Proyecto Cómputo Estadístico

Modelado de Tópicos con Asignación Latente de Dirichlet (LDA) y Muestreo de Gibbs

Autores: David Muro Campa, Gabriel Alejandro Aguilar Farrera

Maestría en Cómputo Estadístico, CIMAT Sede Monterrey

1 de diciembre de 2023

Resumen: El proyecto se enfoca en el modelado de tópicos utilizando el método de Asignación Latente de Dirichlet (LDA) y Muestreo de Gibbs para analizar y extraer tópicos latentes de reseñas de aerolíneas. La base de datos utilizada es la colección de reseñas de Skytrax Airline Reviews a la cual se le ha aplicado un proceso de limpieza y preprocesamiento de texto para transformar las reseñas en un formato apto para el modelado. Se exploraron y compararon dos formatos de corpus diferentes: El modelo Tf-idf y la representación de Bolsa de palabras (Bag of Words), analizando cómo estas representaciones influyen en la modelación de los tópicos. Posteriormente, se hizo un análisis del número de tópicos con Escalamiento Multidimensional (MDS) para finalmente llevar a cabo una evaluación comparativa de algoritmos de clasificación con los tópicos generados, obteniendo mejores resultados con el algoritmo de clasificación de regresión logística.

1. Introducción

1.1. Antecedentes

Desde su publicación en 2003, Latent Dirichlet Allocation (LDA) de Blei *et al.* [1] ha convertido al modelado de tópicos, un subcampo del aprendizaje automático aplicado a todo, desde la lingüística computacional [2] hasta la bioinformática [3] y la ciencia política [5], en uno de los paradigmas más populares y exitosos tanto para el aprendizaje supervisado como no supervisado.

El análisis de sentimientos, en sus primeras etapas, se centraba principalmente en enfoques lingüísticos y psicológicos para comprender las emociones humanas a través del lenguaje. Sin embargo, con el advenimiento de la era digital y el surgimiento de las redes sociales, la cantidad masiva de datos generados por usuarios en línea llevó al desarrollo de métodos computacionales para analizar y comprender estas expresiones emocionales a gran escala.

En sus primeras fases, el análisis de sentimientos se centró en técnicas básicas de minería de texto, identificando palabras clave y patrones simples para clasificar el contenido como positivo, negativo o neutral. Estos métodos, aunque útiles, carecían de sofisticación para capturar matices emocionales o contextuales.

El surgimiento del aprendizaje automático y, específicamente, de los modelos basados en el procesamiento de lenguaje natural (NLP), marcó un hito crucial en el análisis de sentimientos. La

aplicación de algoritmos como el LDA (Latent Dirichlet Allocation) de Blei *et al.* en 2003, introdujo una forma más avanzada de comprender y clasificar el texto, permitiendo la identificación de tópicos latentes en grandes conjuntos de datos textuales.

2. Objetivos

- Implementar el método de Asignación Latente de Dirichlet con Muestreo de Gibbs
- Hacer el modelado de tópicos para la base de datos Skytrax Airline Reviews
- Establecer el número más adecuado de tópicos mediante Escalamiento Multidimensional
- Utilizar un modelo de clasificación con las representaciones dadas por LDA

3. Marco Teórico

3.1. Latent Dirichlet Allocation (LDA)

Formalmente, definimos los siguientes términos:

- Una *palabra* es la unidad básica de los datos. Cada palabra se representa con un vector que tiene un elemento igual a 1 y los demás igual a 0. Por ejemplo, sea una palabra la i-ésima del vocabulario, entonces el vector w tendrá valor de 1 en la i-ésima componente y 0 de todas las demás.
- Un documento es una secuencia de N palabras y se denota por $\mathbf{w} = (w_1, w_2, ..., w_N)$ donde w_n es la n-ésima plabara en la secuencia.
- Un corpus, es una colección de M documentos denotado por $\mathcal{D} = \mathbf{w}_1, .\mathbf{w}_2, ..., \mathbf{w}_M$

La Asignación Latente de Dirichlet es un modelo generativo probabilístico de un corpus. La idea básica es que los documentos están representados como mezclas aleatorias de tópicos latentes, donde cada tópico está caracterizado por una distribución de palabras.

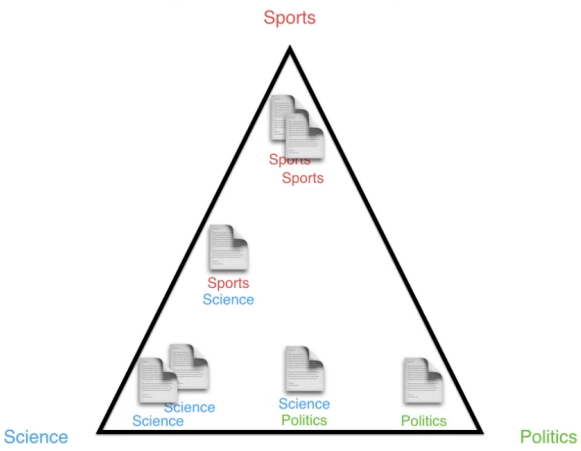


Figura 1: Asignación Latente de Dirichlet para el caso de un corpus de 7 documentos y 3 tópicos latentes.

