OBSERVATORIO
DE TURISMO DE BOGOTÁ
IDT

Instituto Distrital de Turismo

¿CÓMO ESTÁ LA OFERTA HOTELERA EN BOGOTÁ?

Análisis de sentimientos, según Booking

ALCALDÍA MAYOR
DE BOGOTÁ D.C.

BOGOTÁ

OBSERVATORIO
DE TURISMO DE BOGOTÁ
IDT

CARLOS FERNANDO GALÁN PACHÓN
Alcalde Mayor de Bogotá

MARÍA DEL PILAR LÓPEZ URIBE
Secretaria de Desarrollo Económico

ANDRÉS SANTAMARÍA GARRIDO
Director General
Instituto Distrital de Turismo

IVONNE ADRIANA MARTÍNEZ ZAPATA
Subdirectora
Inteligencia y Gestión de Tecnologías de la Información

LUIS FERNANDO PINEDA GONZÁLEZ
DIEGO ALEJANDRO RODRÍGUEZ
Investigadores

Contacto
Instituto Distrital de Turismo
www.idt.gov.co
Bogotá D.C., 2024

CONTENIDO

1. INTRODUCCIÓN 4
2. ANTECEDENTES DE CIENCIA DE DATOS APLICADA AL SECTOR TURISMO 5
3. METODOLOGÍA DE RECOLECCIÓN DE LOS DATOS 11
4. RESULTADOS 15
5. CONCLUSIONES 39
INDICE DE TABLAS, FIGURAS Y GRÁFICOS 39
BIBLIOGRAFIA 42

1. INTRODUCCIÓN

El Instituto Distrital de Turismo (IDT), a través del Observatorio de Turismo de Bogotá (OTB), consolida, sistematiza, analiza, investiga y divulga la información del sector turístico de Bogotá. Este observatorio no solo actúa como un centro de documentación que respalda los procesos de investigación y diagnóstico en el sector, sino que también se articula con niveles de información local, regional, nacional e internacional para generar una visión integral de la actividad turística en Bogotá.

En el 2021, con el objetivo de utilizar los datos no convencionales para generar análisis integrales del sector, se publicó la primera versión del "ESTUDIO DE CIENCIA DE DATOS APLICADA AL SECTOR TURISMO, PERCEPCIÓN DE LOS ATRACTIVOS TURÍSTICOS EN BOGOTÁ". En este documento, se extrajo la información de 50 mil reseñas para 185 atractivos turísticos que se encuentran en la ciudad de Bogotá, logrando obtener información de las características, condiciones, falencias y opciones de mejora de los atractivos turísticos de la ciudad a partir de la percepción de los visitantes nacionales e internacionales.

A partir de este documento, se propuso para el 2024, utilizando nuevamente la técnica de extracción de datos "Web Scraping", junto con el Procesamiento de Lenguajes Naturales (NPL por sus siglas en inglés) y el Análisis de Sentimiento, crear un nuevo documento que permitiera conocer la oferta de hoteles de la ciudad y la calificación de sus servicios a partir de las reseñas escritas por diferentes usuarios en la plataforma Booking.

Es así como este estudio, a partir de la información de cerca de 438.000 calificaciones y 418.000 reseñas[1] de 927 hoteles en Bogotá, busca comprender la percepción de los turistas sobre la infraestructura hotelera de la ciudad, identificando desde la perspectiva de la demanda (los usuarios) las características, condiciones, deficiencias y fortalezas de estos prestadores de servicios, que son un elemento fundamental de la cadena de turismo en la ciudad. Así mismo, estos resultados serán insumos para construir acciones que mejoren la oferta hotelera local.

[1] Una misma calificación puede contener de cero a dos reseñas, una donde se mencionan los aspectos positivos y otra donde se mencionan los negativos.

El presente documento está dividido en cinco secciones, iniciando por la presente introducción; la segunda sección, realiza una revisión bibliográfica sobre los antecedentes de la aplicación de herramientas de ciencia de datos en el sector turismo; la tercera, detalla la metodología utilizada para la construcción de la base de datos y el raspado de web; en la cuarta, se analizan los resultados, comparando las diferentes localidad de la ciudad; finalmente, el quinto capítulo presenta las principales conclusiones del estudio.

2. ANTECEDENTES DE CIENCIA DE DATOS APLICADA AL SECTOR TURISMO

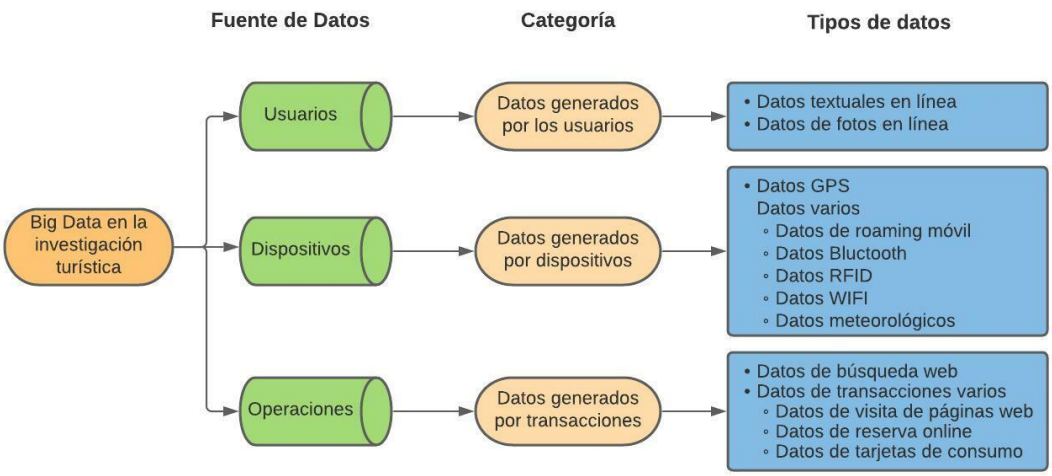
La literatura existente sobre el uso de Big Data y la aplicación de Ciencia de Datos en el Sector Turismo indica que estos temas no son novedosos. En efecto, existe una gran cantidad de trabajos y artículos científicos que han aplicado diversas técnicas a la investigación turística.

Li et al. (2018) ofrece una revisión exhaustiva de la literatura sobre los diferentes tipos de Big Data en la investigación turística. Este estudio generó un marco analítico de referencia, clasificando en tres categorías según el origen de los datos: (1) Datos generados por usuarios: que incluyen datos textuales y fotográficos; (2) datos generados por dispositivos electrónicos, recopilados de forma pasiva mediante GPS, itinerancia móvil, Bluetooth, RFID[2], WIFI, entre otros y; (3) datos generados a partir de transacciones que registran las operaciones realizadas por el usuario, como los datos de búsqueda web, visitas a páginas web, reservas en línea y transacciones con tarjetas débito y crédito.

Como se puede ver en la Figura 1 y conforme al origen de los datos utilizados en este documento, los datos textuales en línea (reseñas en la plataforma de Booking), provienen de turistas que pernoctan o se hospedan en hoteles ubicados en Bogotá, resaltando el marco analítico del presente estudio.

Figura 1. Big Data en la investigación turística

[2] Identificación por radiofrecuencia.



Fuente: Adaptado al español de Li et al. (2018)

En la última década, el sector del turismo y el análisis de la oferta hotelera han comenzado a adoptar enfoques de Big data. A partir de una revisión de la literatura sobre estudios y artículos que utilizan datos textuales en línea (reseñas), se ha identificado un número significativo de investigaciones que recopilan los datos para analizar la percepción que los turistas tienen sobre los hoteles.

Como se puede ver en la tabla 1, entre los estudios seleccionados, se evidencia que las plataformas más destacadas para realizar reservas de alojamiento y recopilación de reseñas han sido Trip Advisor, Expedia, Airbnb, Agoda y Pegipegi. En los últimos años, el metabuscador de viajes "Booking", que ofrece más de 28 millones de unidades de alojamiento, se ha convertido en uno de los buscadores de reservas de alojamiento más importantes y, por tanto, en uno de los repositorios de reseñas relacionadas con el hospedaje más significativos.

Por su parte, existe una gran variedad de estudios respecto al área geográfica elegida por los investigadores de los estudios identificados, pues van desde ciudades específicas, a municipios, regiones o países completos.

De igual manera, se ha encontrado que la metodología predominante para la recopilación de los datos es el "Web Scraping" y el "Web Crawler" (rastreador web), en lugar de la adquisición directa de bases de datos de proveedores especializados. Esto se debe a la facilidad que ofrecen estas técnicas para recopilar grandes volúmenes de información de las páginas web.

En cuanto al número total establecimientos o reseñas analizados en los estudios seleccionados, estos van desde 21 hoteles y cerca de 500 reseñas en el estudio de Moro et al. (2017) hasta más de 25.000 hoteles y aproximadamente 266.000 reseñas en el de Guo et al. (2017).

Tabla 1. Trabajos en el que se utilizan las reseñas de los usuarios para el análisis la oferta alojamiento.

Estudio	Fuente de los Datos	Área Geográfica	Método Extracción de los Datos	Tipo de reseñas	Número Total establecimientos o de Reseñas[3]
Flôres Limberger et al. (2014)	TripAdvisor	Brasil	No Disponible	Hoteles	660 reseñas
Berezina et al. (2015)	TripAdvisor	Saratoga (Estados Unidos)	Online robot developed	Hoteles	2.510 reseñas
Xiang et al. (2015)	Expedia	Las 100 ciudades más grandes de EE. UU.	Web crawler / Rastreador web	Hoteles	10.537 hoteles 60.648 reseñas
Guo et al. (2017)	TripAdvisor	100 ciudades en 16 países	Web crawler / Rastreador web	Hoteles	25.670 hoteles 266.544 reseñas
Moro et al. (2017)	TripAdvisor	Strip, Las Vegas	No Disponible	Hoteles	21 hoteles 504 reseñas
Moro et al. (2018)	TripAdvisor	Portugal	Web scraping	Hoteles	967 establecimientos
Cheng & Jin (2019)	Inside Airbnb[4]	Sydney	Obtenidos de Inside Airbnb	AirBNB	170.124 reseñas
Adhinugroho et al. (2020)	Agoda y Pegipegi	Indonesia	Web scraping	Hoteles	19.834 establecimientos

Elaboración: Subdirección de Inteligencia y Tecnologías de la Información – IDT.

[3] Se maneja esta clasificación por número de establecimientos o número de reseñas, ya que no se cuenta con una información homogénea en este campo entre todos los artículos seleccionados.

[4] Inside Airbnb es un proyecto que proporciona datos sobre el impacto de Airbnb en las comunidades residenciales. https://insideairbnb.com/about/

Consideraciones legales sobre Web Scraping:

La práctica del web scraping, que implica la recolección automatizada de datos desde sitios web, plantea interrogantes sobre su legalidad, especialmente en términos de privacidad, propiedad intelectual y el uso adecuado de la información extraída.

En Colombia, la normatividad relevante para el web scraping incluye principalmente la Ley 1581 de 2012 sobre protección de datos personales. Además, también es pertinente la Ley 23 de 1982 sobre derechos de autor, teniendo en cuenta que establece que los contenidos de los sitios web pueden estar protegidos por derechos de autor. Así mismo, se debe considerar la Ley 527 de 1999, pues esta *define y regula el acceso y uso de los mensajes de datos, el comercio electrónico y las firmas digitales.*

Según SANABRIA (2021), existen dos tipos de usuarios de sitios web: aquellos que visitan los sitios para adquirir bienes y/o servicios y aquellos que solo desean navegar por ellos. Además, la Cámara de Comercio de Bogotá señala:

> "*En Bogotá D. C., es costumbre mercantil que todo visitante de un sitio de Internet o una página web, por el hecho de navegar en la misma, se obliga a respetar los términos y condiciones sobre su uso, contenidos y normas de propiedad intelectual*". (Cámara de Comercio de Bogotá, 2021, párr. 4)

De esta manera, la normatividad vigente señala que los usuarios interesados en adquirir bienes y/o servicios se aplicará lo definido en el Artículo 14 de la Ley 527 de 1999[5]. Por su parte, para aquellos que simplemente navegan el sitio web, se aplicará la teoría de la aceptación por la mera navegación.

Por lo tanto, los propietarios de sitios web pueden incluir en sus términos y condiciones cláusulas que prohíben el web scraping para evitar el uso no autorizado de sus datos. Además, los términos y condiciones pueden establecer limitaciones para asegurar la integridad del sitio web y garantizar una experiencia adecuada para todos los usuarios.

[5] ARTÍCULO 14. FORMACIÓN Y VALIDEZ DE LOS CONTRATOS. En la formación del contrato, salvo acuerdo expreso entre las partes, la oferta y su aceptación podrán ser expresadas por medio de un mensaje de datos. No se negará validez o fuerza obligatoria a un contrato por la sola razón de haberse utilizado en su formación uno o más mensajes de datos. (párr. 33)

En algunos casos, estas limitaciones están relacionadas con un número máximo de peticiones (consultas), ya que pueden colapsar el sitio web. Para prevenir estos problemas, los propietarios pueden implementar herramientas como las validaciones por códigos captcha (Completely Automated Public Turing test to tell Computers and Humans Apart) o bloqueo por accesos masivos, con el fin evitar las prácticas de web scraping y/o uso indebido del sitio web.

Así, aunque el propietario del sitio web implemente o no estas herramientas, todas estas limitantes y prohibiciones se consideran usos indebidos y pueden acarrear consecuencias legales. Sin embargo, Sanabria (2021) destaca que, si el listado de usos indebidos no forma parte de los términos y condiciones del sitio web, las reclamaciones del propietario carecerían de soporte legal.

En resumen, aunque el web scraping en Colombia no está directamente regulado por una legislación específica, debe implementarse de una manera ética y responsable. Es fundamental cumplir con las leyes de protección de datos, respetar los derechos de propiedad intelectual y seguir las políticas de uso de los sitios web de los cuales se extraen los datos. De lo contrario, podrían surgir riesgos legales relacionados con la privacidad, los derechos de autor y el uso indebido de la información.

