UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Departamento de Estadística, Investigación Operativa Aplicadas y Calidad
Universitat Politècnica de València

Análisis de Ciudades a Través de su Actividad en Redes Sociales

TRABAJO FIN DE MÁSTER

Máster Universitario en Ingeniería de Análisis de Datos, Mejora de Procesos y Toma de Decisiones

Autor: Liliber Alvarez Ramos

Tutores: Ángeles Calduch Losa
Elena del Val Noguera
Miguel Rebollo Pedruelo

Curso 2016-2017

Agradecimientos

Primero agradecer a Dios por permitirme culminar esta etapa de mi vida en la que he adquirido muchísimos conocimientos nuevos.

A mis tutores: Elena, Ángeles y Miguel por haberme guiado y brindado todo el apoyo necesario durante esta etapa final del máster.

A mis padres, que aún sin estar presentes siempre me han brindado todo el apoyo necesario para cumplir las metas que me propongo.

A mis profesores del máster, porque sin su dedicación no hubiera adquirido todos los conocimientos necesarios para enfrentarme a este proyecto final de máster.

Finalmente a mis compañeros de piso, porque siempre estuvieron a mi lado apoyándome durante todo este año que convivimos juntos.

Resum

L'anàlisi dels temes sobre els quals es parla en una ciutat es realitza per a conèixer les opinions que expressen els usuaris sobre situacions quotidianes o esdeveniments que ocorren en determinat moment. Açò pot servir a diferents sectors de l'àmbit empresarial per a conèixer opinions sobre l'impacte de campanyes publicitàries, també a institucions de l'estat que estiguen interessades a saber què expressen les persones sobre algun tema relacionat amb la gestió governamental o amb campanyes polítiques.

Per a la detecció automàtica del nombre de temes es van utilitzar dues tècniques, una aplicant la mitjana harmònica i l'altra utilitzant la distribució de freqüència dels *hashtags* trobats en la base de dades de la xarxa social sota estudi. Després, per mitjà de l'algorisme Latent Dirichlet Allocation s'obtenen els grups de paraules pertanyents a cada tema. D'acord amb els resultats obtinguts es comprova que l'algorisme per a la detecció dels temes és fiable, ja que en la seua majoria els termes de cada grup estan relacionats entre si.

Paraules clau: LDA, mineria de text, xarxes socials, temes, dades geolocalitzades, ciutats

Resumen

El análisis de los temas sobre los cuales se habla en una ciudad se realiza para conocer las opiniones que expresan los usuarios sobre situaciones cotidianas o eventos que ocurren en determinado momento. Esto puede servir a diferentes sectores del ámbito empresarial para conocer opiniones sobre el impacto de campañas publicitarias, también a instituciones del estado que estén interesadas en saber qué expresan las personas sobre algún tema relacionado con la gestión gubernamental o con campañas políticas.

Para la detección automática del número de temas se utilizaron dos técnicas, una aplicando la media armónica y la otra utilizando la distribución de frecuencia de los *hashtags* encontrados en la base de datos de la red social bajo estudio. Después, por medio del algoritmo Latent Dirichlet Allocation se obtienen los grupos de palabras pertenecientes a cada tema. De acuerdo con los resultados obtenidos se comprueba que el algoritmo para la detección de los temas es fiable, ya que en su mayoría los términos de cada grupo están relacionados entre sí.

Palabras clave: LDA, text mining, redes sociales, temas, información geoposicionada, ciudades

Abstract

The analysis of the topics that are spoken in a city is done to know the opinions expressed by users about everyday situations or events that occur at any given time. This can be useful to different sectors of the business environment to get opinions on the impact of advertising campaigns, also to state institutions that are interested in knowing what people are expressing about issues related to government management or political campaigns.

For the automatic detection of the number of subjects, two techniques were used, the harmonic mean and the frequency distribution of the hashtags found in the database of the social network under study. Then, using the algorithm Latent Dirichlet Allocation the groups of words belonging to each topic are obtained. According to the obtained results it is verified that the algorithm for the detection of the subjects is reliable because the majority of the terms of each group are related to each other.

Key words: LDA, text mining, social networks, topics, geolocated data, cities

Índice general

Índice de figuras

Índice de tablas

CAPÍTULO 1
Introducción

El crecimiento y la diversificación de los recursos informáticos, asociado a la popularización de la web, ha propiciado un cambio en los hábitos de comunicación por parte de los ciudadanos. A nivel mundial las redes sociales, como son `Facebook` [1], `Twitter` [2], `LinkedIn` [3], `Instagram` [4], entre otras, forman parte importante de la comunicación, tanto que se espera que el número de usuarios en todo el mundo llegue a unos 2.95 mil millones en 2020 (alrededor de un tercio de la población de toda la tierra). La región con mayor tasa de penetración en las redes sociales es América del Norte, donde alrededor del 60 por ciento de la población tiene al menos una cuenta social (Gordon, 2017).

Las redes sociales, permiten a los usuarios mantenerse en contacto con otras personas y expresar sus opiniones acerca de diferentes temas (del Val et al., 2015b). Los usuarios de las mismas varían desde personas regulares a celebridades, representantes de empresas, políticos e incluso presidentes de países. Por lo tanto, es posible recopilar mensajes de texto de usuarios de diferentes grupos sociales (Pak and Paroubek, 2010). Todo lo antes mencionado ha hecho que estos medios se conviertan en una potente fuente de datos, por lo que suelen ser utilizados para conocer opiniones de usuarios individuales, controlar actividades, analizar campañas de políticos, estudiar el impacto de campañas publicitarias, entre otros (del Val et al., 2016b; Vivanco et al., 2017; del Val et al., 2015a, 2016a).

Como en Twitter se publican un gran número de textos diariamente, la información recogida puede ser bastante grande. Con los datos recopilados y mediante las herramientas adecuadas, se pueden realizar muchas investigaciones acerca de los temas que se tratan en diferentes ciudades y realizar comparaciones entre las mismas, con el objetivo de identificar posibles temas en común. Existen herramientas, tales como el software R Studio, que por medio de una serie de librerías permite trabajar una gran cantidad de datos con el fin de analizarlos y tomar decisiones con los resultados obtenidos.

El presente trabajo se realiza con la finalidad de investigar cuáles son los principales temas sobre los que se discute en una ciudad. Para dichos fines se utilizarán tuits generados en algunas ciudades y se analizará la información contenida

[1] https://www.facebook.com/
[2] https://twitter.com/
[3] https://es.linkedin.com/
[4] https://www.instagram.com/

en los mismos, para poder observar posibles formas de agrupar informaciones sobre un mismo tema. Para la agrupación de palabras se utilizará el algoritmo Latent Dirichlett Allocation, el cual permite identificar la información de los temas latentes en grandes colecciones de documentos (Ponweiser, 2012).

1.1 Motivación

Las redes sociales generan grandes cantidades de datos cada día, debido a que las personas tratan de descubrir lo que está sucediendo en el mundo y de compartir información al instante acerca de lo que opinan sobre algún tema o bien para conectarse con amigos o familiares (Chaffey, 2017).

Concretamente, esta información se puede aprovechar para entender las necesidades y opiniones de las personas acerca de diversos temas. Su uso para realizar investigaciones y tomar decisiones se está incrementando porque los usuarios suelen compartir sus opiniones acerca de diferentes temas del día a día (Pak and Paroubek, 2010).

Para cualquier entidad del gobierno de un país o una institución académica o de otro índole, sería interesante descubrir qué expresan las personas acerca de las gestiones o bien cuáles son los temas que más abundan relacionados con el bienestar de la sociedad, puesto que con esta información se pueden tomar ciertas decisiones, ya sean para mejorar mediante nuevos planes de gestión o bien para mantener los actuales.

Para todo esto se utilizan técnicas estadísticas que permitan recabar la información necesaria, que sea capaz de responder a las interrogantes que se plantean, para que los resultados que se obtengan estén avalados, dado que la misma proporciona un soporte científico a las observaciones que se realizan.

1.2 Objetivos

El objetivo general de este trabajo es detectar de manera automática cuáles son los principales temas de los que se habla en una ciudad. Para ello, se utilizarán datos geoposicionados de Twitter y técnicas estadísticas.

Para alcanzar este objetivo, se plantean los siguientes subobjetivos:

- Extraer y almacenar información geoposicionada
- Preprocesar los datos, para eliminar el contenido no necesario
- Analizar / evaluar los diferentes métodos estadísticos para la detección automática del número de temas
- Clasificar la información para la detección de los temas en una ciudad

1.3 Estructura del Documento

La memoria se encuentra dividida de la siguiente manera:

Capítulo 2

En este capítulo se realizará una breve introducción acerca de la red social Twitter, también se comentarán algunos trabajos relacionados con el análisis del contenido de los tuits y finalmente se planteará el aporte que hará este trabajo con respecto a lo que ya se ha investigado.

Capítulo 3

En este capítulo, por un lado se expondrán técnicas útiles para la extracción del número de temas más relevantes (k) y por otro el algoritmo Latent Dirichlet Allocation (LDA), el cual es necesario para la detección de los grupos de palabras más relevantes según el valor de (k).

Capítulo 4

En este capítulo se describirá el proceso de detección de las palabras más relevantes alrededor de cada tema y todas las fases llevadas a cabo para su obtención.

Capítulo 5

En el capítulo cinco, se mostrarán y comentarán todos los resultados obtenidos a consecuencia de la aplicación de los algoritmos.

Capítulo 6

El capítulo seis muestra las conclusiones y lineas de trabajo futuras.

CAPÍTULO 2
Estado del Arte

La propuesta que se presenta en este trabajo está relacionada con el tratamiento de información de redes sociales, concretamente Twitter y con técnicas estadísticas para la detección de temas. En este capítulo se describen trabajos que han utilizado la red social Twitter como fuente de información y trabajos que se han centrado en el análisis del contenido de los tuits.

2.1 Twitter

Twitter es una aplicación web que combina aspectos de redes sociales, mensajería instantánea y blogs en un modo de comunicación rápido, sencillo y conveniente. Permite a los usuarios registrados publicar actualizaciones cortas de estado, mensajes, noticias, enlaces, fotos y videos, conocidos como "tuits" (Richardson, 2015).

Según afirma Wang et al. (2017), Twitter es la plataforma tipo microblog abierta más popular, tiene aproximadamente 310 millones de usuarios activos mensuales. Mediante esta nueva forma de socialización, los usuarios tienen la posibilidad de publicar tuits sobre distintos aspectos de su vida cotidiana, desde el desarrollo profesional hasta actualizaciones personales y familiares. De acuerdo con Ma et al. (2014), Twitter originalmente fue diseñado para ser utilizado con servicios de mensajería de texto de teléfonos móviles, la brevedad del formato y la restricción a 140 caracteres por cada tuit crea un canal de comunicación informal y económico.

La información personal que se divulga a través de Twitter, en la sección de perfil de usuario, es reducida, opcional y breve, por lo general solo se coloca el nombre, la ubicación, una breve biografía de 160 caracteres y una dirección web (en caso de que se posea alguna).

Desde su creación en 2007, se ha desarrollado más allá del alcance de una aplicación de redes sociales, al punto de convertirse en una plataforma de noticias, comentarios, opiniones, marketing, activismo político, fotos compartidas, documentación de eventos, conversaciones, entre otros aspectos.

El acceso a los pensamientos, intenciones y actividades de millones de usuarios en tiempo real ha creado un potente canal para entender lo que está pasando en el momento en internet en cualquier lugar del mundo.

De acuerdo con estudios realizados por Ma et al. (2014), este servicio de mensajería cubre un gran número de diversos temas, tal como se ha mencionado, incluyendo comentarios sobre diversos temas de actividades personales, política y muchos otros. Debido a su corta extensión, los usuarios recurren a introducir en el texto *hashtags* (es decir, palabra clave prefijada con el símbolo #), lo cual los hace más llamativos.

Los *hashtags* han demostrado ser muy efectivos para organizar la información en Twitter. Estos mejoran la información y la búsqueda de los tuits, así como facilitan la interacción social. De acuerdo con una investigación que realizaron en Ma et al. (2014), el 58 % de los usuarios de Twitter utilizan *hashtags* en sus comentarios, por lo que este se ha convertido en una característica clave en muchas redes sociales como Telegram, FriendFeed, Facebook, Instagram, entre otras redes.

2.2 Trabajos Relacionados con Análisis de Tuits

Desde hace algunos años se han venido realizando diversas investigaciones relacionadas con los comentarios que exponen las personas en las redes sociales. En esta sección se abordarán algunos artículos en los que se han analizado tuits, algunos con objetivos relacionados con los de este proyecto.

El primero, trabajado por Adnan et al. (2014), tiene como objetivo proporcionar una comparación del uso de Twitter entre diferentes ciudades del mundo. Se eligieron tuits geoposicionados (es decir con información de latitud y longitud) de 15 ciudades, las cuales fueron elegidas con respecto al número de tuits que desde estas se enviaban.

Su análisis fue basado en la creación de gráficos temporales de la actividad de las 15 ciudades bajo estudio, con los cuales se pudieron identificar horas de actividad alta y baja. Esto lo realizaron en un período de tiempo determinado.

Con los mapas de calor pudieron medir la intensidad de la actividad de Twitter en términos de horas del día y día del año, con lo que determinaron diferentes patrones generales de los comportamientos de los usuarios en las diferentes ciudades. Los mismos incluyen: patrones semanales de actividad (horas de sueño versus horas de vigilia versus tiempo de trabajo); tiempos de uso altos y bajos para los servicios de redes sociales; cambios estacionales de la actividades. También observaron muchas diferencias entre los patrones de actividad de una ciudad a otra.

Por su parte, Förster et al. (2014), realizaron un monitoreo de las actividades de Twitter relacionadas con 31 ciudades del mundo que manejan gran volumen de información. El análisis se realizó utilizando principalmente métodos estadísticos cuantitativos respaldados por varias investigaciones cualitativas. Estos muestran que la actividad de los tuits relacionados con las ciudades, varía de una ciudad a otra. Factores como la tasa de desarrollo de los teléfonos inteligentes, número de turistas, etc, influyen en la cantidad de tuits que se producen en o alrededor de una ciudad, aunque no ocurre así para todas las ciudades. Los temas de los que se habla en la red social, están principalmente orientados a eventos o relacionados con deportes y política. En este artículo presentan un enfoque para analizar cuantitativamente el comportamiento de Twitter en distintas ciudades

para encontrar diversos indicadores de cómo se pueden clasificar las actividades de Twitter en todas ellas y cómo varían entre una y otra.

Dentro de sus investigaciones encontraron que aproximadamente solo el 6 % de las personas difunden información sobre su ubicación y que los datos muestran una distribución sesgada, en la que pocos usuarios producen gran cantidad de tuits y muchos solo lo hacen ocasionalmente. Se basaron en el análisis de los *hashtags* y hallaron que los tuits de algunas ciudades contienen *hashtags* sobre otras ciudades y que estos suelen estar relacionados con eventos específicos, circunstancias políticas, clubes deportivos o de fans, o bien campañas de promoción que coincidían con la fecha en la que fue llevado a cabo su estudio.

