencial y valorativo propio de reseñas sobre el acto de comer. Finalmente, la clase Atractivo concentra secuencias con nombres propios (PROPN), lo que indica una fuerte carga topográfica coherente con la necesidad de referirse explícitamente a sitios, monumentos o zonas turísticas.

Si los comparamos, la divergencia de Jensen-Shannon confirmó dichas diferencias: los valores, aunque bajos en magnitud absoluta, fueron sistemáticamente mayores en los pares donde el contraste discursivo era más evidente como lo fue para *Attractive* vs. *Restaurant*. De esta manera, los análisis cuantitativos (frecuencias, log-odds, JSD) y ejemplos textuales convergen a una misma conclusión: cada clase desarrolla un estilo lingüístico característico, enraizado en la función comunicativa de las reseñas. Esto demuestra que la modelación basada en POS n-gramas no solo captura regularidades formales, sino que también permite interpretar y explicar diferencias discursivas entre categorías de texto.

## Ejercicio #5 |

- Construye representaciones BoW con TF y con TF-IDF.
- Aplica alguna medida estadística como Ji-cuadrada, información mutua o *information gain*.
- Obtén el top 20 de características más importantes en cada representación.
- Analiza diferencias entre ambas representaciones.

El objetivo en este problema es representar los textos como vectores numéricos para después poder aplicar algún análisis estadístico o modelo de ML. En NLP, la Bolsa de Palabras surgió como un primer, y poderoso intento, modelo clásico para ello. Cada documento se representa por un vector donde cada dimensión corresponde a un término y su valor refleja alguna medida de importancia, como el conteo, la frecuencia o el TF-IDF, que describiremos a continuación.

Ahora bien, para determinar qué palabras son más relevantes para distinguir entre clases, podemos aplicar medidas estadísticas como las mencionadas en el cuerpo del problema.

### 0.13. Representaciones BoW: TF y TF-IDF

De acuerdo con el libro “Speech and Language Processing” de Jurafsky & Martin, las representaciones clásicas de texto en NLP se basan en modelos de espacios vectoriales. En ellas, cada documento se convierte en un vector en el que cada dimensión es un término del vocabulario. Para nuestro caso, usaremos *Term Frequency* (Tf) y *Inverse Document Frequency* (IDF).

- **TF:** mide cuántas veces aparece un término en un documento. Puede usarse en crudo como:  $count(t, d)$  o con un suavizado logarítmico:

$$tf_{t,d} = \log(count(t, d) + 1)$$

El ajuste se realiza para evitar que una palabra con, por ejemplo, 100 repeticiones valga 100 veces más que una con una sola repetición.

- **IDF:** por otro lado, *Inverse Document Frequency* mide qué tan raro es cierto término

en nuestro conjunto:

$$idf_t = \log_{10} \frac{N}{df_t}$$

Aquí,  $N$  es el número de documentos y  $df_t$  en cuántos de ellos aparece el término.

- **TF-IDF:** combina las dos aproximaciones anteriores:

$$tfidf(t, d) = tf_{t,d} \cdot idf_t$$

Se podría decir, tomando a Jurafsky como referencia, que TF-IDF equilibra dos fuerzas. Por un lado, las palabras frecuentes en un documento son informativas, pero si son frecuentes en todo el corpus (como es el caso de las *stop words* o términos genéricos que no aportan a la semántica, no discriminan bien.

Siguiendo la misma explicación del libro, nos encontramos con que, acorde a Jurafsky, TF crudo favorece palabras muy frecuentes en un documento, pero puede confundir si esas palabras son también frecuentes en otras clases. Caso contrario al de TF-IDF, este método penaliza términos muy comunes en el corpus, y resaltan los que distinguen a cada documento. Una primera hipótesis puede ser que con TF van a destacar palabras como “hotel”, “restaurante” o “lugar”, mientras que con TF-IDF se le dará peso a términos menos generales.

Además, podemos apoyarnos de estadísticos para la selección de características. Justo las tres opciones que nos da el inciso de la tarea nos pueden aportar para definir qué tan informativo es un término respecto a las clases.

- **Ji-cuadrada  $\chi^2$ :** prueba de independencia entre término y clase. Palabras con una Ji-cuadrada de alto valor aparecen mucho más en cierta clase.
- **Mutual Information (MI):** mide cuánto reduce la incertidumbre sobre la clase el enfocarse en un término.
- **Information Gain (IG):** cuantifica la reducción de entropía de las clases al usar un término como característica.

### Ji-Cuadrada $\chi^2$

- Evalúa si la ocurrencia de un término y la pertenencia a una clase son independientes.
- Si son independientes, el término no ayuda a predecir la clase.
- Si no lo son, el término es discriminativo.

En términos más formales, para cada palabra  $t$  y clase  $c$ , se construye una tabla  $2 \times 2$ :

	Clase $c$	No clase $c$
Contiene $t$	A	B
No contiene $t$	C	D

$$\chi^2(t, c) = \frac{N(AD - BC)^2}{(A + B)(C + D)(A + C)(B + D)}$$

donde  $N = A + B + C + D$ .

Un  $\chi^2$  alto indica que el término  $t$  aparece desproporcionadamente en la clase  $c$ .

Detrás del formalismo, es como si se estuviera preguntando: “¿esta palabra ocurre mucho más de lo esperado en esta clase?”. Si en reseñas de Hoteles la palabra **spa** aparece 80 % de las veces, pero casi nunca en Restaurante, su  $\chi^2$  será alto → es un fuerte discriminador.

### Mutual Information (MI)

- Mide la reducción de incertidumbre sobre la clase al observar un término.
- Se inspira en la teoría de la información de Shannon.

Formalmente hablando:

$$I(t; c) = \sum_{t' \in \{0,1\}} \sum_{c' \in \{0,1\}} p(t', c') \log \frac{p(t', c')}{p(t')p(c')}$$

donde:

- $t' = 1$  si el término aparece,  $t' = 0$  si no.
- $c' = 1$  si el documento pertenece a la clase,  $c' = 0$  si no.

Por lo que MI mide cuánto saber que la palabra aparece (o no aparece) nos informa sobre la clase. Si la palabra está distribuida de manera muy distinta entre clases, la MI será alta. La palabra **buffet** en Restaurante. Si aparece, casi seguro es esa clase → alta MI.

## Information Gain (IG)

Esta es una métrica muy utilizada en algoritmos como árboles de decisión. Se pregunta: “¿cuánta incertidumbre sobre la clase se reduce si consideramos la presencia o la ausencia de este término?”. El formalismo es el siguiente:

$$IG(t) = H(C) - H(C|t)$$

$H(C)$ : entropía de la distribución de clases.

$H(C|t)$ : entropía condicional al observar el término.

**Entropía:**

$$H(C) = - \sum_c p(c) \log p(c)$$

De esa forma, IG valora términos que partitionan bien los datos por clase. Si al dividir por la presencia de una palabra los documentos quedan mucho más “puros” en cuanto a clase, entonces la ganancia de información es alta. La palabra **recepción** podría dividir reseñas: cuando aparece, casi siempre son de la categoría Hotel.

## Invarianza de *Information Gain* frente a Ponderación TF y TF-IDF

Durante el proceso de selección de características, se observó que la métrica de IG produce un ordenamiento de características idéntico para TF y para TF-IDF. Esta invarianza no es coincidencia, pues es algo completamente provocado por la construcción matemática de *Information Gain* en el caso de NLP.

El cálculo de IG se basa en la reducción de la entropía de las clases  $H(C)$  al conocer la presencia o ausencia de un término  $t$ . Su formulación es la siguiente:

$$IG(C, t) = H(C) - H(C|t)$$

La clave aquí está en la entropía condicional:  $H(C|t)$ , que se calcula a partir de la probabilidad de presencia del término  $P(t)$ , y la probabilidad de su ausencia,  $P(\neg t)$ . Estas probabilidades se derivan de la matriz documento-término  $X$ , donde  $x_{i,j}$  es el valor del término  $j$  en el documento  $i$ :

$$P(t_j) = \frac{|\{i | x_{ij} > 0\}|}{N}$$

La condición de existencia de un término en un documento es, por tanto,  $x_{ij} > 0$ . Esta evaluación binaria es el núcleo de la invarianza.

- Para la matriz TF, el valor  $x_{ij}$  es el conteo  $TF_{ij}$ . La condición  $x_{ij} > 0$  es verdadera si y solo si  $TF_{ij} \geq 1$ .

- Para la matriz TF-IDF, el valor  $x_{ij} = TF_{ij} \times IDF_j$ . Dado que  $IDF_j$  es una constante positiva, para términos no presentes en el documento, la condición  $x_{ij} > 0$  se cumple si y solo si  $TF_{ij} > 0$

Dado que la condición de presencia define al mismo subconjunto de documentos para ambas representaciones, todos los componentes de la formulación de IG son numéricamente idénticos.

Podemos ilustrar esto con una analogía. Las representaciones TF y TF-IDF son como dos esquemas de ponderación distintos para evaluar la importancia de un término. Sin embargo, la métrica de IG, en su implementación estándar, no utiliza la magnitud de esta ponderación. En su lugar, aplica un umbral binario para determinar si una característica está "presente" (valor  $> 0$ ) o "ausente" (valor  $= 0$ ). Este proceso es análogo a convertir una calificación numérica detallada en un simple resultado de "Existe" o "No Existe".

## Flujo de trabajo

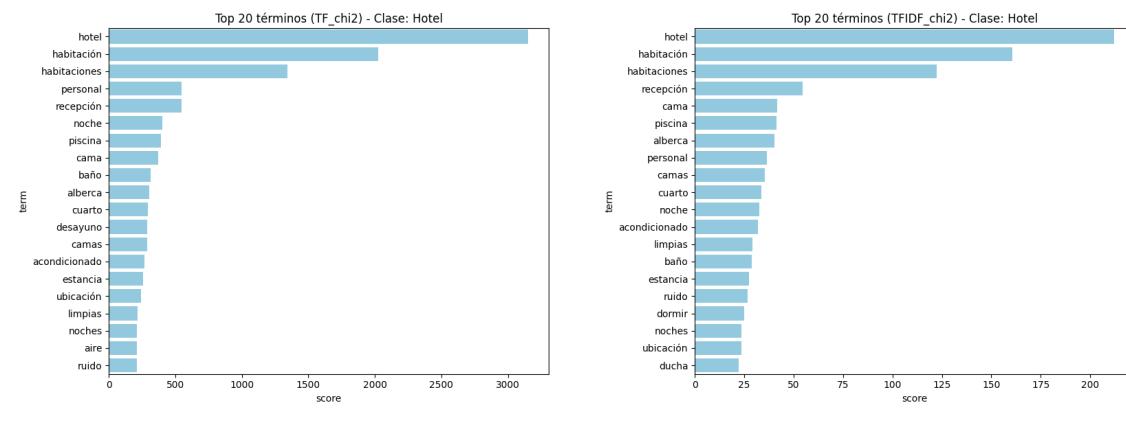
Nuestro código hace lo siguiente:

- Construcción de representaciones BoW** → mediante los vectorizadores CountVectorizer y TfidfVectorizer, se generan las matrices dispersas de términos para las dos variantes: **TF** y **TF-IDF**.
- Selección de características por clase (one-vs-rest)** → para cada clase en la columna especificada, se crean etiquetas binarias y se aplican las métricas a continuación:
  - Chi-cuadrada** → implementada con chi2 de sklearn.feature\_selection, mide la asociación estadística entre la presencia de un término y la pertenencia a la clase.
  - Información Mutua (MI)** → implementada con mutual\_info\_classif, cuantifica la reducción de incertidumbre de la clase al observar un término.
  - Information Gain (IG)** → implementada manualmente en la función compute\_ig, calcula la disminución de entropía de las clases al condicionar por la presencia/ausencia de cada término.
- Ranking de términos relevantes** → los resultados de cada métrica se ordenan y se exportan en archivos .csv, mostrando los top-n términos más discriminativos para cada clase y representación (TF y TF-IDF).
- Visualización** → si se activa la opción -plots, se generan gráficas de barras (plot\_top) con los términos más relevantes por clase y métrica, guardadas como imágenes .png.

- **Salida final** → en la carpeta indicada por `-save-dir` se almacenan todas las tablas y gráficos generados, organizados por métrica, representación y clase.