Queremos un modelo que asigne una probabilidad de pertenecer a determinado tópico a cada elemento de un corpus (Figura 1).

LDA asume lo siguiente para cada documento $\mathbf{w}$ en un corpus $\mathcal{D}$

1. $N \sim$ Poisson(λ)

2. $\boldsymbol{\theta} \sim \text{Dir}(\boldsymbol{\alpha})$
3. Para cada una de las N palabras w_n:
    a) Elige un tópico $z_n \sim \text{Multinomial}(\boldsymbol{\theta})$
    b) Elige una palabra w_n de $p(w_n|z_n, \boldsymbol{\beta})$, una probabilidad multinomial condicionada en el tópico z_n

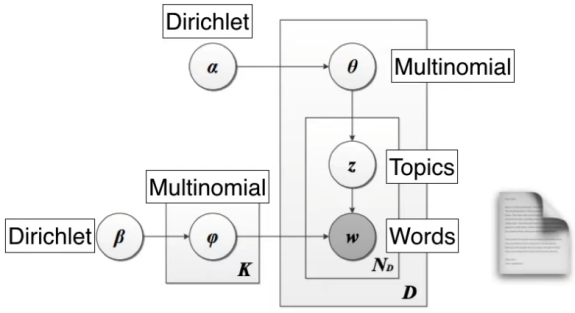


Figura 2: Supuestos de LDA gráficamente.

Estos supuestos se traducen en cómo asume LDA que se construyen los documentos de un corpus como se aprecia en la Figura (2).

Una variable aleatoria Dirichlet θ_d con parámetros $\alpha_1, ..., \alpha_k > 0$ tiene una densidad de probabilidad

$$p(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1} \tag{1}$$

El soporte de una distribución Dirichlet es el simplex estándar. Por ejemplo, en la siguiente Figura podemos apreciar como se ve la densidad para distintos parámetros en un 2-símplex

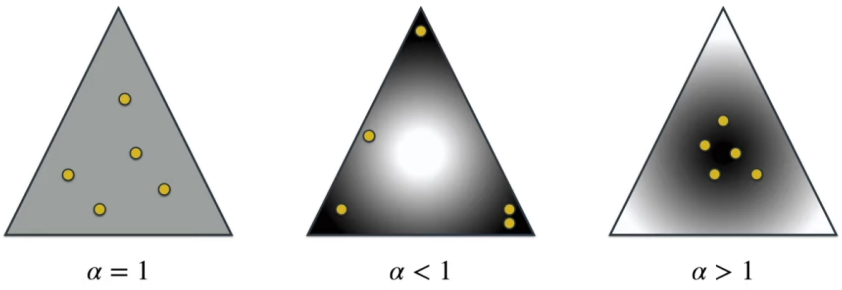


Figura 3: Densidad de probabilidad de la Distribución de Dirichlet para el caso de soporte en un 2-símplex.

Debido a esto la distribución de Dirichlet para $\alpha < 1$ es una distribución conveniente para asignar tópicos a documentos de un corpus. Además, es de la familia exponencial, tiene estadísticos suficientes finitos y es conjugada a la distribución multinomial. Estas propiedades facilitan la inferencia y estimación de parámetros mediante el muestreo de Gibbs.
LDA asume el modelo generativo de un documento bajo la siguiente Ecuación (2)

$$P(\boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{j=1}^{M} P(\theta_j; \alpha) \prod_{i=1}^{K} P(\phi; \beta) \prod_{t=1}^{N} P(Z_{j,t}|\theta_j) P(W_{j,t}|\phi_{Z_{j,t}}) \tag{2}$$

Las dos primeros términos en el lado derecho son distribuciones de Dirichlet, la primera correspondiente a como se distribuyen los documentos en los tópicos y la segunda a como se distribuyen los tópicos en el vocabulario. Los dos siguientes términos son distribuciones multinomiales, la primera selecciona un tópico dada la probabilidad de que un documento pertenezca a cierto tópico y la segunda el vocabulario dada la probabilidad de que cada tópico contenga ciertas palabras. Estas probabilidades son obtenidas mediante sus parámetros que son optimizados mediante algoritmos como Muestreo de Gibbs con el fin de encontrar los mejores valores de tal manera que dado un documento original, LDA pueda acercarse a reproducirlo.

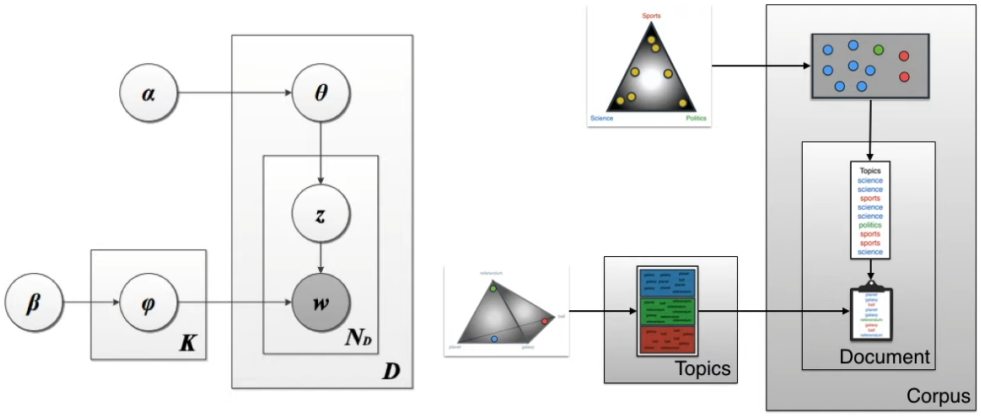


Figura 4: Generación de documentos con LDA.