Otro estudio en el que se analizaron tuits geoposicionados, fue el de Rios and Lin (2013), en este al igual que el primer artículo analizado, se crearon visualizaciones de mapas de calor con el fin de evidenciar la intensidad de la actividad de Twitter en términos de tiempo del día y día del año. Por un lado evidenciaron el ritmo de las actividades en las grandes ciudades, vieron ciclos diurnos de vigilia y sueño, ciclos semanales de trabajo y también grandes cambios estacionales en el comportamiento e incluso patrones de actividad que se derivaban de las prácticas religiosas. Encontraron grandes diferencias en esos patrones en diferentes partes del mundo, reflejando las diferencias culturales y las innumerables formas en que se usa Twitter. Por otro lado, en la segunda parte de su análisis, trataron esos patrones de actividad como "huellas dactilares"de cada ciudad y realizaron un análisis de agrupamiento para cuantificar las similitudes entre ciudades individuales y grupos de ciudades. Esto lo realizaron con el fin de comprender cómo se comportan los usuarios de Twitter en diferentes partes del mundo. Encontraron agrupaciones de ciudades de un mismo país, también algunos ciclos diurnos y otros semanales, cambios estacionales y otros cambios en el patrón de comportamiento a gran escala.

En cuanto a temas relacionados con el análisis de la infomación contenida en tuits, en los que no se toma en cuenta la ubicación del usuario, se han realizado muchos trabajos, algunos en los que el análisis se centra en los *hashtags* utilizados en cada tuit y otros en los que se estudian temas específicos.

El trabajo de Berrocal et al. (2016), el cual se centra en analizar temas emergentes en las redes sociales, con el fin de conocer las opiniones que expresan los usuarios individuales, controlar actividades y actos de asociaciones, analizar las campañas de los políticos o estudiar el impacto de campañas publicitarias por parte de las empresas. Para la detección de dichos temas utilizaron el algoritmo Latent Dirichllett Allocation.

Por otro lado en artículos en donde el objetivo es estudiar los *hashtags* de los tuits, tal como el de Ma et al. (2014), lo que buscaban era relaciones entre los temas de los tuits y su correspondiente *hashtag*. En este se llegó a la conclusión de que los *hashtags* siguen una distribución *Power Law*, lo que significa que la mayoría de los *hashtags* son utilizados por pocos usuarios, mientras que la minoría son extremadamente populares por lo que aparecen en muchos tuits.

A diferencia de todos los trabajos estudiados, en este proyecto se pretende investigar cuáles son los principales temas sobre los que se habla en una ciudad. Para dichos fines se utilizan tuits geoposicionados de diferentes ciudades y se analiza la información contenida en los mismos mediante técnicas estadísticas.

En los artículos anteriores solo se analizaba el contenido de los tuits, principalmente con el fin de estudiar patrones de comportamiento, algunos basándose en las palabras contenidas en los tuits, otros en los *hashtags*, pero no se enfocaban en discutir si estos temas guardaban alguna relación al pasar de una ciudad a otra con el fin de realizar comparaciones entre las mismas. Finalmente, la tabla 2.1 hace una comparación de este proyecto con los trabajos mencionados en esta sección, en donde se observan las características más relevantes que cada uno toma en cuenta en el momento de realizar el análisis del contenido de los tuits.

	Características				
Trabajos	Análisis del contenido de tuits	Geoposicionados	Hashtags	Palabras	Temas
Adnan et al. (2014)	✓	✓		✓	
Förster et al. (2014)	✓	✓	✓		✓
Rios and Lin (2013)	✓	✓			
(Berrocal et al., 2016)	✓		✓	✓	✓
(Ma et al., 2014)	✓		✓		✓
Esta propuesta	✓	✓	✓	✓	✓

Tabla 2.1: Trabajos relacionados con el análisis de tuits comparados con esta propuesta

CAPÍTULO 3
Modelado Estadístico

En este capítulo se comentarán las diferentes técnicas utilizadas para extraer de forma automática el número de temas más relevantes del contenido de los tuits y a su vez el algoritmo Latent Dirichlet Allocation (LDA), por medio del cual se logra la extracción de los temas y el contenido de los mismos.

3.1 Notación y Terminología

En este trabajo se van a tratar las siguientes terminologías, como son: palabras, documentos, corpus y temas. Su definición sería:

- Una *palabra* es una unidad básica, definida como un ítem de un vocabulario, y la denotaremos por w.
- Un *documento* es una secuencia de N palabras.
- Un *corpus* es una colección de M documentos.
- Un *tema* es un conjunto de palabras que guardan una relación entre sí, y lo denotaremos por z, siendo K el número de temas.

3.2 Técnicas para la Extracción Automática del Número de Temas Relevantes

En esta sección serán abordadas las dos aproximaciones utilizadas para la extracción automática del número de temas más relevantes. Una de ellas mediante la Media Armónica y la otra por medio de los *hashtags* mencionados en los tuits.

3.2.1. Método de la Media Armónica

La media armónica es un método para determinar el número de temas más relevantes en el conjunto de textos. Este método fue aplicado por primera vez por Griffiths y Steyvers en su enfoque bayesiano de 2004 para encontrar el número óptimo de temas, se ha utilizado desde entonces en una variedad de análisis

(Griffiths et al., 2005; Wallach, 2006) debido a su simplicidad y relativa eficiencia computacional.

"En este caso se utilizan como datos las palabras w dentro del corpus, y el modelo se especifica por el número de temas, K, por lo que se pretende calcular la probabilidad de $P(w \mid K)$. Este cálculo se complica porque requiere la suma de todas las asignaciones posibles de palabras a los temas z, es decir: $p(w \mid K) = \int p(w \mid z, K)\, p(z)\, dz$. Sin embargo $p(w|K) = \int p(w|z,K)p(z)dz$, se puede aproximar calculando la media armónica de un conjunto de valores de $p(w \mid z, K)$ cuando z se aproxima a partir de la probabilidad a posteriori $p(z \mid w, K)$. En este caso, el algoritmo de muestreo que se va a utilizar es el de Gibbs, porque proporciona las muestras necesarias y el valor de $p(w \mid z, K)$ se puede calcular a partir de la siguiente ecuación":

$$P(w \mid z) = \left(\frac{\Gamma(V\beta)}{\Gamma(\beta)^V}\right)^K \prod_{k=1}^{K} \frac{\prod_V \Gamma\left(n_k^{(w)} + \beta\right)}{\Gamma\left(n_k^{(\cdot)} + V\beta\right)} \tag{3.1}$$

"En la que $n_k^{(w)}$, es el número de veces en el que una palabra w es asignada a un tema k en el vector de asignaciones z, y $\Gamma(\cdot)$ es la función Gamma estándar" (Griffiths and Steyvers, 2004).

El algoritmo de muestreo de Gibbs, antes mencionado, es un método típico de Markov Chain Monte Carlo (MCMC) y fue inicialmente propuesto para la restauración de imágenes por Geman and Geman (1984). Los métodos MCMC, se utilizan para resolver el problema de obtención de muestras de complejas distribuciones de probabilidad mediante el uso de números aleatorios. El muestreo de Gibbs (también conocido como muestreo condicional alternativo), lo que hace es simular una distribución de alta dimensión, mediante muestro en subconjuntos de menor dimensión de variables, donde cada subconjunto está condicionado al valor de todos los demás. El muestreo se realiza secuencialmente y continúa hasta que los valores muestreados se aproximan a la distribución objetivo. Aplicado al modelo LDA, se necesita que la probabilidad del tema $z_{a,b}$ sea asignada a $w_{a,b}$, la b-ésima palabra del a-ésimo documento, dada, $z_-(a,b)$, todos los demás temas son asignados a todas las demás palabras, es decir: (Ponweiser, 2012)

$$p(z_{a,b} \mid z_-(a,b), w, \alpha, \beta) \tag{3.2}$$

3.2.2. Extracción de Temas por Medio de la Cantidad de *Hashtags*

Los *hashtags* están presentes en las principales redes sociales, tales como Twitter, Facebook, Instagram, entre otras, lo cual es una fuerte evidencia de su importancia debido a que facilitan la difusión de la información en las redes sociales. Las personas suelen utilizar *hashtags* como palabras claves del contenido de su mensaje, de forma que facilite la transmisión del mismo. En el caso de Twitter, gran parte de los tuits de un usuario común son sus intereses o actividades personales (p.ej.,música, deportes, comida, viajes). Como Twitter es una red social en tiempo real, muchos de los tuits y sus *hashtags* asociados están relacionados con eventos recientes o en curso (Ma et al., 2014).

En Twitter los *hashtags* son utilizados para clasificar mensajes, propagar ideas y también para promover temas y personas específicas. Permiten a los usuarios crear comunidades de personas interesadas en el mismo tema, facilitándoles encontrar y compartir informaciones relacionadas Cunha et al. (2011).

En este trabajo se pretende determinar el número de temas más relevantes de acuerdo con la cantidad de *hashtags* más frecuentes publicados por los usuarios, debido a que una de las características de estos es que suelen ser utilizados para promover diversos temas, tal como se mencionó anteriormente.

Para determinar el número de temas, primero se va a comprobar si los *hashtags* siguen una distribución *Power Law*. Esto significa que la mayoría de los *hashtag* son utilizados por pocos usuarios, mientras que la minoría es extremandamente popular, es decir, muchos usuarios los utilizan en sus tuits.

De acuerdo con Newman (2005), cuando la probabilidad de medir un determinado valor de una cantidad varía inversamente como una potencia de ese valor, se dice que la cantidad sigue una *Power Law*, también conocida como ley del Zipf o distribución de Pareto. La *Power Law* aparece ampliamente en física, biología, economía, finanzas, informática, entre otras. Cuando los datos tienen el comportamiento de una *Power Law* aparecen como una linea recta al momento de graficarlos en escala logarítmica.

Estadísticamente, una cantidad x sigue una *Power Law* si se extrae de una distribución de probabilidad:

$$p(x) \propto x^{-\alpha}, \tag{3.3}$$

donde α es un parámetro constante de la distribución, conocido como exponente o parámetro de escala. Este parámetro de escala se encuentra normalmente en el intervalo 2 <α <3, aunque pueden haber ciertas excepciones ocasionales en las que se salgan del rango.

En un artículo de Clauset et al. (2009), describen cómo analizar si los datos siguen una distribución *Power Law*, mediante los siguientes pasos:

- Estimar los parámetros xmin y α del modelo *Power Law*.
- Calcular la bondad de ajuste entre los datos y la *Power Law*. Si resulta un p-valor mayor que 0.1, se acepta la hipótesis de que los datos siguen una *Power Law*, de lo contrario se rechaza.
- Comparar la *Power Law* con una hipótesis alternativa a través de una prueba de verosimilitud. Para cada alternativa, si la razón de verosimilitud calculada es significativamente distinta de cero, entonces su signo indica si la alternativa es favorecida sobre una *Power Law* o no.

Como en este caso se trabaja con variables discretas, los valores que tome x deben estar dentro de un conjunto de valores discretos. La distribución de probabilidad es la siguiente:

$$p(x) = Pr(X = x) = Cx^{-\alpha} \tag{3.4}$$

Como la distribución difiere de cero, debe haber un límite inferior $x_{min} > 0$ en el comportamiento de la *Power Law*. Cuando se calcula la constante normalizada, se

obtiene que:

$$p(x) = \frac{x^{-\alpha}}{\zeta(\alpha, x_{min})} \tag{3.5}$$

en donde

$$\zeta(\alpha, x_{min}) = \sum_{n=0}^{\infty} (n + x_{min}) \tag{3.6}$$

es la función Zeta o Hurwitz generalizada (Hawking, 1977).

Los estudios de las distribuciones empíricas que siguen una *Power Law* usualmente suelen dar algunas estimaciones del parámetro de escala α y en algunas ocasiones también el límite inferior en la región de escala x_{min}. La herramienta más utilizada para esta tarea es el histograma. Para esto se toma el logaritmo de ambos lados de la ecuación 3.3, se tiene que la distribución *Power Law* obedece a $lnp(x) = \alpha lnx + c$, lo que implica que sigue una recta en una gráfica doble logarítmica, como se observa en la figura 3.1. El parámetro de escala de esta distribución viene dado por la pendiente de la recta, normalmente esa pendiente se extrae realizando una regresión lineal por mínimos cuadrados en el logaritmo del histograma. Este procedimiento se remonta al trabajo realizado por Pareto en la distribución de riquezas a finales del siglo XIX (Clauset et al., 2009; Pareto, 1964).

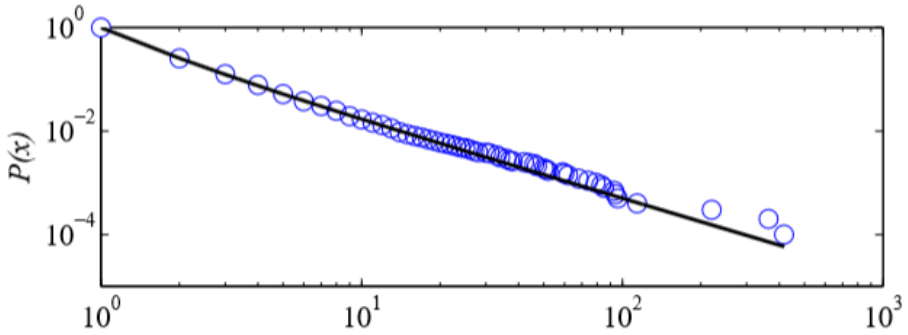


Figura 3.1: Distribución Power Law

3.3 Técnicas para la Detección de Grupos de Palabras más Relevantes de acuerdo con el Número de Temas

La técnica implementada para la extracción de los temas más relevantes en el contenido de los tuits es el algoritmo Latent Dirichlet Allocation (LDA). Es una técnica de aprendizaje no supervisado, es decir, no se conoce a priori el objetivo buscado (no hay clases predefinidas y puede que no se conozca el número de grupos). Su objetivo es identificar la información de los temas latentes en grandes colecciones de documentos.

Este algoritmo utiliza un enfoque de "bolsa de palabras", lo que significa que las palabras en un documento son intercambiables y por lo tanto su orden no es importante. Cada documento se representa como una distribución de probabilidad sobre algunos temas, mientras que cada tema se representa como una distri-

bución de probabilidad sobre un número de palabras, como se explica en Hong and Davison (2010).

Básicamente, el modelo LDA consiste en extraer muestras de las distribuciones Dirichlet y Multinomial. La distribución de Dirichlet es la generalización multivariada de la distribución beta que se ha utilizado en estadística bayesiana para modelar creencias y la distribución Multinomial es una generalización de la distribución Binomial, la cual expresa la probabilidad de observar un recuento de dos o más eventos independientes, dado un número de sorteos y unas probabilidades fijas por resultado, las cuales suman uno. En este caso los resultados son términos y temas.