## Resultados

### Hotel



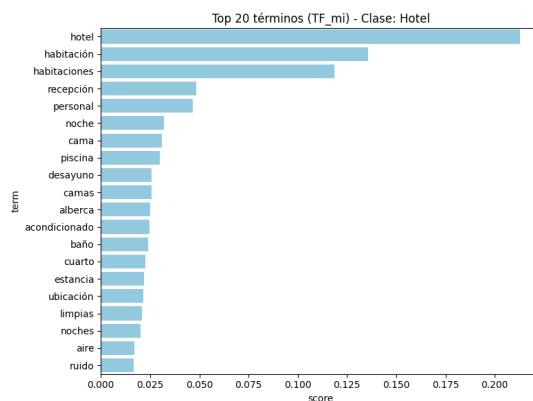
(a) Top 20 términos TF por  $\chi^2$  para la clase Hotel.

(b) Top 20 términos TF-IDF por  $\chi^2$  para la clase Hotel.

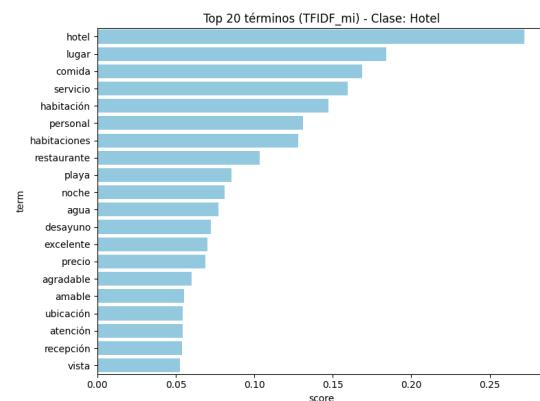
**Ji-cuadrada  $\chi^2$ :** Al contrastar las características seleccionadas con la métrica de Ji-cuadrada, podemos observar diversas cosas de interés. Primero, ambas implementaciones coinciden en 18 términos altamente discriminativos para la clase Hotel. La lista incluye conceptos centrales a la experiencia hotelera, como lo son: “hotel”, “habitación”, “recepción”, “piscina” o “personal”.

La diferencia principal en la figura ?? radica en los términos que TF considera importantes, pero que para TF-IDF no lo son. Con la implementación TF se destacan palabras como: “desayuno” y “aire”. Pero, por ejemplo, en el caso de TF-IDF, desayuno no aparece en el ranking. Es bastante seguro mencionar que esto puede estar sucediendo debido a que “desayuno” es un término que seguramente también podemos encontrar en la categoría de Restaurante. Dicho motivo disminuye el peso de la palabra para la clase actual.

En este caso, TF-IDF no cambia demasiado en comparación con solo TF. Sin embargo, los sutiles detalles como el de “desayuno” permiten identificar el contraste entre ambas aproximaciones. Al menos en esta sección, podemos ver que TF-IDF le da pesos similares a los términos identificados por TF, pero sí reordena algunos de ellos.



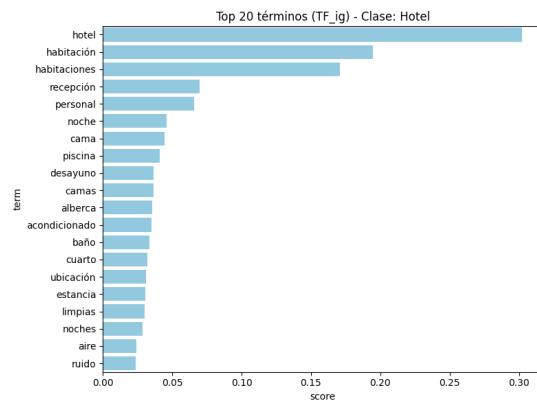
(a) Top 20 términos TF por Información Mutua (IM) para la clase Hotel.



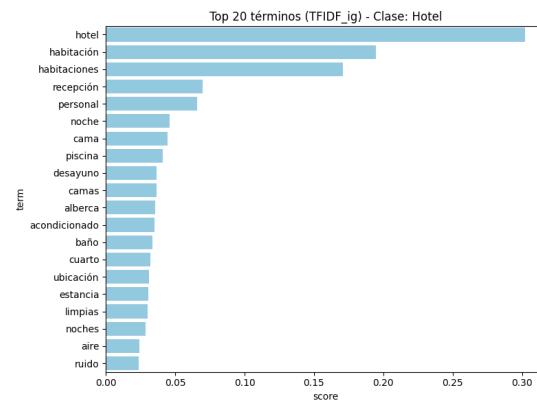
(b) Top 20 términos TF-IDF por Información Mutua (IM) para la clase Hotel.

**Información Mutua (IM):** Ahora, hablando de *Mutual Information*, al medir la reducción de incertidumbre, muestra una sensibilidad mayor a la representación utilizada. En lugar de empatar en casi todos los términos como sucedía con  $\chi^2$ , en este caso solo hay empate con ocho palabras: “desayuno”, “habitaciones”, “habitación”, “hotel”, “noche”, “personal”, “recepción”, “ubicación”. Estos son los términos más informativos cuando ponderamos el peso de ambos enfoques.

Ahora, TF tiene una gran cantidad de términos que describen características del hotel y experiencias dentro del mismo, como lo son: “piscina”, “cama”, “baño”, “ruido” y “alberca”. Debido a la alta frecuencia de este tipo de cosas o de situaciones en hoteles, TF las captura y considera como de gran aporte informativo. Por otra parte, el análisis de TF-IDF revela que este método se queda con palabras que le dan más peso a la condición con la que recibieron un servicio con palabras como: “precio”, “amable”, o “excelente”.



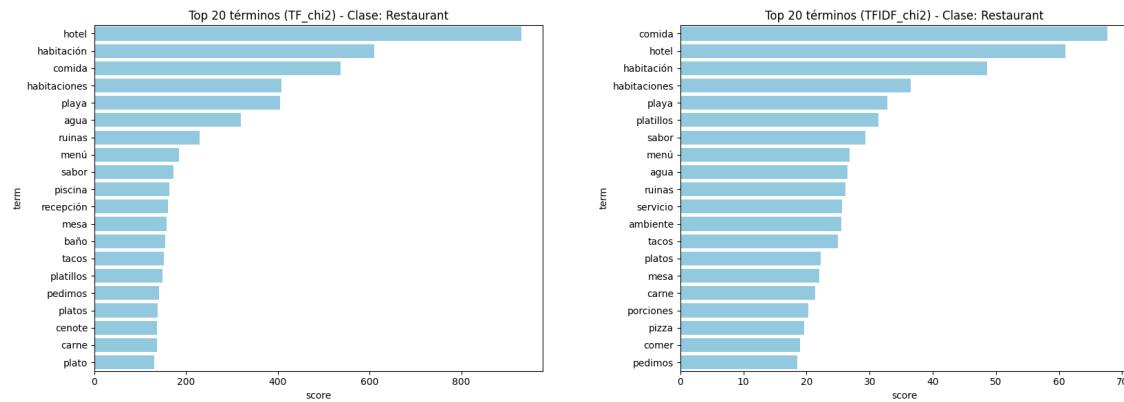
(a) Top 20 términos TF por Information Gain (IG) para la clase Hotel.



(b) Top 20 términos TF-IDF por Information Gain (IG) para la clase Hotel.

**Information Gain (IG):** Para el caso de *Information Gain*, nos encontramos con que tanto para TF como TF-IDF, los resultados son los mismos. Esto tiene concordancia con lo mencionado en el apartado de Invarianza de Information Gain frente a Ponderación TF y TF-IDF 0.13.

## Restaurante



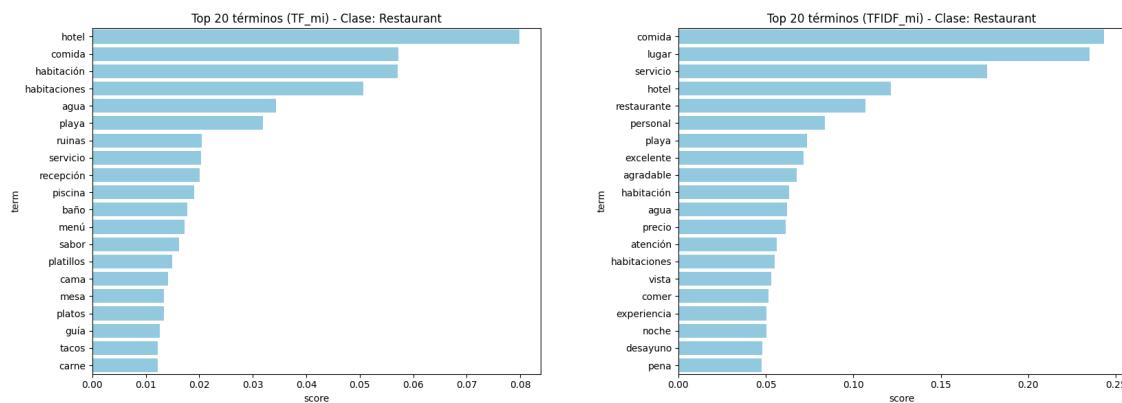
(a) Top 20 términos TF por  $\chi^2$  para la clase Restaurante.

(b) Top 20 términos TF-IDF por  $\chi^2$  para la clase Restaurante.

**Ji-cuadrada  $\chi^2$ :** Al contrastar las características seleccionadas con la métrica Ji-cuadrada para la clase Restaurante, se observan diferencias notables que ilustran la ventaja de usar TF-IDF sobre TF. Ambas implementaciones coinciden en 15 términos, incluyendo palabras centrales para la experiencia culinaria como "comida", "platillos", "menú", "sabor" y "tacos".

La divergencia principal radica en los términos que TF considera importantes debido a su alta frecuencia, pero que TF-IDF logra identificar como poco específicos. La implementación con TF resalta palabras muy asociadas a un contexto hotelero o turístico, tales como "piscina", "recepción", "baño" y "cenote". Esto sugiere que muchas reseñas de restaurantes provienen de establecimientos ubicados dentro de hoteles o cerca de atracciones, y TF, al ser una medida de frecuencia simple, captura este contexto sin discriminar su relevancia temática.

En cambio, TF-IDF demuestra su eficacia al penalizar esos términos comunes y promover palabras mucho más discriminativas y exclusivas de la clase Restaurante. En su top 20 aparecen conceptos como "servicio", "ambiente", "porciones", "pizza" y "comer". Estos términos son semánticamente más puros y describen directamente la experiencia gastronómica, demostrando que TF-IDF filtra con éxito el "ruido" de temas adyacentes (como Hotel o Atractivo) para quedarse con las características verdaderamente informativas de la clase.



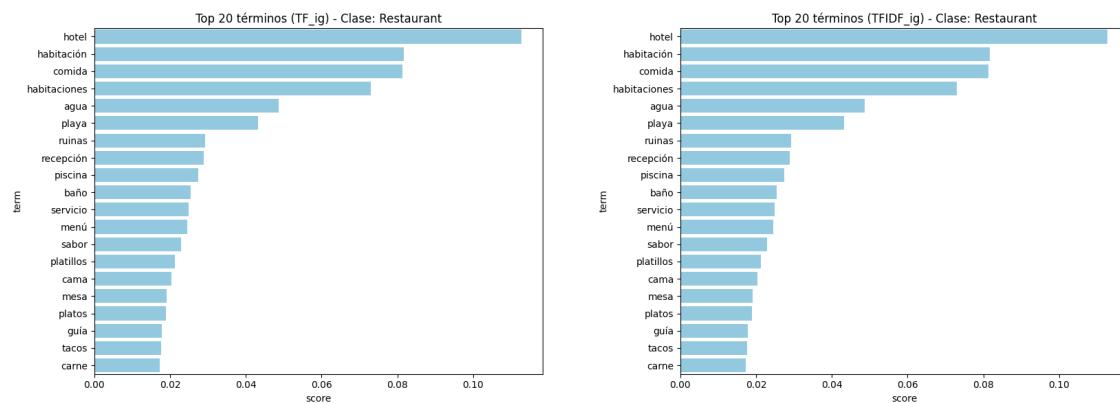
(a) Top 20 términos TF por Información Mutua (IM) para la clase Restaurante.

(b) Top 20 términos TF-IDF por Información Mutua (IM) para la clase Restaurante.

**Información Mutua (IM):** Al utilizar Información Mutua, la diferencia entre las ponderaciones TF y TF-IDF se vuelve aún más pronunciada, revelando una alta sensibilidad de esta métrica. En este caso, ambas implementaciones solo coinciden en 12 términos, entre los que se encuentran palabras clave como "restaurante", "comida", "servicio" y "platillos".

El enfoque con TF tiende a seleccionar términos que describen la calidad del servicio y la experiencia de manera general. Palabras como "atención", "excelente", "buena", "personal", "amables" y "bueno" dominan la lista. Si bien son informativas, estas palabras no describen el restaurante por su comida, sino por la calidad de su servicio, algo que podría ser común en reseñas de muchas otras categorías.