En este sentido, el ejercicio realizado no viola la ley, teniendo en cuenta que:

1. **Conformidad con los términos y condiciones de Booking.com.**

- La sección A14 sobre derechos de propiedad intelectual, especifica que los usuario son pueden monitorear, copiar, raspar/rastrear, descargar, reproducir o usar contenido de la plataforma para propósitos comerciales sin el permiso por escrito de Booking.com o sus licenciantes.
- Además, Booking.com monitoreo las visitas a su plataforma, bloqueando a cualquier persona o sistema autorizado que sospechen de:
    - Realizar una cantidad irrazonable de búsquedas.
    - Usar dispositivos o software para recopilar precios y otra información.
    - Causar una tensión indebida en su plataforma.

En el presente ejercicio, no se incumplen estas reglas, ya que el raspado de datos se efectúa con fines no comerciales, sino para investigaciones que aportan a la elaboración de políticas públicas del sector por parte del Instituto Distrital de Turismo.

De esta manera, la información se descargó en lotes para minimizar solicitudes consecutivas a la plataforma, evitando generar visitas sospechosas o bloqueos, como se menciona en los términos y condiciones. Así mismo, tampoco se ejecutaron procedimientos que comprometieran la seguridad del sitio, como hackeos o rompimientos de códigos CAPTCHA, lo que significa que no se incurrió en un delito de acceso abusivo a un sistema informático introducido en el Código Penal por la Ley 1273 de 2009.

2. **Legalidad del web scraping en el marco colombiano.**

Según Sanabria, "En el caso particular del web scraping como herramienta para la extracción y análisis de datos sobre tendencias de mercado y otros factores …

… las empresas utilizan esta técnica para llevar a cabo estudios sobre la población, el mercado y diversos fenómenos, así como para generar proyecciones acerca de las conductas observadas en su entorno económico, apoyando su gestión en el marco del ordenamiento jurídico colombiano, conforme al Artículo 333 de la Constitución Política:

"La actividad económica y la iniciativa privada son libres, dentro de los límites del bien común. Para su ejercicio, nadie podrá exigir permisos previos ni requisitos, sin autorización de la ley. La libre competencia económica es un derecho de todos que supone responsabilidades." (Corte Constitucional de Colombia, 2016).

Estos principios se alinean con la técnica de web scraping, ya que la mayoría de las interacciones entre los usuarios, proveedores de sitios web y navegantes ocurren en una fase precontractual, en la que el usuario no siempre está obligado a aceptar los términos y condiciones de uso del sitio en todos los casos.

En resumen, el ejercicio realizado no infringe de forma alguna los términos y condiciones del sitio web booking.com, ni el marco legal en Colombia.

3. METODOLOGÍA DE RECOLECCIÓN DE LOS DATOS

La metodología utilizada para la recolección de datos fue el Web Scraping, también conocido como "recolección web", "extracción de datos web" o "minería de datos web". Esta técnica se define como "la construcción de un agente para descargar, analizar y organizar datos de la web de manera automatizada" (Vanden Broucke & Baesens, 2018).

El web scraping permite acceder a información no estructurada en sitios web que carecen de API pública o cuyo acceso a los datos está restringido. La estructura de estos datos suele estar compuesta por código HTML, aunque también pueden presentarse en formatos como XML o JSON. El agente de web scraping interactúa con la página web igual que un ser humano, pero de manera automatizada, lo que facilita la recolección masiva de datos.

Además, el web scraping "… es frecuentemente utilizado como una herramienta de análisis de mercado y, por ello, diversos autores concuerdan en que el uso del web scraping junto con las técnicas asociadas al Big Data ha traído grandes beneficios a este sector …" Sanabria (2021).

El proceso de web scraping planteado se divide principalmente en 4 pasos:

1. **Recolección de la lista de hoteles.**
   Se realiza un raspado inicial para conformar una lista de hoteles disponibles en la plataforma Booking.com, bajo una configuración de fechas determinada por el usuario. Esta lista, denominada "marco de hoteles", comprende 927 alojamientos ubicados en distintas localidades de Bogotá.

2. **Extracción de información básica de cada hotel.**
   En esta fase se recopila información básica de cada hotel del marco, incluyendo el enlace del hotel, su nombre, dirección, ubicación geográfica (longitud y latitud), horas de check-in y check-out, idiomas hablados y evaluaciones sobre diversos aspectos del hotel (como, por ejemplo, la calidad del servicio, limpieza, confort, relación calidad-precio, ubicación y conexión Wi-Fi. Además, se obtiene información detallada sobre los servicios disponibles en cada alojamiento.

3. **Consulta de información sobre las habitaciones disponibles.** Se realiza una consulta adicional para obtener detalles sobre las habitaciones ofrecidas. Esto incluye el tipo de habitación (como apartamentos, estudios, casas, bungalows, chalets y lofts), características específicas de las habitaciones (por ejemplo, tamaño, servicios, tipo de cama y baño), capacidad de huéspedes, precios y tarifas, así como la disponibilidad y precio del desayuno.

4. **Extracción de reseñas de los huéspedes.** El paso final consiste en la extracción de reseñas de huéspedes publicadas una vez finalizada su estancia. Cada huésped puede dejar una calificación general y dos reseñas detalladas, una resaltando los aspectos positivos y otra mencionando los negativos. Esta información se recopila desde octubre de 2021 hasta octubre de 2024.

Finalmente, la información recolectada se almacena en una base de datos estructurada, lo que facilita su análisis posterior. Las reseñas descargadas provienen de huéspedes de 187 países y están escritas en más de 20 idiomas diferentes.

Estructura de la información:

A partir de la técnica ya señalada, se prosiguió a descargar, recopilar y estructurar información extraída de Booking.com para la conformación de 11 variables, entre las que resaltan el nombre del hotel, país de origen del visitante, número de noches, fecha del hospedaje, fecha de reseña, entre otras, como se puede ver en la tabla 2.

Tabla 2. Estructura de los datos que se extrajeron

Variable	Explicación
Link del Hotel	Corresponde con el link del hotel en el sitio web de booking
Nombre del Hotel	Contiene el Nombre del Hotel
País de origen	País de residencia del turista
Número de noches	Número de noches que se hospedo el turista
Fecha de hospedaje	Fecha en la cual el turista se hospedo en el hotel

Grupo de Viaje	Grupo de viaje o personas con las que turista se hospedo en el hotel
Fecha de reseña	Fecha en la cual el turista escribió la reseña
Título Reseña	El título que el turista que se hospedo escribió para la reseña
Calificación Reseña	La calificación del hotel que reporto el turista (rango de 1 a 10)
Texto Reseña Positiva	El texto completo de la reseña positiva del turista que se hospedo
Texto Reseña Negativa	El texto completo de la reseña negativa del turista que se hospedo

Elaboración: Subdirección de Inteligencia y Tecnologías de la Información – IDT.

Una vez realizada la descarga de la información, el proceso de preparación de los datos para el análisis incluyó las siguientes etapas:

1. **Depuración de comentarios vacíos:** Se identificaron y eliminaron comentarios que no aportaban valor al análisis, como aquellos sin contenido (solo puntuaciones) y comentarios en las columnas de reseñas positivas y negativas que contenían únicamente palabras como "todo", "nada", "bien", "excelente" o "sin comentarios", entre otros. La tabla 3 muestra los resultados en la columna de datos depurados.

2. **Limpieza de cadenas de texto:** Se realizó una limpieza de los textos al eliminar todos los caracteres no alfanuméricos y de puntuación, tanto en las reseñas positivas como negativas.

3. **Generación de diagramas de redes de concurrencia:** Esta técnica, utilizada en minería de textos y análisis de datos, estudia las relaciones entre términos que aparecen juntos en los documentos. Este enfoque se usa frecuentemente en el análisis de redes semánticas y extracción de temas.

4. **Modelos grandes de lenguaje (LLM, modelos de IA de texto generativos)**: Se incorporaron reseñas en diferentes idiomas categorizando los comentarios en temáticas y sentimientos. Este procedimiento fue necesario porque se detectó una mezcla de comentarios positivos y

negativos en las columnas correspondientes, es decir, comentarios positivos en la columna de reseñas negativas y viceversa.

5. **Análisis adicionales usando los datos clasificados**: Se realizaron análisis descriptivos mediante tablas cruzadas (**tablas de contingencia**) y nuevamente se generaron diagramas de concurrencia para combinar categorías de variables de interés, con el fin de identificar posibles acciones de mejora por zonas.

En cuanto al total **de calificaciones descargadas**, se obtuvieron 438.213, de las cuales 236.936 estaban en la columna de reseñas positivas y 181.438 en las negativas. Sin embargo, después del proceso de limpieza de comentarios, estas cifras se redujeron a 223.388 y 156.900 respectivamente., como se puede ver en la tabla 3.

Tabla 3. Calificaciones totales extraídas a los hoteles

	Totales	Depuradas	Porcentaje
Reseñas positivas	236.936	223.388	54,1%
Reseñas negativas	181.438	156.900	41,4%
Calificaciones a hoteles	438.213		

Elaboración: Subdirección de Inteligencia y Tecnologías de la Información – IDT.

En cuanto a la clasificación de las reseñas por idioma, el 71,2% se encuentran en español, mientras que cercade un 20% se han realizado en Inglés, francés y portugués, como se ve en la tabla 4.

Tabla 4. Idioma de las reseñas por tipo

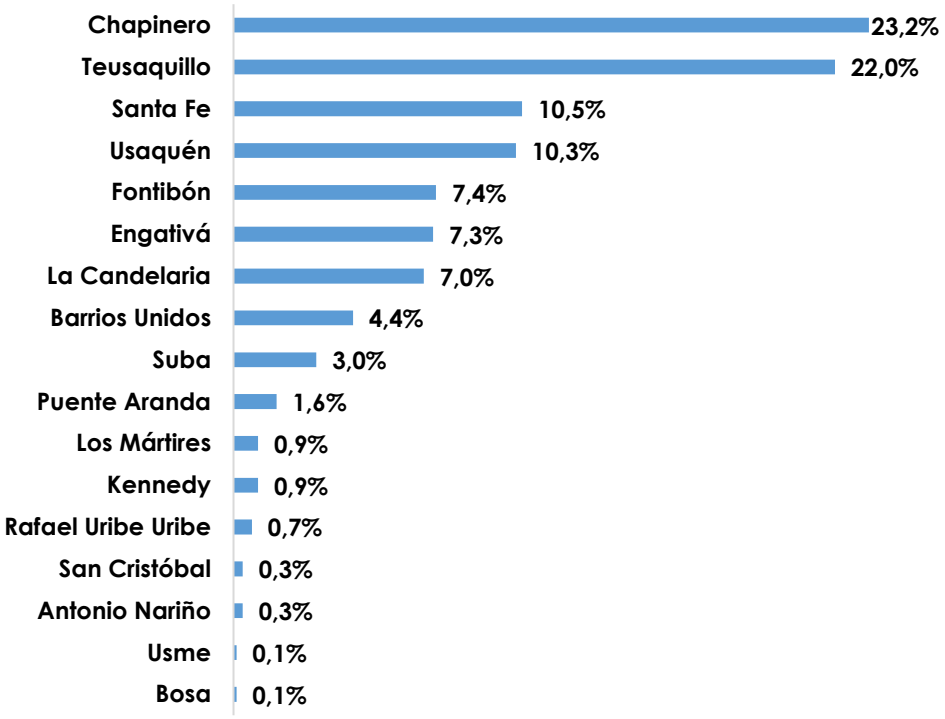
Idioma	Reseñas positivas	Reseñas negativas	Porcentaje
Español	159.267	111.347	71,2%
Inglés	31.701	22.455	14,2%
Otros	18.090	13.314	8,3%
Francés	7.242	4.951	3,2%
Portugués	7.085	4.832	3,1%

Elaboración: Subdirección de Inteligencia y Tecnologías de la Información – IDT.

4. RESULTADOS

Para el cálculo de los resultados se utilizó la información de 927 hoteles que tienen calificaciones y reseñas en la plataforma de Booking y se encuentran ubicados en Bogotá. Esta ciudad está conformada por 20 localidades y, como se observa en la gráfica 1, en 17 de ellas se encuentra ubicado alguno de los hoteles identificados. Así mismo, las cifras muestran que el 23% de los hoteles están ubicados en la localidad de Chapinero, mientras que la localidad de Teusaquillo cuenta con un 22%, Santa Fe con el 10,5% y Usaquén con el 10,3%. Es decir, cerca de 2 de cada 3 hoteles están ubicados en estas cuatro localidades. Por su parte, no se identificaron hoteles que contaran con calificaciones y/o reseñas en las localidades de Ciudad Bolívar, Tunjuelito y Sumapaz.

Gráfica 1. Porcentaje de Hoteles por localidad (2024)

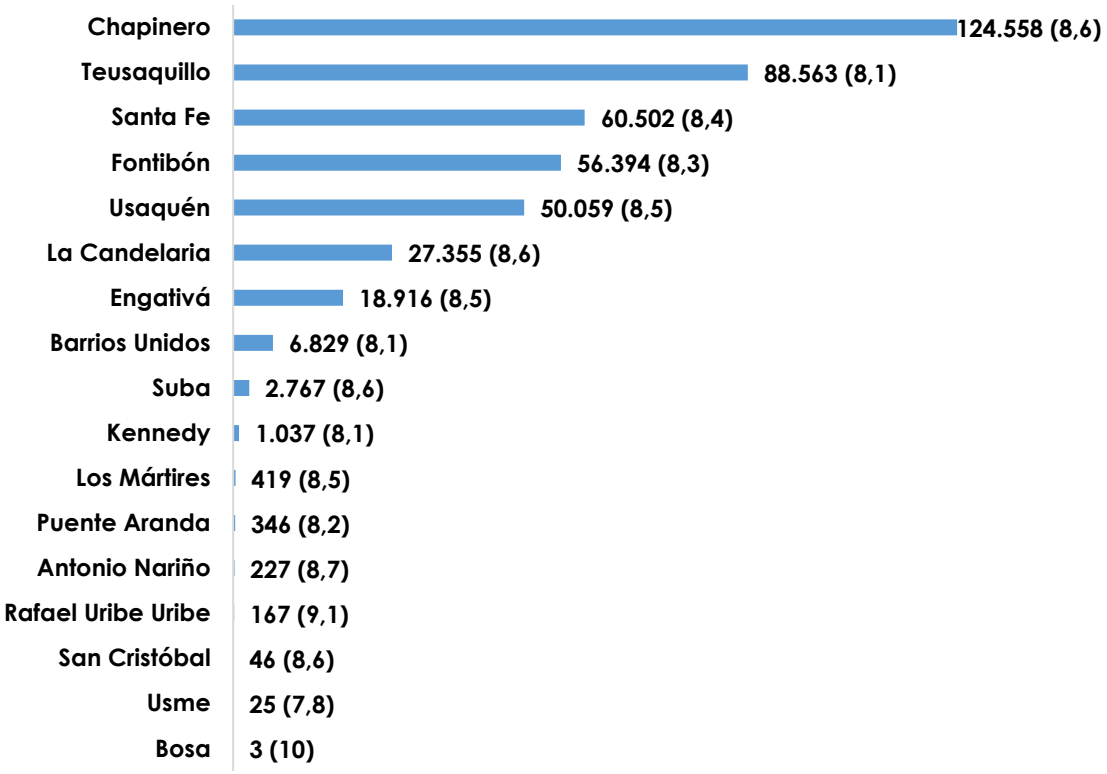


Fuente: Booking 2024.
Elaboración: Subdirección de Inteligencia y Tecnologías de la Información – IDT.