LDA define el siguiente proceso generativo para cada documento d en el corpus:

1. Elige $N \sim$ Poisson (ξ)
2. Elige $\theta \sim$ Dir (α)
3. Para cada una de las N palabras w_n:
   a) Elige un tema $Z_n \sim$ Multinomial (θ)
   b) Elige una palabra w_n desde $p(w_n|Z_n, \beta)$, que es la probabilidad multinomial condicionada sobre el tema Z_n

Una variable aleatoria Dirichlet θ k-dimensional, puede tomar valores en el rango de (k-1)-simplex, en donde simplex es un vector-k dentro de la variable θ aleatoria que se incluye en la expresión (k-1)-simplex si $\theta_i \geq 0, \sum_{i=1}^{k} \theta_1 = 1$, y tiene la siguiente densidad de probabilidad en este simplex:

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} ... \theta_k^{\alpha_k - 1} \tag{3.7}$$

En la ecuación 3.7, el parámetro α es un vector-k con componentes $\alpha_i > 0$, y donde $\Gamma(x)$ es la función Gamma. Dados los parámetros α y β, la distribución conjunta de una mezcla de temas θ, un conjunto N de temas z y un conjunto N de palabras w, están dadas por:

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta), \tag{3.8}$$

donde $p(z_n|\theta)$ es simplemente θ_i para un único i tal que $z_n^i = 1$. Integrando sobre θ y sumando sobre z se obtiene la distribución marginal de un documento:

$$p(w|\alpha, \beta) = \int p(\Theta|\alpha) (\prod_{n=1}^{N} \sum_{z_n} p(z_n|\Theta) p(w_n|z_n, \beta)) d\theta \tag{3.9}$$

Finalmente, tomando el producto de la probabilidad marginal de un solo documento, se tiene la probabilidad marginal de un corpus:

$$p(|\alpha,\beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha)(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}|\theta_d)p(w_{dn}|z_{dn,\beta}))d\theta_d \tag{3.10}$$

Es importante distinguir el LDA de un modelo simple de *clustering* Dirichlet multinomial. Un modelo de *clustering* clásico incluye un modelo de dos niveles, donde el Dirichlet es muestreado una sola vez para el corpus, una variable de *clustering* multinomial es seleccionada una vez para cada documento en el corpus, y un conjunto de palabras son seleccionadas para el documento condicional en la varibale *cluster*. Como sucede en muchos modelos *clustering*, estos restringen un documento a ser asociado a un único tema. Sin embargo, LDA incluye tres niveles y los nodos del tema son muestreados varias veces dentro del documento. Bajo el modelo LDA los temas pueden ser asociados con múltiples temas (Blei et al., 2003).

Para entender con más claridad los antes expuesto, a continuación se presenta un ejemplo sencillo del funcionamiento del LDA.

Suponemos que se tienen las siguientes oraciones:

- Me gusta comer brócoli y bananas.
- Esta mañana me desayuné un batido de espinaca y banana.
- Los gatos y los perros son adorables.
- Mi hermana adoptó un perro ayer.
- Mira ese lindo hámster comiendo brócoli.

Dadas las oraciones anteriores, los elementos con los que trabaja el LDA son:

- Documento: cada una de las oraciones contenidas en el ejemplo, en este caso hay cinco documentos.
- Corpus: formado por el conjunto de los cinco documentos.
- Palabras: Cada uno de los ítem que conforman los documentos, sin tomar en cuenta las palabras vacías, es decir, las que por sí solas no tienen ningún significado.

En este caso, lo que hace el LDA es descubrir automáticamente los temas que contienen el grupo de oraciones. Por ejemplo, si se le pidieran dos temas al algoritmo, realizaría un proceso como el siguiente:

- Oraciones 1 y 2: 100 % del Tema A
- Oraciones 3 y 4: 100 % del Tema B
- Oración 5: 60 % del Tema A y 40 % del Tema B

Los temas quedarían distribuidos de la siguiente manera:

- Tema A: 30 % brócoli, 15 % bananas, 10 % desayuno, 10 % comiendo... (en este caso se puede interpretar que este tema esta relacionado con la categoría de comida).
- Tema B: 20 % gatos, 20 % perros, 20 % adorables, 15 % hámster... (en este caso se interpretaría como una categoría de animales).

En este capítulo se describieron las dos aproximaciones que serán utilizadas para la extracción del número de temas más relevantes, dentro del corpus de los tuits (media armónica y *hashtags*). También se realizó una breve descripción del funcionamiento del algoritmo LDA, el cual se interpretó de forma más clara con el ejemplo de las cinco oraciones en el que se observó la obtención de dos temas.

CAPÍTULO 4
Metodología de Análisis de Tuits

En este capítulo se describirá todo el proceso realizado para la obtención de los temas. Esto incluye la obtención de los tuits geoposicionados, el software utilizado para el tratamiento de los mismos, el preprocesado y las técnicas estadísticas aplicadas hasta la obtención de los temas.

4.1 Descripción General del Proceso de Detección de Palabras Relevantes Alrededor de los Temas

Para realizar el análisis de los tuits, se utiliza el software R Studio, el cual por medio de diferentes librerías permite la extracción de la información necesaria de manera que se pueda estudiar el objetivo principal de este trabajo, que como ya se ha mencionado, es detectar de forma automática los temas principales de los que se habla en una ciudad.

R Studio es un lenguaje de programación interpretado, de distribución libre, bajo licencia GNU (Licencia Pública General) y se mantiene en un ambiente para el cómputo estadístico y gráfico. Las herramientas de R pueden ser extendidas mediante paquetes provistos por el equipo central de R, o por contribuyentes que publican sus paquetes mediante la red CRAN (Comprehensive R Archive Network) (Santana Sepúlveda and Mateos Farfán, 2014).

A continuación se muestra un diagrama (figura 4.1), en el que de forma comprimida se expone todo el proceso llevado a cabo para la obtención de forma automática de los principales temas que se habla en una ciudad. Para la consecución del mismo, se hace uso de una serie de paquetes en R, los cuales se irán detallando a medida que se vaya explicando cada paso en las subsecciones siguientes.

4.2 Obtención/Extracción de Información de Twitter

Twitter tiene bastante información acerca de lo usuarios, como es: nombre, fecha de creación del tuit, contenido del tuit, coordenadas del tuit (ésta solo cuando

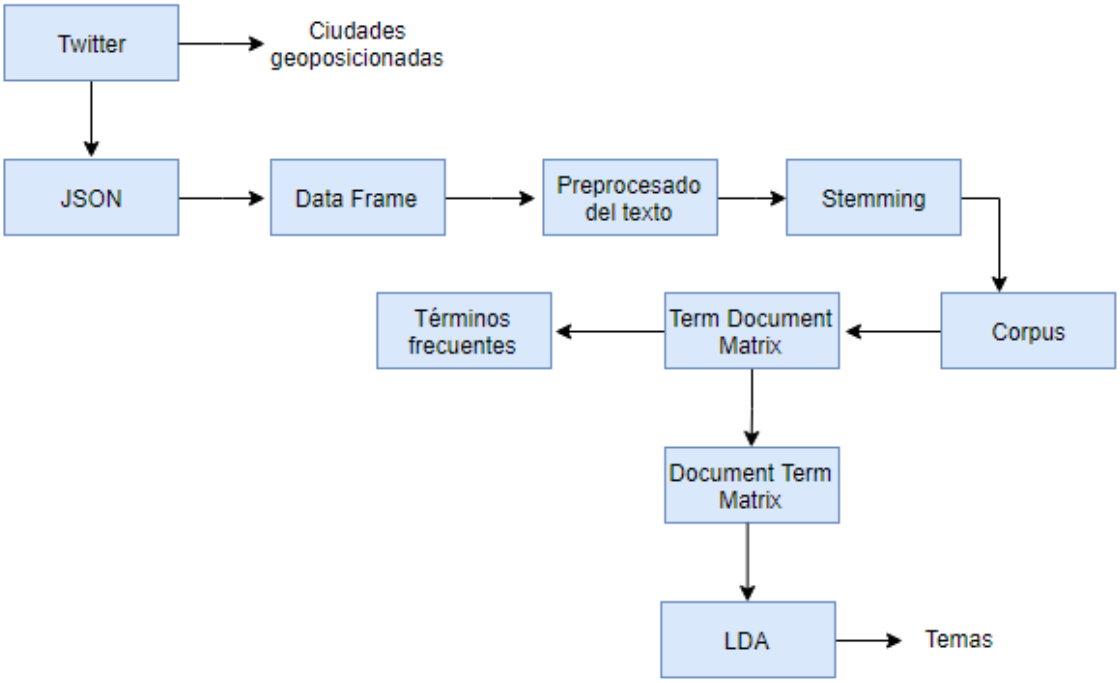


Figura 4.1: Diagrama de análisis del proceso

el usuario activa la opción de coordenadas), entre otras. En este trabajo solo se utilizan los campos que contengan datos relacionados con el objetivo que se desea lograr. Concretamente, en la tabla 4.1 se muestran los campos analizados.

Campo	Descripción
id	ID del tuit
text	Texto del tuit
coordinates	Información relativa a las coordenadas desde donde se generó el tuit (longitud y latitud)
user.name	Información relativa al usuario que emitió el tuit
created_at	Fecha y hora de cuando se creó el tuit

Tabla 4.1: Campos de los tuits utilizados en formato JSON

Para la extracción de datos de Twitter se ha utilizado el API que proporciona Twitter [1]. El API de Twitter es un sistema para que desde otros softwares se puedan conectar al servicio de Twitter y puedan recuperar información. Para poder utilizar el API es necesario disponer de unas credenciales. Concretamente api_key, api_secret, access_token, y access_token_secret.

```
library(twitteR)

  api_key <- "valor correspondiente"
  api_secret <- "valor correspondiente"
  access_token <- "valor correspondiente"
  access_token_secret <- "valor correspondiente"
```

[1] https://dev.twitter.com/overview/api

```
7
8   setup_twitter_oauth(api_key,api_secret, access_token, access_token_
        secret)
```

Listing 4.1: Conexión con el API de Twitter

Primero se conecta con la base de datos en donde estos están almacenados. En este caso es una base de datos MongoDB [2]. Se le especifica la colección de datos de la ciudad bajo interés, el nombre de la base de datos y la dirección de la máquina donde está alojada la misma. Como se trata de una consulta geoposicionada, en el código se especifica información de latitud, longitud y radio. Por otro lado, en la consulta también se especifican los campos de los tuits que nos interesa recuperar, en este caso son: *text*, *id*, *coordinates*, *user.name*, *geo y created_at*. Finalmente se especifica el número máximo de tuits que se quieren recuperar.

El resultado de la consulta se almacena en un fichero con formato JSON. En estos tipo de estudios es muy utilizado porque mantiene la jerarquía de los datos, característica necesaria en estos casos porque un tuit contiene muchos campos y se requiere el cumplimiento de la jerarquía de los mismos. Este fichero nos evitará tener que hacer consultas de manera repetitiva a la base de datos.

```
1  library(mongolite)
2  library(leaflet)
3  library(magrittr)
4
5  m <- mongo(collection = "tweets",  db = "EstadosUnidos", url = "mongodb
       ://dir.del.servidor:puerto")
6  #consulta geoposicionada a la base de datos:
7  #query --> creamos el filtro de la consulta
8  #geo.coordinates --> es el campo que vamos a utilizar para filtar
9  #geoWithin --> es un operador de mongo para buscar los documentos que
       esten dentro de la forma espacial especificada
10 #center --> indica que vamos a hacer busquedas en una circunferencia.
       Le pasamos latitud, longitud, radio.
11 #fields --> indicamos que campos del tweet nos interesan. En este caso
       son: text, id, coordinates, user.name, geo, created_at
12 #limit --> indicamos el numero de tweets que queremos recuperar como
       maximo
13 #la linea lo que hace es buscar que tweets geoposicionados de la base
       de datos EstadosUnidos estan dentro de la circunferencia con centro
        [39.7391500, -104.9847000] y radio 0.2
14
15 query <- m$find(query = '{"geo.coordinates": {"$geoWithin":{ "$
       centerSphere": [ [51.509865, -0.118092], 0.001] }}}', fields = '{"
       text" : true, "user.name":true, "geo.coordinates":true, "created_at
       ":true}', limit = 1000)
16
17 library(jsonlite)
18 #convertimos el resultado al formato json
19 file <-toJSON(query)
20 #escribimos el resultado al fichero ciudad.json
21 writeLines(file, "ciudad.json")
```

Listing 4.2: Extracción de tuits de la base de datos

[2] https://www.mongodb.com/es

El formato de datos JSON es un formato de texto ligero para el intercambio de datos. Un tuit en formato .JSON se representa como un conjunto de propiedades y sus correspondientes valores.


```
{
  "_id": {
    "$oid": "575a11a480e55946bcddeb3f"
  },
  "text": "@crazypedros best pizza in #Manchester aye? #Parklife #
     prebeers #tequila #pabst @ Crazy Pedros https://t.co/ZccKZlcMP6",
  "id": {
    "$numberLong": "741072966662905858"
  },
  "coordinates": {
    "type": "Point",
    "coordinates": [
      -2.2496709,
      53.4812461
    ]
  },
  "geo": {
    "type": "Point",
    "coordinates": [
      53.4812461,
      -2.2496709
    ]
  },
  "created_at": {
    "$date": "2016-06-10T01:02:18.000Z"
  }
}
```


Listing 4.3: Información de un tuit en formato JSON

Cuando queramos recuperar la información de una ciudad desde R, sólo tendremos que cargar el fichero .JSON que nos interese y a partir de ahí crear una estructura de tipo *dataframe* que almacene la información de cada uno de los tuits.

Para comprobar que los tuits son correctos, podemos pintarlos en el mapa (ver Figuras 4.2 y 4.3).

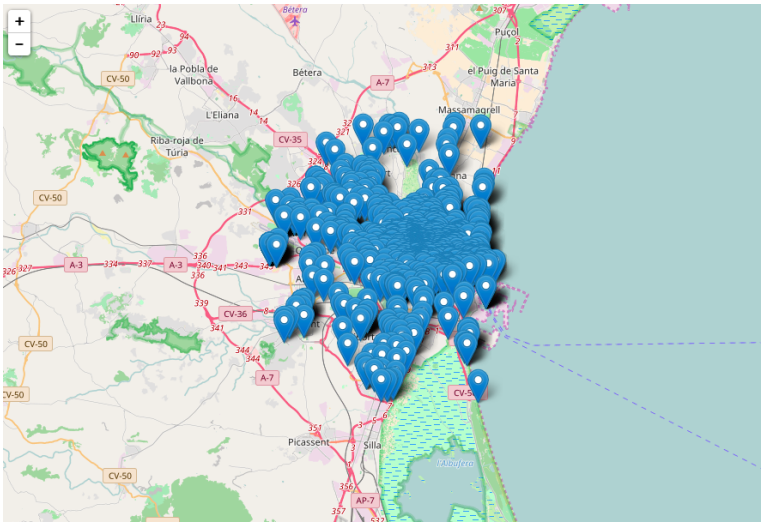


Figura 4.2: Mapa de Valencia con los tuits geoposicionados durante un intervalo de tiempo.