Por el contrario, TF-IDF transforma radicalmente la selección. Elimina casi todos los adjetivos de servicio generales y los sustituye por un vocabulario puramente culinario. Términos como "pizza", "porciones", "pasta", "postre", "sopa", "ensalada", "vino" y "desayuno" emergen con fuerza. Esto demuestra que TF-IDF, al medir la especificidad de un término, permite a Información Mutua identificar las características que describen el menú y la oferta gastronómica del restaurante, ofreciendo un perfil mucho más concreto y relevante que el obtenido con TF.



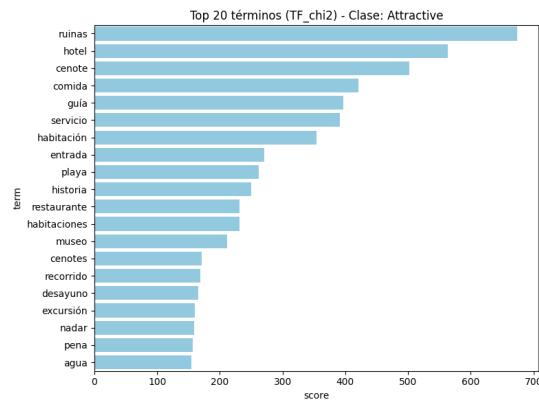
(a) Top 20 términos TF por Information Gain (IG) para la clase Restaurante.

(b) Top 20 términos TF-IDF por Information Gain (IG) para la clase Restaurante.

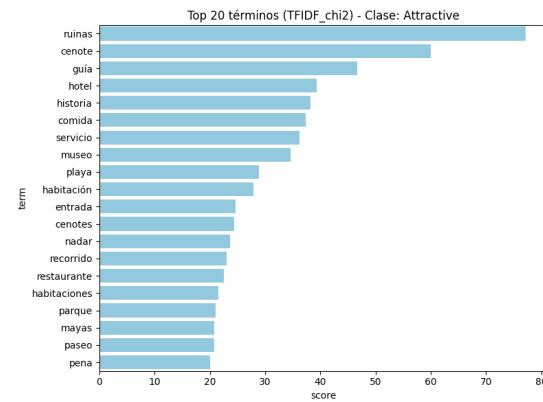
**Information Gain (IG):** Para la métrica Information Gain, se confirma el comportamiento teórico esperado. Al igual que ocurrió en el análisis de la clase Hotel, los resultados obtenidos con la ponderación TF y TF-IDF son exactamente los mismos.

Ambos gráficos muestran idénticos términos con el mismo puntaje y en el mismo orden. Esto se debe a la naturaleza del cálculo de Information Gain, que se basa en la presencia o ausencia de un término en los documentos de una clase, y no en su frecuencia ponderada (como TF) o su especificidad (como TF-IDF). Por lo tanto, la transformación de pesos de TF a TF-IDF no altera el resultado final de esta métrica.

## Atractivo



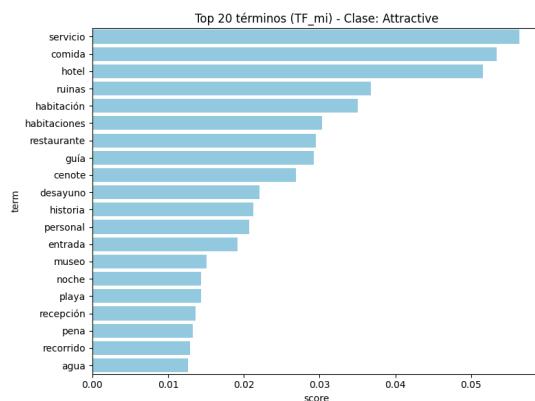
(a) Top 20 términos TF por  $\chi^2$  para la clase Atractivo.



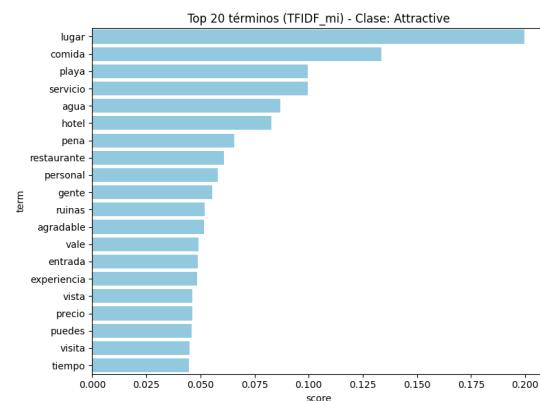
(b) Top 20 términos TF-IDF por  $\chi^2$  para la clase Atractivo.

**Ji-cuadrada  $\chi^2$ :** La representación basada en frecuencia simple (TF) selecciona términos de muy alta frecuencia en el corpus general, pero que no son exclusivos de los atractivos turísticos. Palabras como "hotel", "comida", "playa" y "agua" dominan su top, indicando que muchas reseñas sobre atractivos se dan en contextos de vacaciones donde también se habla de alojamiento y comida.

En contraste, TF-IDF penaliza esos términos generales y resalta un vocabulario mucho más específico y descriptivo de lugares de interés. Su lista incluye palabras como "ruinas", "cenote", "zona", "arqueológica", "pirámide" y "museo". Estos términos son inequívocamente relevantes para la clase Atractivo.



(a) Top 20 términos TF por Información Mutua (IM) para la clase Atractivo.

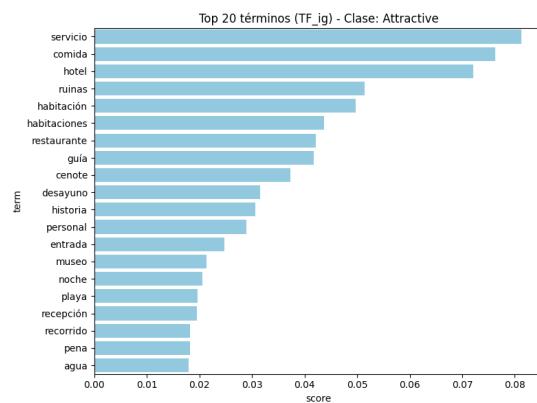


(b) Top 20 términos TF-IDF por Información Mutua (IM) para la clase Atractivo.

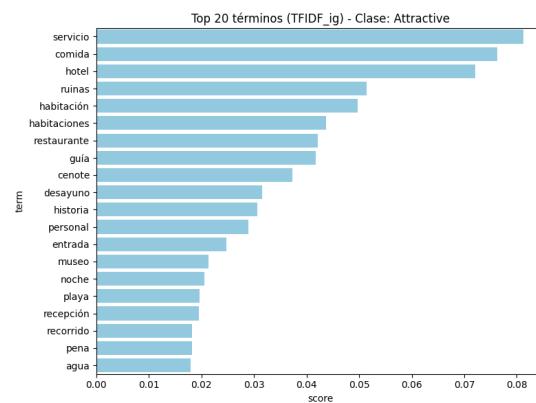
**Información Mutua (IM):** La representación TF selecciona un conjunto de términos muy genéricos y positivos que podrían aplicar a casi cualquier categoría. Palabras como "excelente", "agradable", "amable", "recomendable" y "buena" dominan el ranking. Si bien estos adjetivos describen una experiencia positiva, carecen de especificidad temática y no informan sobre la naturaleza del atractivo turístico en sí.

La selección de TF-IDF es radicalmente diferente y mucho más informativa. Elimina casi todos los adjetivos genéricos y los reemplaza por sustantivos que describen directamente los atractivos. Términos como "cenote", "ruinas", "museo", "zona", "parque", "historia" y "arqueológica" emergen como los más importantes. Este vocabulario es temáticamente puro y describe sin ambigüedad el tipo de lugar que se está reseñando.

TF-IDF permite que la métrica de Información Mutua alcance su máximo potencial. Al filtrar las palabras comunes y de sentimiento general, TF-IDF le presenta a IM un conjunto de términos donde la presencia de una palabra reduce drásticamente la incertidumbre sobre si la reseña trata de un Atractivo.



(a) Top 20 términos TF por Information Gain (IG) para la clase Atractivo.



(b) Top 20 términos TF-IDF por Information Gain (IG) para la clase Atractivo.

**Information Gain (IG):** Sigue lo mismo que en los casos anteriores.

## Ejercicio #6 | Bigramas

Repite el ejercicio anterior, pero utilizando bigramas de palabras. Compara resultados y discute si los bigramas aportan mayor discriminación semántica.

### N-gramas

Acorde con lo expuesto por Jurasfky en “Speech and Language Processing”, sea un documentno o secuencia de palabras:

$$w_1, w_2, \dots, w_n$$

Un n-grama es una secuencia de  $n$  palabras consecutivas:

$$(w_i, w_{i+1}, \dots, w_{i+n-1})$$

Para el caso de bigramas:

$$(w_i, w_{i+1}) \quad \forall i = 1, \dots, n - 1$$

**Unigramas (n=1)** Son simplemente cada una de las palabras (o *tokens*) de la secuencia. Pensemos en la siguiente frase:

«Bilbo Bolsón er aun hobbit muy respetable».

- ("Bilbo")
- ("Bolsón")
- ("era")
- ("un")
- ("hobbit")
- ("muy")
- ("respetable")

**Bigramas (n=2)** Son todas las secuencias de dos palabras consecutivas,  $(w_i, w_{i+1})$ .

- $(w_1, w_2)$ : ("Bilbo", "Bolsón")
- $(w_2, w_3)$ : ("Bolsón", "era")
- $(w_3, w_4)$ : ("era", "un")
- $(w_4, w_5)$ : ("un", "hobbit")
- $(w_5, w_6)$ : ("hobbit", "muy")
- $(w_6, w_7)$ : ("muy", "respetable")

Mientras más palabras agreguemos al n-grama, más contexto local podemos capturar.

**Probabilidad de bigrama** Usando el modelo de lenguaje clásico, Jurafsky plantea que, a través de bigramas, la probabilidad de la secuencia se aproxima como:

$$p(w_1^n) \approx \prod_{i=1}^n P(w_i | w_{i-1})$$

donde la probabilidad condicional se estima con máxima verosimilitud:

$$P(w_i | w_{i-1}) = \frac{C(w_{i-1}, w_i)}{C(w_{i-1})}$$

En este caso,  $C(w_{i-1}, w_i)$  es el número de veces que aparece el bigrama  $(w_{i-1}, w_i)$  en el corpus. Por otro lado,  $C(w_{i-1})$  es el número de veces que aparece la palabra  $w_{i-1}$ .

A diferencia de lo que se puede llegar a pensar, una aproximación BoW con bigramas no modela la probabilidad de la secuencia completa, sino que construye un vector donde cada dimensión corresponde a un bigrama del vocabulario. De la forma:

$$\vec{d} = (x_{(w_1, w_2)}, x_{(w_2, w_3)}, \dots)$$

## Flujo de trabajo

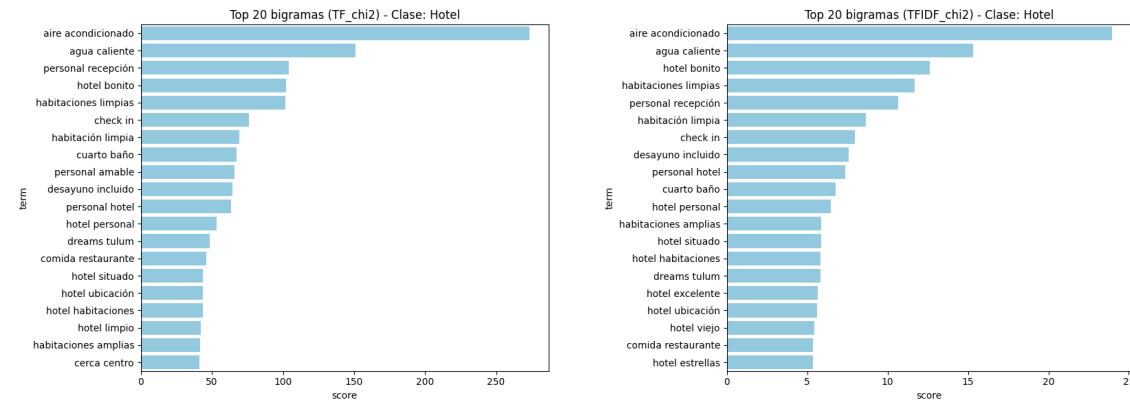
Nuestro código hace lo siguiente:

- **Construcción de representaciones BoW con bigramas** → mediante los vectorizadores CountVectorizer y TfidfVectorizer configurados con ngram\_range=(2, 2), se generan matrices dispersas donde cada dimensión corresponde a un **bigrama** (par de palabras consecutivas). Se consideran dos variantes: **TF** (frecuencia bruta de bigramas) y **TF-IDF** (frecuencia-inversa de documento de bigramas).