La localidad de Chapinero, que cuenta con el mayor número de hoteles, también registra el mayor número de calificaciones, superando las 124.500, con una puntuación promedio de 8,6. Le sigue la localidad de Teusaquillo,

con más de 88.500 calificaciones y un promedio de 8,1. La tercera, Santa Fe, cuenta con aproximadamente 60,500 calificaciones y un promedio de 8,4, como se puede ver en la gráfica 2. En total, las tres localidades representan el 62,4% de calificaciones recopiladas para el estudio.

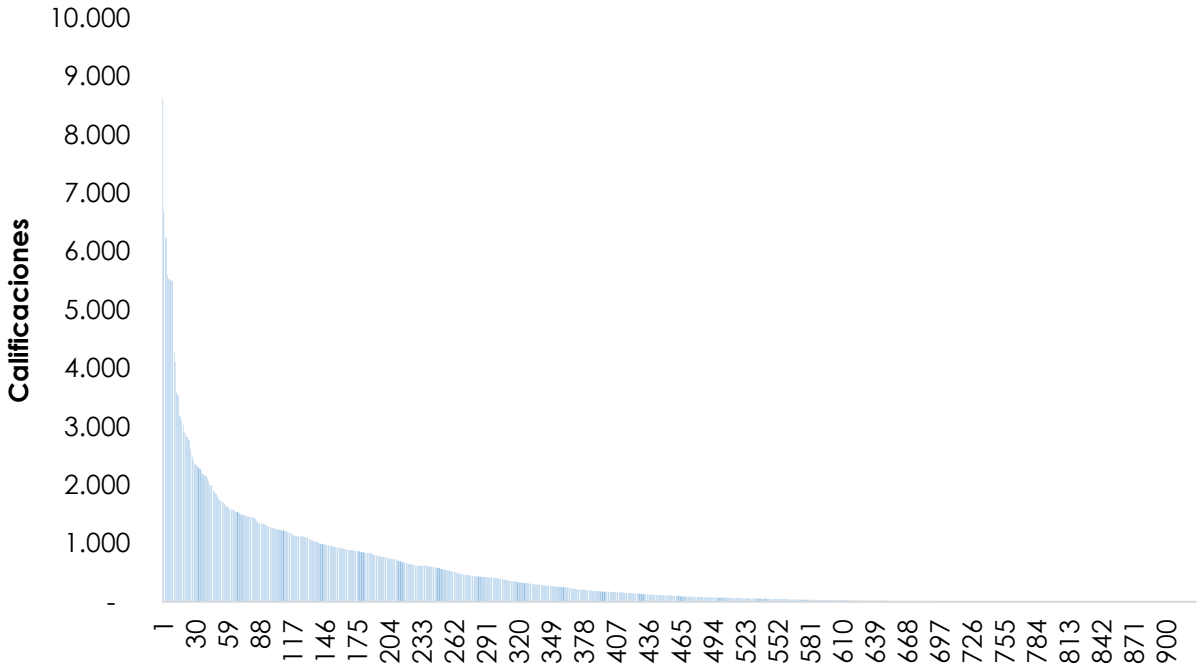
Gráfica 2. Número y promedio de calificaciones por localidad (2021-2024)



Fuente: Booking.
Elaboración: Subdirección de Inteligencia y Tecnologías de la Información – IDT.

Al analizar la distribución de calificaciones por hotel, se identificó que el hotel con más calificaciones tiene alrededor de 8.600, representando el 2% del total. Además, los 25 hoteles con más calificaciones concentran el 25% del total, los primeros 50 hoteles reúnen aproximadamente el 37% y los 100 con más calificaciones suman más del 54% del total de calificaciones descargadas. Estos resultados muestran una alta concentración de calificaciones (más del 50%) en aproximadamente el 10% de los hoteles.

Gráfica 3. Número de calificaciones por hotel (2021-2024)

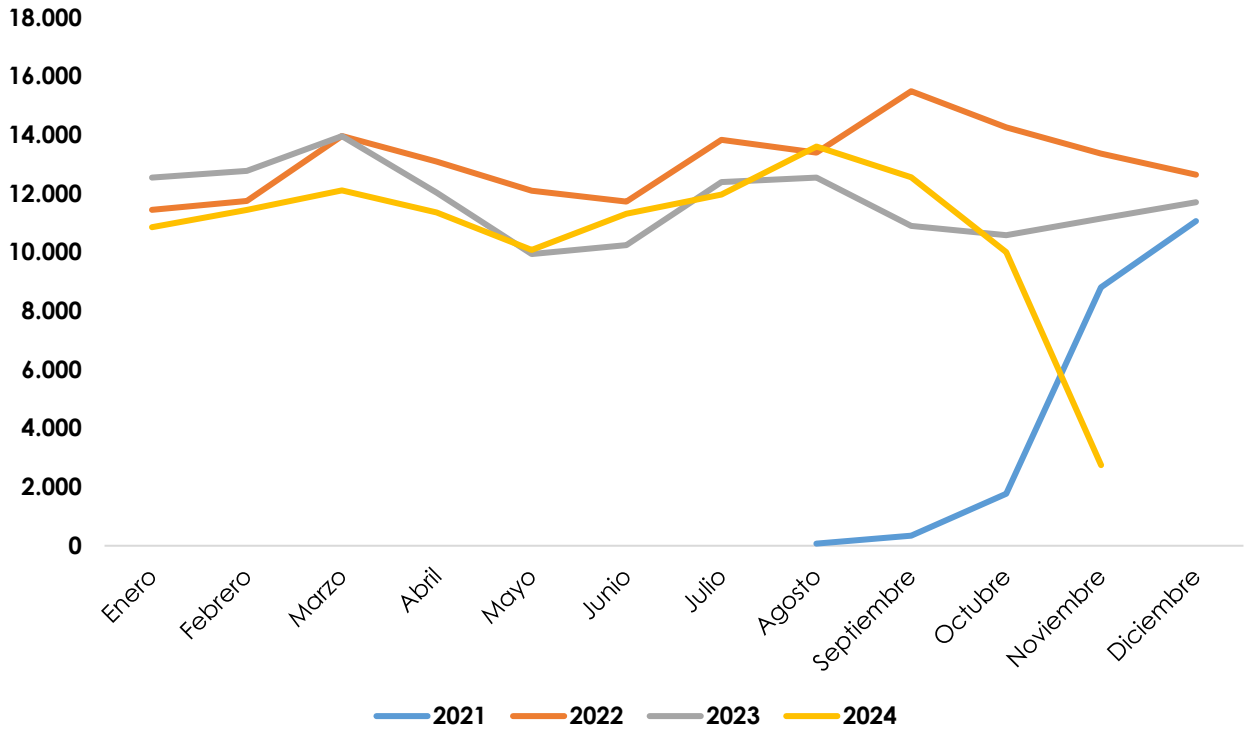


Fuente: Booking.
Elaboración: Subdirección de Inteligencia y Tecnologías de la Información – IDT.

En cuanto a los meses de hospedaje de los turistas que califican en la plataforma, se observa que, desde diciembre de 2021[6] hasta octubre de 2024, en promedio se hospedaron mensualmente cerca de 12.000 turistas. Por su parte, septiembre de 2022 fue el mes con el mayor número de turistas, llegando a más de 15.500. mientras que en mayo de 2023 y 2024 se registró el menor número, donde cerca de 10.000 turistas indicador haberse hospedado en hoteles de la ciudad.

[6] La extracción y recolección de calificaciones y datos textuales en línea del metabuscador de viajes "Booking", se realizó con corte a la primera semana de noviembre e incluye calificaciones y reseñas escritas entre octubre de 2021 y noviembre de 2024. Sin embargo, algunas calificaciones y reseñas de octubre de 2021 corresponden a estancias en hoteles ubicados en Bogotá desde agosto del 2021, mostrando un rezago de entre uno y dos meses entre el hospedaje y la realización de la calificación.

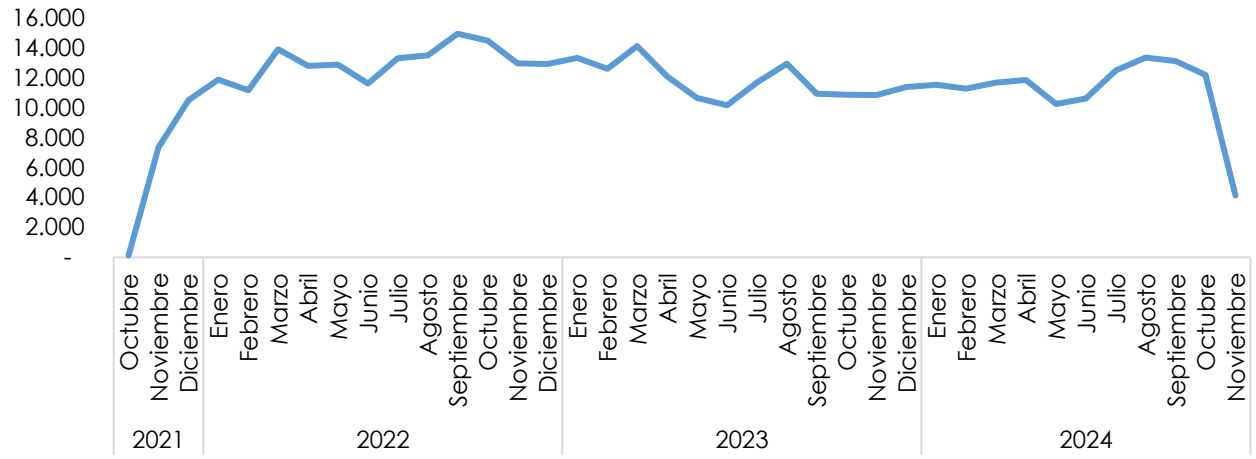
Gráfica 4. Número de hospedajes realizados por mes (2021-2024)



Fuente: Booking.
Elaboración: Subdirección de Inteligencia y Tecnologías de la Información – IDT.

Al analizar el comportamiento de las calificaciones de los hoteles de Bogotá a lo largo del tiempo, se observa que los últimos cuatro meses de 2021, mostraron un incremento mensual significativo en el número de calificaciones, resultado esperado por la reactivación del sector post pandemia. Así mismo, para los años 2022, 2023 y hasta octubre de 2024, las cifras se han mantenido entre 10.000 y 15.000 calificaciones mensuales.

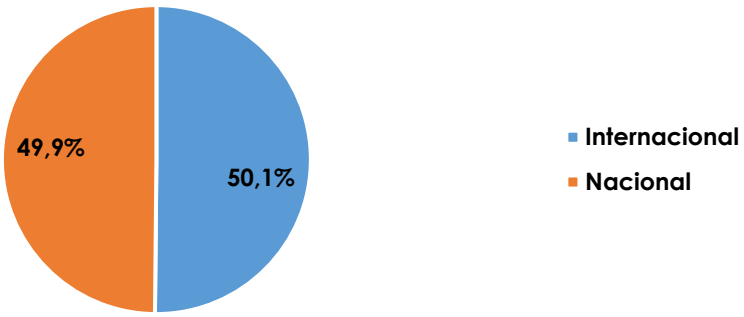
Gráfica 5. Número de calificaciones realizadas por mes (2021-2024)



Fuente: Booking.
Elaboración: Subdirección de Inteligencia y Tecnologías de la Información – IDT.

En cuanto al origen de los turistas que califican y reseñan los hoteles de Bogotá, aproximadamente el 50% son de origen nacional y el 50% de origen internacional. Sin embargo, comparando con los patrones de turismo identificados mediante la operación estadística de "viajeros en Bogotá" para visitantes (turistas y excursionistas), el porcentaje de turistas de origen nacional que se hospedan en la ciudad es mayor. Esto podría sugerir un menor uso de esta plataforma de reserva por parte de turistas nacionales.

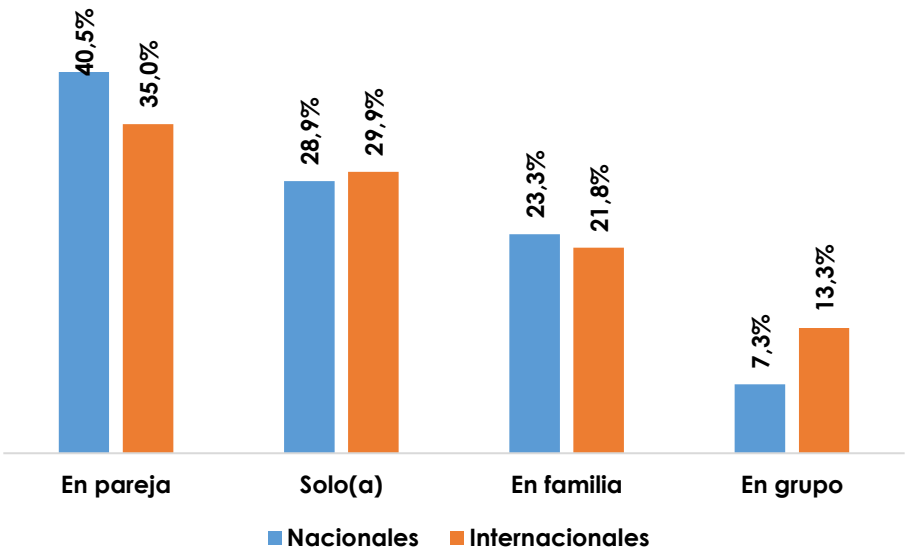
Gráfica 6. Origen de los turistas (2021-2024)



Fuente: Booking.
Elaboración: Subdirección de Inteligencia y Tecnologías de la Información – IDT.