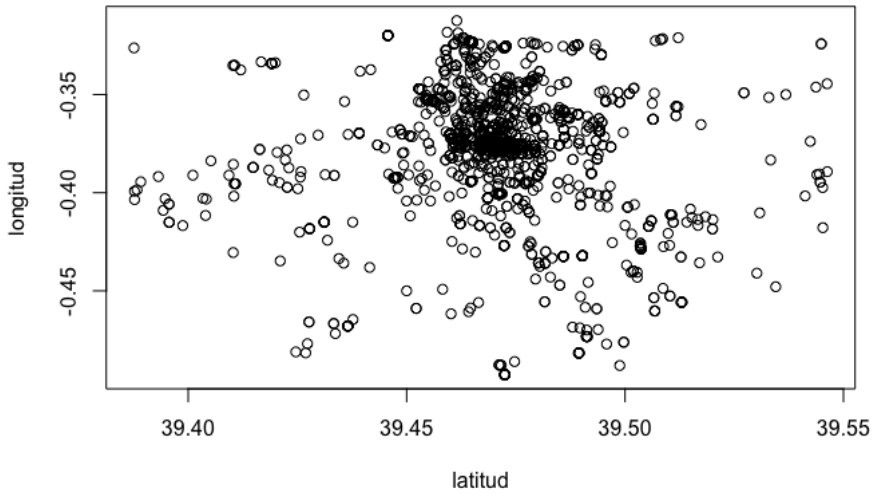


Figura 4.3: Coordenadas de los tuits geoposicionados en Valencia durante un intervalo de tiempo.

```
json_string <- sprintf("[%s]", paste(readLines("ValenciaGeoTime_Trump.
    json"), collapse=""))
cad1 <- "\\}\\{"
cad2 <- "\\},\\{"
cadena <- gsub(cad1, cad2, json_string)

testdf <- fromJSON(cadena)

#extraer las coordenadas de los tweets
coordinates <- testdf$geo
coordinatesdf <- as.data.frame(coordinates)
longitud <- sapply(coordinatesdf$coordinates, "[[", 2)
latitud <- sapply(coordinatesdf$coordinates, "[[", 1)
```

```
plot(latitud,longitud)

#pintar en el mapa
library(leaflet)
library(magrittr)
df<-data.frame(longitud, latitud)
leaflet(data = df) %>% addTiles() %>%
  addMarkers(~longitud, ~latitud)
```

Listing 4.4: Código para posicionar los tuits con coordenadas dentro de Valencia.

La tablas 4.2 y 4.3 muestran los tuits recopilados por ciudad con su correspondiente fecha de creación y los tuits obtenidos de algunas ciudades durante un tiempo específico, respectivamente.

Ciudad	Fecha incio	Fecha fin	Tweets
London	2015-08-29 17:49:32	2015-08-31 16:05:58	10000
Chicago	2015-11-03 18:59:22	2015-11-09 06:22:07	10000
NewYork	2015-11-02 19:38:38	2015-11-05 03:04:35	10000
Los Ángeles	2015-11-03 17:59:19	2015-11-06 16:46:07	10000
Washington	2015-11-03 18:59:24	2015-11-13 03:18:43	10000
San Francisco	2015-11-03 17:59:54	2015-11-14 04:15:25	10000
Phoenix	2015-11-03 17:59:23	2015-11-17 19:28:35	10000
Dallas	2015-11-03 17:59:31	2015-11-15 06:09:37	10000
Denver	2015-11-03 17:59:43	2015-11-19 17:31:06	10000

Tabla 4.2: Tuits recopilados por ciudad

Ciudad	Fecha incio	Fecha fin	Tweets
London	2017-05-01	2017-05-14	10820
Chicago	2017-05-01	2017-05-19	226
Nueva York	2017-05-01	2017-05-19	501
Washington	2017-05-01	2017-05-19	255

Tabla 4.3: Tuits recopilados en algunas ciudades durante un tiempo específico

4.3 Creación del Corpus, Preprocesado del Texto y Term Document Matrix (TDM)

La primera parte del proceso es la lectura de los datos desde el formato JSON, para lo cual se utiliza el paquete de R, `"jsonlite"` [3]. Este paquete funciona como generador y analizador de archivos en formato JSON. Ofrece flexibilidad, robustez y herramientas de alto rendimiento para trabajar con este tipo de archivos en R.

El algoritmo LDA trabaja con un corpus de las palabras contenidas en el texto y para que este funcione correctamente debe estar libre de caracteres o elementos

[3] https://cran.r-project.org/web/packages/jsonlite/index.html

extraños. Usualmente el contenido de los tuits está integrado por muchos caracteres que para el propósito del LDA son innecesarios y lo que hacen es agregar ruido al modelo.

Con paquete de R, `"tm"` [4] se construye el corpus y se realiza la limpieza del texto. Durante el preprocesado se deben realizar las siguientes operaciones:

- Eliminar las URL (p.ej. www.ejemplo.com), los *hashtags* (p.ej. #), nombres de usuarios (p.ej. @usuario) y términos especiales de Twitter (p.ej. RT").
- Eliminar las palabras vacías (p.ej. ahí , mío, lo, la, sin, tal...)
- Signos de puntuación
- Convertir las palabras a minúsculas

Otra parte del preprocesado incluye la lematización o *"stemming"* de las palabras del corpus, en este caso se utilizará el término en inglés por estar comúnmente aceptado. Este proceso consiste en reducir las formas derivadas de algunas palabras a una forma de base común, es decir hallar el lema (*stem*). La tabla 4.4 muestra un ejemplo del proceso de *"stemming"*.

Palabras	***lema***
Perro Perros Perrito Perrote	Perr

Tabla 4.4: Ejemplo de *stemming*

En la tabla 4.5 se puede observar un ejemplo de los textos de cinco tuits antes y después de ser preprocesados.

Después de haber realizado el proceso anterior, con el corpus se procede a crear una matriz con los términos y sus frecuencias (*Term Document Matrix*). Luego la transpuesta de ésta (Document Term Matrix) es utilizada para el cálculo de los temas. Todo este proceso se realiza mediante el paquete de R "tm".

En las tablas 4.6 y 4.7 se observa un ejemplo de cómo se muestran los términos y los documentos en cada una de las formas de la matriz de frecuencias.

[4] https://cran.r-project.org/web/packages/tm/tm.pdf

Antes del Preprocesado	**Después del Preprocesado**
"Who else is excited for coat weather? Can't wait for autumn! @Balenciaga @ Stanhope Gardens https://t.co/P8e4WbcGvp"	els excit coat weather cant wait autumn stanhop garden
"Just posted a photo @ Natural History Museum, London https://t.co/zZZC4WAQct"	just post photo natur histori museum london
Rawr @ Natural History Museum, London https://t.co/7TiHQYyN4y"	rawr natur histori museum london
"#legoland #windsor @ Legoland Windsor Theme Park https://t.co/NppYEClAI6"	legoland windsor legoland windsor theme park
"Last morning shift tomorrow !!"	last morn shift tomorrow

Tabla 4.5: Tuits antes y después del preprocesado

	Docs									
Terms	1	11	2	3	4	5	6	7	8	9
autumn	1	0	0	0	0	0	0	0	0	0
cant	1	1	0	0	0	0	0	0	0	0
coat	1	0	0	0	0	0	0	0	0	0
els	1	0	0	0	0	0	0	0	0	0
excit	1	0	0	0	0	0	0	0	0	0
garden	1	0	0	0	0	0	0	0	0	0
histori	0	0	1	1	0	0	0	0	0	0
london	0	0	1	1	0	0	0	0	1	0
museum	0	0	1	1	0	0	0	0	0	0
natur	0	0	1	1	0	0	0	0	0	0

Tabla 4.6: Ejemplo de *Term Document Matrix*

	Terms									
Docs	autumn	cant	coat	els	excit	garden	histori	london	museum	natur
1	1	1	1	1	1	1	0	0	0	0
11	0	1	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	1	1	1	1
3	0	0	0	0	0	0	1	1	1	1
4	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	1	0	0
9	0	0	0	0	0	0	0	0	0	0

Tabla 4.7: Ejemplo *Document Term Matrix*

4.4 Aplicación de Técnicas Estadísticas para la Estimación del Número de Temas

En esta sección se explicarán las dos aproximaciones utilizadas para determinar el número óptimo de temas. La primera de ellas es mediante el uso de las palabras del corpus utilizando la media armónica y la otra utilizando los *hashtags* contenidos en los tuits de las diferentes ciudades.

Para la aproximación mediante la media armónica, el primer paso es calcular la matriz *Term frequency–Inverse document frequency* (tf-idf), la cual es una medida estadística utilizada para evaluar la importancia de una palabra para un documento en el corpus. La importancia aumenta proporcionalmente al número de veces que una palabra aparece en el documento, pero está compensada por la frecuencia de la palabra en el corpus, por eso se utiliza para su cálculo el *Document Term Matrix*. El peso tf-idf (ver fórmula 4.1), está compuesto por dos términos, el primero calcula la frecuencia normalizada del témino (tf). El número de veces que una palabra aparece en un documento, dividida por el número total de palabras en ese documento (ver fórmula 4.2); el otro término, es la frecuencia inversa del documento (idf), la cual se calcula como el logaritmo del número de documentos en el corpus dividido por el número de documentos donde aparece el término específico (ver fórmula 4.3) (Ramos et al., 2003).

$$tfidf\left(t,d,D\right) = tf\left(t,d\right) \times idf\left(t,D\right) \tag{4.1}$$

$$tf\left(t,d\right) = \frac{f\left(t,d\right)}{max\left\{f\left(w,d\right) : w \in d\right\}} \tag{4.2}$$

$$idf\left(t,D\right) = log\frac{|D|}{|\{d \in D : t \in d\}|} \tag{4.3}$$

El proceso anterior se consigue en R con la librería `"slam"` [5] (se utiliza para trabajar con matrices), siguiendo el código mostrado a continuación:

```
library(slam)
term_tfidf <- tapply(dtm$v/row_sums(dtm)[dtm$i], dtm$j, mean) * log2(
    nDocs(dtm)/col_sums(dtm > 0))
summary(term_tfidf)
#quedarse con los terminos con tfidf >= Mediana
reduced.dtm <- dtm[,term_tfidf >=  1,1710]
dtm <- dtm[row_sums(reduced.dtm) > 0,]
inspect(dtm)
summary(col_sums(dtm))
```

Listing 4.5: Código para calcular la tf-idf.

Después del proceso anterior se obtiene una matriz tf-idf (ver tabla 4.8), de acuerdo con la cantidad de términos del documento.

[5] https://cran.r-project.org/web/packages/slam/slam.pdf

Término	TF-IDF
1	1.1643849
2	0.8788539
3	0.9863548
4	1.0982206
5	1.1554447
6	1.3043255
7	1.4764125
8	1.0676319
9	0.8507002

Tabla 4.8: Ejemplo term frequency - inverse document frequency

Luego se obtienen los estadísticos de la tabla 4.9, de los que interesa la mediana porque con este valor se reduce la matriz tf-idf. Esto se logra descartando todos los términos que tengan el tf-idf por debajo de la mediana.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.4883	1.232	1.476	1.838	1.898	13.29

Tabla 4.9: Estadistísticos obtenidos de la matrix tf-idf

Cuando se tiene la matriz tf-idf final, entonces se procede a calcular la media armónica.

```
library(topicmodels)
harmonicMean <- function(logLikelihoods, precision = 2000L) {
  llMed <- median(logLikelihoods)
  as.double(llMed - log(mean(exp(-mpfr(logLikelihoods, prec = precision
      ) + llMed))))
}

# Para encontrar el mejor valor de k para nuestro corpus, lo hacemos
    sobre una
# secuencia de temas con valores diferentes para k

seqk <- seq(2, 20, 1) # cantidad de temas a analizar (2-20), de uno en
    uno.
burnin <- 50
iter <- 50
keep <- 20
system.time(fitted_many <- lapply(seqk, function(k) topicmodels::LDA(
    dtm, k = k,
method = "Gibbs",control = list(burnin = burnin, iter = iter, keep =
    keep) ))

# extraer los logaritmos para cada tema
logLiks_many <- lapply(fitted_many, function(L)  L@logLiks[-c(1:(burnin
    /keep))])

library(Rmpfr) #Multiple Precision Floating-Point Reliable
# computar la media armonica
hm_many <- sapply(logLiks_many, function(h) harmonicMean(h))

```

```
#Graficar la media armonica para cada tema
library(ggplot2)
ldaplot <- ggplot(data.frame(seqk, hm_many), aes(x=seqk, y=hm_many)) +
    geom_path(lwd=1.5) +
  theme(text = element_text(family= NULL),
        axis.title.y=element_text(vjust=1, size=14),
        axis.title.x=element_text(vjust=-.5, size=14),
        axis.text=element_text(size=16),
        plot.title=element_text(size=16)) +
  xlab('Numero de Temas') + ylab('Media Armonica') +
  annotate("text", x = 5, y = -5000, label = paste("La cantidad optima
     de temas es", seqk[which.max(hm_many)])) +
  ggtitle(expression(atop("Latent Dirichlet Allocation, Analisis en
     base a datos de Twitter", "")))

ldaplot
```

Listing 4.6: Código para calcular la media armónica.

Finalmente se genera un gráfico como el 4.4, donde se observa que a medida que aumenta el número de temas (*topics*) llega un punto en el que la curva deja de aumentar y comienza a decaer, ahí es donde se encuentra el número aproximado de temas de la colección de documentos.

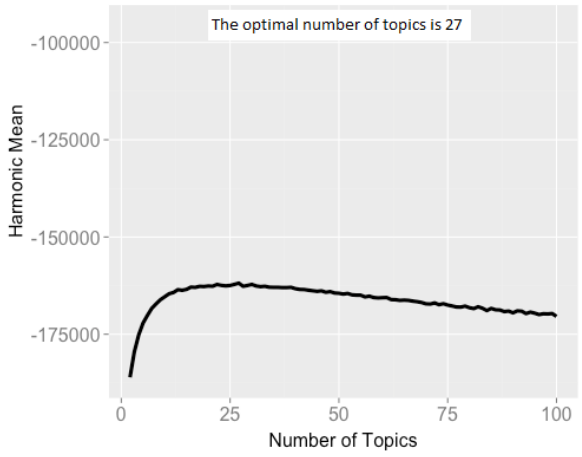


Figura 4.4: Cálculo del número de temas mediante la media armónica.

El otro método implementado para estimar el número de temas es mediante la frecuencia de los *hashtags* en la colección de documentos. Para esto, el primer paso es comprobar que los mismos siguen una distribución *Power Law*, para así poder encontrar un punto de corte que divida los *hashtags* más frecuentes de los menos frecuentes. Este punto de corte se puede aproximar aplicando la Ley de Pareto, en donde se tendría que un 20 por ciento de los *hashtags* es más frecuente en los tuits y el 80 por ciento restante es menos frecuente.

En este proceso, en vez de tomar todo el texto del documento solo se eligen los *hashtags* y se analiza su frecuencia. Para la extracción de los *hashtags* se utiliza la librería `"stringr"` [6].

```
library(stringr)
tweetsHashtag <- txt[grep("#", txt)]
```

[6]https://cran.r-project.org/web/packages/stringr/stringr.pdf

```
for(i in 1:length(tweetsHashtag)) {
  print(tweetsHashtag[i])
  tweet <- tweetsHashtag[i]
  tags <- str_extract_all(tweet, "#[:alpha:]+")
  vector <- c(vector, tags)
}
print(vector)
vector <- unlist(vector,recursive=TRUE)
vector <-tolower(vector)
vector <- vector[-1]
summary(vector)
```

Listing 4.7: Código para extraer los *hashtags*.