- **Selección de características por clase (one-vs-rest)** → para cada clase de la columna especificada en el dataset, se codifican etiquetas binarias y se aplican las siguientes métricas estadísticas sobre los bigramas:
  - **Ji-cuadrada ( $\chi^2$ )** → implementada con `chi2` de `sklearn.feature_selection`, evalúa la dependencia entre la presencia de un bigrama y la pertenencia a una clase.
  - **Información Mutua (MI)** → implementada con `mutual_info_classif`, mide cuánto reduce la incertidumbre sobre la clase la observación de un bigrama específico.
  - **Information Gain (IG)** → implementada manualmente en la función `compute_ig`, cuantifica la reducción de entropía de la distribución de clases al condicionar por la presencia o ausencia de un bigrama.
- **Ranking de bigramas relevantes** → los resultados de cada métrica se ordenan de mayor a menor, y se exportan en archivos `.csv`, mostrando los top-n bigramas más discriminativos para cada clase y representación (TF y TF-IDF).
- **Visualización** → si se activa la opción `-plots`, se generan gráficas de barras (`plot_top`) con los bigramas más relevantes por clase y métrica, guardadas como imágenes `.png`.
- **Salida final** → en la carpeta indicada por `-save-dir` se almacenan todas las tablas y gráficos generados, organizados por métrica, representación y clase. La carpeta por defecto es `out_bigrams`.

### 0.13.1. Resultados

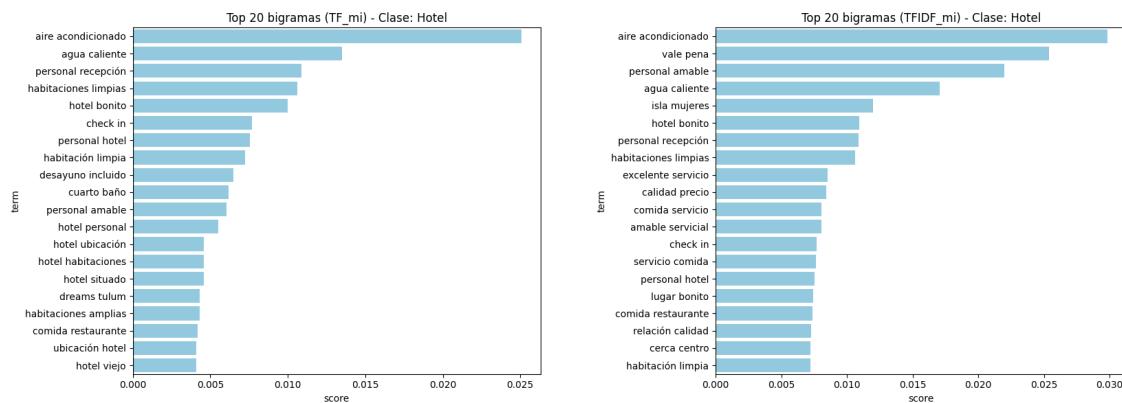
#### Hotel



(a) Top 20 bigramas TF por  $\chi^2$  para la clase Hotel.

(b) Top 20 bigramas TF-IDF por  $\chi^2$  para la clase Hotel.

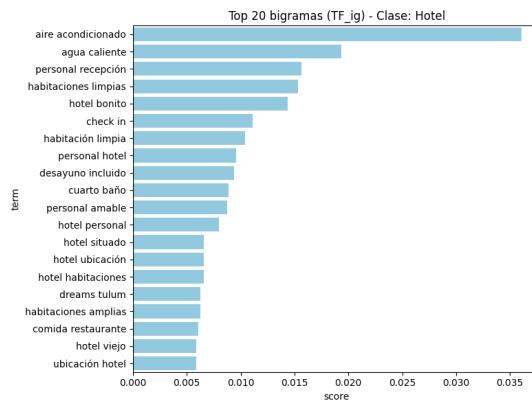
**Ji-cuadrada  $\chi^2$ :** Para el caso de los bigramas, nos encontramos con que “aire acondicionado” fue la cadena ancla que aparece tanto para TF como para TF-IDF. En general, los primeros cinco puestos del ranking fueron los mismos bigramas, aunque con un orden ligeramente distinto. Esto refleja que, al utilizar el contexto inmediato, lo encontrado por TF es muy similar a lo encontrado por TF-IDF. Además, el solapamiento entre los Top-20 es alto y el orden relativo entre los términos coincidentes es muy similar.



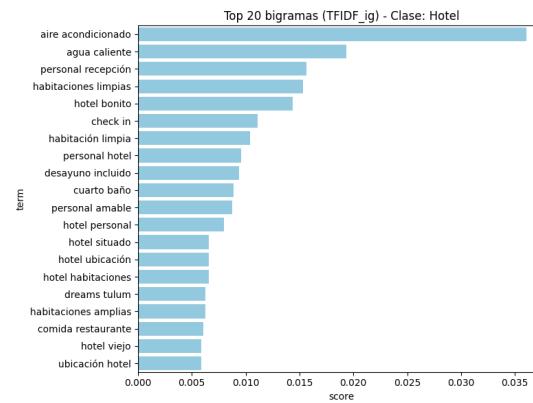
(a) Top 20 bigramas TF por Información Mutua (IM) para la clase Hotel.

(b) Top 20 bigramas TF-IDF por Información Mutua (IM) para la clase Hotel.

**Información Mutua (IM):** Para el caso de *Mutual Information*, aunque “aire acondicionado” conserva el primer lugar tanto para TF como para TF-IDF, el solapamiento entre los top 20 del ranking se reduce drásticamente a comparación del caso previo. El orden entre los términos compartidos muestra solo una concordancia moderada.



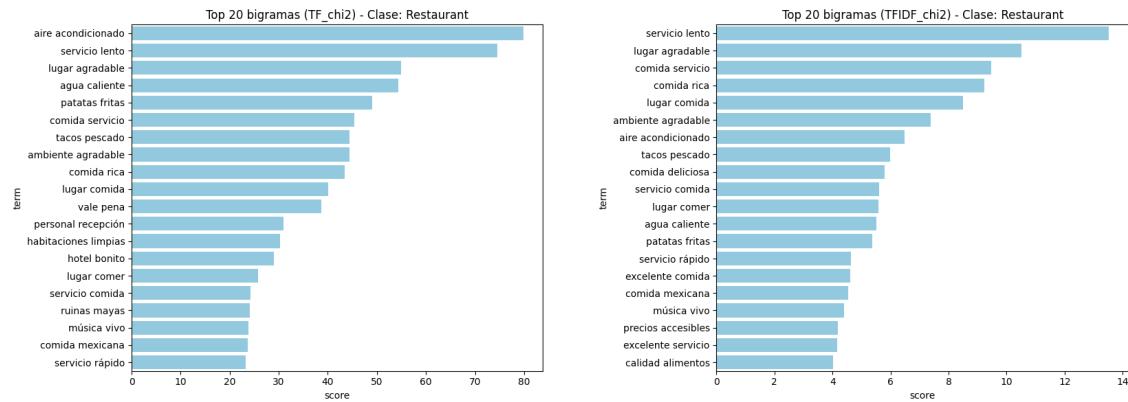
(a) Top 20 bigramas TF por Information Gain (IG) para la clase Hotel.



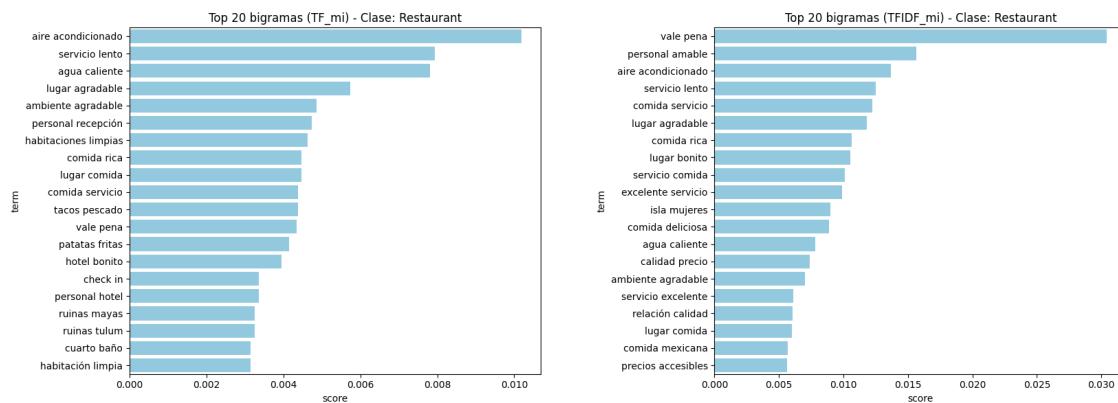
(b) Top 20 bigramas TF-IDF por Information Gain (IG) para la clase Hotel.

**Information Gain (IG):** En Information Gain los resultados con TF y con TF-IDF son exactamente iguales: el Top-20 coincide por completo, el orden relativo es idéntico y los puntajes de cada bigrama son numéricamente iguales. Esto es coherente con la invarianza de IG frente a la ponderación (Sec. 0.13): al basarse en la presencia/ausencia del término, IG no depende de los pesos TF o TF-IDF. De nuevo, “aire acondicionado” encabeza el ranking.

## Restaurante

(a) Top 20 bigramas TF por  $\chi^2$  para la clase Restaurante.(b) Top 20 bigramas TF-IDF por  $\chi^2$  para la clase Restaurante.

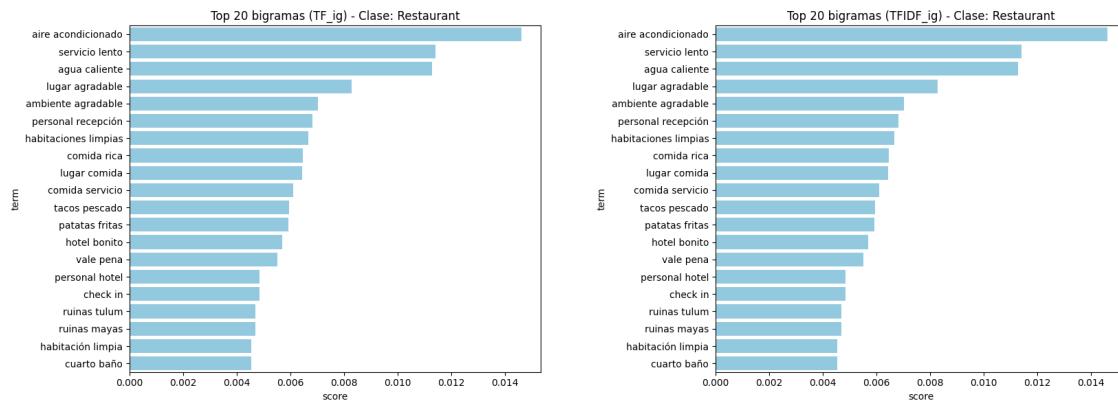
**Ji-cuadrada  $\chi^2$ :** Para Restaurante, la selección por  $\chi^2$  muestra un solapamiento alto entre TF y TF-IDF, aunque con reordenamientos moderados. TF resalta bigramas frecuentes y a veces transversales a varias clases, como fórmulas de servicio genéricas, mientras que TF-IDF penaliza esos patrones y favorece combinaciones más específicas de la experiencia gastronómica, como “servicio rápido” que gana posiciones. En conjunto, los bigramas aportan contexto y la ponderación TF-IDF afina la discriminación semántica removiendo ruido global.



(a) Top 20 bigramas TF por Información Mutua (IM) para la clase Restaurante.

(b) Top 20 bigramas TF-IDF por Información Mutua (IM) para la clase Restaurante.

**Información Mutua (IM):** En Información Mutua para Restaurante, el solapamiento entre TF y TF-IDF cae y el orden relativo casi no concuerda. Esto confirma que MI es especialmente sensible a la ponderación: TF-IDF atenúa bigramas genéricos o transversales, como “servicio rápido” y promueve aquellos cuya presencia es más informativa para la clase como lo es “mala atención”. En suma, con bigramas, TF-IDF afina la señal semántica y MI la refleja con cambios notables en el ranking.