En relación con el grupo de viaje o compañía de los turistas durante su estancia, se observa que los turistas nacionales principalmente se hospedan con su pareja (40,5% de las calificaciones), seguidos de quienes viajan solos (28,9%) y, en tercer lugar, aquellos que viajan en familia (23,3%). Para los turistas internacionales, la tendencia es similar, pero destaca un porcentaje más alto de aquellos que indican viajar en grupo (13,3%), lo que representa una diferencia de 5 puntos porcentuales respecto al mismo grupo de turistas nacionales.

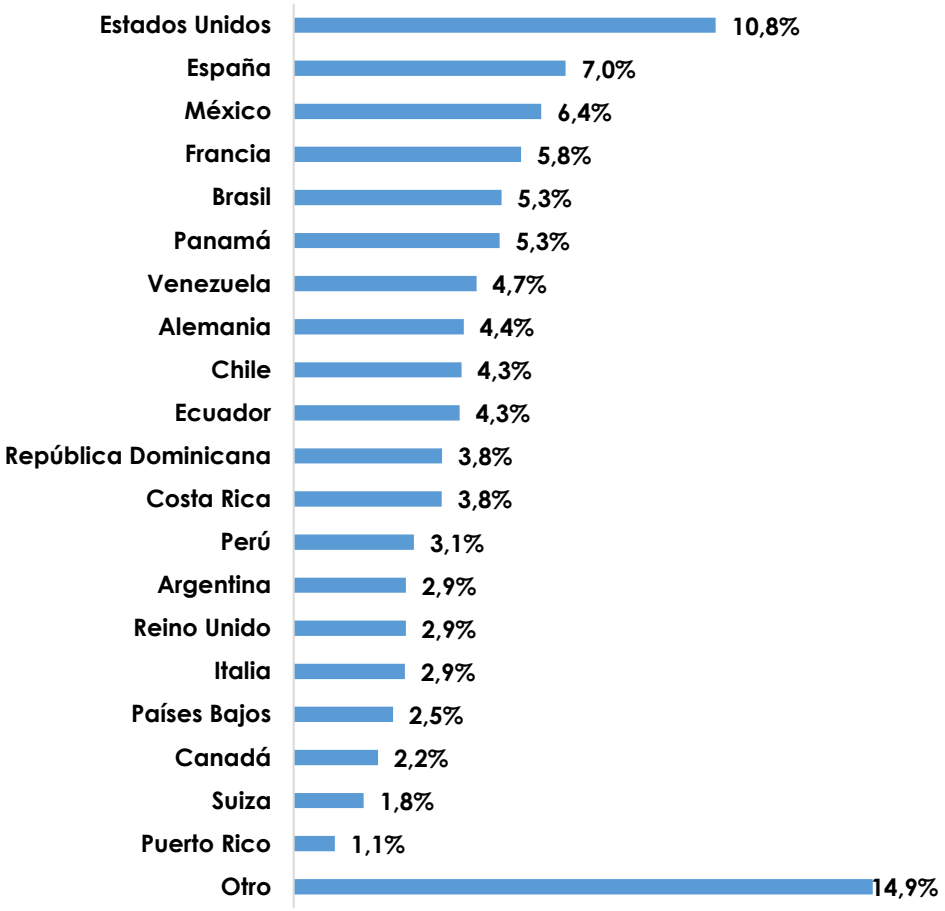
Gráfica 7. Grupo de Viaje de los turistas (2021-2024)



Fuente: Booking.
Elaboración: Subdirección de Inteligencia y Tecnologías de la Información – IDT.

Ahora bien, en cuanto al análisis sobre el origen de los turistas y su hospedaje en la oferta hotelera de Bogotá, la gráfica 8 muestra los veinte principales orígenes internacionales. En primer lugar, se encuentran los turistas provenientes de Estados Unidos, que representan un 10,8%, seguidos por los turistas españoles, con un 7%, y mexicanos, con un 6,4%. Los primeros 20 países comprenden el 85,1% del total de calificaciones recopiladas para el estudio, dejando un 14,9% correspondiente al resto de países de origen de los turistas que se hospedan en Bogotá.

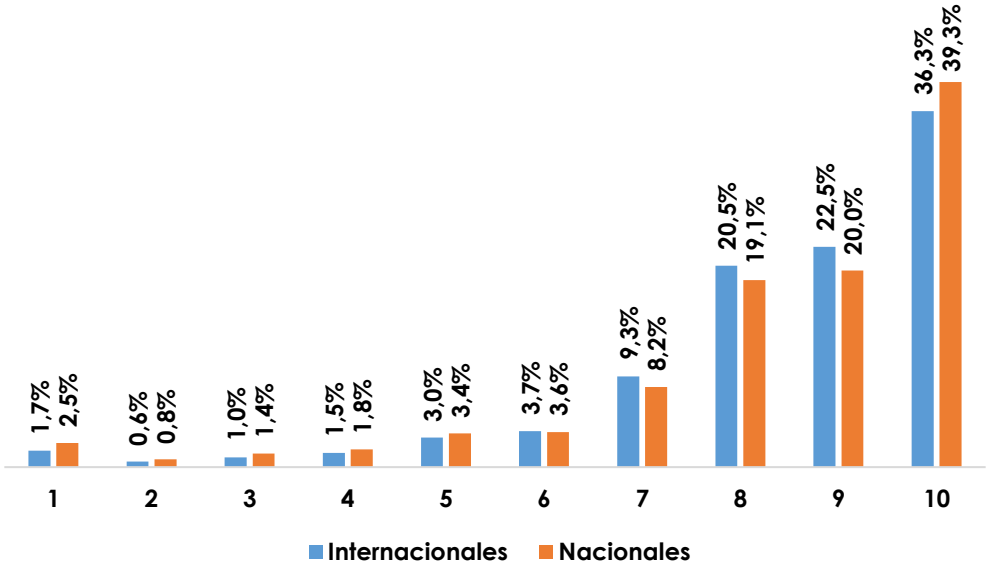
Gráfica 8. Principales países de origen de los turistas (2021-2024)



Fuente: Booking.
Elaboración: Subdirección de Inteligencia y Tecnologías de la Información – IDT.

A partir del análisis de las calificaciones de los usuarios a cada atractivo turístico, se construyeron diferentes categorías para medir sus resultados. Por ejemplo, valores menores o iguales a tres, se clasifican como negativas; valores de cuatro a siente como neutrales y; valores mayores a 7, como positivas. Los resultados de esta clasificación indican que más del 78% de las calificaciones son positivas para turistas tanto nacionales como internacionales, mientras que 17% son neutras y menos del 5% son negativas, lo cual indica un desbalance en la base de datos.

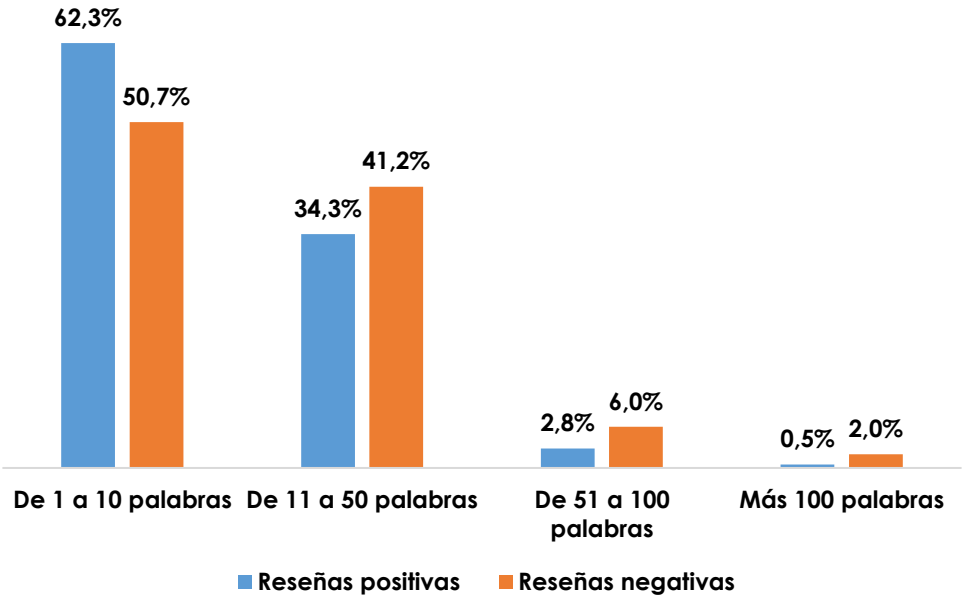
Gráfica 9. Clasificación de las reseñas por origen (2021-2024)



Fuente: Booking.
Elaboración: Subdirección de Inteligencia y Tecnologías de la Información – IDT.

Respecto a la longitud de las reseñas, evaluada por el número de palabras que contiene cada una, se observa que un 62,3% las reseñas positivas tienen entre 1 y 10 palabras, mientras que para las reseñas negativas, este porcentaje es de 50,7%. Por el contrario, el 34,3% de reseñas positivas tienen entre 11 a 50 palabras, mientras que n las reseñas negativas este porcentaje aumenta a 41,2%, sugiriendo que las reseñas negativas suelen ser más detalladas.

Gráfica 10. Longitud de las reseñas por tipo (2021-2024)



Fuente: Booking.
Elaboración: Subdirección de Inteligencia y Tecnologías de la Información – IDT.

Finalmente, se realizó un análisis del texto de las reseñas utilizando técnicas de Procesamiento de Lenguaje Natural (NLP). Mediante redes de coocurrencia, se identificaron diferentes aspectos (problemáticas y fortalezas) asociados a las reseñas positivas y negativas según las distintas variables presentes en la base de datos. La tabla 5 presenta el cruce de variables propuesto para los análisis iniciales de redes de coocurrencia, así como el total de reseñas por tipo.

Las redes de coocurrencia son herramientas utilizadas para proporcionar una visualización gráfica de las relaciones potenciales entre personas, organizaciones, conceptos u otras entidades representadas en el material escrito. En el análisis de textos de este estudio, la coocurrencia puede interpretarse como un indicador de proximidad semántica, revelando patrones y análisis estadísticos dentro de un idioma.

Tabla 5. Variables para los análisis de redes de coocurrencia

	Reseñas		Reseñas Totales	
Variable	**Negativo**	**Positivo**	**Positivas**	**Negativas**
Origen	Nacional	Nacional	93.274	65.918
	Internacional	Internacional	65.995	45.430
Idioma	Ingles	Ingles	31.701	22.455
Localidad	Chapinero	Chapinero	43.348	29.791
Grupo Viaje	En pareja	En pareja	59.447	42.823
Calificación	Intervalo de: 8 a 10	Intervalo de: 1 a 3	68.635	6.467
Año	2024	2024	47.588	32.901

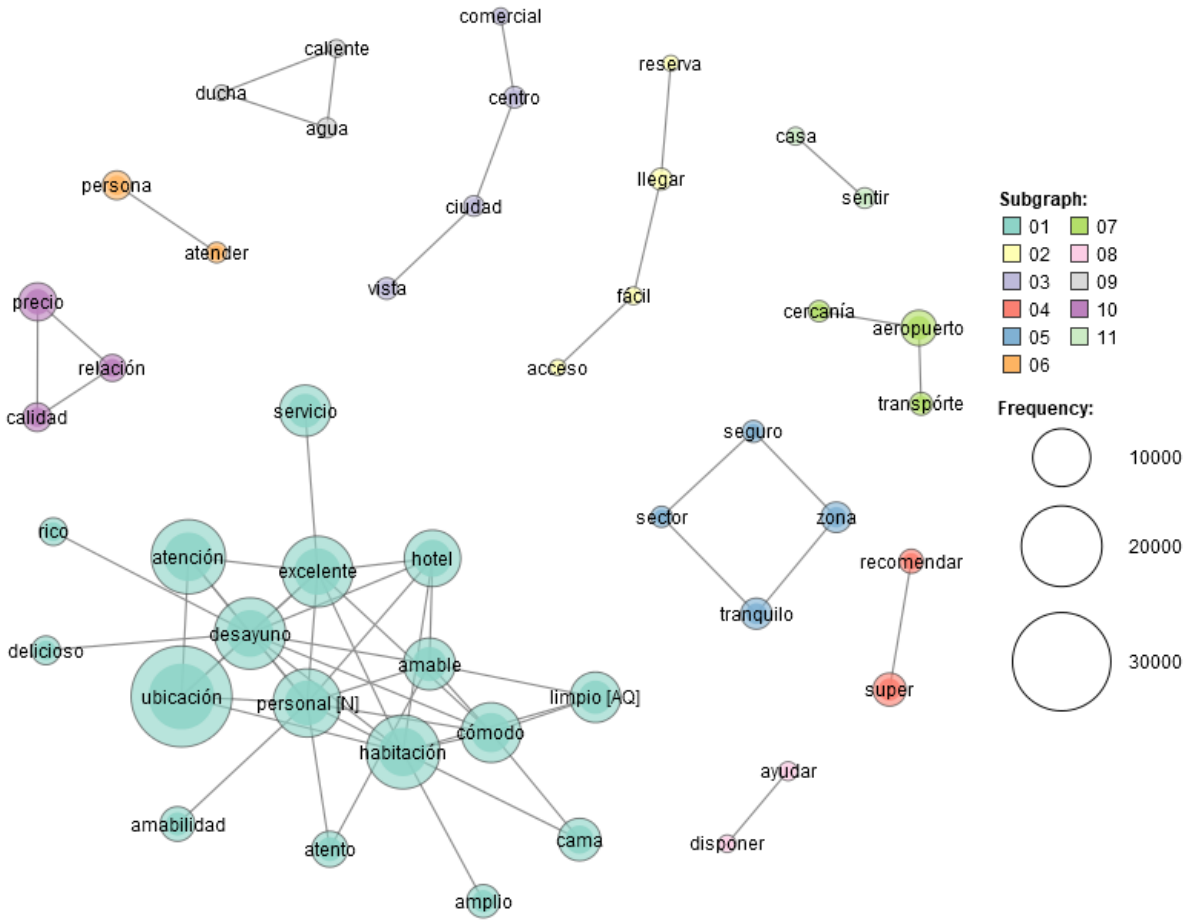
Elaboración: Subdirección de Inteligencia y Tecnologías de la Información – IDT.