Una vez se tienen los *hashtags*, se procede a crear el corpus y después la *Term Document Matrix* con el objetivo de inspeccionar los términos más frecuentes. Después se verifica si la frecuencia de los *hashtags* siguen una distribución *Power Law*. Para realizar este proceso se utiliza el paquete de R `"poweRlaw"` [7], mediante el cual se realiza una prueba de bondad de ajuste, usando el método de remuestreo `bootstrapping` [8]. Este método se utiliza para aproximar la distribución en el muestreo de un estadístico. Aproxima el sesgo o la varianza de un análisis estadístico y permite construir intervalos de confianza o realizar contrastes de hipótesis sobre parámetros de interés, en este caso se van a generar p-valores para cuantificar la veracidad de la hipótesis. Si el p-valor es menor que 0.1, los datos indican que debemos rechazar que siguen una distribución *Power Law*.

H0: Los datos siguen una distribución *Power Law*
H1: Los datos no siguen una distribución *Power Law*

```
library("poweRlaw")
# Datos
data <- as.vector(as.matrix(df[,2]))

# Parametros
xmax <- 1000
simulations <- 50

# Ajustar a los datos el xmax
data_xmax = subset(data, data<xmax)

# Ajustar los datos a una power law
fit_object = displ$new(data_xmax)

# Estimar los parametros xmin y alpha
est = estimate_xmin(fit_object)
fit_object$setXmin(est)

xmin <- est$xmin
alpha <- est$pars

plot(fit_object) # graficar las frecuencias de las palabras
lines(fit_object, col="red", lwd=2) # graficar la linea del ajuste

```

[7] https://cran.r-project.org/web/packages/poweRlaw/poweRlaw.pdf
[8] https://pdfs.semanticscholar.org/0be0/d781305750b37acb35fa187febd8db67bfcc.pdf

```
# bootstrap para obtener el parametro de incertidumbre
bs = bootstrap(fit_object, no_of_sims = simulations, threads=6)
xmin_sd <- sd(bs$bootstraps[, 2])
alpha_sd <- sd(bs$bootstraps[, 3])

# Bondad de ajuste para obtenre el p-valor.
bs_p = bootstrap_p(fit_object, no_of_sims = simulations, threads = 6)
p_value <- bs_p$p

# Imprimir los resultados
cat("xmin =", xmin, "+/-", xmin_sd, "\n", file="outfile.txt")
cat("alpha =", alpha, "+/-", alpha_sd, "\n", file="outfile.txt", append
    =TRUE)
cat("p_value =", p_value, "\n", file="outfile.txt", append=TRUE)
cat("p_value>0.1 se acepta la hipotesis de que sigue una power-law", "\
    n", file="outfile.txt", append=TRUE)
```

Listing 4.8: Código para la validar si los datos siguen una distribución Power Law.

Luego de aceptar que las frecuencias de los *hashtags* siguen una *Power Law*, se procede a identificar aplicando la ley de Pareto cuáles son los *hashtags* que se encuentran dentro del 20 por ciento de los más frecuentes que producen el 80 por ciento de un impacto significativo, para de acuerdo a esa cantidad elegir el número de temas que posteriormente se van a extraer con el LDA. Para aplicar la ley de Pareto, se utiliza el paquete de R, `"qcc"` [9], el cual permite crear el diagrama de Pareto de forma automática y observar las frecuencias acumuladas de los *hashtags* más utilizados por los usuarios de Twitter en un tiempo específico.

4.5 Latent Dirichlet Allocation (LDA)

La extracción de los temas es el último paso del proceso. Se realiza después que se tiene la cantidad de temas, que es un dato de entrada calculado con los métodos vistos en la sección 4.4.

Para extraer los temas se utiliza el paquete de R `"topicmodels"` [10], el cual permite, después de haber creado el corpus y la *Document Term Matrix*, determinar de forma automática los temas en los que se agrupan las diferentes palabras de los documentos. Cabe destacar que con esta técnica una palabra puede estar contenida en más de un tema. Tambíen que el LDA no dice cuál es el tema sino qué términos pertenecen a cada uno de los temas.

```
library(topicmodels)
lda <- LDA(dtm, k = 12) # encontrar el numero de temas = k
term <- terms(lda, 5) # primeros 5 terminos de cada topic
(term <- apply(term, MARGIN = 2, paste, collapse = ", "))
```

Listing 4.9: Código para extraer los temas mediante LDA.

Finalmente se obtiene una tabla como la 4.10 con las palabras asociadas a un mismo tema.

[9] https://cran.r-project.org/web/packages/qcc/qcc.pdf
[10] https://cran.r-project.org/web/packages/topicmodels/topicmodels.pdf

Tema 1	música, cerveza, diversión
Tema 2	fútbol, estadio, valencia
Tema 3	comida, salud, familia

Tabla 4.10: Ejemplo de temas generados con LDA

En este capítulo se describió todo el proceso, desde lo más simple hasta lo más complejo, para la obtención de los temas contenidos en los tuits geoposicionados en las ciudades de los Estados Unidos y Europa. A lo largo del mismo se trataron varios ejemplos con el fin de poder comprender el proceso interno llevado a cabo por el código creado en R Studio.

CAPÍTULO 5
Aplicación de la Metodología y Resultados

Este capítulo se inicia con la presentación de los resultados obtenidos a raíz de la aplicación de cada una de las metodologías estudiadas para la extracción automática de la cantidad de temas y luego se presentan los resultados del algoritmo LDA, utilizado para la agrupación de palabras relacionadas a un mismo tema.

Al final se realiza una comparación con los resultados obtenidos sobre los temas encontrados para cada una de las ciudades bajo estudio.

5.1 Extracción del Número de Temas (k)

El primer paso es realizar la conexión con el servicio API de Twitter para iniciar la descarga de los tuits geoposicionados de las siguientes ciudades: Chicago, Dallas, Denver, Las Vegas, Londres, Los Ángeles, Nueva York, Phoenix, San Francisco y Washington. Para cada una de ellas se descargan unos 10.000 tuits.

Los primeros resultados están relacionados con las palabras más frecuentes en cada una de las ciudades, lo cual se logra utilizando el paquete de R "tm". Se han denominado más frecuentes aquellas que su frecuencia es igual o mayor a 300. La visualización de estas palabras servirá para tener una visión panorámica de cuáles podrían ser los temas de los que más se habla en una ciudad.

En las figuras desde la 5.1 hasta la 5.10 se observan cuáles son las palabras que con mayor frecuencia fueron utilizadas por los usuarios en el tiempo en que fueron generados los tuits en cada ciudad. En Chicago (5.1), Dallas (5.2), Denver (5.3), Phoenix (5.7) y Washington (5.9), el término más frecuente es el nombre de la propia ciudad, sin embargo en las demás es el término trabajo (job), siguiéndole *hire*, que en español se traduce como contratar. Esto puede suponer dos cosas: que las personas tienden a publicar temas relacionados con búsqueda de trabajo o que las empresas realizan publicaciones relacionadas con oportunidades de contratación.

Otro término que llama mucho la atención y se menciona en muchas ciudades es "*Veteran*", el cual se traduce como `veterano` [1], esto se debe a que el 11 de

[1] http://dle.rae.es/?id=bi1RtCG

noviembre se conmemora el día de los veteranos en los Estados Unidos, lo cual coincide con la fecha en la que fueron tomados los datos para este trabajo, tal como se aprecia en la tabla 4.2.

También el término "amp" es mencionado en todas las ciudades. Éste, según investigaciones realizadas, proviene de la palabra *Accelerated Mobile Pages*, la cual hace referencia a un código de programación que permite la creación de sitios web y anuncios bastante rápidos y de alto rendimiento en dispositivos y plataformas de distribución. En algunos casos esta palabra es un *hashtag* que hace alusión a *amplifier*, que traducido al español es "amplificador", lo cual está relacionado con música o más bien con instrumentos musicales.