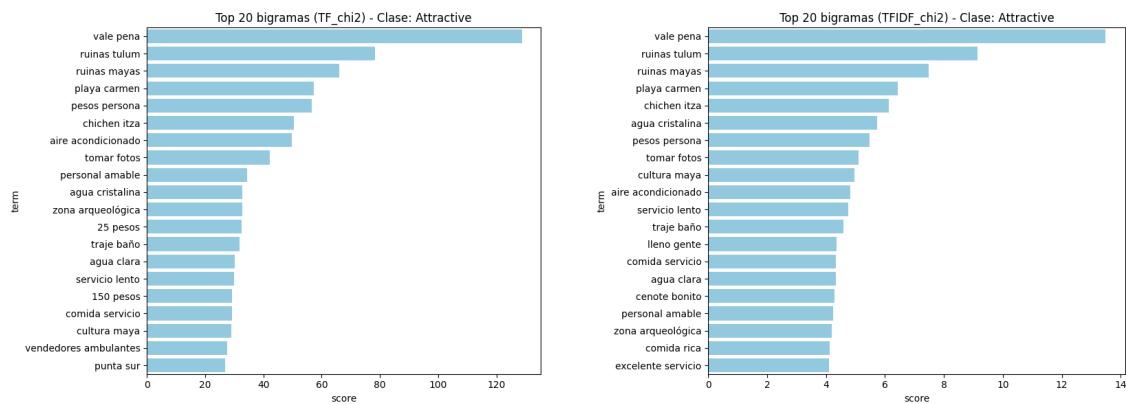
(a) Top 20 bigramas TF por Information Gain (IG) para la clase Restaurante.

(b) Top 20 bigramas TF-IDF por Information Gain (IG) para la clase Restaurante.

**Information Gain (IG):** En Information Gain los resultados con TF y con TF-IDF son exactamente iguales: el Top-20 coincide por completo, el orden relativo es idéntico y los puntajes de cada bigrama son numéricamente iguales. Esto es coherente con la invarianza de IG frente a la ponderación (Sec. 0.13): al basarse en la presencia/ausencia del término, IG no depende de los pesos TF o TF-IDF.

## Atractivo

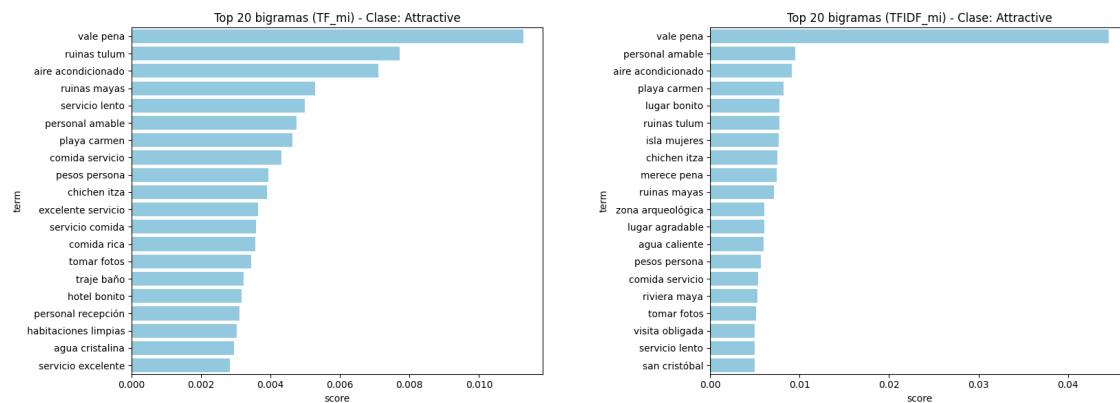
### Atractivo



(a) Top 20 bigramas TF por  $\chi^2$  para la clase Atractivo.

(b) Top 20 bigramas TF-IDF por  $\chi^2$  para la clase Atractivo.

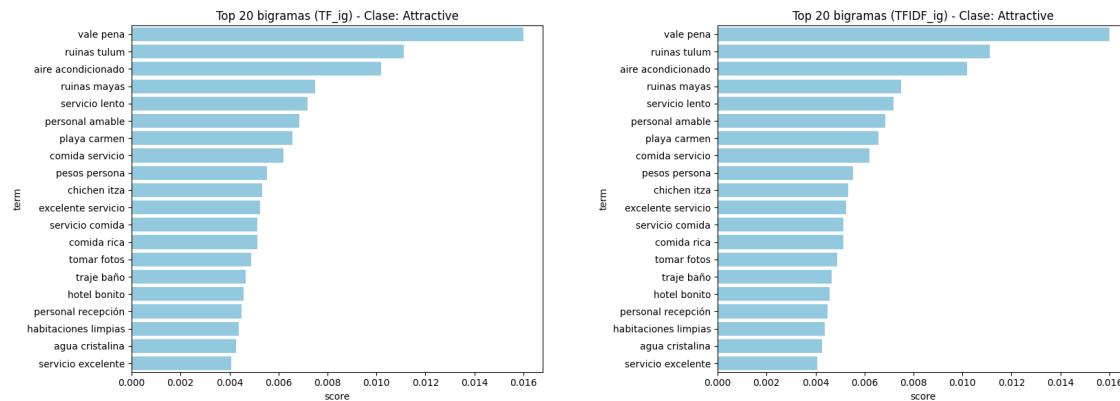
**Ji-cuadrada  $\chi^2$ :** En el caso de Atractivo, la selección realizada por Ji-cuadrada con bigramas presenta un solapamiento alto entre TF y TF-IDF, y una concordancia de orden notable. Los bigramas temáticamente puros como “ruinas mayas” se mantienen en los primeros lugares en ambas ponderaciones, mientras que las combinaciones más genéricas como “zona arqueológica” o “sitio arqueológico” tienden a perder posiciones bajo la estructura de TF-IDF.



(a) Top 20 bigramas TF por Información Mutua (IM) para la clase Atractivo.

(b) Top 20 bigramas TF-IDF por Información Mutua (IM) para la clase Atractivo.

**Información Mutua (IM):** En Información Mutua para Atractivo, el solapamiento entre TF y TF-IDF es más bien moderado y el orden relativo coincide solo parcialmente. TF-IDF promueve bigramas que condensan información temática, como “zona arqueológica”; y modera aquellos de uso amplio o poco distintivo. En consecuencia, con bigramas, MI refleja con claridad el sesgo de especificidad de TF-IDF, mientras que TF tiende a priorizar combinaciones frecuentes aunque menos exclusivas de la clase.



(a) Top 20 bigramas TF por Information Gain (IG) para la clase Atractivo.

(b) Top 20 bigramas TF-IDF por Information Gain (IG) para la clase Atractivo.

**Information Gain (IG):** En Information Gain los resultados con TF y con TF-IDF son exactamente iguales: el Top-20 coincide por completo, el orden relativo es idéntico y los puntajes de cada bigrama son numéricamente iguales. Esto es coherente con la invariancia de IG frente a la ponderación (Sec. 0.13): al basarse en la presencia/ausencia del término, IG no depende de los pesos TF o TF-IDF.

### 0.13.2. Aportación semántica de los bigramas

Lo primero que me gustaría discutir es lo que entendemos por semántica. Cuando se habla de semántica, se habla de capturar el significado y relaciones que van más allá de la ocurrencia aislada de palabras. De manera formal, queremos n-gramas que presenten composición de sentido y que resulten una herramienta discriminante para la clase de objetivos que trabajemos, siempre bajo las métricas definidas, como TF, TF-IDF,  $\chi^2$ , etc.

Si lo describimos de manera más práctica, un bigrama aporta semántica adicional cuando logra codificar la polaridad que un unígrafo no alcanza a discutir. Por ejemplo, no es lo mismo quedarse solo con “caro”, que con “poco caro” o “muy caro”. Además, se reduce la ambigüedad, pues cuando en un 1-grama tenemos solo la palabra “agua”, en un bigrama podemos tener “agua caliente”.

Este enfoque es la manifestación más directa del famoso principio del lingüista J.R. Firth: “*You shall know a word by the company it keeps*”. Los bigramas capturan la “compañía” inmediata de una palabra. Sin embargo, este método se limita a los vecinos adyacentes. Esto nos lleva a una pregunta natural: ¿podemos modelar esta “compañía” de una forma más profunda y generalizada? La respuesta a esta pregunta da paso a la siguiente generación de representaciones textuales.

## Ejercicio #7 | Word2Vec y analogías

- Entrena un modelo Word2Vec sobre el corpus.
- Realiza al menos cinco analogías interesantes y discute resultados.

### 0.13.3. Distancia coseno

Llevando el principio de Firth a un nivel superior de abstracción, surgen modelos como Word2Vec, que aprenden representaciones distribucionales de palabras en un espacio vectorial de baja dimensión. En lugar de simplemente contar co-ocurrencias locales como los n-gramas, Word2Vec trabaja sobre la Hipótesis Distribucional (Harris, 1954; Firth, 1957) de una manera más sofisticada.

La idea central sigue siendo la misma, el significado de una palabra puede inferirse por los contextos en los que aparece, pero ahora, el "contexto" o la "compañía" de una palabra se captura en un vector denso que codifica relaciones semánticas complejas con cientos de otras palabras del vocabulario. Cuando trabajamos con Word2Vec, cada palabra se representa como un vector denso en  $\mathbb{R}^d$ . La geometría de este espacio captura relaciones semánticas y sintácticas. De ese modo, palabras similares estarán cerca, distancia que suele medirse con la distancia coseno.

La distancia coseno es la métrica más común para evaluar similitud entre vectores de palabras en Word2Vec. Si  $u, v \in \mathbb{R}^d$  son embeddings de dos palabras, la similitud coseno se define como:

$$\cos(\theta) = \frac{u \cdot v}{\|u\| \|v\|} \quad \in [-1, 1]$$

Si  $\cos(\theta) \approx 1$ , entonces los vectores apuntan a la misma dirección, i.e. las palabras son semánticamente similares. Si  $\cos(\theta) \approx 0$ , entonces los embeddings son ortogonales, no hay relación aparente. Así, si  $\cos(\theta) \approx -1$ , entonces los vectores apuntan a direcciones opuestas, i.e. el modelo los detecta como antónimos o definidos bajo contextos contrastantes.

Así, las relaciones se preservan en forma de operaciones vectoriales. Uno de los ejemplos más clásicos al respecto es:

$$\text{vec("rey")} - \text{vec("hombre")} + \text{vec("mujer")} \approx \text{vec("reina")}$$

Jurafsky menciona que Word2Vec se entrena como una red neuronal. Para ello se apoya de *Continuous Bag of Words*, predice la palabra central  $w_t$  a partir del contexto; y de

*Skip-gram*, predice palabras del contexto a partir de la central  $w_t$  (hay una estructura de autoencoder ahí). El modelo se entrena para maximizar la log-verosimilitud. Así, el modelo en esencia aprende una matriz de embeddings que aproxima la co-ocurrencia de palabras en contexto.

#### 0.13.4. Analogías Semánticas

El objetivo es encontrar una palabra  $w_4$  que complete la analogía " $w_1$  es a  $w_2$  como  $w_3$  es a  $w_4$ ". Esto se traduce en la siguiente operación vectorial, donde  $w$  denota el vector de embedding para la palabra  $w$ :

$$w_2 - w_1 \approx w_4 - w_3$$

Despejando para el vector objetivo  $w_4$ , obtenemos la expresión utilizada para los cálculos:

$$w_4 \approx w_3 - w_1 + w_2$$

A continuación, se presentan los resultados de varias de estas operaciones. Para cada una, se calcula el vector resultante y se buscan las palabras en el vocabulario cuyos vectores tengan la mayor similitud de coseno con dicho resultado.

#### 0.13.5. Resultados

##### 1. Operación: horrible – comida + deliciosa

- rica: 0.8328
- rico: 0.7926
- volveremos: 0.7856
- ambiente: 0.7845
- auta: 0.7840

##### 2. Operación: caro – precio + barato

- razonable: 0.7808
- comparado: 0.7794
- accesible: 0.7692
- alto: 0.7552

- lugares: 0.7542

### 3. Operación: malo – servicio + excelente

- sa°per: 0.8547
- impecable: 0.8454
- gracias: 0.8435
- relajado: 0.8367
- rica: 0.8356

### 4. Operación: ciudad – tulum + playa

- arena: 0.7776
- bucear: 0.7741
- pasamos: 0.7707
- privada: 0.7642
- complejo: 0.7583

### 5. Operación: antiguo – queretaro + moderno

- **Error:** La palabra 'queretaro' no fue encontrada en el vocabulario del modelo (Out Of Vocabulary - OOV).

Los resultados obtenidos con Word2Vec validan de gran manera la capacidad de este tipo de embedding para capturar relaciones semánticas complejas a través de operaciones vectoriales. La efectividad de este enfoque no reside en una compresión lógica del lenguaje, sino en las propiedades geométricas del espacio vectorial semántico aprendido a partir de correlaciones estadísticas en el corpus que se utilizó para el procesamiento.