*Para todas las variables, exceptuando la del idioma el análisis se realiza utilizando las reseñas escritas en español.

Este conjunto de palabras identificado para los distintos cruces de variables y tipo de reseñas se ilustran en 14 redes de coocurrencia (Figura 2 a 14). En general, se identificó que en las reseñas positivas las palabras "ubicación", "tranquilidad y seguridad", "atención y servicio" son de gran relevancia entre los turistas nacionales e internacionales. Por otro lado, las reseñas negativas para estos dos grupos de turistas, destacan palabras relacionadas con la habitación como "ruido", "tamaño", "comodidades del baño" y "limpieza".

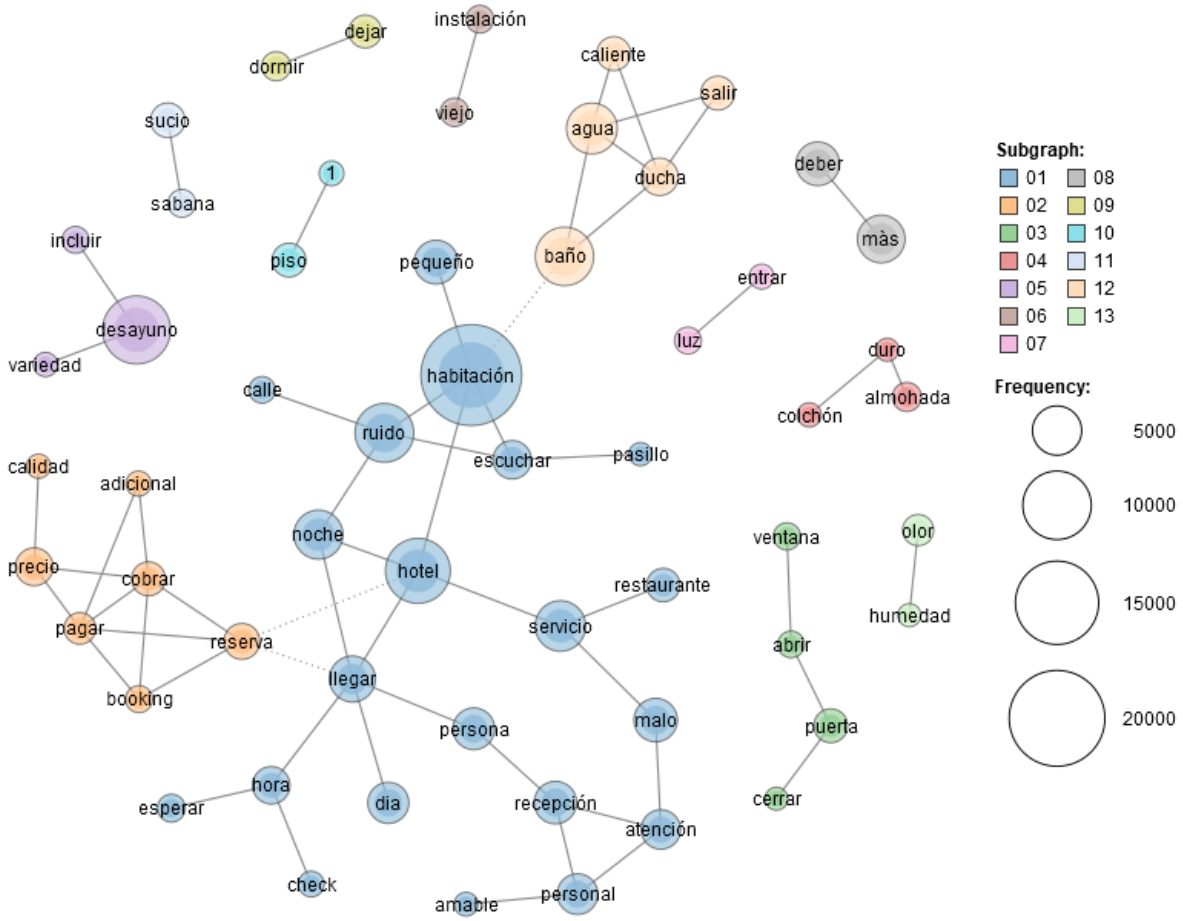
Cuando se analiza específicamente las reseñas de los hoteles ubicados en la localidad de Chapinero, se evidencia que en las reseñas positivas destaca uno de los atractivos turísticos del sector, como lo es la "Zona T", conocida por sus bares y restaurantes. En cambio, en las reseñas negativas, palabras como "variedad" e "incluir", relacionadas con el "desayuno", son prominentes.

Figura 2. Red de coocurrencia de las reseñas positivas de los turistas de origen Nacional (2021-2024)



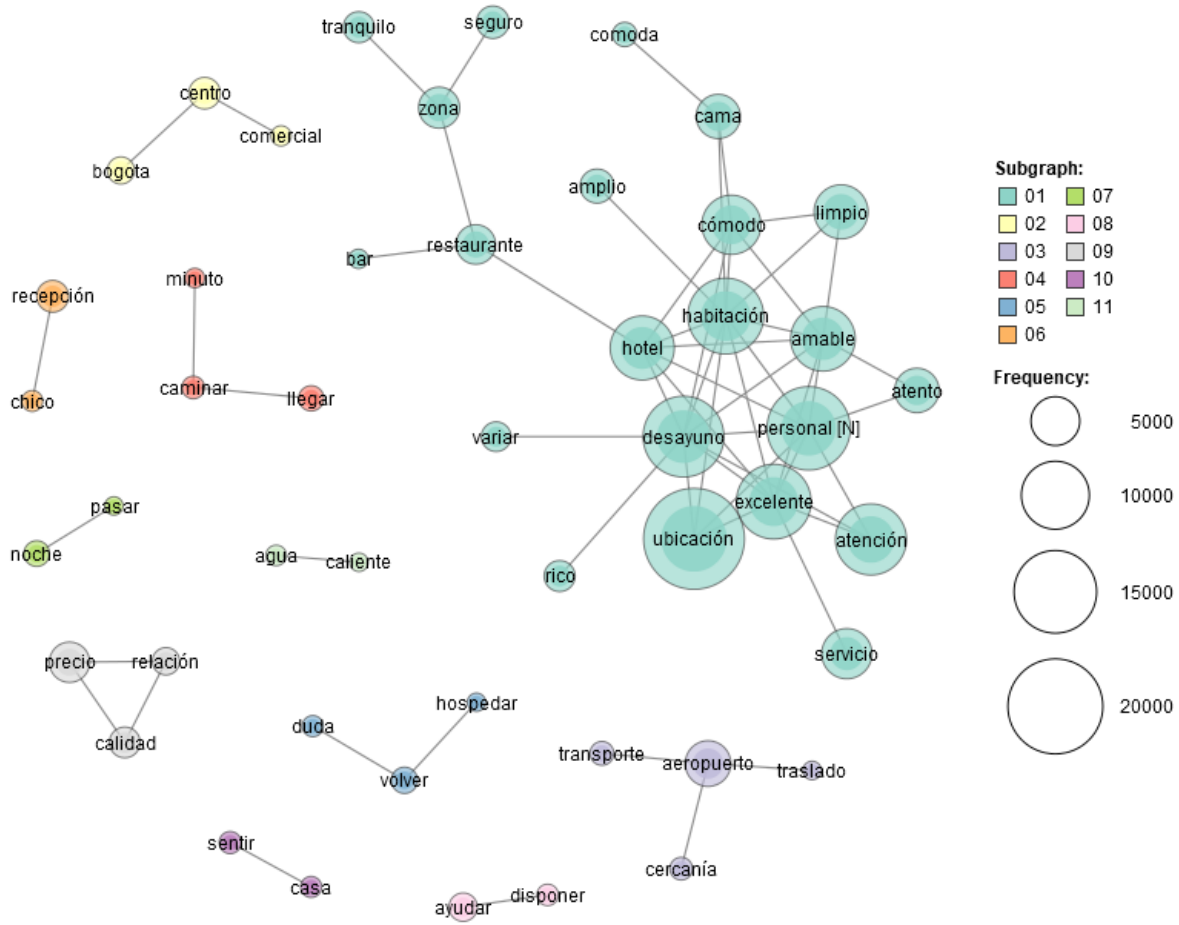
Fuente: Booking. / Elaboración: Subdirección de Inteligencia y Tecnologías de la Información – IDT.

Figura 3. Red de coocurrencia de las reseñas negativas de los turistas de origen Nacional (2021-2024)



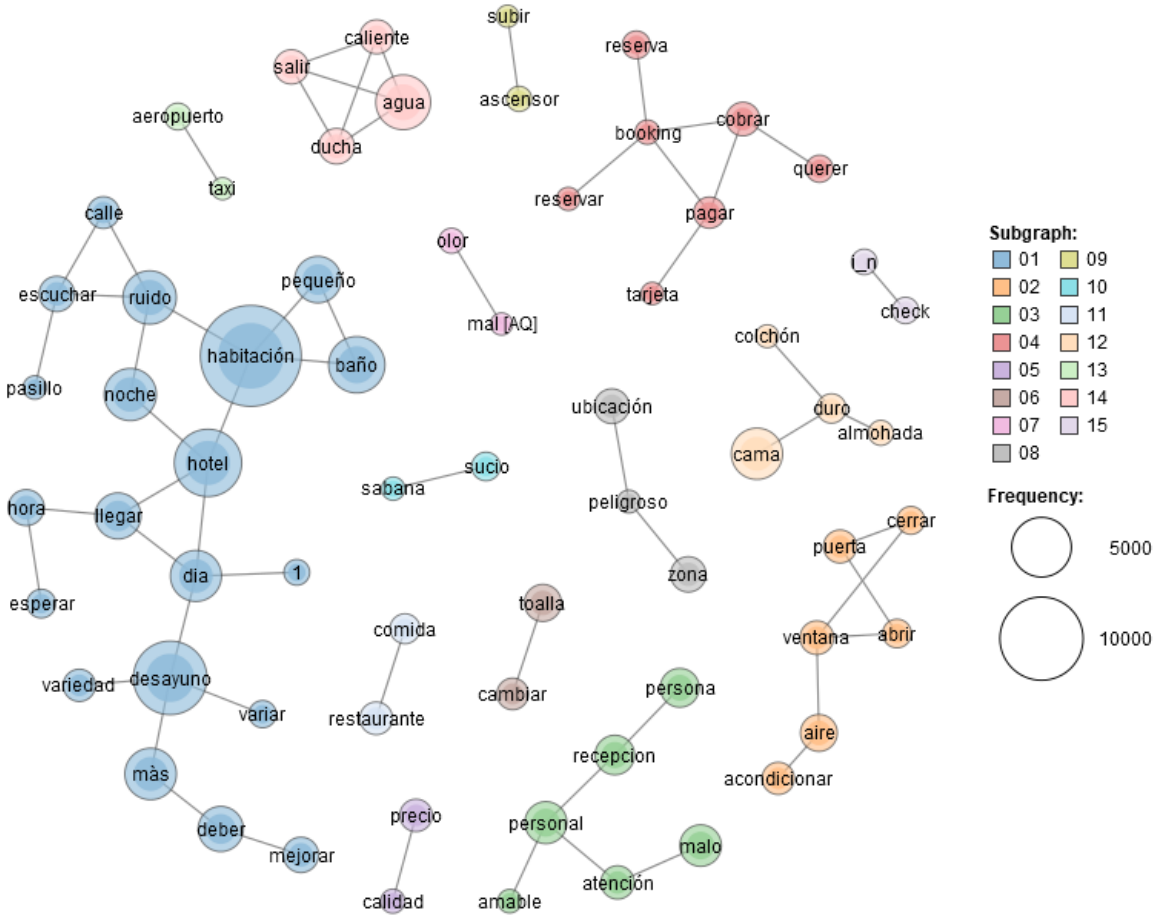
Fuente: Booking. / Elaboración: Subdirección de Inteligencia y Tecnologías de la Información – IDT.