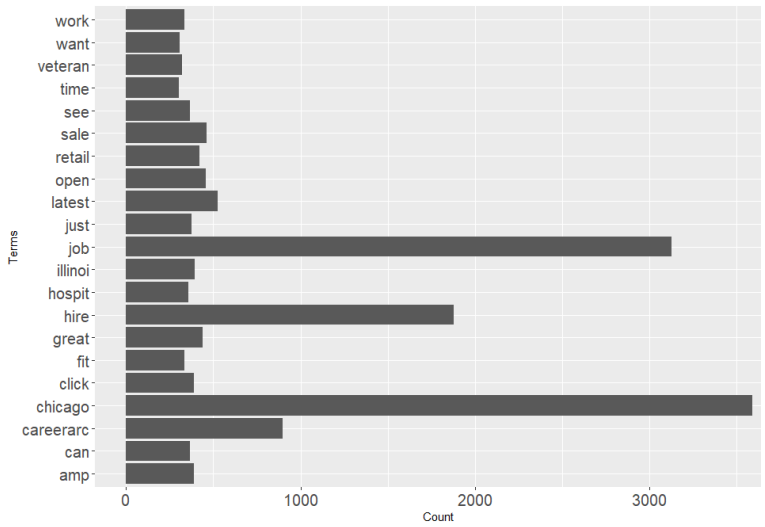


Figura 5.1: Palabras más frecuentes en la ciudad de Chicago

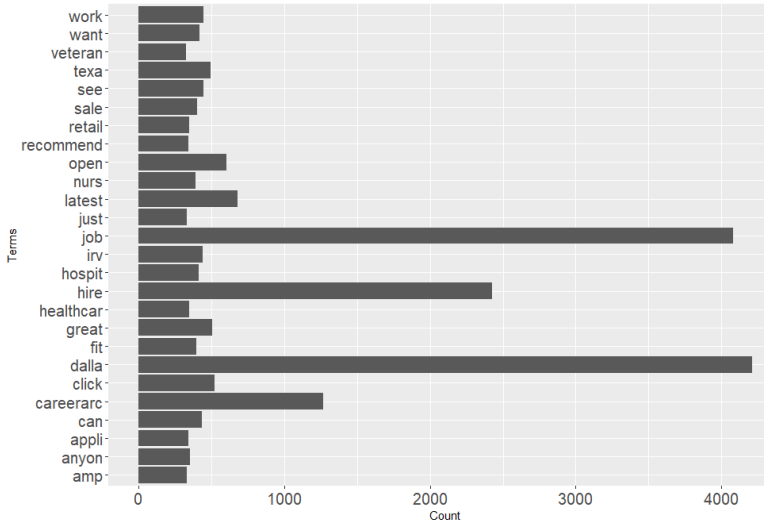


Figura 5.2: Palabras más frecuentes en la ciudad de Dallas

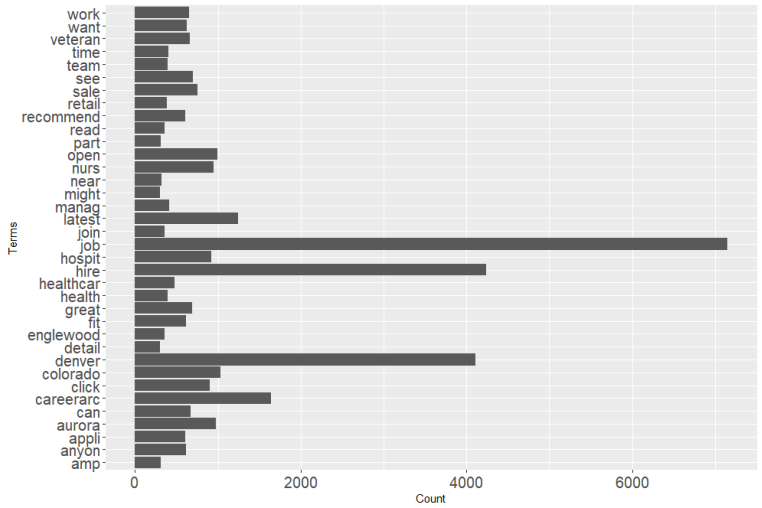


Figura 5.3: Palabras más frecuentes en la ciudad de Denver

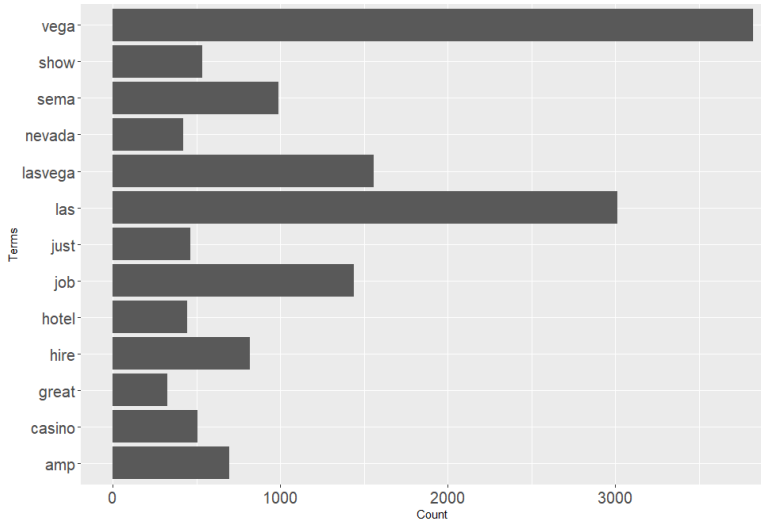


Figura 5.4: Palabras más frecuentes en la ciudad de Las Vegas

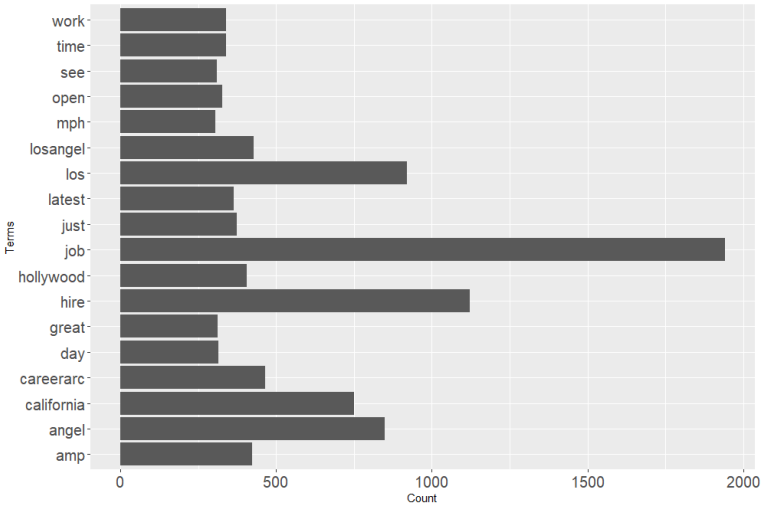


Figura 5.5: Palabras más frecuentes en la ciudad de Los Ángeles

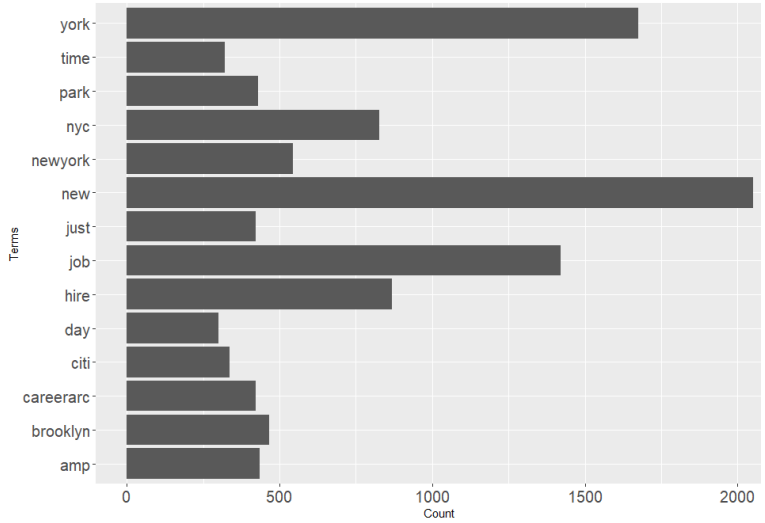


Figura 5.6: Palabras más frecuentes en la ciudad de Nueva York

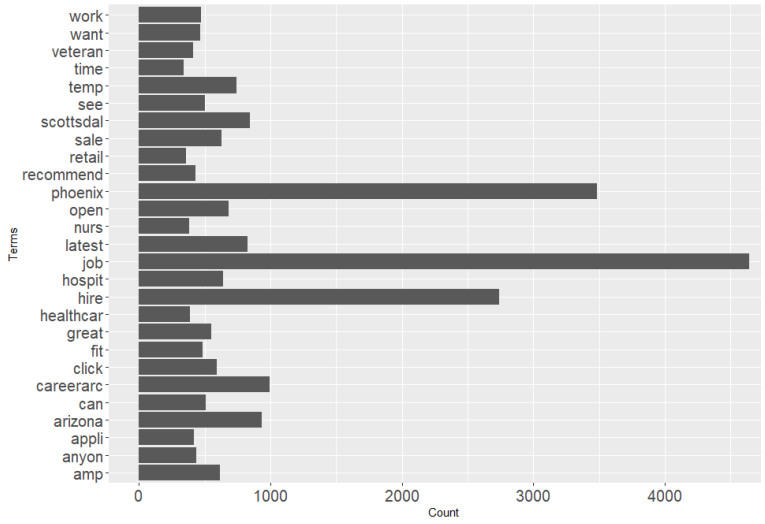


Figura 5.7: Palabras más frecuentes en la ciudad de Phoenix

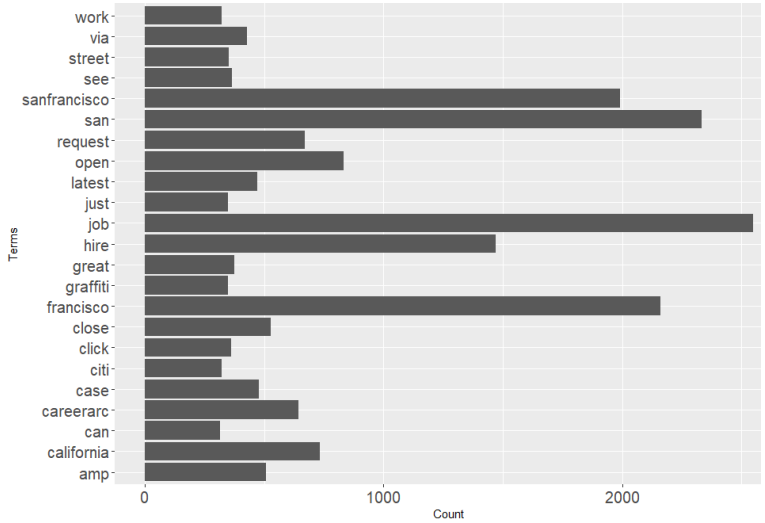


Figura 5.8: Palabras más frecuentes en la ciudad de San Francisco

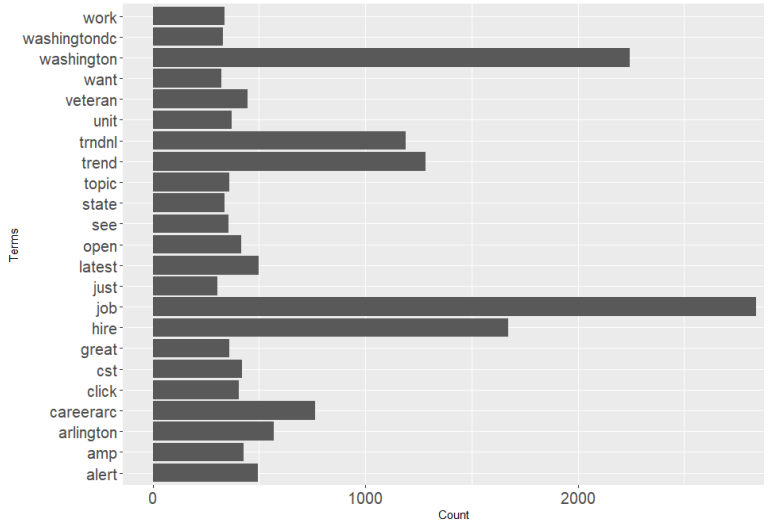


Figura 5.9: Palabras más frecuentes en la ciudad de Washington

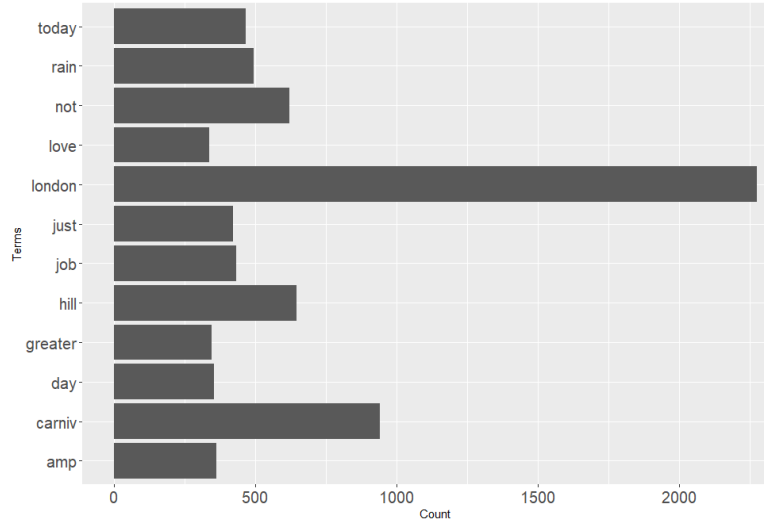


Figura 5.10: Palabras más frecuentes en la ciudad de Londres

Después de realizado el análisis preliminar, se procede a mostrar los resultados obtenidos para determinar el número de temas utilizando la media armónica. De acuerdo con este método, observando la gráfica, el número óptimo de temas se encuentra en el punto en donde la curva de la gráfica se estabiliza o comienza a disminuir. En este caso se observa que el menor número de temas ha sido para las ciudades de Los Ángeles 5.15 y San Francisco 5.18 con 13 temas, sin embargo para Nueva York se ha registrado el mayor número, 19; las demás ciudades se encuentran en ese rango de 13 a 19 temas.