Así, la operación ciudad – tulum + playa sirve como un excelente caso de estudio. En este caso, el vector de desplazamiento (playa – tulum) puede interpretarse como la dirección semántica que encapsula el concepto de “playa” en el contexto de “Tulum”. Entonces, al aplicar este mismo vector al punto que representa “ciudad”, se busca un término análogo. Lo que obtuvimos fue “arena”, término que concentra la mayor similitud coseno con 0.77.

El modelo no devuelve otra ciudad costera, sino el componente fundamental y estadísticamente asociado a la palabra “playa”.

Sin embargo, es crucial nota como la “calidad” del resultado depende de las relaciones aprendidas. Mientras que las analogías de atributos como “precio” (“barato” vs. “caro”) y “servicio” (“excelente” vs. “malo”) fucnionan bien, otras operaciones arrojan resultados menos directos, aunque también útiles para el estudio.

Por ejemplo, la operación barato – hotel + caro no devuelve un sinónimo de “barato”, sino términos como “limpieza”, “instalaciones” y “servicios”. Esto nos podría estar indicando que, en nuestro corpues, el eje semántico “caro-barato” para hoteles está fuertemente correlacionado con discusiones sobre la calidad de las instalaciones, siendo una preocupación fundamental para los alojamientos. Es decir, se discute si el precio y su accesibilidad está relacionada con su limpieza o presentación.

De igual forma, el último ejemplo es interesante, pues nos encontramos con un error tipo *Out-Of-Vocabulary* (OOV) para “queretaro” que demuestra una limitante en la práctica: palabras que no están dentro del vocabulario. Esta restricción demuestra que no se puede hacer ninguna asociación semántica válida ya que es una palabra sin peso, no existe en el corpus por lo que no se puede operar con ella.

## Ejercicio #8 | Embeddings de documento y clusterización

- Calcula embeddings de documentos como el promedio de Word2Vec.
- Aplica K-means con  $k = 5$
- Reporta los cinco textos más cercanos al centroide de cada clúster.
- Discute si los clústers se alinean con las etiquetas originales.

### 0.13.6. Word2Vec Promediado

Cuando queremos construir un embedding que represente el contenido de un documento completo, una de las formas más simples a las que podemos recurrir es el promedio de los vectores de las palabras que lo conforman. A esta forma de trabajo se le conoce como average embedding o “embedding promedio”. La construcción matemática es la siguiente.

Sea un documento  $D$  con palabras  $w_1, w_2, \dots, w_n$ , Word2Vec asigna a cada palabra  $w_i$  un vector  $v(w_i) \in \mathbb{R}^d$ . El embedding del documento se define como:

$$v(D) = \frac{1}{n} \sum_{i=1}^n v(w_i)$$

Es decir, se suman todos los vectores de las palabras del documento, y se divide la suma entre el número total de palabras, normalizando así su longitud. En ciertos casos, como en TF-IDF, se puede dar más importancia a ciertas palabras si se utilizan pesos:

$$v(D) = \frac{1}{\sum_{i=1}^n \alpha_i} \sum_{i=1}^n \alpha_i \cdot v(w_i)$$

En este caso, la variable  $\alpha_i$  corresponde al peso de la palabra  $w_i$ .

Detrás de todo el formalismo anterior, lo que está sucediendo es que Word2Vec organiza palabras de forma que los términos con significados similares tienen vectores cercanos. Ahora, al promediar, estamos creando un centroide cargado de semántica de todas las palabras del documento. Esto funciona como una especie de resumen, si un documento menciona mucho la palabra “fútbol”, “portero”, “balón”, “referí”, su vector promedio seguro ubicará dicho documento en un tópico de Fútbol o Deportes, dentro del espacio de Word2Vec.

Sin embargo, este enfoque tiene ciertas limitaciones. La primera de ellas es que se pierde el orden de las palabras, i.e. no distingue entre “medio pollo frío” y “pollo medio frío”. Además, palabras muy frecuentes y poco informativas pueden distorsionar el embedding promediado si no se procesan.

### 0.13.7. Resultados

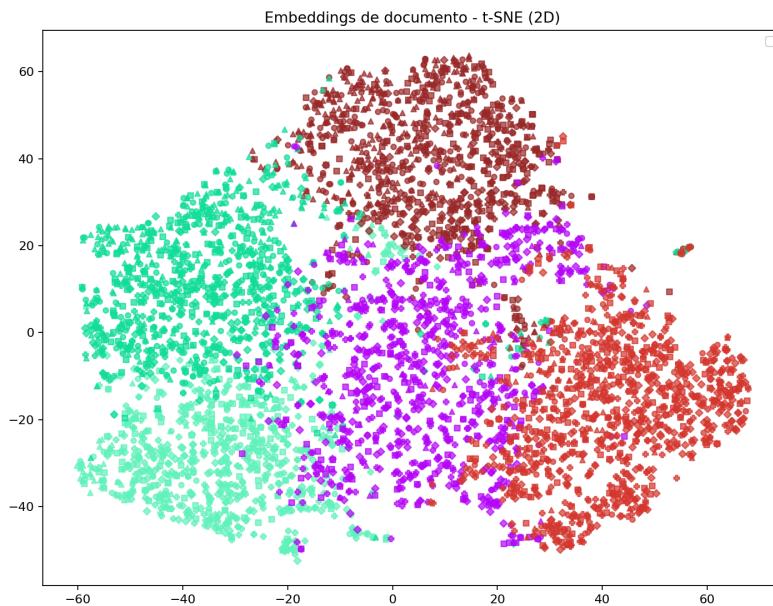


Figura 29: K-means sobre la columna Polarity del corpus proyectado con t-sne.

**Polarity** En esta sección se reportan los cinco documentos más cercanos al centroide de cada clúster, junto con la distribución de etiquetas originales de Polarity. Esto permite observar cómo se separan las reseñas positivas (valores altos) y negativas (valores bajos) en el espacio de embeddings.

## Clúster 0

### Distribución de Polarity:

- 5.0: 287
- 4.0: 195

- 3.0: 162
- 2.0: 104
- 1.0: 90

**Documentos representativos:**

1. Comida abundante, tortillas a mano, ambiente agradable y familiar. Recomiendo mejorar la calidad en el servicio fuimos en Domingo y deja mucho que decir; pasa tiempo y nadie te hace caso.
2. El lugar está céntrico y bonito. La comida de muy buen sabor. Probamos el chile en nogada que nos pareció muy bueno y la paella está buena. Lamentablemente el servicio me pareció un poco deficiente pero no desanimo ...
3. El servicio, desde que llegas es de primera. La comida es deliciosa, todas las personas que te atienden hacen que tu visita sea placentera. Tienen un espacio para niños funcional. La carta de vinos es excelente. En gene...
4. La comida exquisita, la decoración bellísima, el cambio de iluminación a partir de las 18:30 le da otra sensación, el servicio atentísimo sin caer en excesos y los precios justos. ¡Excelente opción para una delici...
5. El restaurante se ve padre; la carta crea expectativas altas pero la comida no está tan rica y el precio es elevado si se compara con otros restaurantes. La atención de la mesera dejó mucho que desechar, no nos ponía ate...

## Clúster 1

**Distribución de Polarity:**

- 2.0: 320
- 1.0: 293
- 3.0: 220
- 4.0: 148
- 5.0: 75

**Documentos representativos:**

1. Llegamos a este hotelito porque nos pareció lindo, de hecho, el exterior del hotel está muy coqueto. El problema son las habitaciones, nunca en toda mi historia de viajes, me había tocado dormir en una cama tan terri...
2. Me quedé en este hotel nueve días con esposo y mascota (labrador negro, cinco años, excellentemente bien educado) y realmente creo que pudimos haber sido más felices en otro espacio. La atención de la gerencia y per...
3. Fue buena estancia, el hotel está limpio y cómodo, las habitaciones son espaciosas y por lo menos las de hasta atrás del hotel tienen una vista impresionante del peñón. Hubo cosas que me molestaron en mi estancia, ...
4. Una decepción. Leyendo comentarios pensé que iba a un lugar confortable. No buscaba lujo, por supuesto, pero sí algo habitable. Dejé el hotel una noche antes aunque la estaba abonada porque nos morímos de frío! Habita...
5. En realidad presume de ser uno de los hoteles más tradicionales de la zona, de mayor calidad y servicio, en realidad no es así, no niego que es pintoresco, y su antigüedad te remonta al pasado, a quienes nos gusta eso ...

## Clúster 2

### Distribución de Polarity:

- 1.0: 254
- 2.0: 248
- 3.0: 199
- 5.0: 155
- 4.0: 136

### Documentos representativos:

1. Fuimos a Tango basado en las críticas y no falla. Nunca habíamos estado en Ajijic antes y ojalá hubiéramos tenido más tiempo para explorar. Había cuatro de nosotros para la cena y nos recibieron con una señorita m...
2. Lo siento por hacer una queja entre todas las buenas críticas, pero tuvimos una manera otra experiencia de lo que es otro lugar descrito. De todas las buenas críticas, decidimos tener nuestra cena navideña en Lola V...

3. La comida estaba bien, parte todo correcto. Pero anuncian “menú del día” fuera, en cuanto te sientas te dicen hay no. No te tratan a los extranjeros como los locales. El camarero parecía estar demasiado ocupado y no e...
4. No estoy seguro de todo el bombo! Llegamos alrededor de las 23:00 un domingo, comparándola el lugar estaba lleno! Le pedí al camarero si hay una mesa para 4 personas, él groseramente respondió “tienes...
5. Nos recomendaron este restaurante por alguien viviendo en la isla y no nos decepcionó. Como íbamos a hacer nuestro asiento fuera, un hombre que estaba dejando recomendó el pargo relleno, así que fue una de nuestras ...

## Clúster 3

### Distribución de Polarity:

- 5.0: 354
- 4.0: 282
- 3.0: 170
- 2.0: 108
- 1.0: 68

### Documentos representativos:

1. Es el lugar perfecto! Se encuentra en medio de la naturaleza, con habitaciones y cabañas fabricadas con bambú, todo muy limpio, el personal es amable y siempre está al pendiente de todo, la comida es deliciosa, rec...
2. Es un muy buen lugar para hospedaje por trabajo. El hotel cuenta con espacios para reuniones y tiene los servicios adecuados para realizar trabajo en aula o en espacios al aire libre. La comida es rica, aunque podrían...
3. Durante nuestro viaje ida y vuelta que se mantuvo por 5 noches en el hotel romántico Santo Domingo resultó ser un oasis en Yucatán. Lo recomendaría a todo el mundo. Izamal es un pu...
4. Estuve aquí con un grupo el pasado mes de marzo y los 20 de nosotros entramos en el espacio, todos nos aventuramos con los ojos bien abiertos, teniendo en la espectacular decoración y jardines. Velas flotantes en una plan...

5. Dreams Tulum es enorme, pero funciona. El salón VIP es una gran ventaja allí, así que reserve en consecuencia. El barman, Epitacio, en el salón es un día radiante y estupenda persona. La playa es una ...

## Clúster 4

### Distribución de Polarity:

- 4.0: 339
- 5.0: 329
- 3.0: 249
- 2.0: 120
- 1.0: 95

### Documentos representativos:

1. Si has estado en Chichen Itza, entonces no te quedarás impresionado. La única ventaja verdadera es la vista al mar (puede ser hecho en otro lugar) y clase adicional de historia (que es interesante, pero se pueden leer ...)
2. Luego de visitar las ruinas de Tulum decidimos emprender el camino hasta Playa Paraíso. Son más de 2km de caminata o puedes tomar un taxi (\$mex 70). El lugar es paradisíaco solo que bastante pequeño para nuestro gus...
3. Si has visitado Palenque, Chichen Itza o Tehotihuacán y estás esperando lo mismo de Tulum, lamento decirte que no lo tendrás. La vista al mar y la brisa fresca son lo único que vale la pena del lugar. Las ruinas están ...
4. Eso que fui en temporada baja y aún así la cantidad de turistas hace que pierda el encanto. De todas maneras al ser las únicas ruinas mayas en el mar pues... mantiene su atractivo.
5. Dos advertencias: todo el sitio es significado en monocultivos que visité en el verano, que es como una sauna húmeda con un fuerte sol que azotaba sin cesar. No recomendaría una visita en verano para niños, ancianos, ...

## 0.14. Análisis

Para el caso de Polarity, las etiquetas en general mantienen un equilibrio entre reseñas negativas y positivas.

### ■ Cluster 0

- Positivos:  $76.8\% \cdot \text{Media} \approx 3.58 \rightarrow \text{positivo}$
- Muestras: “comida”, “restaurante”, “servicio” → huele a Restaurantes.

