Sistemas de recomendación, restaurantes y reseñas: más allá de la media

[iStock]/Getty Images.

Al elegir un restaurante se tiene una gran variedad de opciones, muchas de ellas desconocidas. Para decidir a dónde ir es común usar portales de reseñas, que, a partir de filtros, recomiendan aquellos que cumplen los criterios del consumidor y que en el mejor de los casos se acoplan a sus gustos. Sin embargo, este enfoque normalmente carece de personalización, y se limita en general a ordenar los resultados descendientemente por una calificación media. En mi TFM muestro cómo usando sistemas de recomendación basados en la información textual y numérica de las reseñas es posible tener un grado de personalización que beneficia tanto a consumidores como restaurantes.

Juan Carlos Ruiz | *Alumno de Afi Escuela*

Los consumidores prestan cada vez más atención a las reseñas en línea antes de tomar una decisión de compra o de consumo. El 91% de las personas leen reseñas de manera ocasional o regular, y el 84% confía en las reseñas online al mismo nivel que en recomendaciones personales (Gather-Up, 2018). Sin embargo el 50% de los consumidores no escribe reseñas ni califica a los restaurantes (Gather-Up, 2018), y este desbalance entre la producción y el consumo de reseñas dificulta la construcción de una imagen clara de la calidad y la experiencia de un restaurante, y la limita a los puntos de vista de los usuarios que sí reseñan.

Sistemas de recomendación

Los sistemas de recomendación manejan tres tipos de objetos: ítems, usuarios y transacciones. Los ítems son cualquier elemento por recomendar, en este caso, los restaurantes. Luego está el usuario, a quien se ofrece las recomendaciones. Y tercero, las transacciones, que son la unión entre usuarios e ítems, que en este contexto son las visitas a restaurantes de un consumidor.

Generalmente, los sistemas de recomendación se clasifican en dos tipos: Content-Based Filtering (filtrado basado en contenidos).los cuales le ofrecen al usuario ítems similares a los que consumió en el pasado; **y Collaborative-Filtering (filtrado Colaborativo)**, los cuales ofrecen al usuario ítems que personas con gustos similares al suyo evaluaron de manera positiva en el pasado.

A pesar de la amplia investigación en sistemas de recomendación, en la academia, el uso de técnicas de procesamiento del lenguaje natural (NLP) en idiomas distintos al inglés junto con sistemas de recomendación es bastante limitado. Debido a lo anterior, decidí desarrollar y evaluar un sistema de recomendación basado en procesamiento del lenguaje natural enfocado en el contexto de restaurantes.

Acoté este proyecto al contexto a la **industria de restaurantes de Bogotá (Colombia)**. En Colombia, la aparición del COVID-19 y las consecuentes medidas preventivas de cuarentena obligatoria tuvieron un fuerte impacto en la industria de los restaurantes y bares. **De febrero a marzo de 2020 este sector presentó una variación de -33% en los ingresos percibidos (DANE, 2020)**. Además, **a pesar de que la industria de restaurantes aporta el 4% al PIB de Colombia y cerca del 6% de los empleos del país** (León, 2016), **solo el 40% de los restaurantes llega a los cinco años** (Nuñez, 2018).

Fuentes y obtención de los datos

En todo ejercicio de Data Science, uno de los factores más importantes (sino el más importante) es la calidad de los datos. Así, la fiabilidad de las reseñas fue un factor clave, pues si las reseñas no reflejan la realidad del restaurante (siendo posible un sesgo hacia opiniones positivas o negativas) las recomendaciones no serían correctas.

El 94% de los consumidores evalúa las reseñas de Tripadvisor como más fiables, rigurosas, útiles y descriptivas. Y el 90% indica que las reseñas de Tripadvisor coinciden con las experiencias reales en los restaurantes, comparado con un 31% de Google y 18% de Facebook (Influences on Diner Decision-Making Survey, 2018). Es por esto que decidí obtener las reseñas y la información propia de los restaurantes de los restaurantes de TripAdvisor usando Web Scrapping[1].