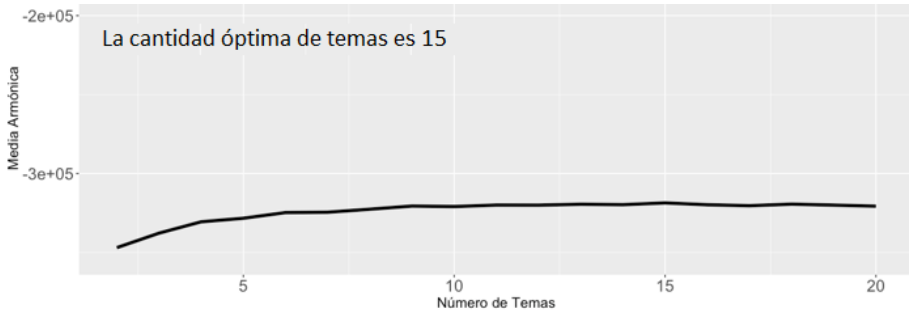


Figura 5.11: Cantidad óptima de temas para la ciudad de Chicago

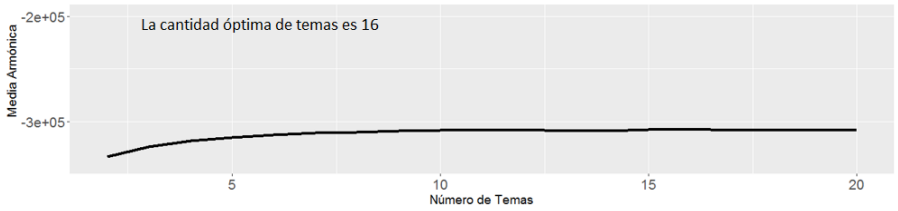


Figura 5.12: Cantidad óptima de temas para la ciudad de Dallas

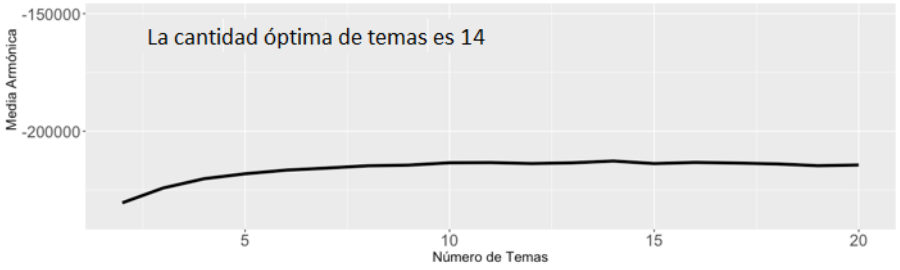


Figura 5.13: Cantidad óptima de temas para la ciudad de Denver

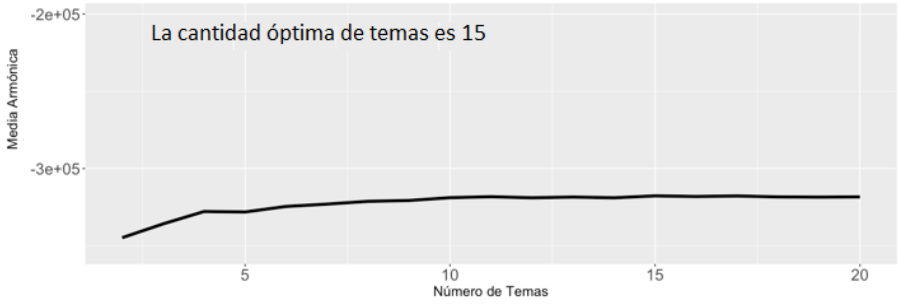


Figura 5.14: Cantidad óptima de temas para la ciudad de Las Vegas

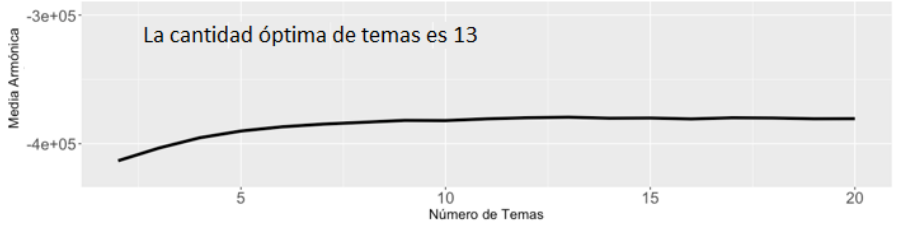


Figura 5.15: Cantidad óptima de temas para la ciudad de Los Ángeles

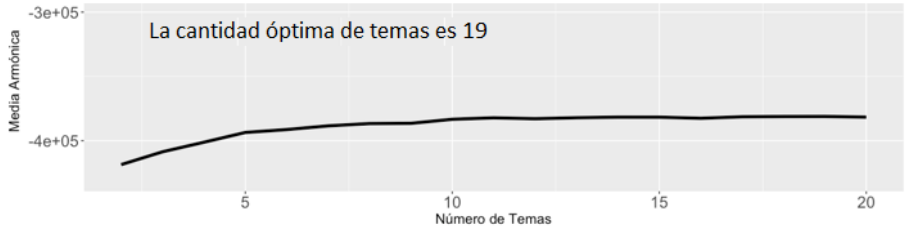


Figura 5.16: Cantidad óptima de temas para la ciudad de Nueva York

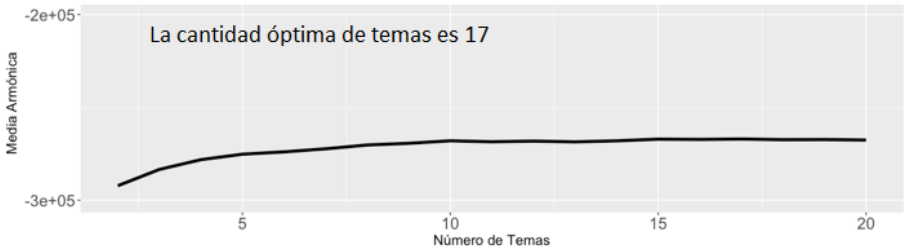


Figura 5.17: Cantidad óptima de temas para la ciudad de Phoenix

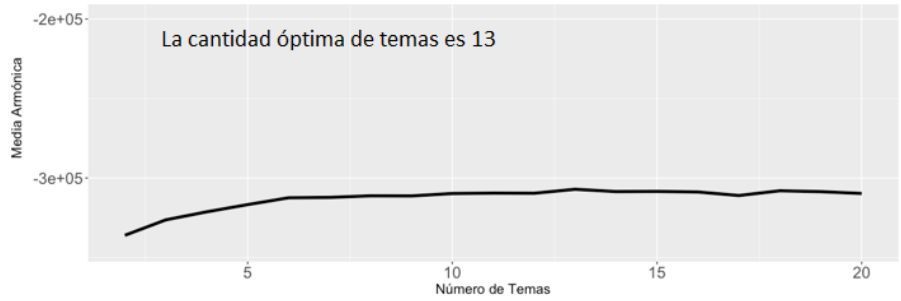


Figura 5.18: Cantidad óptima de temas para la ciudad de San Francisco

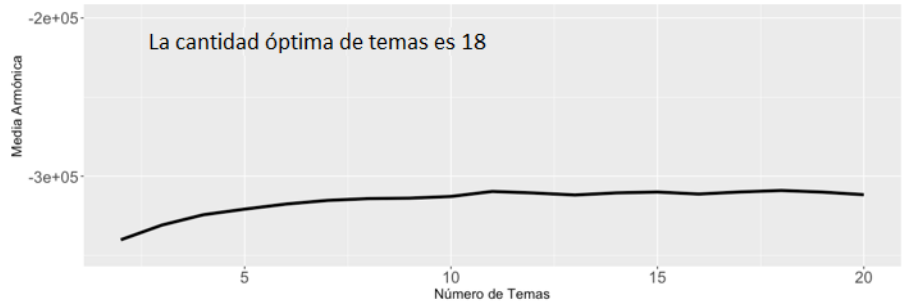


Figura 5.19: Cantidad óptima de temas para la ciudad de Washington

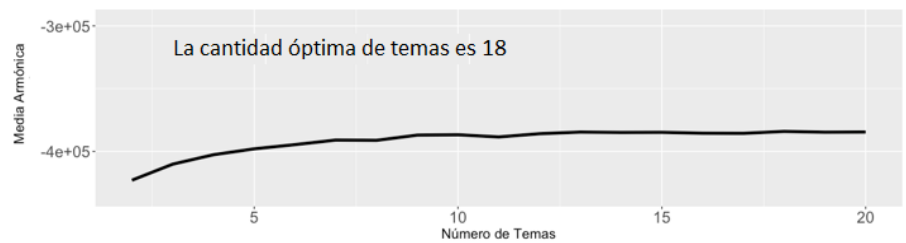


Figura 5.20: Cantidad óptima de temas para la ciudad de Londres

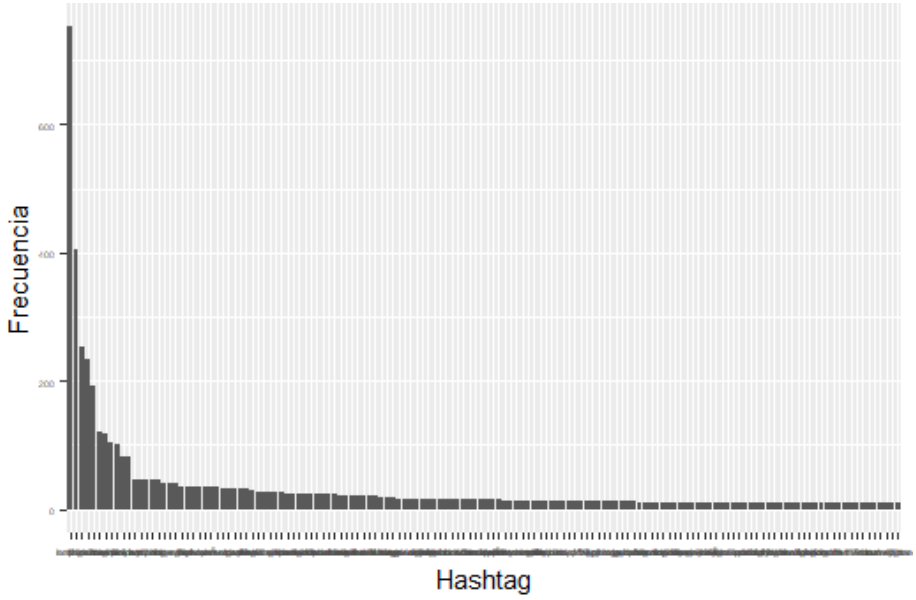
A continuación se presentan los resultados obtenidos de acuerdo con el número de *hashtags* y sus frecuencias para determinar la cantidad de temas. Para lograr el objetivo de este trabajo, se han tomado los *hashtags* cuya frecuencia sea mayor que uno, aquellos que tienen frecuencia igual a uno no resultarían relevantes. En ese mismo sentido, cuando existan más de quince palabras con la misma frecuencia, se elige esa como punto de corte y se trabaja con los que tengan un valor igual o superior a ese. Lo que se pretende es quitar el ruido introducido por palabras poco utilizadas. Luego, con los *hashtags* elegidos se comprueba la hipótesis de que estos se distribuyen como una *Power Law*, para después determinar los que contienen el número de temas requeridos para la agrupación de las palabras, esto último aplicando la Ley de Pareto.

En los resultados mostrados desde la figura 5.21 hasta la 5.30 se observa que los p-valores calculados para las frecuencias de los *hashtags* son mayores que 0.1, los datos nos indican que no se rechaza la hipótesis nula de que su distribución sigue una *Power Law*. Otro criterio que confirma lo dicho anteriormente son los valores obtenidos para el parámetro alpha, el cual debe encontrarse entre los valores 2 y 3, aunque en ciertas ciudades ha resultado ligeramente por debajo de 2, no obstante, como bien se había descrito en el capítulo 3 en ciertas ocasiones este valor podría estar fuera de ese rango.

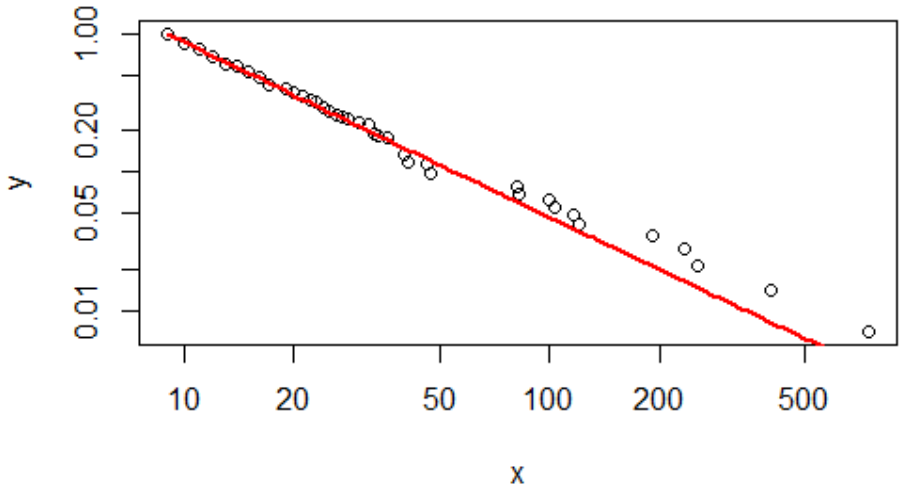
Después de comprobar que los datos se distribuyen como un *Power Law*, se procede a determinar el porcentaje de *hashtags* más importantes dentro de cada ciudad. En las figuras desde la 5.31 hasta 5.40 se presentan los diagramas de Pareto para cada ciudad, en los que aparecen el 20 por ciento de los *hashtags* más importantes para cada corpus. La tabla 5.1 muestra el número de temas para cada ciudad, los cuales son equivalentes al número de *hashtags* más importantes de acuerdo con los diagramas de Pareto.

Ciudad	**Hashtags**	
	Total	**Más importantes**
Chicago	4367	10
Dallas	3481	10
Denver	2965	12
Las Vegas	5639	35
Los Ángeles	6534	43
Nueva York	5918	39
Phoenix	3549	12
San Francisco	4400	19
Washington	3846	15
Londres	4712	60

Tabla 5.1: *Hashtags* más importantes por ciudad



(a) Histograma de frecuencias de los *hashtags* en Chicago

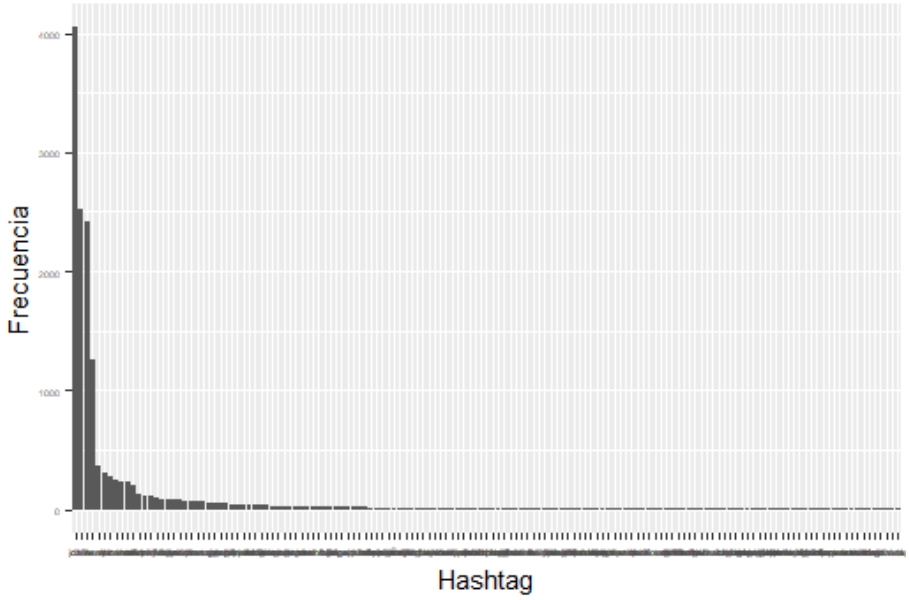


(b) Frecuencias en escala logarítmica

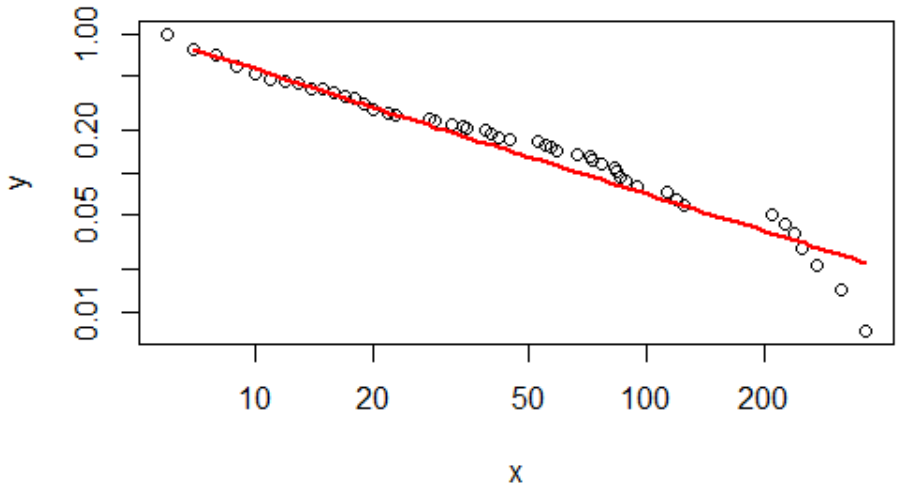
```
xmin = 9 +/- 0.4218521
alpha = 2.241616 +/- 0.08510036
p_value = 0.62
p_value>0.1 indica que es plausible la hipotesis power-law
```

(c) Parámetros de la *Power Law*

Figura 5.21: Resultados *Power Law* para la ciudad de Chicago



(a) Histograma de frecuencias de los *hashtags* en Dallas

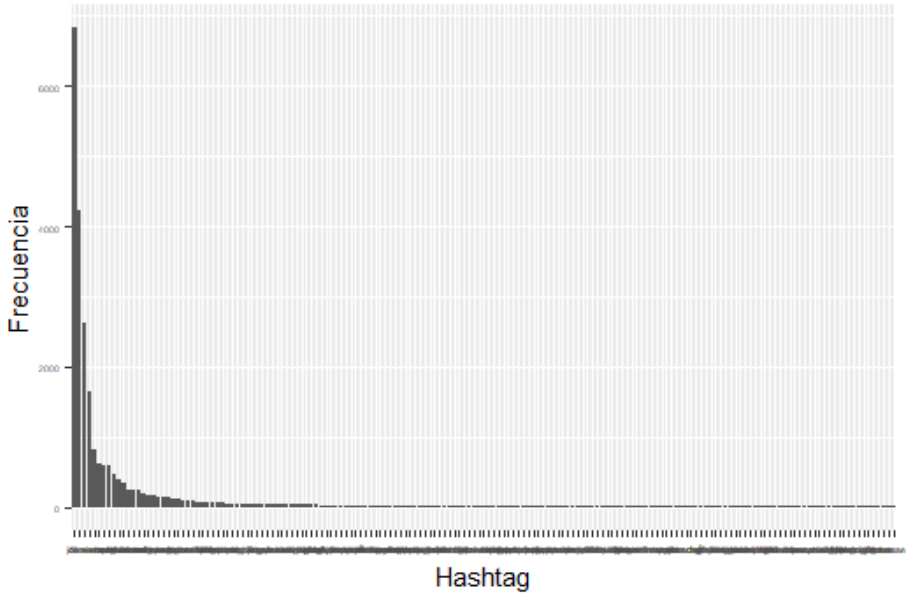


(b) Frecuencias en escala logarítmica

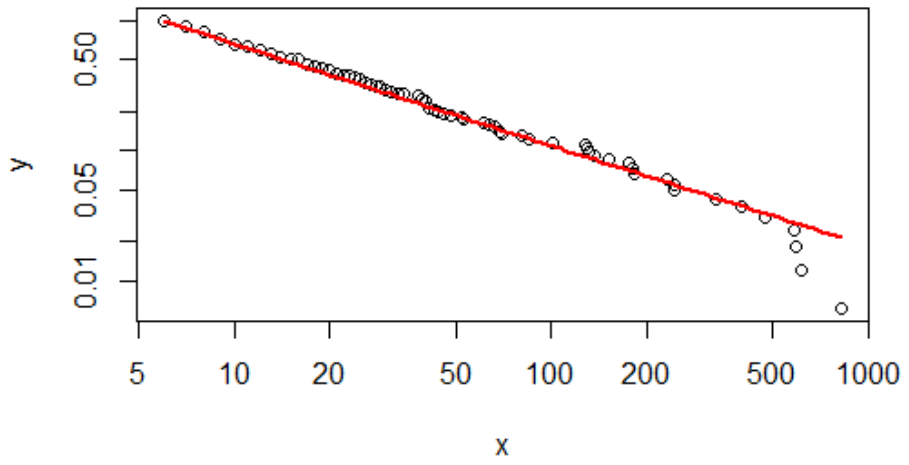
```
xmin = 7 +/- 8.429879
alpha = 1.881147 +/- 0.1869987
p_value = 0.18
p_value>0.1 indica que es plausible la hipotesis power-law
```