Si se puede reproducir un documento entonces quiere decir que LDA ha encontrado los tópicos correctamente, y de ahí se pueden extraer los tópicos y utilizarlos en aplicaciones específicas.

3.2. Gibbs Sampling

Como vimos en el ejemplo anterior, el problema de asignar tópicos a documentos es una tarea fácil de resolver para los humanos ya que conocemos el significado de las palabras. Pero una computadora no puede hacer los mismo ya que no conoce el significado de las palabras, pero lo que si conoce es cuantas veces aparece cada palabra en todo un corpus completo. Entonces bien, el objetivo de Gibbs sampling es asignar un tópico (color) a todas las palabras dentro de un corpus de manera que cada palabra y cada documento sea lo más mono-tópico (monocromático) posible, esto se puede ver mejor en las Figuras. (5 y 6):

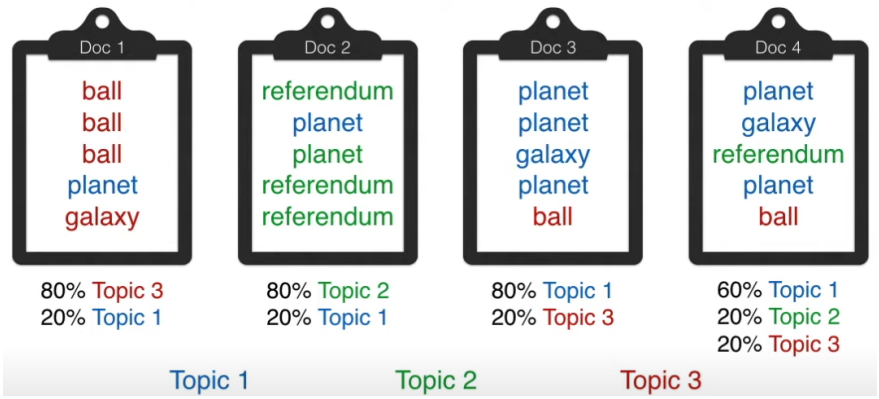


Figura 5: **Propiedad 1:** En su mayor parte, cada documento trata de un sólo tópico (color) en particular.

Ahora bien, si extraemos todas las palabras del corpus anterior observamos que en su mayoría cada palabra trata de un tópico (color) en particular.

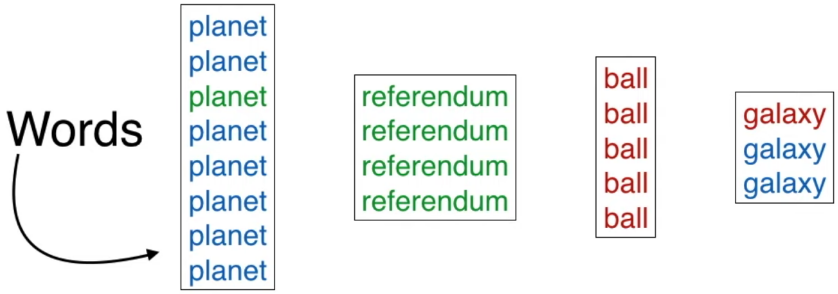
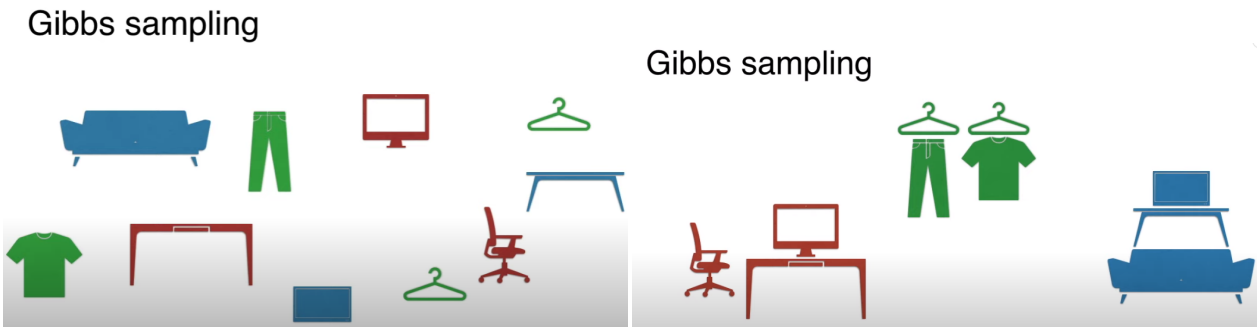


Figura 6: **Propiedad 2:** En su mayor parte, cada palabra trata de un sólo tópico (color) en particular.