Análogamente, sin datos correctamente organizados, con duplicados, o con errores, se llegaría a conclusiones erróneas. Por lo que se hizo un proceso de limpieza para pasar de los datos en bruto a un modelo de datos (Figura 1). Entre las modificaciones realizadas destacan: la **eliminación de duplicados y valores nulos; la corrección y unificación de valores con distinta ortografía; y el uso del API de Google Maps para enriquecer la información geográfica de los restaurantes, y para completar información faltante en algunos de ellos**.

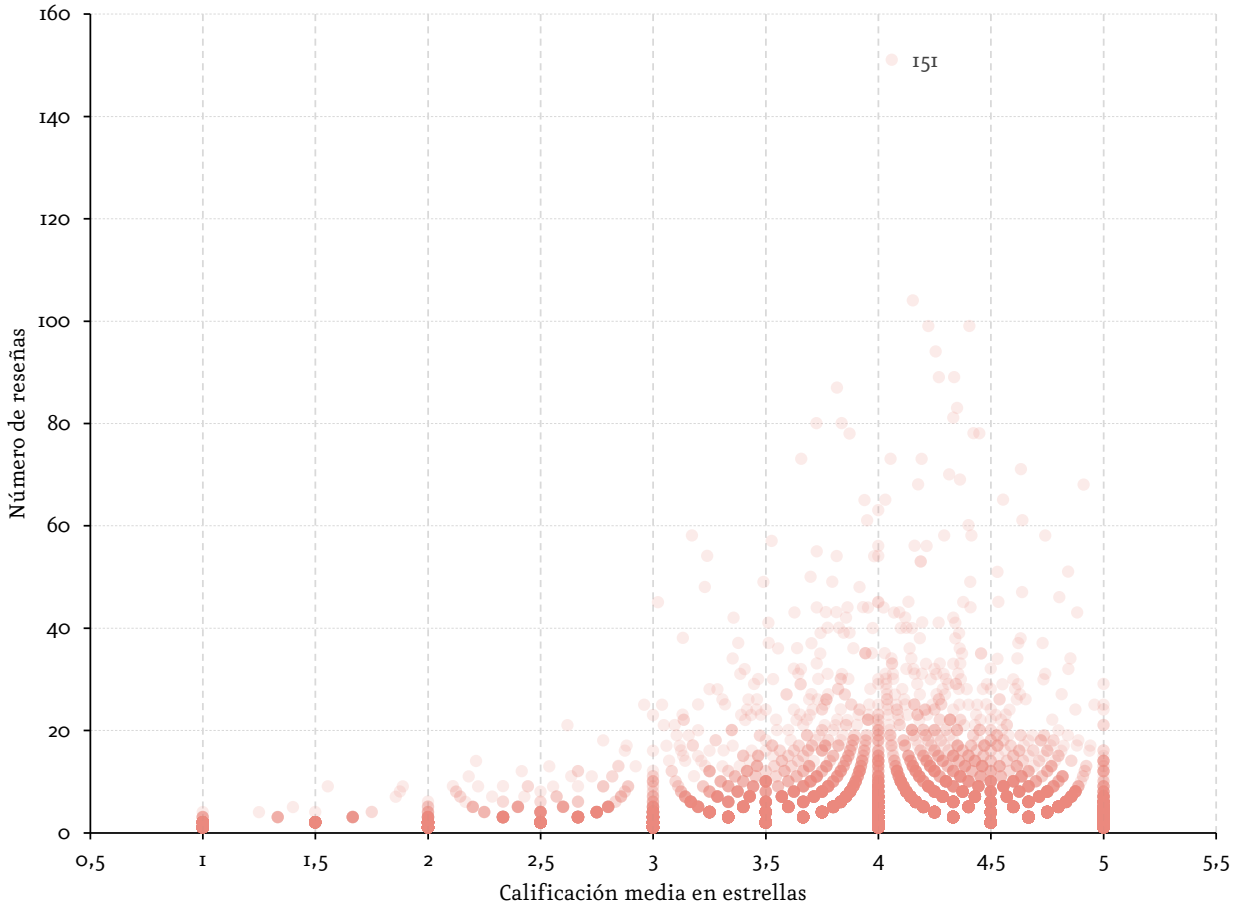
Por su parte, las reseñas, al ser textos libres, se catalogan como información no estructurada. Para darles la estructura necesaria para que fuesen usadas para el modelado se automatizó un proceso de limpieza que cambió el texto a minúsculas, corrigió la ortografía, reemplazó caracteres repetidos innecesarios y eliminó stop-words[2]. Además, se usó el modelo CoreNLP (Manning et. al. 2014) para hacer lematización del texto[3]. **Como resultado de este proceso se obtuvo un set de datos con 2130 restaurantes, 92.024 reseñas y 42.449 usuarios**.