Figura 4. Red de coocurrencia de las reseñas positivas de los turistas de origen Internacional (2021-2024)



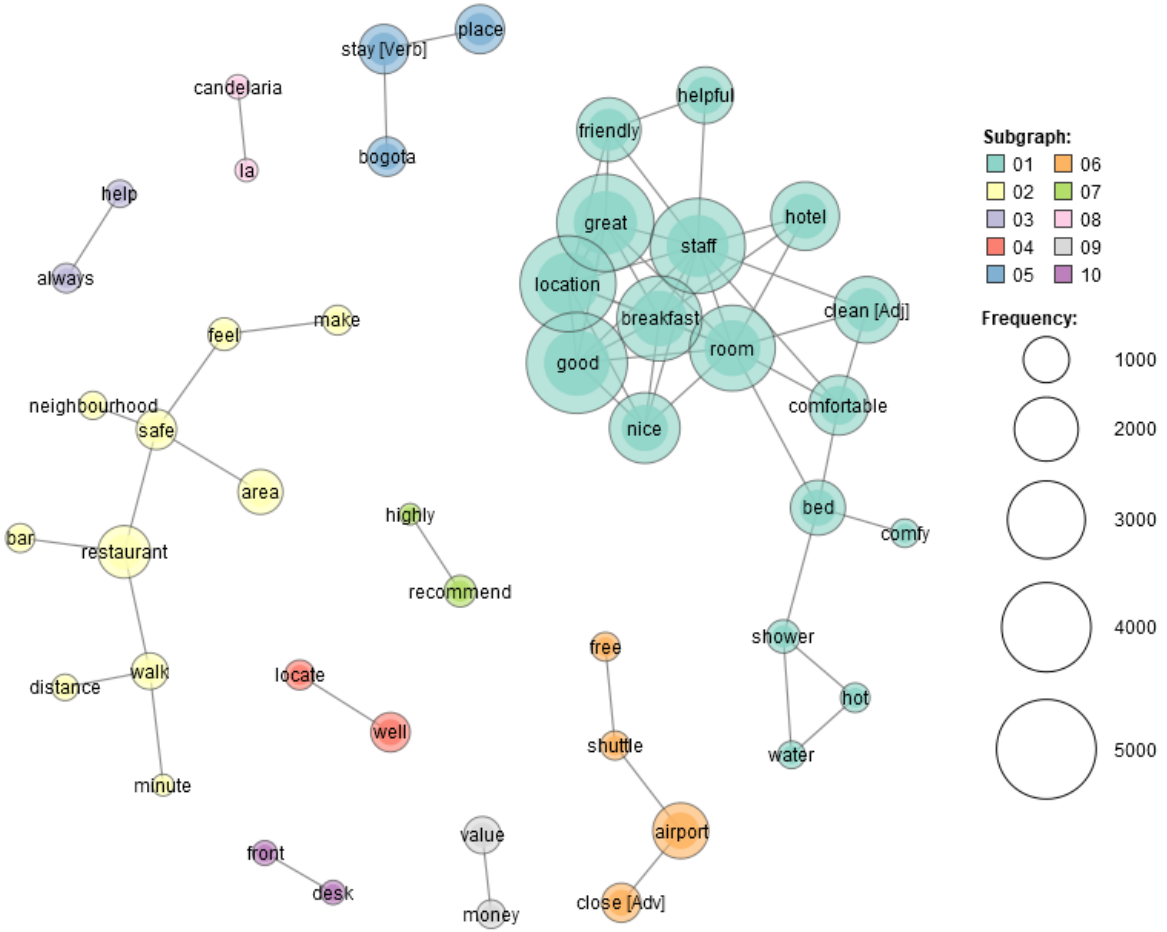
Fuente: Booking. / Elaboración: Subdirección de Inteligencia y Tecnologías de la Información – IDT.

Figura 5. Red de coocurrencia de las reseñas negativas de los turistas de origen Internacional (2021-2024)



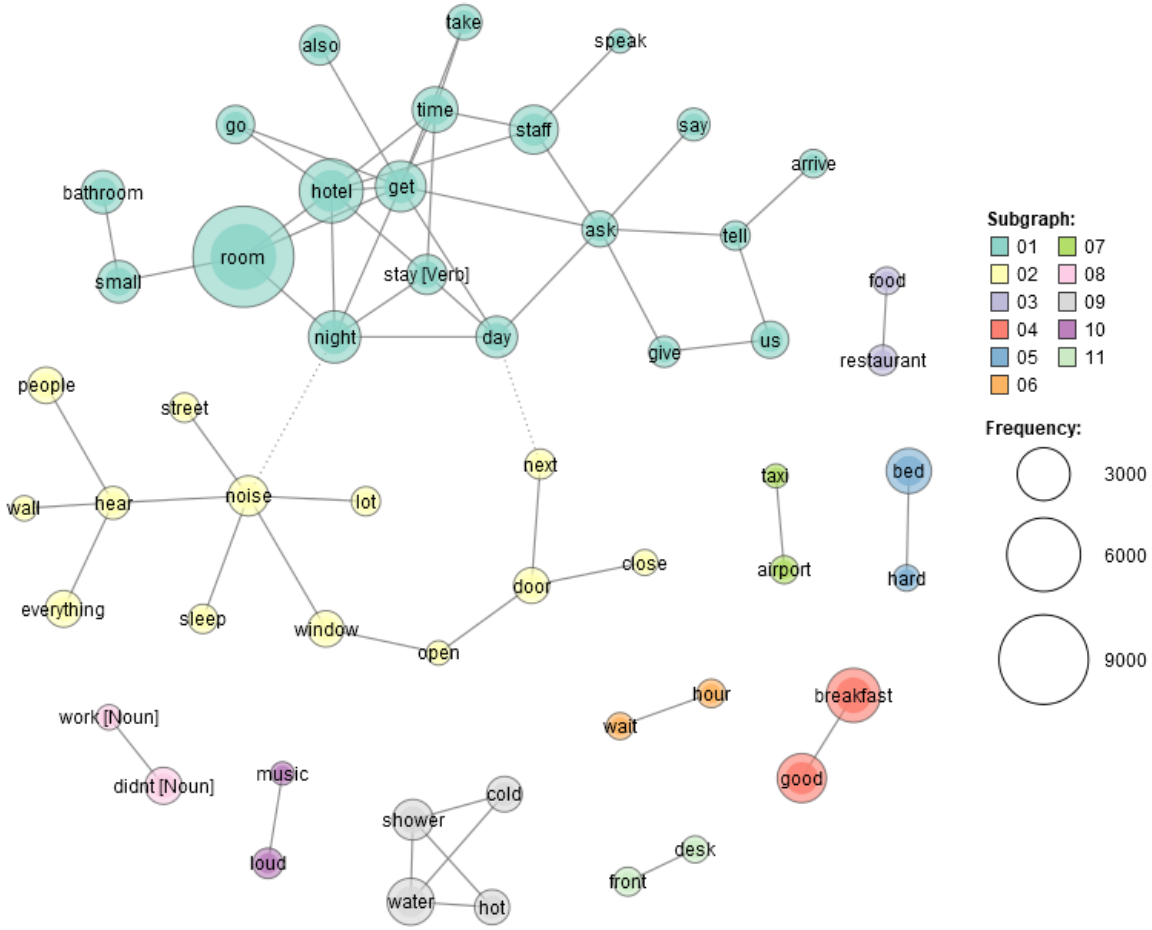
Fuente: Booking. / Elaboración: Subdirección de Inteligencia y Tecnologías de la Información – IDT.

Figura 6. Red coocurrencia de las reseñas positivas en ingles de los turistas (2021-2024)



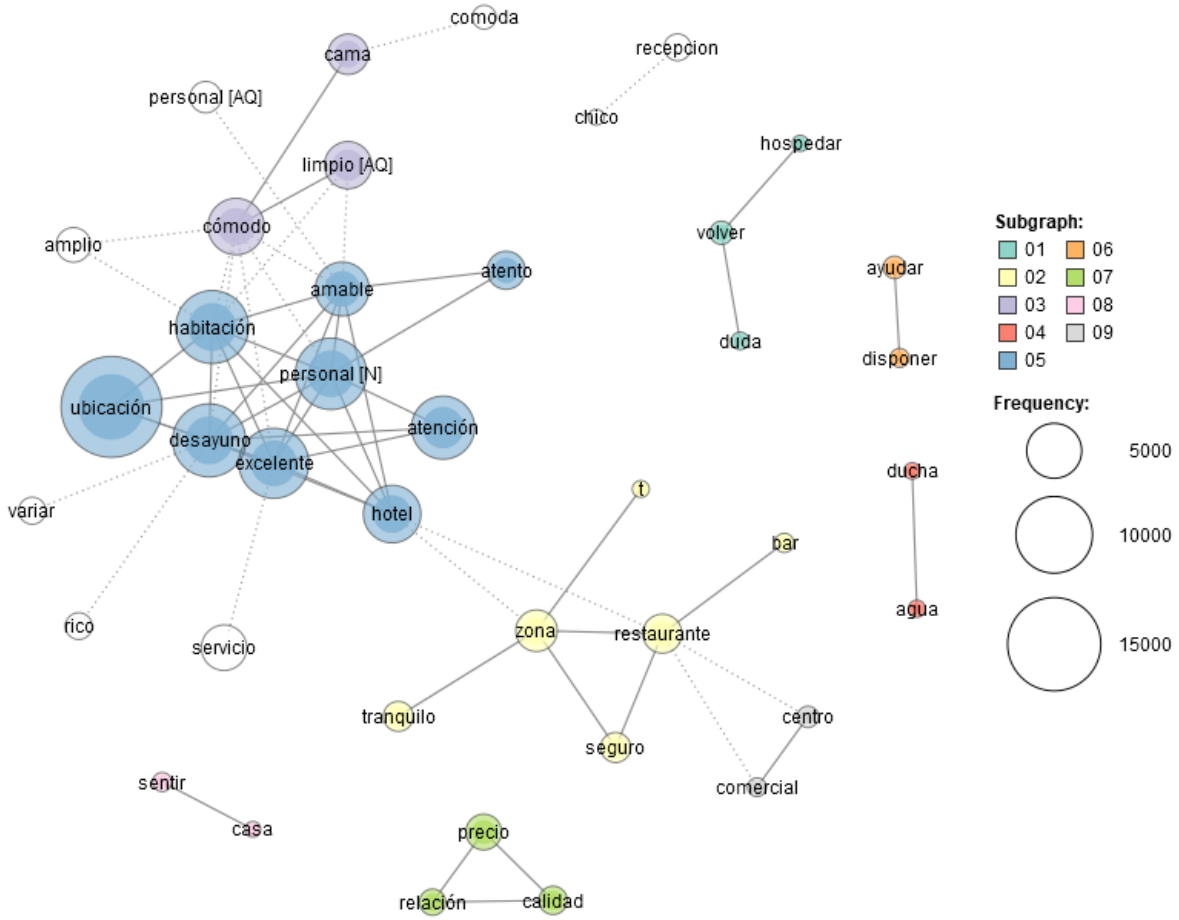
Fuente: Booking. / Elaboración: Subdirección de Inteligencia y Tecnologías de la Información – IDT.

Figura 7. Red de coocurrencia de las reseñas negativas en ingles de los turistas (2021-2024)



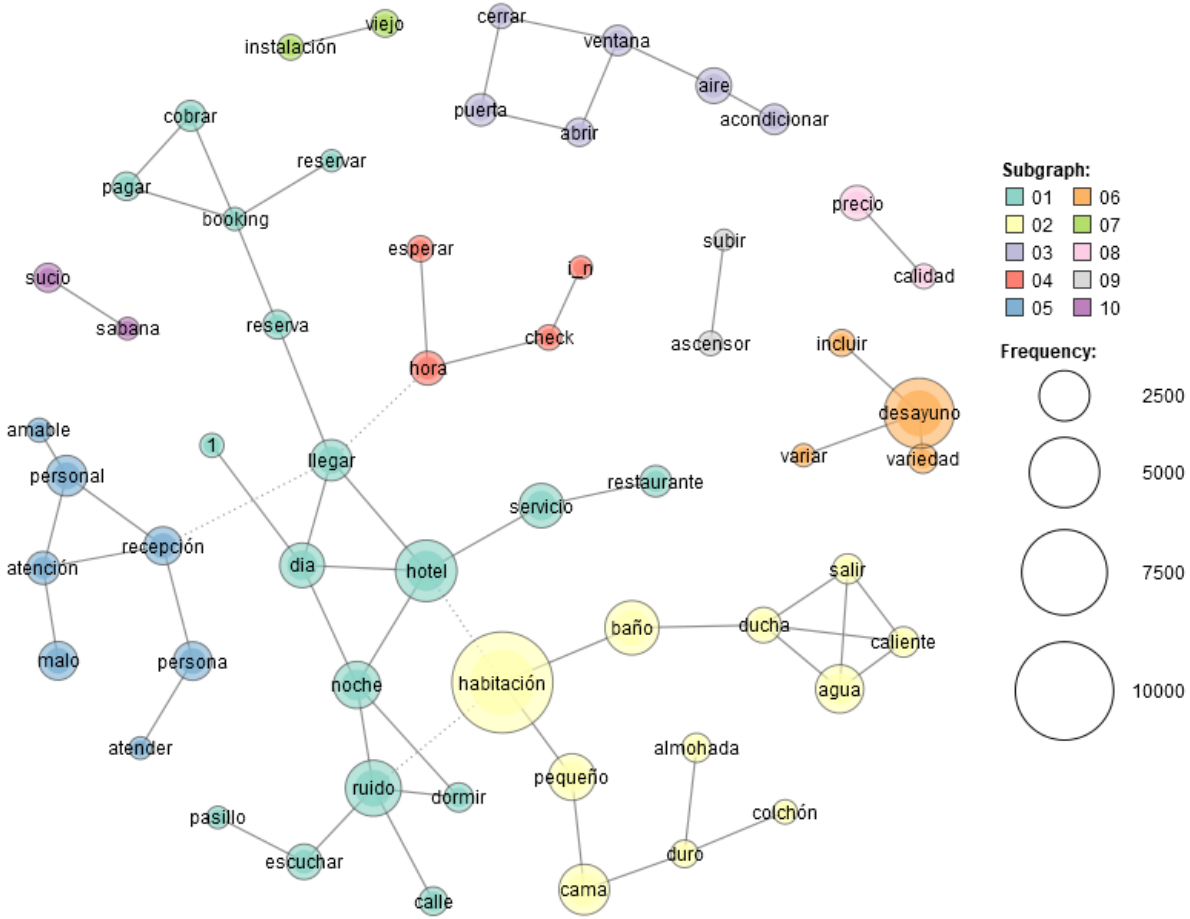
Fuente: Booking. / Elaboración: Subdirección de Inteligencia y Tecnologías de la Información – IDT.