### ■ Cluster 1

- Positivos:  $42.0\% \cdot \text{Media} \approx 2.42 \rightarrow \text{negativo}$
- Muestras: “hotel”, “habitaciones”, “cama”, “gerencia” → muy probablemente Hoteles con quejas.

### ■ Cluster 2

- Positivos:  $49.4\% \cdot \text{Media} \approx 2.69 \rightarrow \text{mixto/ligeramente negativo}$
- Muestras: “mesas”, “camarero”, “extranjeros”, “menú del día” → suena a Restaurantes con experien-

## Análisis de Clusters

cias encontradas.

### ■ Cluster 3

- Positivos:  $82.1\% \cdot \text{Media} \approx 3.76 \rightarrow \text{muy positivo}$
- Muestras: “hotel”, “habitaciones”, “cabañas”, “personal”, “playa” → Hoteles (experiencias muy buenas).

### ■ Cluster 4

- Positivos:  $81.0\% \cdot \text{Media} \approx 3.61 \rightarrow \text{positivo}$
- Muestras: “ruinas”, “Tulum”, “Playa Paraíso”, “vista al mar” → Atractivos turísticos / sitios arqueológicos y playas.

Entonces, por Polarity, mis clusters se separan de forma muy clara. Tenemos al Cluster 1 como negativo uno que concentra reseñas negativas, mientras que los Cluster 3 y Cluster 4 son muy positivos, mientras que Cluster 0 es positivo y Cluster 2 es más neutro. Sin embargo, creemos que una mejor clusterización sería la de Type . Si nos apoyamos con esa categoría, pensamos que podríamos hacer una separación más evidente. Veámoslo a continuación.

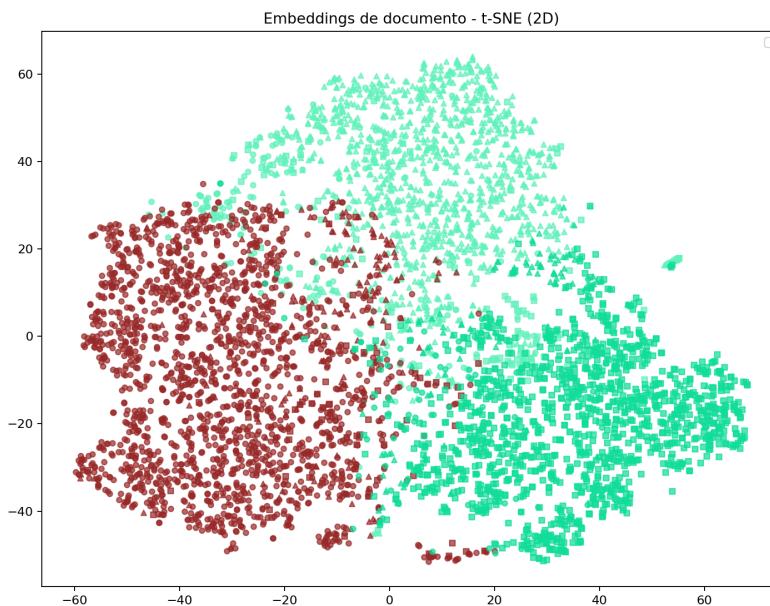


Figura 30: K-means sobre la columna Type del corpus proyectado con t-sne.

**Type** En esta sección se reportan los cinco documentos más cercanos al centroide de cada clúster, junto con la distribución de etiquetas originales (Type). Esto permite identificar qué tipo de reseñas (Hotel, Restaurante, Atractivo) dominan en cada grupo.

## Clúster 0

### Distribución de Type:

- Hotel: 1200
- Restaurant: 176
- Attractive: 118

### Documentos representativos:

1. Fue buena estancia, el hotel está limpio y cómodo, las habitaciones son espaciosas y por lo menos las de hasta atrás del hotel tienen una vista impresionante del peñón. Hubo cosas que me molestaron en mi estancia, ...
2. Este “hotel” es mitad hotel y mitad hostal. Por el precio, esperaba que todo estuviera impecable y que las áreas comunes fueran como un hotel y no como un albergue. Las habitaciones no estaban ordenadas, el inodoro ni s...

3. Obtuvimos lo que esperábamos y fuimos bien atendidos desde que hicimos la reservación. El lugar es grande y en medio de la naturaleza te permite caminar por senderos entre árboles y las dos albercas para refrescarse. ...
4. Una decepción. Leyendo comentarios pensé que iba a un lugar confortable. No buscaba lujo, por supuesto, pero sí algo habitable. Dejé el hotel una noche antes aunque la estaba abonada porque nos morimos de frío! Habita...
5. Nunca habíamos estado en un hotel resort, este fue un regalo al final de nuestro viaje y no sino a otros resorts! Una vez dicho esto, las habitaciones eran amplias, los jardines eran preciosos y el personal era muy amab...

## Clúster 1

### Distribución de Type:

- Restaurant: 1753
- Hotel: 187
- Attractive: 73

### Documentos representativos:

1. Me parece un lugar mágico y acogedor. Atendido por sus dueños, el chef Eugenio y su esposa que son muy amables y ofrecen un excelente servicio. La comida de primera tanto las pastas como la pizza, elaborada en horno d...
2. Un lugar agradable, si tienes suerte de que te toque el chef o ciertos cocineros la comida estará rica, hay unos cocineros suplentes en domingo que son muy malos. El servicio es muy lento, ve con tiempo. Los meseros distra...
3. El servicio era muy lento y la comida no era muy buena. Se veía hermosa pero no era fresco y le faltaba sabor. Fue muy decepcionante. Además, los insectos eran terribles. El único restaurante durante la semana donde l...
4. Una grata experiencia! La atención de Alberto marcó diferencia, elegimos comer en la terraza, muy agradable! La comida deliciosa (recomendación de Alberto). Si hay un punto de mejora es la carne, estaba un poco dura, ...
5. Es 100 % comida regional, el lugar es rústico, ni siquiera hay menú, pero todo lo que probé riquísimo y el servicio es cálido y familiar. El detallazo fue que como esa región es fría (íbamos a la playa en bikinis ...

## Clúster 2

### Distribución de Type:

- Attractive: 1261
- Hotel: 124
- Restaurant: 108

### Documentos representativos:

1. Luego de visitar las ruinas de Tulum decidimos emprender el camino hasta Playa Paraíso. Son más de 2km de caminata o puedes tomar un taxi (\$mex 70). El lugar es paradisíaco solo que bastante pequeño para nuestro gus...
2. Amplia playa llena de sargazo pudriéndose adentro del agua y en la playa. HORRIBLE. Si estás por ir a la Riviera Maya plantéatelo seriamente, ya no es el destino paradisíaco que en algún momento fue, las playas de toda ...
3. Un bello parque natural con unas cascadas muy bellas y amplio espacio para senderismo y diversas actividades. Muy recomendable para diversas personas e intereses, desde un paseo en familia, viaje en pareja o aventura in...
4. Sin duda es un lugar con una magia especial, es fascinante ver y escuchar la historia de una cultura prehispánica como la que aquí se desarrolló... Un imperdible de México... Una maravilla que atrapa... Se recomie...
5. Tengo 25 años hospedándome en Posada de la Misión y me ha sorprendido que descubrieron una mina prehispánica debajo del edificio principal que tiene como 80 años. Es impresionante ver cómo la beta está a menos de 15...

## 0.15. Análisis

Si bien los clústers muestran cierta separación en términos de polaridad, el análisis con la variable Type ofrece resultados mucho más claros: los grupos se alinean de manera natural con hoteles, restaurantes y atractivos turísticos. Esto evidencia que el promedio de embeddings de Word2Vec permite capturar no sólo la valoración positiva o negativa de una reseña, sino también su tema principal, logrando una clusterización semánticamente coherente.

## Ejercicio #9 | Clasificando con Partición 70/30

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

- Sin preprocessamiento.
- Con minúsculas.
- Con minúsculas y stemming/lematización.
- Con minúsculas, stemming y filtrando palabras con frecuencia mínima de 10.

Compara métricas como accuracy, F1-macro, matriz de confusión y discute si el preprocessamiento es importante.

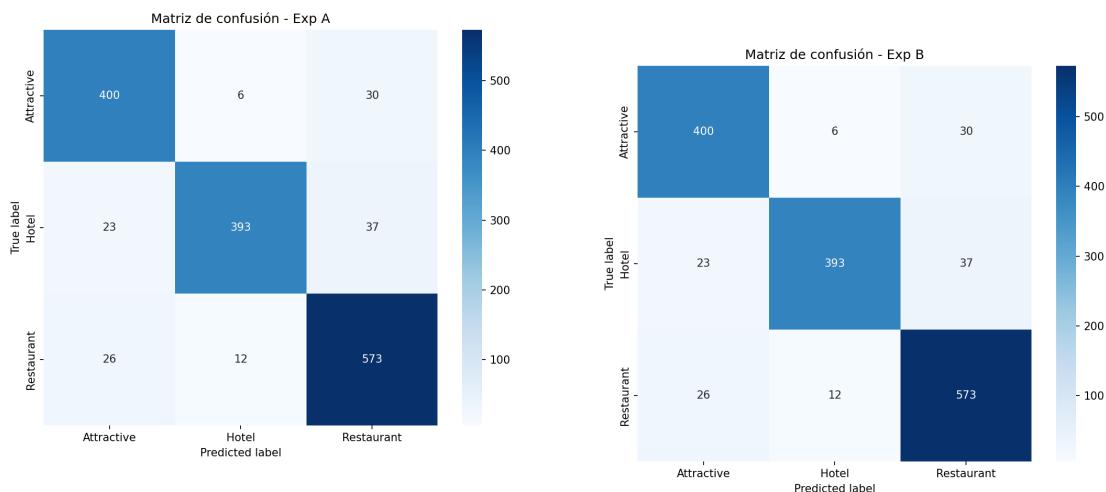
### 0.15.1. Resultados

Utilizando SVM, para los cuatro casos mencionados, nos encontramos con resultados que nos llevan a conclusiones interesantes. La primera de ellas es que nuestro corpus parece ser uno bastante limpio y separable, en el sentido de que las clases pueden separarse con bastante facilidad para la variable Type. Las matrices de confusión en la figura ?? refuerzan dicha hipótesis, pues en cada uno de nuestros cuatro casos se muestra un comportamiento bastante sólido con apenas unos cuantos errores.

Sumado a los resultados de las matrices de confusión, podemos agregar lo encontrado con las métricas de evaluación.

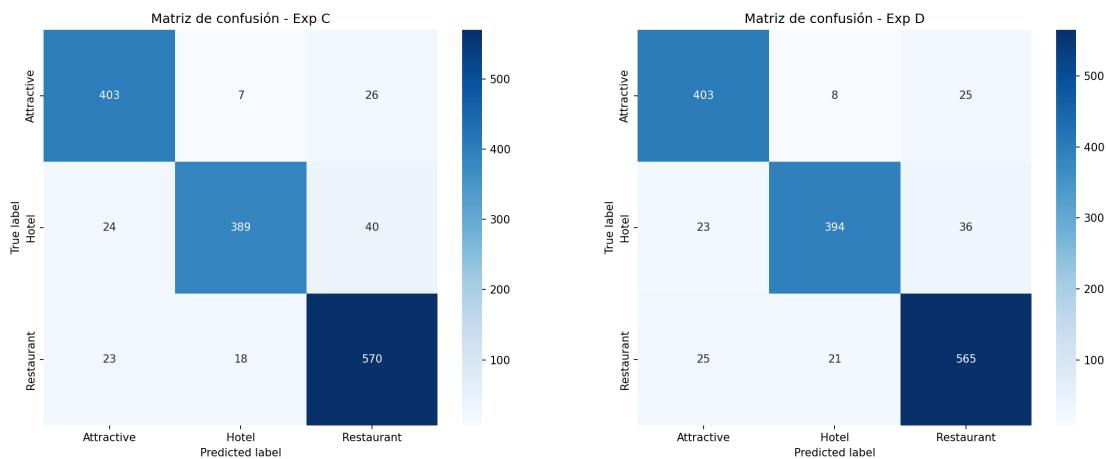
- **Experimento A (sin preprocessamiento):** Accuracy = 0.9107, F1-macro = 0.9099
- **Experimento B (minúsculas):** Accuracy = 0.9107, F1-macro = 0.9099
- **Experimento C (minúsculas + lematización):** Accuracy = 0.9080, F1-macro = 0.9071
- **Experimento D (minúsculas + lematización + filtrado min\_df=10):** Accuracy = 0.9080, F1-macro = 0.9072

Cada una de las métricas demuestra solidez al momento de hacer la clasificación de documentos por categoría (Hotel, Restaurante o Atractivo). Esto nos indica que, aún sin limpieza, el corpus es lo suficientemente bueno como para realizar clasificación de texto.