1 Método que simula el comportamiento humano en una página, y que de manera masiva obtiene la información de una página web. La obtención de datos de Tripadvisor se limitó a las 300 primeras reseñas por restaurante para no violar ninguna ley de propiedad de los datos, y no se obtuvo informa

2 Palabras sin significado como artículos, pronombres, preposiciones, etc. Y que no agregan valor al análisis. Para esto se usó la librería NLTK.

3 La lematización consiste en reemplazar las formas flexionadas de una palabra por su lema correspondiente (como se encontraría en un diccionario) basado en su significado. Por ejemplo, de «Excelentísimo» por «excelente», o «Comimos» por «comer».

Número de reseñas de los consumidores en función de su puntuación



La mayoría de las reseñas asignadas por los usuarios son de 5 o 4 estrellas. Es decir, la mayoría tiene experiencias satisfactorias con los restaurantes que visita o tiende a escribir reseñas cuando tiene una experiencia positiva. De 2.130 restaurantes posibles, en promedio, un usuario solo habrá evaluado dos de ellos. Esto deja en evidencia una marcada situación de data-sparsity[4]. De hecho, de la totalidad de posibles parejas (usuario-restaurante) tan solo el 0.1% están presentes.

Al analizar las palabras más usadas en cada calificación de reseñas de manera comparativa se ve que **conforme aumenta el número de estrellas, aumentan los calificativos positivos, especialmente los relacionados con 'recomendado' y 'delicioso'**. Mientras que **las reseñas 1 o 2 estrellas hacen más énfasis en el servicio y la atención**.

Palabras más usadas en reseñas de 1 estrella

Palabras más usadas en reseñas de 5 estrellas

4 Si se contempla una matriz de dos dimensiones en la que en un lado están los restaurantes y en el otro los usuarios, se tiene que se cubren muy pocas de todas las posibles combinaciones.

División de datos y modelado

El conjunto de variables (features) de los restaurantes se construyó con base en el texto pre procesado: luego de agrupar por los niveles de puntuación (de 1 a 5 estrellas) se aplicó el método TF-IDF normalizado, se seleccionaron para cada nivel las 100 palabras más importantes y se hizo un perfil por restaurante usando una representación de bag-of-words sobre las 500 palabras más importantes. Para hacer una estimación correcta del desempeño de cada modelo, se dividió el dataset en dos particiones: Entrenamiento (80%) y Prueba(20%)[5]. De los modelos probados destacan los siguientes:

Singular Value Descomposition

Es un **método basado en álgebra lineal que permite la reducción de dimensionalidad**. Se basa en la factorización de matrices, y no hace uso de features textuales. Se enmarca entre las técnicas de filtrado colaborativo, y utiliza una matriz en la que cada fila representa un usuario, cada columna un ítem y los elementos de esta matriz son las calificaciones.