Para evitar el hecho de que palabras poco relevantes (su, ya, pero, ...) se asocien a más de un tópico se pueden aplicar técnicas como tf-idf para eliminar palabras poco informativas.

Ejemplo de Gibbs sampling: El muestro de Gibbs se puede ver como el procedimiento de organizar tu habitación, en donde se hacen dos supuestos muy importantes: El primero es que todos los objetos (cama, sillón, pantalones, etc.) están en el lugar correcto (esto obviamente es falso) y el segundo es que no sabes donde deben ir los objetos pero si sabes donde deben de ir con respecto a otros (e.g. la computadora va sobre la mesa). Entonces bien, en cada iteración se elije un objeto al azar y se coloca donde debe ir con respecto a otro, tal que al final se llega a algo como se muestra en la Fig. (7)



(a) (b)

Figura 7: Izquierda: Antes de Gibb samplig. Derecha: Después de Gibb sampling.

Entonces, volviendo al ejemplo original tenemos documentos (corpus) y ese corpus tiene palabras cuyo tópico desconocemos por lo que al inicio le asignamos un tópico aleatorio a cada palabra (ver Fig. 8)

Figura 8: Tópicos aleatorios para cada palabra dentro del corpus. Esto hace referencia a la habitación desordenada (ver Fig. 7a)

Ahora bien, la idea es ir mejorando la elección de tópicos de cada palabra y por ende la asignación de tópico de cada documento (**Propiedad 1** y **Propiedad 2**). Como se dijo antes, suponemos que desde el inicio todas las palabras tienen su tópico correcto (aunque esto es falso), entonces bien tomamos primera palabra ball del Documento 1 (Fig. 8) y nos hacemos las siguientes preguntas (al elegir una palabra no la tomamos en cuenta):
Pregunta 1:

1. ¿Cuántas veces aparece Tópico 1 en Doc 1?
   Respuesta: $2 + \alpha$
2. ¿Cuántas veces aparece Tópico 2 en Doc 1?
   Respuesta: $0 + \alpha$
3. ¿Cuántas veces aparece Tópico 3 en Doc 1?
   Respuesta: $2 + \alpha$

Pregunta 2:

1. ¿Cuántas veces aparece ball en Tópico 1?
   Respuesta: $0 + \beta$
2. ¿Cuántas veces aparece ball en Tópico 2?
   Respuesta: $1 + \beta$

3. ¿Cuántas veces aparece ball en Tópico 3?

   Respuesta: 3 + α

Donde α y β son parámetros de la dist. de Dirichlet (Ec. 2). Entonces bien, la probabilidad de que ball sea del Tópico 1 dadas las veces que Tópico 1 aparece en Doc 1 es igual al producto de (2 + α) * (0 + β) y de manera análoga para Tópico 2 y Tópico 3. Esto se puede ver mejor en el siguiente diagrama (ver Fig. 9):

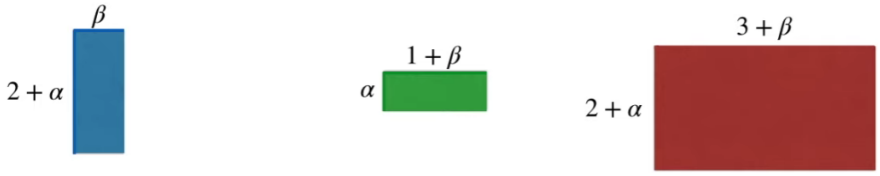


Figura 9: De manera que ball ahora es ball, es decir, pertenece al Tópico 3 ya que tiene una pribabilidad (área) mayor.

Por lo que el corpus de la Fig. (8) se actualiza de la sig. forma (ver Fig. 10)

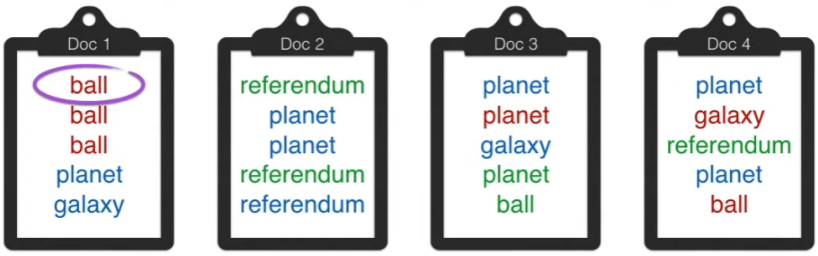


Figura 10: Actualización de tópicos de las palabras qe conforman al corpus.