(a) Matriz de confusión (a) Sin preprocessamiento.

(b) Matriz de confusión (b) Con minúsculas.



(c) Matriz de confusión (c) Con minúsculas y stemming/lematización.

(d) Matriz de confusión (d) Con minúsculas, stemming/lematización y filtro de frecuencia.

Figura 31: Matrices de confusión de los cuatro experimentos acumulativos.

Además, al utilizar SVM, los pesos se pueden adaptar para darle menos valor a palabras que introducen ruido, como palabras con mayúsculas, acentos o uso poco común. Por eso, convertir a minúsculas o lematizar no cambia demasiado los resultados entre aproximaciones. El lenguaje utilizado en cada una de las categorías es lo suficientemente diverso y distinto entre ellos como para utilizarse de buena manera.

Sería interesante comprobar de manera sólida que las pequeñas diferencias entre accuracy y F1-macro no son estadísticamente relevantes. Pero supongo que solo con mencionarlo es suficiente.

## Ejercicio #10 | LSA con 50 tópicos

- 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.

### 0.16. Latent Semantic Analysis (LSA)

Cuando queremos representar un corpus como una matriz de documento-término, ya sea a través de TF o TF-IDF, tenemos lago de la siguiente manera:

$$X \in \mathbb{R}^{n \times m}$$

Entonces cada fila es un documento, y cada columna es un término. Sin embargo, este tipo de representaciones tienen un problema: son de alta dimensionalidad y presentan *sparsity*. De esa maenra, necesitamos realizar una proyección de documentos a menor dimensión, de tal modo que se preserve la información semántica para, posteriormente, encontrar los tópicos que queramos.

De esa forma, a través de la Descomposición en Valores Singulares, podemos encontrar una representación matemática para nuestra matriz  $X$ . SVD se presente así:

$$X = U\Sigma V^\top$$

Donde  $\Sigma$  contiene los valores singulares  $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r$ . Cada  $\sigma_i^2$  es proporcional a la varianza explicada por la  $i$ -ésima componente. Ahora, al truncar a  $k$  dimensiones, entonces nos quedamos con los  $k$  vectores que capturan mayor varianza de los datos. Justo esto es lo que hace PCA, encontrar direcciones que maximizan la varianza. De forma análoga, en el Análisis Semántico Latente (LSA), cada componente se interpreta como un "tópico" latente. Así, cada tópico representa una constelación de palabras que tienden a aparecer juntas, capturando la estructura semántica principal en los documentos.

De esa manera, si dos palabras nunca coocurren en un documento, entonces su similaridad es igual a 0. Pero si ambas aparecen en contextos similares, como "fútbol" y "portero", sus patrones de coocurrencia están correlacionados. Si después proyectamos con SVD, entonces estas correlaciones indirectas se reflejan en los vectores de  $V_k$ . El resultado será que

palabras sin coocurrencia directa terminan cerca en el espacio latente, siendo semánticamente relacionadas.

Ahora queremos saber cuáles de los 50 tópicos son más útiles para distinguir entre las clases objetivo. Para ello, podemos utilizar Ji-cuadrada ( $\chi^2$ ) como prueba de independencia o Información Mutua (MI). La primera toma cada columna de la matriz de documentos-tópico, se compara con la categoría y se calcula el estadístico, un valor grande nos dice que el tópico está asociado con la clase. Por otro lado, MI mide cuánto reduce la incertidumbre conocer un tópico sobre la clase.

También nos interesa la coherencia de tópicos, i.e qué tan interpretables son estos. Si un tópico es coherente, esperaríamos que las palabras principales estén semánticamente relacionadas, como: “péximo, caro, malo, servicio”. Un tópico incoherente sería algo que mezcla palabras sin relación clara: “legal, Pac-man, sushi, mezquita, llanta”.

Lo que haremos será lo siguiente:

- Aplicar LSA para 50 tópicos sobre la matriz TF-IDF.
- Cada documento se representa como una combinación de esos tópicos latentes.
- Se evaluará qué tan informativos son cada uno de los tópicos respecto a la variable de clase usando Mutual Information o Ji-cuadrada.

## 0.17. Resultados

El cuadro ?? muestra el top 15 de tópicos junto con las cinco primeras palabras por tópico. Podemos observar que tenemos los siguientes tópicos de más peso:

- T2 (MI0.448): hotel, habitación, habitaciones, personal, servicio → hospedaje muy marcado.
- T1 (MI0.343): comida, servicio, lugar, buena, excelente → restauración/servicio.
- T0 (MI0.096): comida, lugar, hotel, servicio, buena → mezcla restauración + hospedaje (más genérico).
- T3 (MI0.073): lugar, hotel, excelente, vale, pena → genérico con sesgo a hospedaje/atracción.
- T14 (MI0.052): cenote, agua, vista, restaurante, excelente → atracciones naturales (cenote/agua/vista).

Una de las primeras conclusiones que podemos observar es que hay tres dominios léxicos dominantes en los primeros lugares tras Mutual Information. Tenemos: “hospedaje” con términos de peso como hotel y habitación, tenemos “restaurante” con términos como comida y servicio, y “atracciones” con términos como cenote, agua o visita.

Tras este análisis rápido de MI, los tópicos más informativos están dominados por léxicos de hospedaje (T2: hotel/habitación/personal), restaurante/servicio (T1: comida/servicio/atención), y atracciones (T14: cenote/agua/vista). Algunos tópicos como T0–T3 muestran términos genéricos de alta frecuencia (lugar, excelente, vale, pena), por lo que los catalogamos como parcialmente coherentes. También se identifica un eje de quejas de servicio (términos como mala, mal, pésimo, lento), consistente con reseñas negativas. En conjunto, el CSV evidencia que los ejes latentes mejor rankeados por MI son semánticamente plausibles para el dominio.

Cuadro 6: Top 15 tópicos más informativos según Mutual Information (5 palabras por topical).

Tópico	Mutual Information	Palabras más representativas
0	0.096	comida, lugar, hotel, servicio, buena
1	0.343	comida, servicio, lugar, buena, excelente
2	0.448	hotel, habitación, habitaciones, personal, servicio
3	0.073	lugar, hotel, excelente, vale, pena
4	0.034	playa, excelente, buena, ambiente, agradable
5	0.026	servicio, playa, mala, mal, meseros
6	0.012	lugar, excelente, mejor, personal, playa
7	0.000	excelente, pena, vale, restaurante, personal
8	0.017	atención, excelente, buena, cenote, agua
9	0.024	buen, servicio, precio, bien, agradable
10	0.012	habitación, pena, vale, servicio, lugar
11	0.028	calidad, mejor, precio, atención, playa
12	0.008	cenote, personal, precio, mejor, comida
13	0.009	bien, habitación, comida, ruinas, excelente
14	0.052	cenote, agua, vista, restaurante, excelente

Podemos visualizar estos resultados también en un gráfico de barras como el que se muestra en la figura ??, con el top 10 de los tópicos.

Salta a la vista que T2 y T1 son los más informativos para nuestro caso. Creo que ambos representan, primero que nada a Hotel en el T2 (aparece el término “hotel”), y a Restaurante en el T1 (aparece el término “comida”). Creo que es bastante claro que la

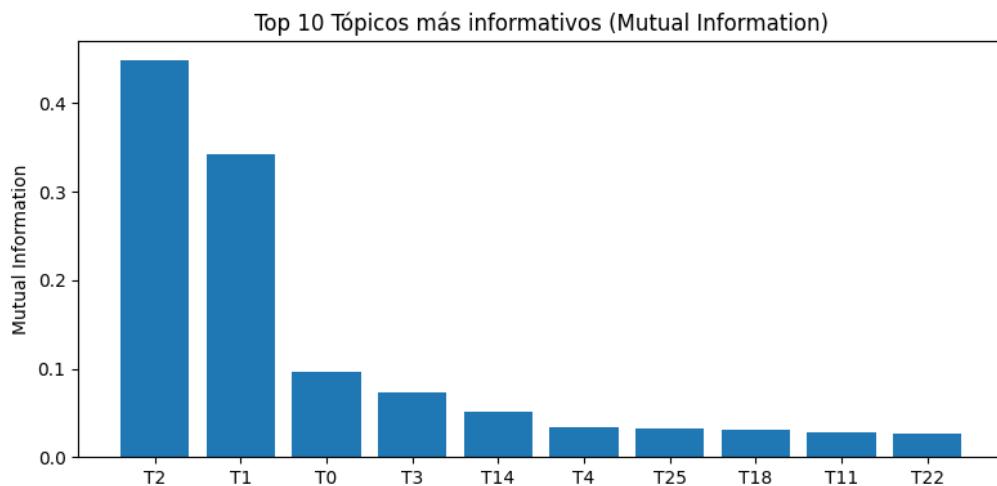


Figura 32: Top 10 de los tópicos.

categoría Attractive es mucho más complicada, pues esta es mucho más general, en ella tenemos cosas como playas o ruinas, además de otro tipo de lugares turísticos, por lo cual hay una mezcla bastante variada de ejemplos. Esto provoca que la categoría Attractive se vuelva un poco más compleja que las otras dos de representar.

Podemos continuar con el análisis a través de un heatmap, como el que se ve en la figura 33.

El heatmap de activación promedio muestra que los tópicos T2 y T1 se asocian diferencialmente con las clases Hotel y Restaurant, respectivamente, mientras que T0 aparece de manera transversal en todas las clases. El tópico T14, aunque semánticamente corresponde a atracciones (cenote, agua, vista), no mostró una activación clara en la clase Attractive, lo que podría indicar ruido o insuficiencia de datos. El resto de los tópicos presentan valores cercanos a cero, confirmando su baja capacidad discriminativa.

Finalmente, podemos ver con unos ejemplos de wordclouds las palabras que aparecen en varios de los tópicos.

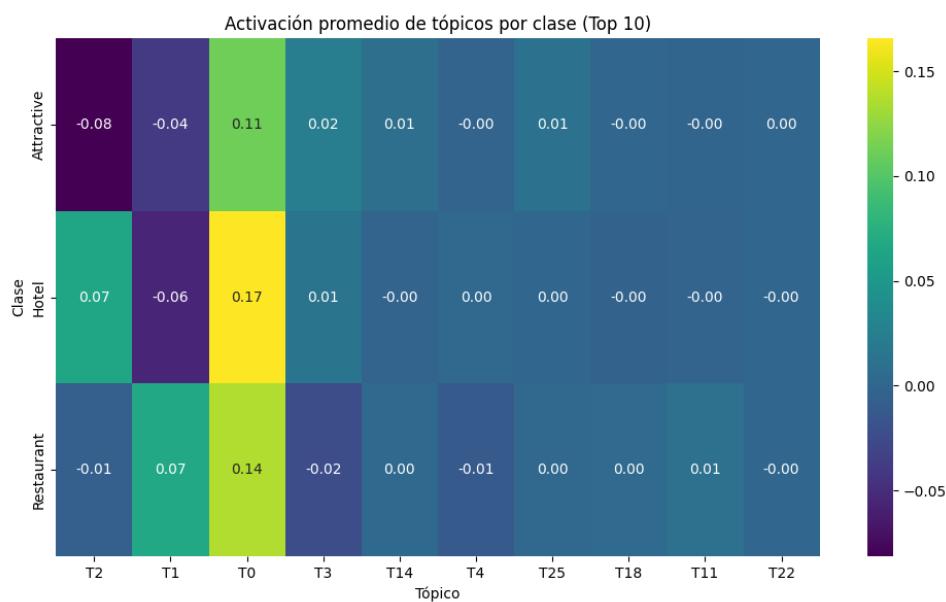
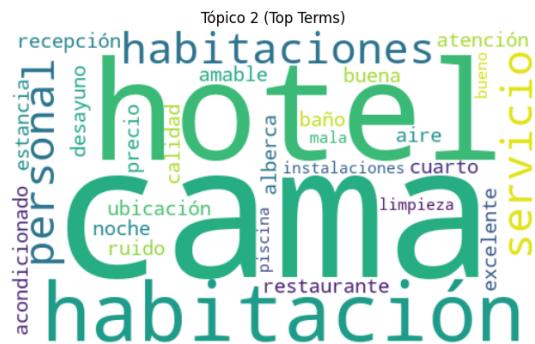


Figura 33: Heatmap de los tópicos.



(a) Wordcloud del Tópico 2.



(b) Wordcloud del Tópico 1.



### (c) Wordcloud del Tópico 0.



(d) Wordcloud del Tópico 3.

Figura 34: Wordclouds de los tópicos.