Figura 8. Red de coocurrencia de las reseñas positivas para los hoteles ubicados en Chapinero (2021-2024)



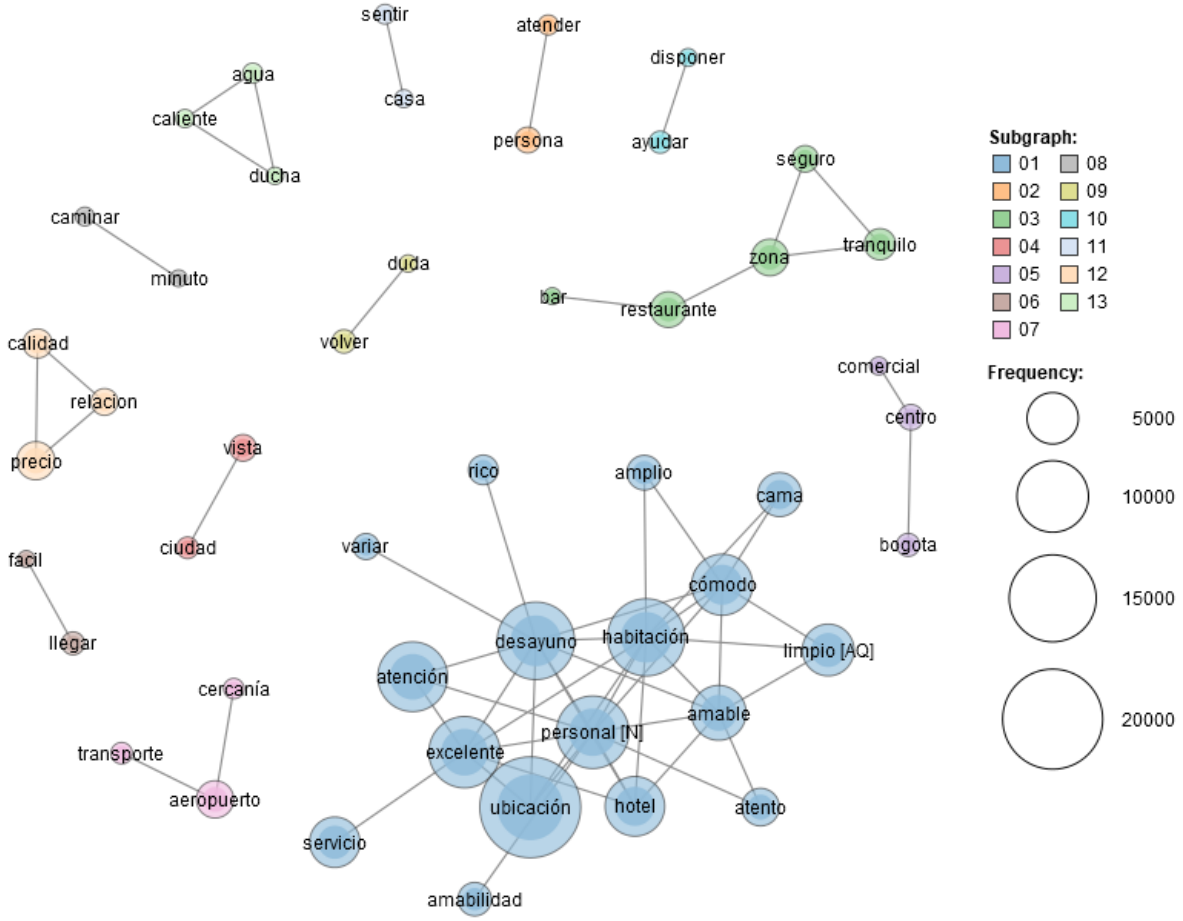
Fuente: Booking. / Elaboración: Subdirección de Inteligencia y Tecnologías de la Información – IDT.

Figura 9. Red de coocurrencia de las reseñas negativas para los hoteles ubicados en Chapinero (2021-2024)



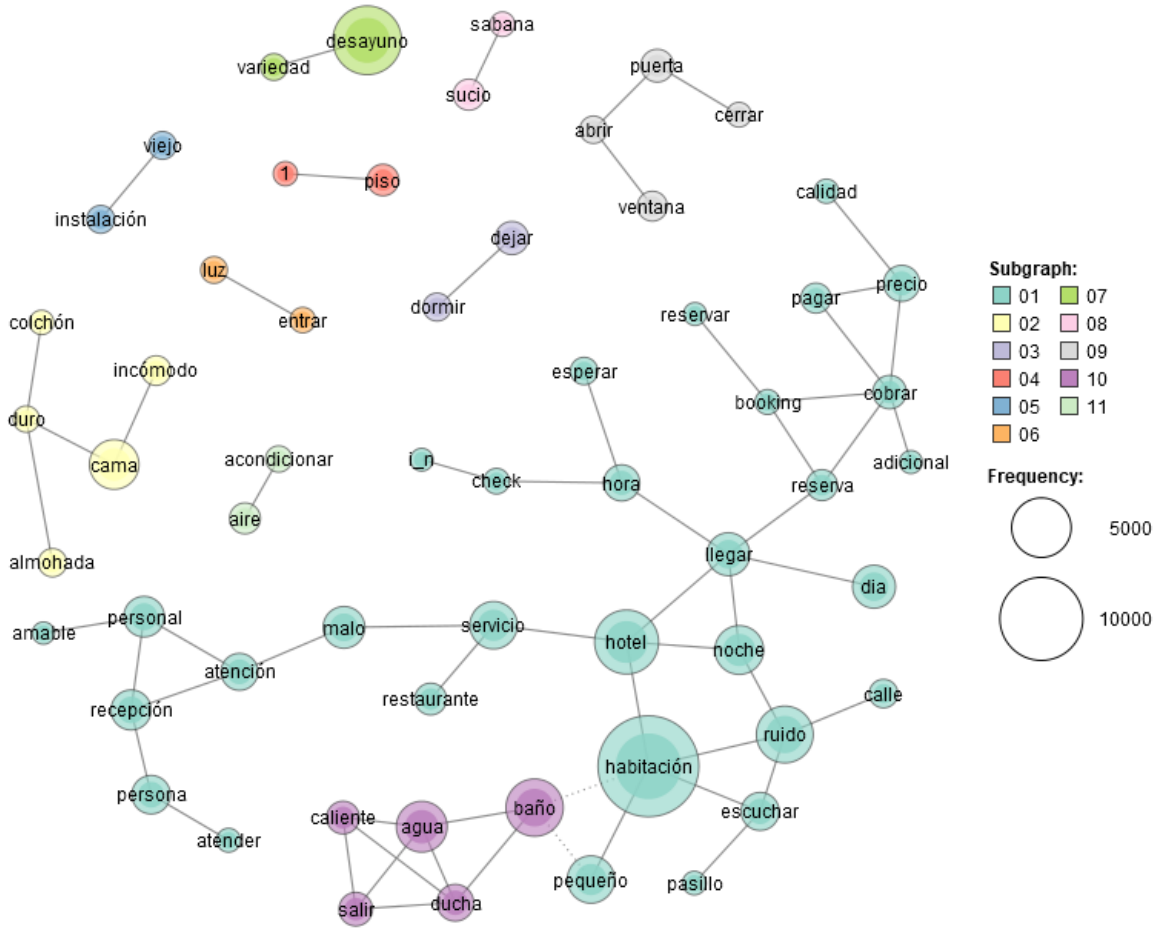
Fuente: Booking. / Elaboración: Subdirección de Inteligencia y Tecnologías de la Información – IDT.

Figura 10. Red de coocurrencia de las reseñas positivas de los turistas que viajan en pareja (2021-2024)



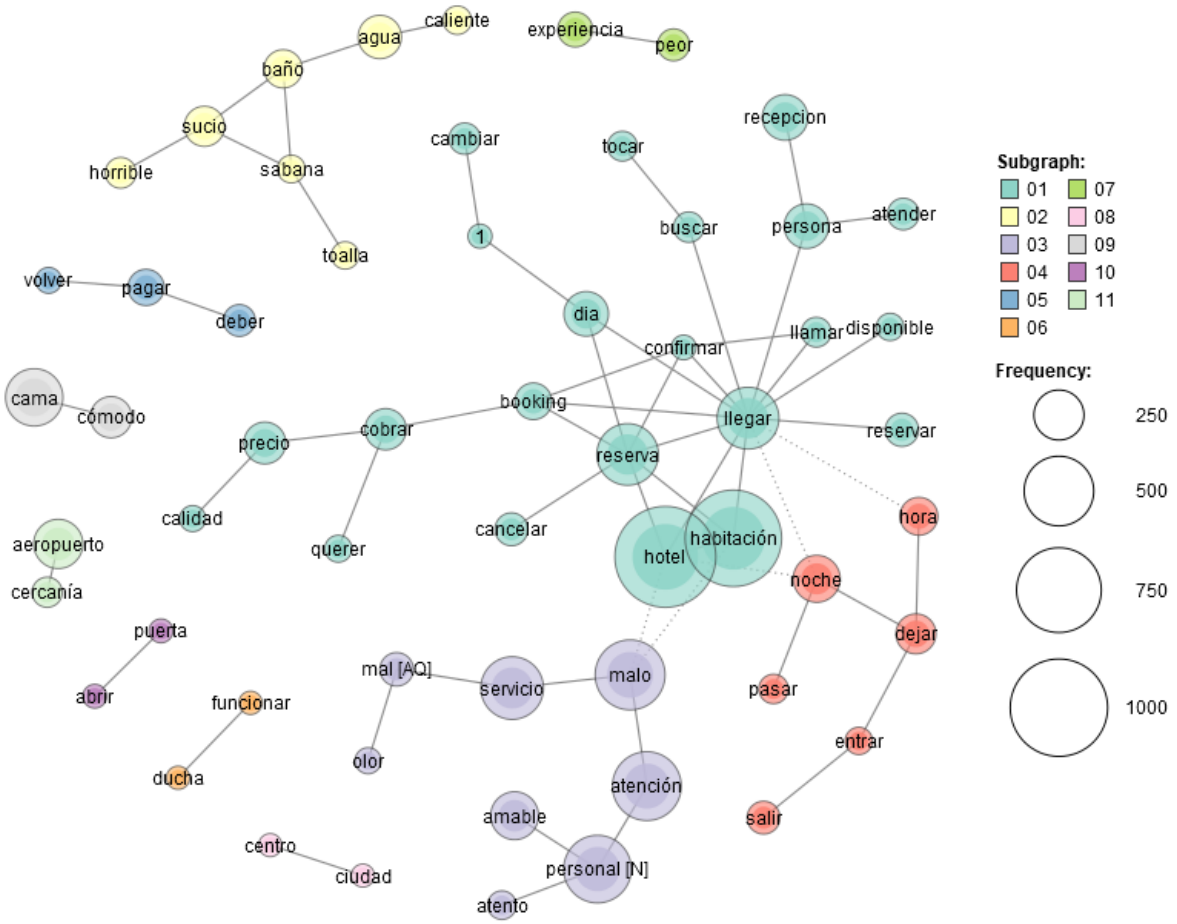
Fuente: Booking. / Elaboración: Subdirección de Inteligencia y Tecnologías de la Información – IDT.

Figura 11. Red de coocurrencia de las reseñas negativas de los turistas que viajan en pareja (2021-2024)



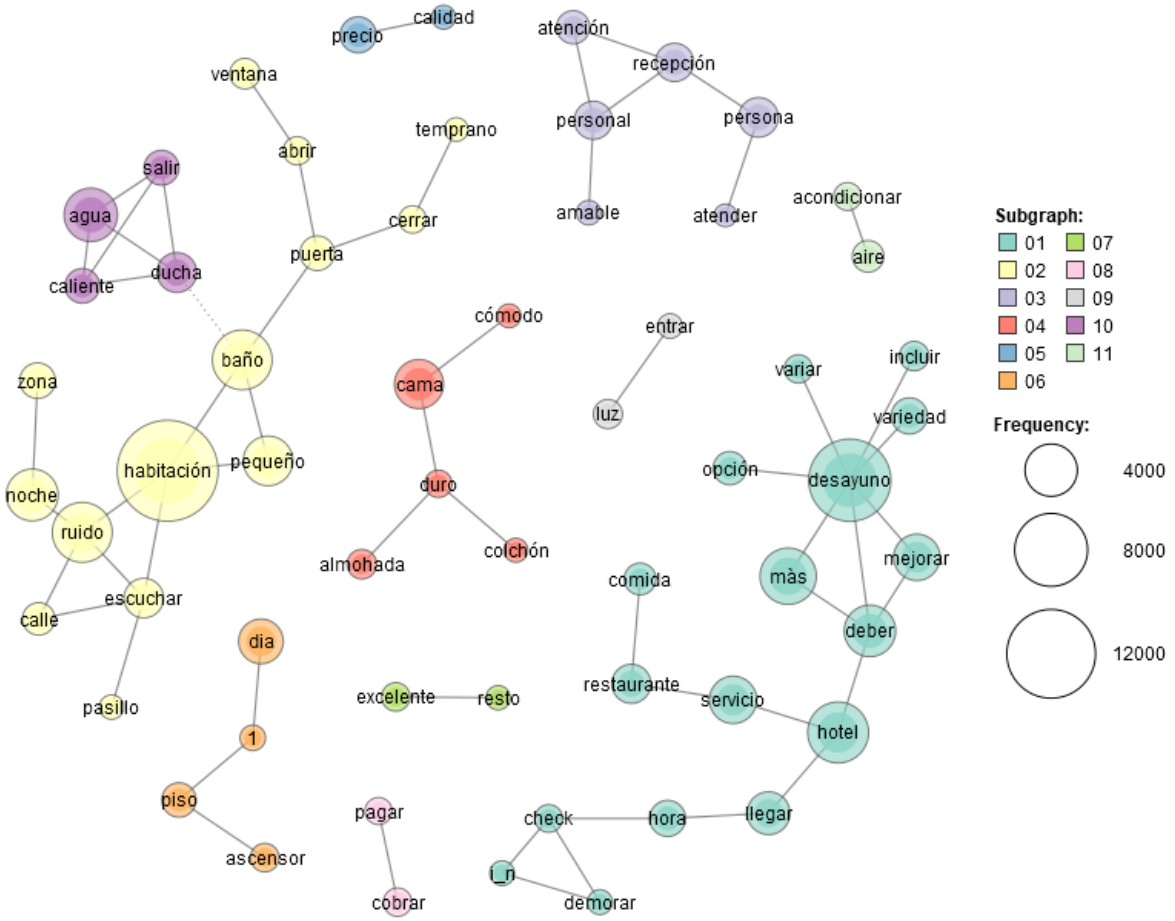
Fuente: Booking. / Elaboración: Subdirección de Inteligencia y Tecnologías de la Información – IDT.