En la Fig. (10) podemos observar que las palabras son más monocromáticas que antes y esto tiene sentido porque estamos coloreando la palabra del color que prevalece en el corpus. Por lo tanto, si repetimos este procedimiento para todas las palabras de cada uno de los documentos (y esto mismo se puede hacer más de una vez) podemos llegar a algo como lo siguiente (ver Fig. 11)

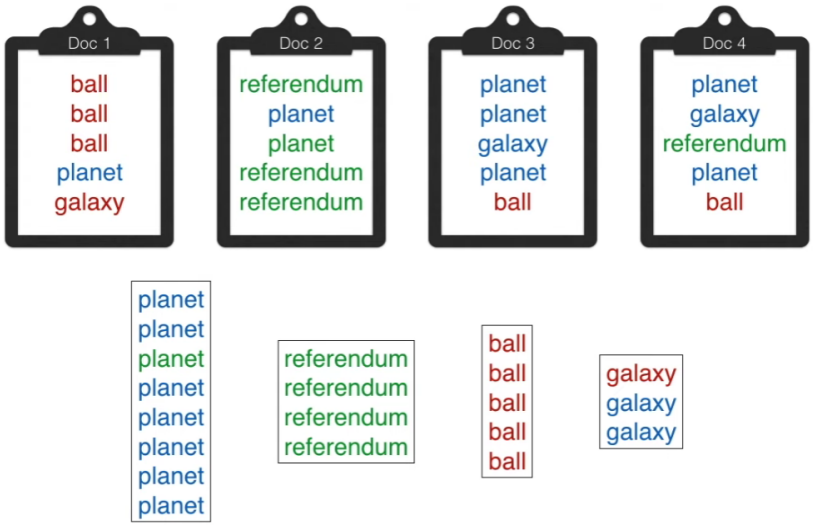


Figura 11: Resultado después de terminar las iteraciones del Gibbs sampling. Oservamos que tanto los documentos como las palabras son mucho más mono-tópicos que al inicio.

En la Fig. (11) podemos observar que cada Documento tiene una asignación de Tópico de la siguiente forma:

1. Doc 1:

   80 % Tópico 3 y 20 % Tópico 1.

2. Doc 2:

   80 % Tópico 2 y 20 % Tópico 1.

3. Doc 3:

   80 % Tópico 1 y 20 % Tópico 3.

4. Doc 4:

   60 % Tópico 1, 20 % Tópico 2 y 20 % Tópico 3

Ahora bien, es cuando interviene la parte humana, ya que tenemos un tópico (color) de cada palabra, podemos juntar todas las palabras de un mismo tópico (color) y nosotros como humanos le damos un nombre de acuerdo al conjunto de palabras, en este caso queda de la siguiente manera (ver Fig. 12):

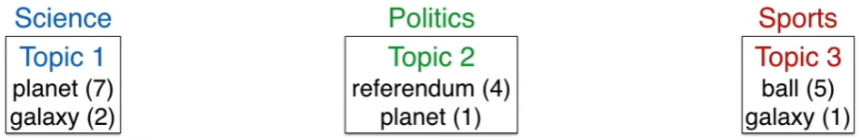


Figura 12: Tópicos asignados por un humano.

4. Metodología

4.1. Descripción de los datos y Tratamiento

La base de datos analizada en este proyecto es *Skytrax Airline Reviews* [8], la cual se puede encontrar de manera libre en la plataforma Kaggle. Dicha base contiene 65,948 reviews. Los datos incluyen reseñas de aerolíneas desde 2006 hasta 2019 para aerolíneas populares de todo el mundo, con preguntas de opción múltiple y texto libre. Los datos fueron recopilados en la primavera de 2019. Las descripciones de las características son brevemente las siguientes:

- **airline**: Nombre de la aerolínea.
- **overall**: Puntuación general dada al viaje entre 1 y 10.
- **author**: Autor del viaje.
- **review_date**: Fecha de la reseña.
- **customer_review**: Reseña de los clientes en formato de texto libre.
- **aircraft**: Tipo de aeronave.
- **traveller_type**: Tipo de viajero (por ejemplo, negocios, ocio).
- **cabin**: Cabina en el vuelo.
- **date_flown**: Fecha del vuelo.
- **seat_comfort**: Clasificación entre 1 y 5 para comodidad del asiento.
- **cabin_service**: Clasificación entre 1 y 5 para servicio de cabina.

- **food_bev**: Clasificación entre 1 y 5 para comida/bebida.
- **entertainment**: Clasificación entre 1 y 5 para entretenimiento.
- **ground_service**: Clasificación entre 1 y 5 para servicio en tierra.
- **value_for_money**: Clasificación entre 1 y 5 para relación calidad-precio.
- **recommended**: Binario, variable objetivo.

Para propósitos de este proyecto sólo estamos interesados en la variable **customer_review** ya que es mediante estas reviews que vamos a obtener el modelado de tópicos.

Para la limpieza de los datos se efectuaron las siguientes operaciones en los datos

- Se eliminarion carácteres y palabras que se agregan por default cuando un usuario hace una reseña (*Trip Verified, not Verified, Verified review*).
- Se eliminarion las reseñas duplicadas
- Se removieron *stop words*
- Tokenización
- Lematización

4.2. Desarrollo del Modelo

Después de la limpieza de datos se procedió a pasar a una representación númerica cada una de las reseñas. Para ello se utilizaron dos enfoques diferentes, uno mediante la representacion Bolsa de Palabras (BoW) y TF-IDF.