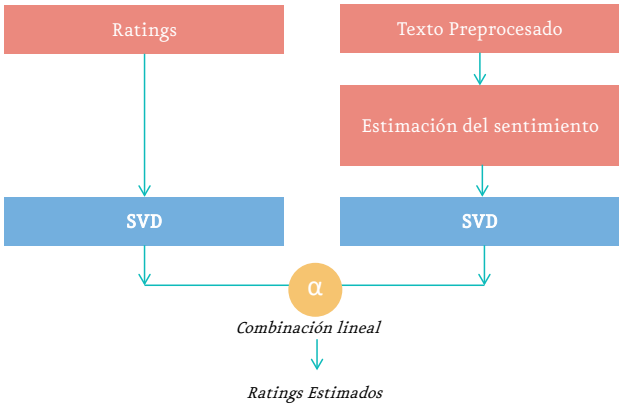
Modelo Light FM

Es el modelo implementado en la librería del mismo nombre, propuesto por Kula (2015). **El modelo aprende embeddings[6] para consumidores y restaurantes de una manera que codifica las preferencias del consumidor sobre los restaurantes**. Este modelo tiene dos características principales: 1) Aprende a partir representaciones de ítems y usuarios. Y 2) Permite computar recomendaciones a usuarios e ítems nuevos.[7]

Modelo de Reseña mixta

Desarrollé un modelo similar al planteado por Pero & Horváth (2013), el cual contempla tanto el sentimiento de la reseña como la calificación de la misma. **Este modelo tiene dos partes: por un lado, primero se estiman los sentimientos de la reseña (positivo o negativo) para crear calificaciones virtuales, y a estas se les aplica el un procedimiento de factorización de matrices** (SVD en este caso)[8].

Figura 28 Modelo de Reseña Mixta



Por otro lado, a las calificaciones dadas por los usuarios se le aplica SVD. Y finalmente, se hace una combinación lineal de las calificaciones predichas por cada una de las matrices para dar una calificación final sobre la cual se ordenan los resultados y se dan las recomendaciones.

$$R\_final = R\_virtual * a + R\_real * (1-a)$$

Métricas

Para medir los modelos y compararlos entre sí, se usaron principalmente tres métricas: RMSE, Average Precision at K y Average Accuracy at K.

La evaluación con **RMSE** funciona de la siguiente manera: el modelo genera predicciones de las valoraciones para un consumidor, y luego se comparan las predicciones contra los valores reales por medio de la fórmula de RMSE. La ventaja de este enfoque es que no cae en el error de penalizar al sistema en caso de que haya recomendaciones de items que el usuario no ha evaluado.

La Precision at K (P@K) es la proporción del top K recomendaciones que son relevantes para un usuario. Por ejemplo, si K =10, sería el porcentaje de los restaurantes que son relevantes que llegan al top 10 (para un usuario dado). Luego, si se hace por usuario una media de la P@K para K valores (K≤100), se consigue la Average Precision at K (APK). Y si luego se hace una media de estas APK entre los usuarios, se obtiene la Mean Average Precision at K.

De manera similar, **la Recall at K (R@K)** es la proporción de los ítems relevantes que llegan al top K. La Average Recall at K se calcula por usuario una media de la R@K para K valores (K≤100), y la Mean Average Recall at K es el promedio estos valores.

5 Debido al data sparsity se decidió tener en Train la mayor cantidad de datos, sin que Test dejase de ser representativo. Sobre el de Train se hizo cross-validation de 5 folds para los algoritmos que requirieron encontrar hiper parámetros. Los hiper parámetros se optimizaron usando grid-search. Los modelos finales fueron entrenados en la totalidad de datos de Train.

6 Representaciones latentes en un espacio de alta dimensión.

7 Los usuarios y los ítems se pueden describir dadas sus features, y estas son conocidas con antelación y representan meta-datos de usuarios y de ítems. Para este caso se tienen en cuenta sólo features de los ítems, dado que no se tiene información de los usuarios. Estas features fueron los vectores resultantes del proceso de Bag-of-words.