Figura 12. Red de coocurrencia de las reseñas positivas de los turistas que calificaron mal al hotel donde se alojaron (1 a 3) / (2021-2024)



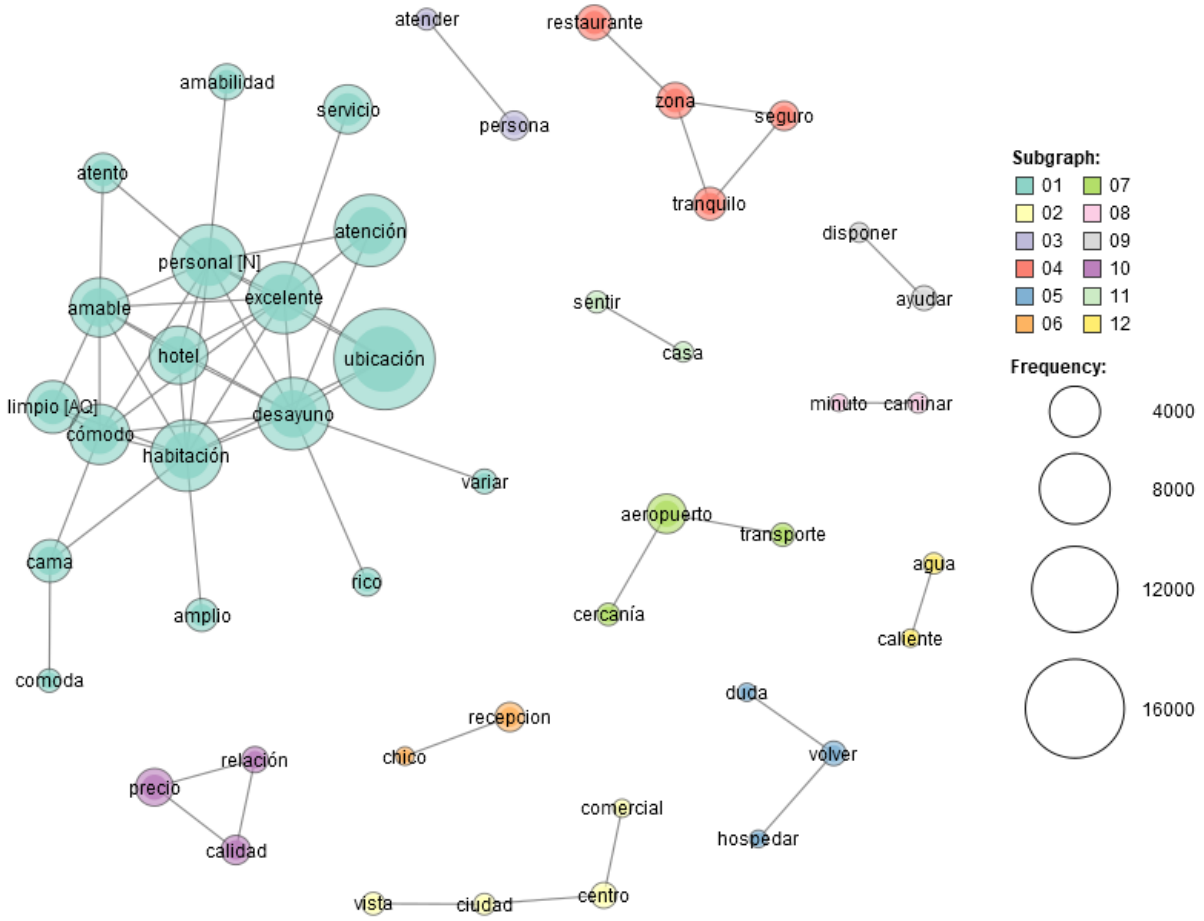
Fuente: Booking. / Elaboración: Subdirección de Inteligencia y Tecnologías de la Información – IDT.

Figura 13. Red de coocurrencia de las reseñas negativas de los turistas que calificaron bien al hotel donde se alojaron (8 a 10) / (2021-2024)



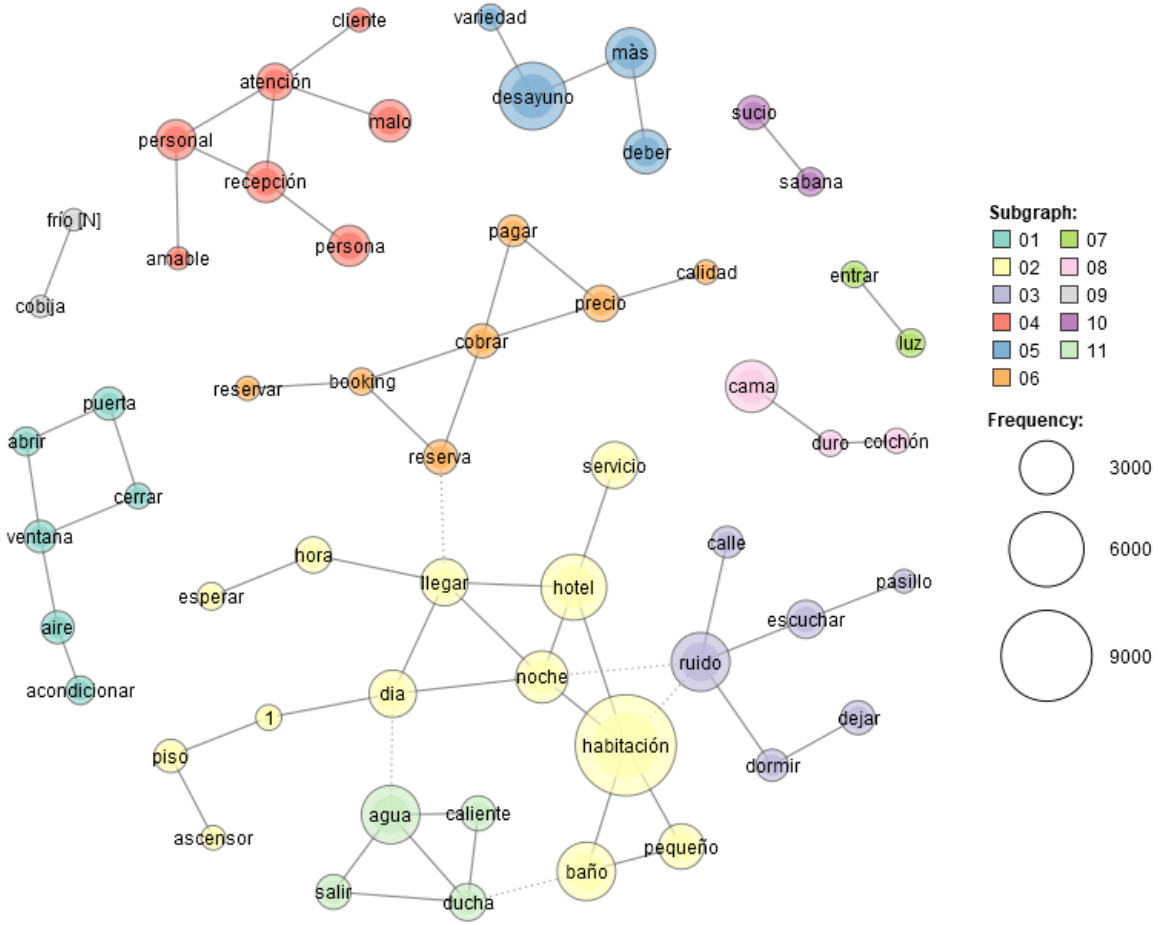
Fuente: Booking. / Elaboración: Subdirección de Inteligencia y Tecnologías de la Información – IDT.

Figura 14. Red de coocurrencia de las reseñas positivas de los turistas (2024)



Fuente: Booking. / Elaboración: Subdirección de Inteligencia y Tecnologías de la Información – IDT.

Figura 15. Red de coocurrencia de las reseñas negativas de los turistas (2024)



Fuente: Booking. / Elaboración: Subdirección de Inteligencia y Tecnologías de la Información – IDT.

5. CONCLUSIONES

- La utilización de herramientas para el Big Data y la ciencia de datos en la investigación turística de Bogotá desempeña un papel fundamental para obtener información primaria. Su aplicación en diversos estudios proporciona información valiosa que no puede obtenerse únicamente mediante formularios tradicionales, generando así datos complementarios a los producidos por el Observatorio de Turismo de Bogotá.

- De los 927 hoteles en Bogotá que tienen calificaciones y reseñas en la plataforma de Booking, la localidad de Chapinero alberga el mayor porcentaje de hoteles (23,2%), seguida de Teusaquillo con un 22%.

- Chapinero cuenta con el mayor número de calificaciones, con más de 124.500 y un promedio de 8,6. Le sigue Teusaquillo, con más de 88.500 calificaciones y un promedio de 8,1.

- Entre 2022 y octubre de 2024, el número de calificaciones de turistas hospedados en la ciudad ha oscilado entre 10.000 y 15.500 mensuales.

- Estados Unidos es el principal país de origen de los turistas internacionales, representando el 10,8% del total, seguido por turistas de España (7%) y México (6,4%).

- Tanto para turistas nacionales como internacionales, más del 78% de las calificaciones son positivas, cerca del 17% son neutras y menos del 5% son negativas.

INDICE DE TABLAS, FIGURAS Y GRÁFICOS

Índice de Tablas

Tabla 1. Trabajos en el que se utilizan las reseñas de los usuarios para el análisis la oferta alojamiento. 7
Tabla 2. Estructura de los datos que se extrajeron 12
Tabla 3. Total de Calificaciones extraídas a los hoteles 14
Tabla 4. Idioma de las reseñas por tipo 14
Tabla 5. Variables que se proponen para los análisis de redes de coocurrencia .. 23

Índice de Figuras

Figura 1. Big Data en la investigación turística 5
Figura 2. Red de coocurrencia de las reseñas positivas de los turistas de origen Nacional (2021-2024) 25
Figura 3. Red de coocurrencia de las reseñas negativas de los turistas de origen Nacional (2021-2024) 26
Figura 4. Red de coocurrencia de las reseñas positivas de los turistas de origen Internacional (2021-2024) 27
Figura 5. Red de coocurrencia de las reseñas negativas de los turistas de origen Internacional (2021-2024) 28
Figura 6. Red coocurrencia de las reseñas positivas en ingles de los turistas (2021-2024) 29
Figura 7. Red de coocurrencia de las reseñas negativas en ingles de los turistas (2021-2024) 30
Figura 8. Red de coocurrencia de las reseñas positivas para los hoteles ubicados en Chapinero (2021-2024) 31
Figura 9. Red de coocurrencia de las reseñas negativas para los hoteles ubicados en Chapinero (2021-2024) 32
Figura 10. Red de coocurrencia de las reseñas positivas de los turistas que viajan en pareja (2021-2024) 33
Figura 11. Red de coocurrencia de las reseñas negativas de los turistas que viajan en pareja (2021-2024) 34
Figura 12. Red de coocurrencia de las reseñas positivas de los turistas que calificaron mal al hotel donde se alojaron (1 a 3) /(2021-2024) 35
Figura 13. Red de coocurrencia de las reseñas negativas de los turistas que calificaron bien al hotel donde se alojaron (8 a 10) /(2021-2024) 36
Figura 14. Red de coocurrencia de las reseñas positivas de los turistas (2024) 37
Figura 15. Red de coocurrencia de las reseñas negativas de los turistas (2024) 38

Índice de Gráficos

Gráfica 1. Porcentaje de Hoteles por localidad (2024) 15
Gráfica 2. Número y promedio de calificaciones por localidad (2021-2024) 16
Gráfica 3. Número de calificaciones por hotel (2021-2024) 17
Gráfica 4. Número de hospedajes realizados por mes (2021-2024) 18
Gráfica 5. Número de calificaciones realizadas por mes (2021-2024) 19
Gráfica 6. Origen de los turistas (2021-2024) 19
Gráfica 7. Grupo de Viaje de los turistas (2021-2024) 20
Gráfica 8. Principales países de origen de los turistas (2021-2024) 21
Gráfica 9. Clasificación de las reseñas por origen (2021-2024) 22
Gráfica 10. Longitud de las reseñas por tipo (2021-2024) 23

BIBLIOGRAFIA

Adhinugroho, Y., Putra, A. P., Luqman, M., Ermawan, G. Y., Takdir, Mariyah, Siti, & Pramana, Setia. (2020). Development of online travel "Web Scraping" for tourism statistics in Indonesia. *Information Research, 25*(4), paper 885.

Alaei, A. R., Becken, S., & Stantic, B. (2019). Sentiment analysis in tourism: capitalizing on big data. Journal of Travel Research, 58(2), 175-191.

Berezina, K., Bilgihan, A., Cobanoglu, C., & Okumus, F. (2015). Understanding satisfied and dissatisfied hotel customers: text mining of online hotel reviews. Journal of Hospitality Marketing & Management, 25(1), 1-24.

vanden Broucke, S., & Baesens, B. (2018). Practical "Web Scraping" for Data Science: Best Practices and Examples with Python. Apress.

Caballero, R., Martín, E., & Riesco, A. (2018). Big Data con Python. Recolección, almacenamiento y proceso. Alfaomega Grupo Editor.

Cheng, M., & Jin, X. (2019). What do Airbnb users care about? An analysis of online review comments. *International Journal of Hospitality Management*, *76*, 58-70.

Flôres Limberger, P., dos Anjos, F. A., de Souza Meira, J. V., & dos Anjos, S. J. G. (2014). Satisfaction in hospitality on TripAdvisor. com: An analysis of the correlation between evaluation criteria and overall satisfaction. Tourism & Management Studies, 10(1).

Guo, Y., Barnes, S. J., & Jia, Q. (2017). Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation. *Tourism management*, *59*, 467-483.

Heydt, M. (2018). Python "Web Scraping" Cookbook: Over 90 proven recipes to get you scraping with Python, microservices, Docker, and AWS. Packt Publishing Ltd.

Li, J., Xu, L., Tang, L., Wang, S., & Li, L. (2018). Big data in tourism research: A literature review. Tourism Management, 68, 301-323.

Mitchell, R. (2018). "Web Scraping" with Python: Collecting more data from the modern web. " O'Reilly Media, Inc.".

Moro, S., Rita, P., & Coelho, J. (2017). Stripping customers' feedback on hotels through data mining: The case of Las Vegas Strip. *Tourism management perspectives*, *23*, 41-52.

Moro, S., Rita, P., Oliveira, C., Batista, F., & Ribeiro, R. (2018). Leveraging national tourist offices through data analytics. International Journal of Culture, Tourism and Hospitality Research.

Sanabria De Luque, J. (2021). Sector privado y libre competencia: implicaciones jurídicas del web scraping. Universidad Externado de Colombia.

Xiang, Z., Schwartz, Z., Gerdes Jr, J. H., & Uysal, M. (2015). What can big data and text analytics tell us about hotel guest experience and satisfaction?. *International journal of hospitality management*, *44*, 120-130.