Se implementó desde cero el método de Asignación Latente de Dirichlet (LDA) con muestreo de Gibbs en Python. Implementar un muestreador de Gibbs colapsado para LDA es relativamente sencillo. Involucra configurar las variables de conteo necesarias, inicializarlas aleatoriamente y luego ejecutar un bucle sobre el número deseado de iteraciones, donde en cada iteración se muestrea un tema para cada instancia de palabra en el corpus. Después de las iteraciones de Gibbs, los conteos pueden ser usados para calcular las distribuciones latentes Dir(α) y Dir (β).

Las únicas variables de conteo requeridas incluyen $n_{d,k}$, el número de palabras asignadas al tema k en el documento d; y $n_{k,w}$, el número de veces que la palabra w es asignada al tema k. Sin embargo, para simplicidad y eficiencia, también mantenemos un conteo continuo de n_k, el número total de veces que cualquier palabra es asignada al tema k. Finalmente, además de las variables obvias como una representación del corpus (w), necesitamos un arreglo z que contendrá la asignación actual del tema para cada una de las N palabras en el corpus. El algoritmo seguido se muestra a continuación:

Input: words $\mathbf{w} \in$ documents $\mathbf{d}$
Output: topic assignments $\mathbf{z}$ and counts $n_{d,k}, n_{k,w}$, and n_k
begin
  randomly initialize $\mathbf{z}$ and increment counters
  foreach *iteration* **do**
    for $i = 0 \rightarrow N - 1$ **do**
      $word \leftarrow w[i]$
      $topic \leftarrow z[i]$
      $n_{d,topic}$-=1; $n_{word,topic}$-=1; n_{topic}-=1
      for $k = 0 \rightarrow K - 1$ **do**
        $p(z = k|\cdot) = (n_{d,k} + \alpha_k)\frac{n_{k,w}+\beta_w}{n_k+\beta\times W}$
      end
      $topic \leftarrow$ sample from $p(z|\cdot)$
      $z[i] \leftarrow topic$
      $n_{d,topic}$+=1; $n_{word,topic}$+=1; n_{topic}+=1
    end
  end
  return $\mathbf{z}$, $n_{d,k}, n_{k,w}, n_k$
end

Figura 13: Algoritmo LDA Muestrador de Gibbs.

5. Resultados

Después de realizar la limpieza de la base de datos y el preporcesamiento se imprimió una nube de palabras para verificar si la limpieza y el preprocesamiento de los datos fue el más adecuado. En la Fig. (14) se puede observar que palabras como *tiempo (time)*, *asiento (seat)*, *vuelo (flight)* o *servicio (service)* son frecuentes/relevantes en las reseñas de los clientes. Ya que no se pueden detectar fácilmente palabras poco informativas se puede asumir que la limpieza y el preprocesamiento del texto es el adecuado.

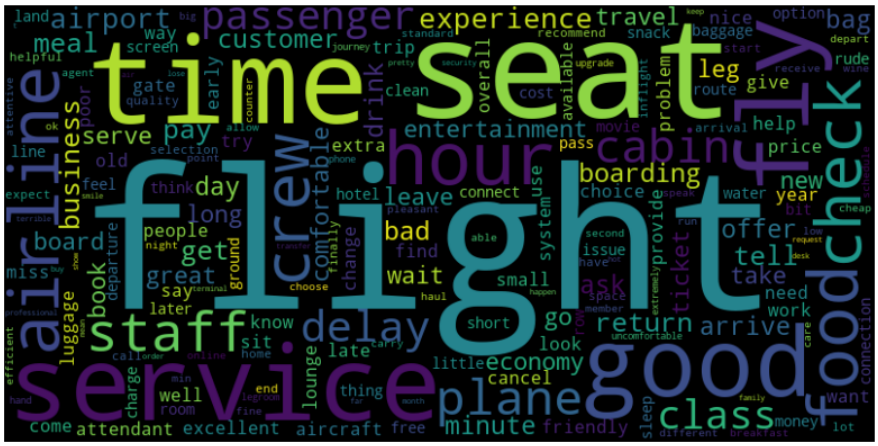
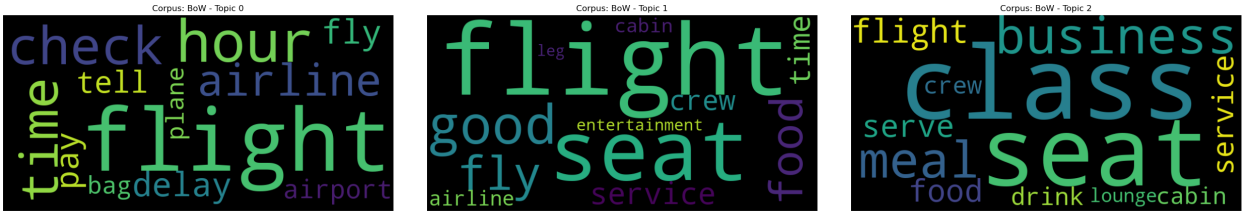


Figura 14: Nube de palabras del las reviews preprocesadas.

Ahora bien, sabemos que Bolsa de palabras (BoW) y TF-IDF (Term Frequency-Inverse Document Frequency) son dos representaciones diferentes de texto utilizadas comúnmente en el procesamiento del lenguaje natural (NLP). Mientras que BoW simplemente cuenta la ocurrencia de palabras en un documento o corpus, TF-IDF asigna pesos a las palabras en función de su frecuencia y su rareza en el corpus completo.