8 Se usó la librería Senti-py que tiene un modelo para la detección de sentimiento en español. Para el método SVD se usó la librería Surprise.

Modelo	Baseline	SVD	LightFM	Reseña mixta (RM)
RMSE – Train CV	1.251	1.03	-	1.002
RMSE – Test	1.258	1.04	-	1.02
Mean Average Precision at K - Test	0.601	0.640	0.695	0.682
Mean Average Recall at K - Test	0.125	0.165	0.232	0.184

El modelo baseline fue un modelo regresivo que contempla una media general, y las desviaciones del usuario y del restaurante.

En términos del RMSE, los modelos SVD y RM logran significativamente mejores resultados que el modelo baseline. El modelo RM al integrar las calificaciones virtuales de los restaurantes y computarlas en conjunto con las del modelo SVD logra un error menor diferenciado por su segundo decimal, lo que para efectos prácticos podría considerarse un resultado igual al del SVD.

La Mean Average Precision at K indica que en promedio el 60% de los ítems recomendados son relevantes para el usuario en el modelo baseline. Esto tiene sentido al considerar que la mayoría de las reseñas tienen calificaciones de 4 o 5 estrellas. **Los modelos LightFM y Reseña Mixta propuestos alcanzan en esta métrica valores de 0.70 y 0.68 respectivamente, superando a SVD.**

A pesar de usar técnicas de NLP no se consiguieron mejoras significativas frente al modelo SVD. Esto puede deberse a varias razones, una de ellas es la data-sparsity del dataset. El dataset solo tiene un 0.1% de las posibles parejas usuario-reseña, lo cual dificulta a cualquier algoritmo el cálculo de recomendaciones, y perjudica las métricas en casos en los que los usuarios tienen un bajo número de reseñas.

Los resultados demuestran que los modelos basados en texto ofrecen una mejora sobre aquellos que solamente tienen en cuenta las calificaciones otorgadas a los restaurantes por parte de los usuarios. Sin embargo, el grado de data-sparsity de los datos es determinante en la consecución de buenas predicciones, pues incluso tras incluir información textual, las mejoras a la hora de hacer recomendaciones son marginales si no existen suficientes reseñas ::

Lecturas recomendadas

Gather-Up. (2018). Online Reviews Study: Restaurants & Reviews. Gather Up. https://gatherup.com/blog/online-reviews-study-restaurants-reviews/

DANE. (04/2020). Encuesta mensual de servicios (EMS). Departamento Administrativo Nacional de Estadística. https://www.dane.gov.co/files/investigaciones/boletines/ems/bol_ems_abril_20.pdf

León, D. (2016, July 10). Restaurantes del país aportan 4% al PIB. Vanguardia. https://www.vanguardia.com/economia/nacional/restaurantes-del-pais-aportan-4-al-pib-CFVL375667

Nuñez, G. E. (2018, December 29). Muchos restaurantes no llegan a los cinco años: Acodres. https://diariolaeconomia.com/fabricas-e-inversiones/item/4130-muchos-restaurantes-no-llegan-a-los-cinco-anos-acodres.html

Pero, Š., & Horváth, T. (2013). Opinion-Driven Matrix Factorization for Rating Prediction. In User Modeling, Adaptation, and Personalization (pp. 1–13). https://doi.org/10.1007/978-3-642-38844-6_1

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55-60.

Hug, N., (2020). Surprise: A Python library for recommender systems. Journal of Open Source Software, 5(52), 2174, https://doi.org/10.21105/joss.02174

Kula, M. (2015). Metadata Embeddings for User and Item Cold-start Recommendations. arXiv Preprint. https://doi.org/arXiv:1507.08439

Chen, L., Chen, G., & Wang, F. (2015). Recommender systems based on user reviews: the state of the art. In User Modeling and User-Adapted Interaction (Vol. 25, Issue 2, pp. 99–154). https://doi.org/10.1007/s11257-015-9155-5