A continuación mostramos la modelación de tópicos con LDA-Gibbs sampling usando un corpus

dado por BoW y otro corpus dado por TF-IDF. En la Fig. (15) podemos observar las 12 palabras más relevantes de cada tópico usando una representación BoW. Dentro de los parámetros del algoritmo LDA elegimos que nos dividiera los conjuntos de palabras en 3 tópicos, más adelante con Escalamiento Multidimensional se verificará que efectivamente 3 tópicos es el número adecuado de tópicos.

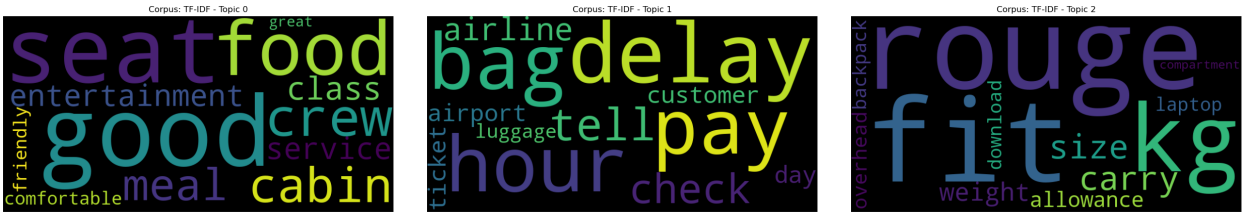


(a) Tópico 0. (b) Tópico 1. (c) Tópico 2.

Figura 15: Modelado de tópicos usando un corpus dado por la representación BoW.

En la Fig. (15) no se puede apreciar una diferencia clara entre tópicos ya que palabras como *vuelo* se repiten en los 3 tópicos y otras más si bien son diferentes pero pueden tener significados similares como *tiempo (time)* o *hora (hour)*. Por lo que se puede concluir que la representación BoW no es una muy buena opción para la modelación de tópicos.

Por otro lado, en la Fig. (16) podemos observar las 12 palabras más relevantes de cada tópico usando una representación TF-IDF.



(a) Tópico 0. (b) Tópico 1. (c) Tópico 2.

Figura 16: Modelado de tópicos usando un corpus dado por la representación TF-IDF.

En la Fig. (16) se puede apreciar con mayor facilidad una diferencia entre tópios. E.g en la Fig. (16a) se pueden apreciar palabras como *asiento (seat)*, *comida (food)*, *alimento (meal)*, *entretenimiento (entertainment)*, etc. las cuales tienen que ver con la experiencia o el trato durante el vuelo, por lo que el tópico 0 puede ser nombrado como **experiencia de vuelo.**

Del mismo modo, en la Fig. (16b) se aprecian palabras como *pagar (pay)*, *bolsa (bag)*, *hora (hour)*, *demora (delay)*, *boleto (ticket)* las cuales tienen que ver con la compra de los boletos y posibes retrasos en el vuelo, por lo que el tópico 1 se puede renombrar como **compra y retrasos del vuelo**.

Finalmente, en la Fig. (16c) se pueden apreciar palaras como *rouge (una marca de aereolínea en específico)*, *fit*, *size*, *carry* o *kg* y todas estas palabras hacen referencia al equipaje o quizá restricciones del mismo, por lo que el tópico 2 se puede renombrar como **equipaje de mano y restricciones de equipaje**.

Una vez que se obtuvieron los tópicos surgen las preguntas: ¿Cuál es el significado del tópico?, ¿Qué tan prevalente es el tópico? y ¿Cómo se relacionan los tópicos entre ellos?. En [4] se desarrolló una técnica que busca responder estas preguntas mediante una visualización como la que vemos en la Figura (17) para nuestros datos.

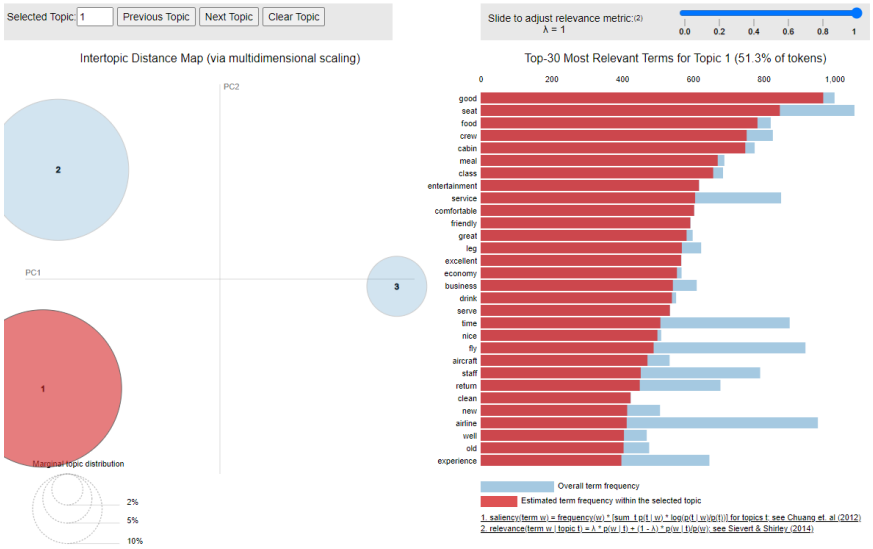


Figura 17: MDS de los tópicos latentes obtenidos.

La importancia de los tópicos se codifica como la importancia, el panel de la derecha muestra barras de cada término que son más útiles para interpretar cada tópico.

La divergencia de Jensen-Shannon es una medida de disimilaridad entre las distribuciones de tópicos de todos los documentos. Esta medida se utilizó para construir una matriz de divergencia. El Escalamiento Multidimensional (MDS) se usa para reducir la dimensionalidad de la matriz de divergencia. Esto nos permitió representar los tópicos en un espacio bidimensional, preservando las relaciones de similitud entre ellos.

La observación de una clara separación entre los tópicos en el espacio bidimensional respalda la elección de tres como un número óptimo de tópicos latentes para el conjunto de datos analizado. Este hallazgo se basa en la premisa de que un número adecuado de tópicos debería reflejar una separación discernible y coherente entre las temáticas abordadas en el corpus.

Posteriormente, se etiquetaron las reseñas según las palabras que se contenían en cada tópico. A continuación se muestra un ejemplo resultado del etiquetado

Reseña	Tópico Asignado
Luton to Bratislava. Cabin crew pleasent with 1 or 2 can't wait to get off. Fast boarding, airport ground crew as rude. Some meals and drinks at reasonable price. Cheap way of getting from point A to B.	Experiencia de vuelo
Luton to Athens. Yesterday, we got here 7:01 and the flight was suppose to take off at 7:20. The receptions did not let us on the plane, we were charged another Â£75. Today 23/03 am here and the plane has been delayed for an hour and all they are saying is that is not our fault, whose is it? The customer service team don't know how to say sorry.	Compra y retrasos del vuelo
We travelled back from Lisbon to Luton on Saturday, we paid for 10kg carry on luggage but once we were boarding the staff started pull passenger's out of line and telling them they have to pay more - their luggage was no bigger then another passenger's and the staff were very rude and aggressive. This is the second time I have been witness to this robbery to vulnerable people who are desperate to get home.	Equipaje de mano y restricciones de equipaje

Cuadro 1: Ejemplo de Etiquetado de Reseñas con Tópicos LDA

Se dividió el conjunto de reseñas etiquetadas en conjuntos de entrenamiento y prueba para desarrollar un modelo de clasificación. Se emplearon dos algoritmos: Regresión Logística y Random Forest.

- Regresión Logística: Se logró un accuracy de 0.84 en el conjunto de prueba, indicando una buena capacidad del modelo para clasificar reseñas en los tres tópicos.
- Random Forest: Se obtuvo un accuracy de 0.79 en el conjunto de prueba, proporcionando una clasificación sólida aunque ligeramente inferior a la regresión logística.

6. Conclusiones

- La representación mediante TF-IDF mostró ser claramente superior a la representación BoW para la modelaciónn de tópicos.
- Se lograrron distinguir claramente 3 tópicos: Experiencia de vuelo, compra y retrasos del vuelo y por último equipaje de mano y restricciones de equipaje.
- La observación de una clara separación entre los tópicos en el espacio bidimensional mediante MDS respalda la elección de tres como un número óptimo de tópicos latentes para el conjunto de datos analizado. Este hallazgo se basa en la premisa de que un número adecuado de tópicos

debería reflejar una separación discernible y coherente entre las temáticas abordadas en el corpus.

- La aplicación exitosa de etiquetas de tópicos a las reseñas, combinada con la precisión alcanzada por el modelo de clasificación, resalta la utilidad de la técnica LDA en la categorización temática. Estos resultados sugieren la viabilidad de implementar este enfoque en sistemas de clasificación automática de reseñas para facilitar la gestión y análisis de grandes conjuntos de datos.

Referencias

[1] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

[2] Boyd-Graber, J., Blei, D., & Zhu, X. (2007, June). A topic model for word sense disambiguation. In Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL) (pp. 1024-1033).

[3] Gerber, G. K., Dowell, R. D., Jaakkola, T. S., & Gifford, D. K. (2007). Automated discovery of functional generality of human gene expression programs. PLoS Computational Biology, 3(8), e148.

[4] Sievert, C., & Shirley, K. (2014, June). LDAvis: A method for visualizing and interpreting topics. In Proceedings of the workshop on interactive language learning, visualization, and interfaces (pp. 63-70).

[5] Gerrish, S. M., & Blei, D. M. (2011, October). Predicting legislative roll calls from text. In Proceedings of the 28th International Conference on Machine Learning, ICML 2011.

[6] Gilks, W. R., Richardson, S., & Spiegelhalter, D. (Eds.). (1995). Markov chain Monte Carlo in practice. CRC press.

[7] Bishop, C. M., & Nasrabadi, N. M. (2006). Pattern recognition and machine learning (Vol. 4, No. 4, p. 738). New York: springer.

[8] Kaggle Skytrax Airline Reviews
https://www.kaggle.com/datasets/efehandanisman/skytrax-airline-reviews