(c) Parámetros de la *Power Law*

Figura 5.22: Resultados *Power Law* para la ciudad de Dallas



(a) Histograma de frecuencias de los *hashtags* en Denver

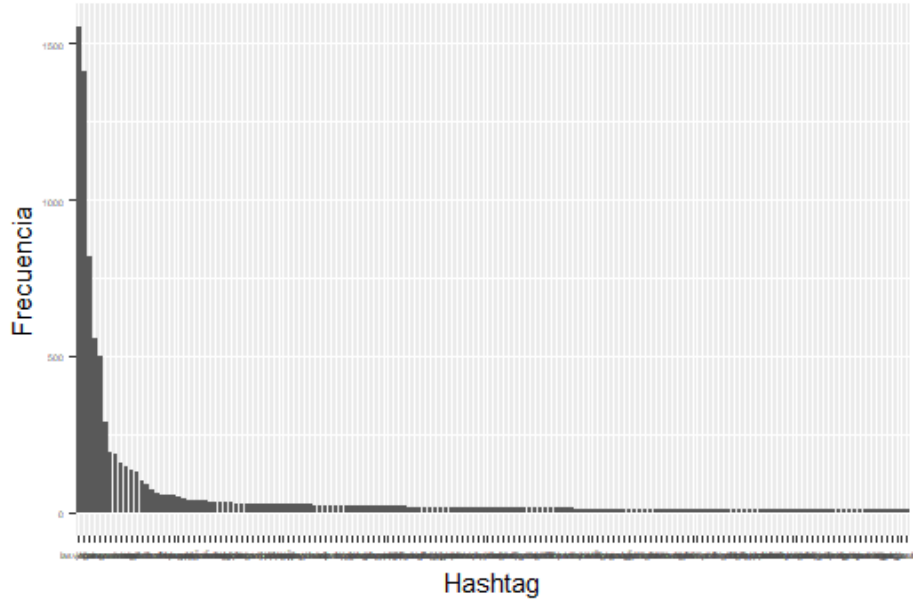


(b) Frecuencias en escala logarítmica

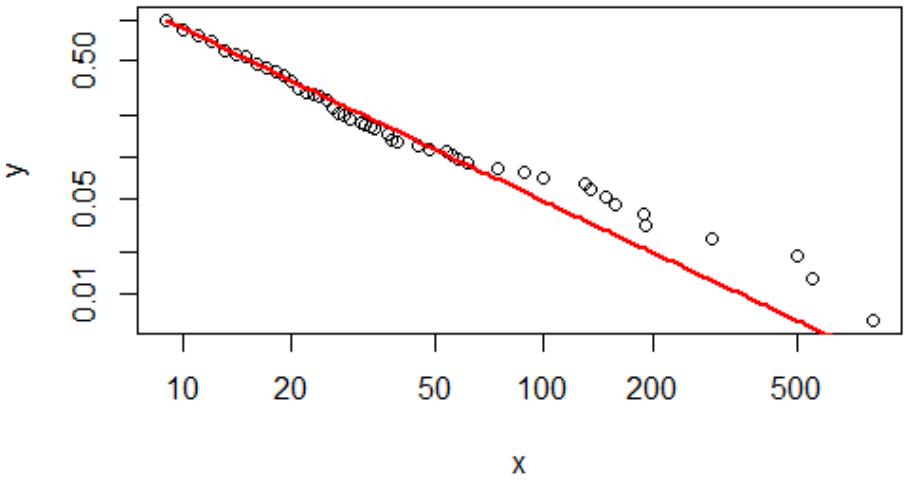
```
xmin = 7 +/- 8.429879
alpha = 1.881147 +/- 0.1869987
p_value = 0.18
p_value>0.1 indica que es plausible la hipotesis power-law
```

(c) Parámetros de la *Power Law*

Figura 5.23: Resultados *Power Law* para la ciudad de Denver



(a) Histograma de frecuencias de los *hashtags* en Las Vegas

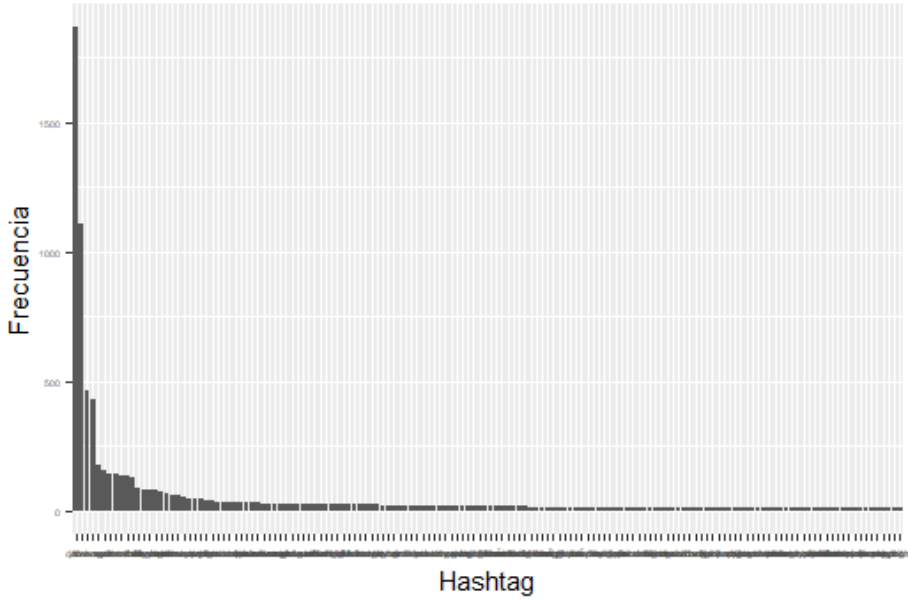


(b) Frecuencias en escala logarítmica

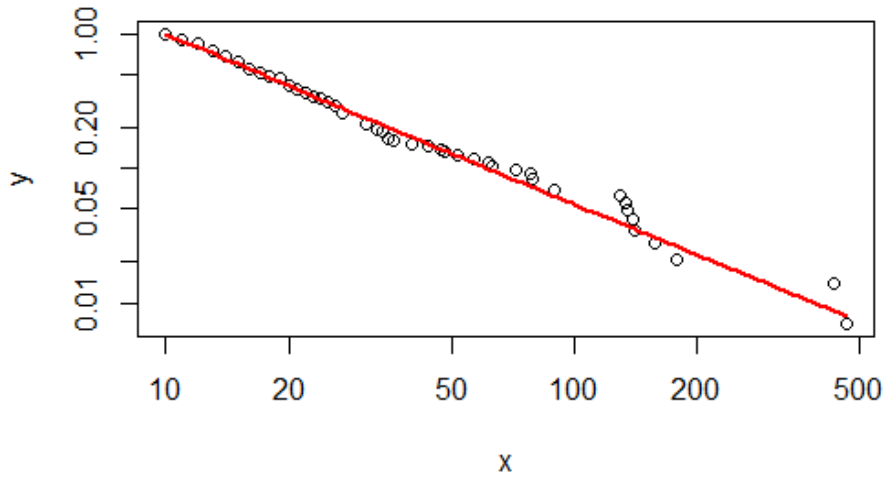
```
xmin = 9 +/- 0.9476071
alpha = 2.239772 +/- 0.09662414
p_value = 0.3
p_value>0.1 indica que es plausible la hipotesis power-law
```

(c) Parámetros de la *Power Law*

Figura 5.24: Resultados *Power Law* para la ciudad de Las Vegas



(a) Histograma de frecuencias de los *hashtags* en Los Ángeles

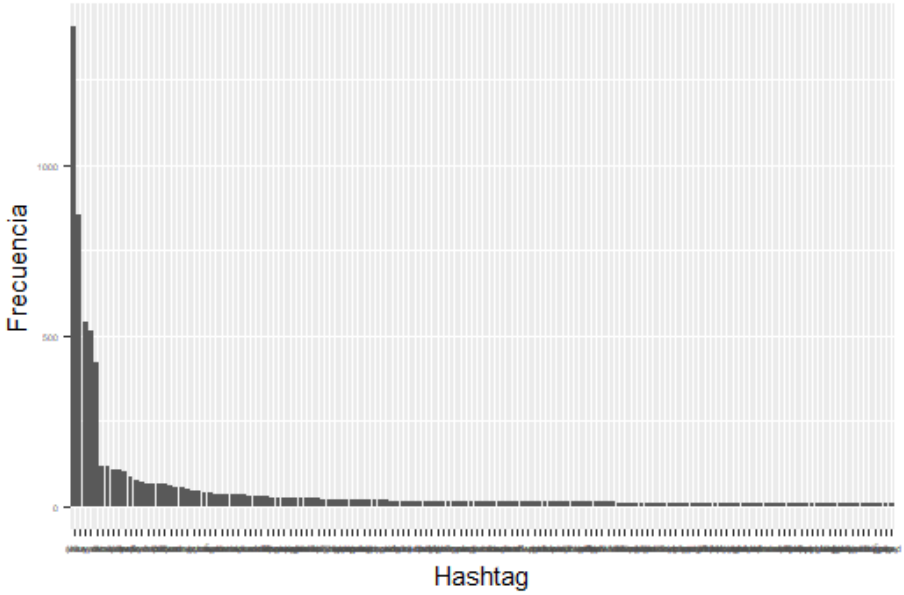


(b) Frecuencias en escala logarítmica

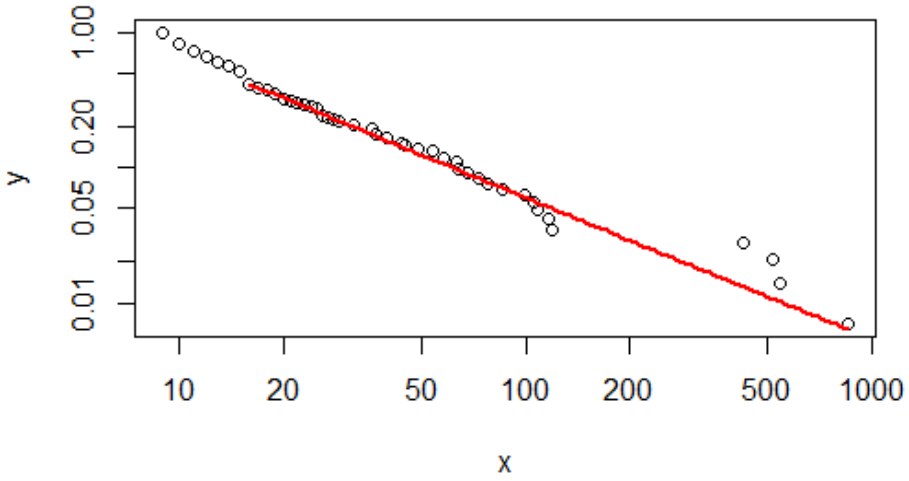
```
xmin = 10 +/- 1.304466
alpha = 2.243898 +/- 0.1132837
p_value = 0.2
p_value>0.1 indica que es plausible la hipotesis power-law
```

(c) Parámetros de la *Power Law*

Figura 5.25: Resultados *Power Law* para la ciudad de Los Ángeles



(a) Histograma de frecuencias de los *hashtags* en Nueva York

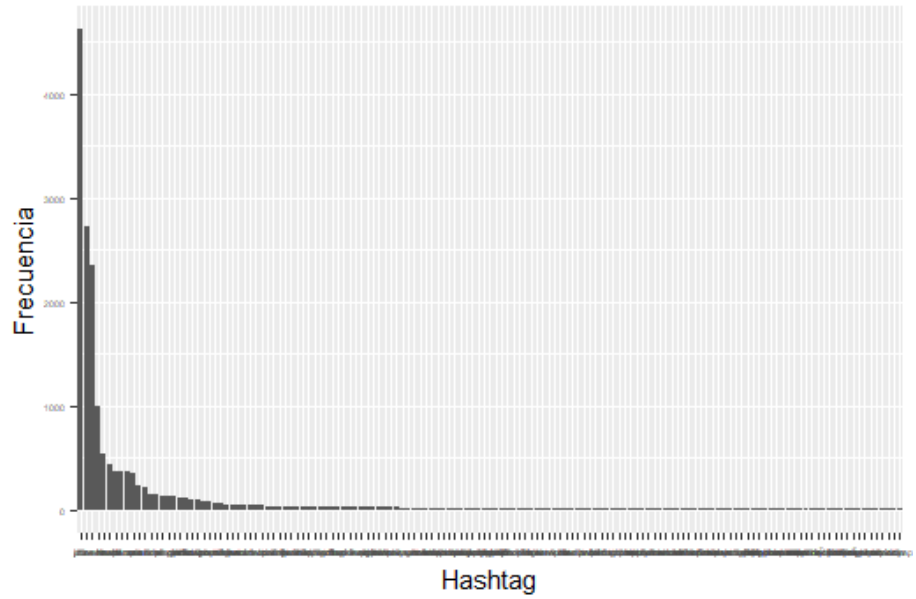


(b) Frecuencias en escala logarítmica

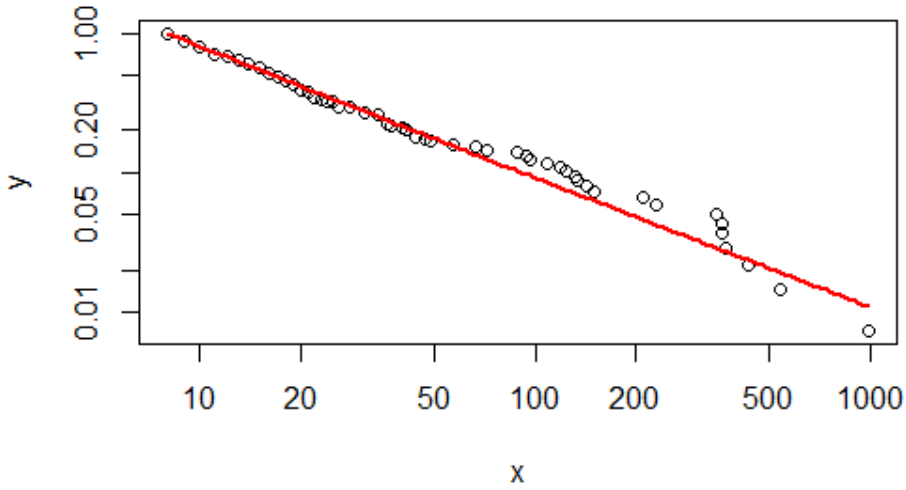
```
xmin = 16 +/- 4.506866
alpha = 2.044696 +/- 0.1328745
p_value = 0.76
p_value>0.1 indica que es plausible la hipotesis power-law
```

(c) Parámetros de la *Power Law*

Figura 5.26: Resultados *Power Law* para la ciudad de Nueva York



(a) Histograma de frecuencias de los *hashtags* en Phoenix

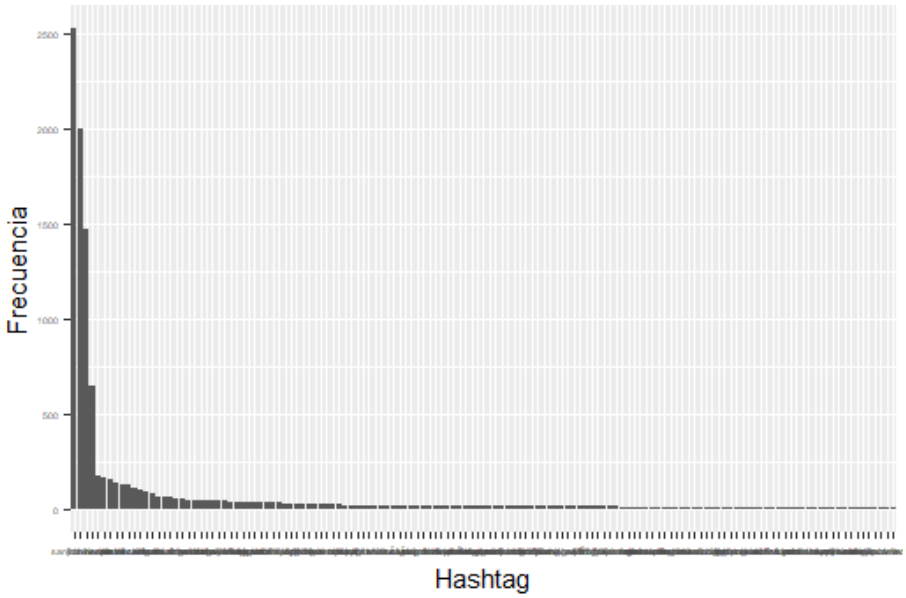


(b) Frecuencias en escala logarítmica

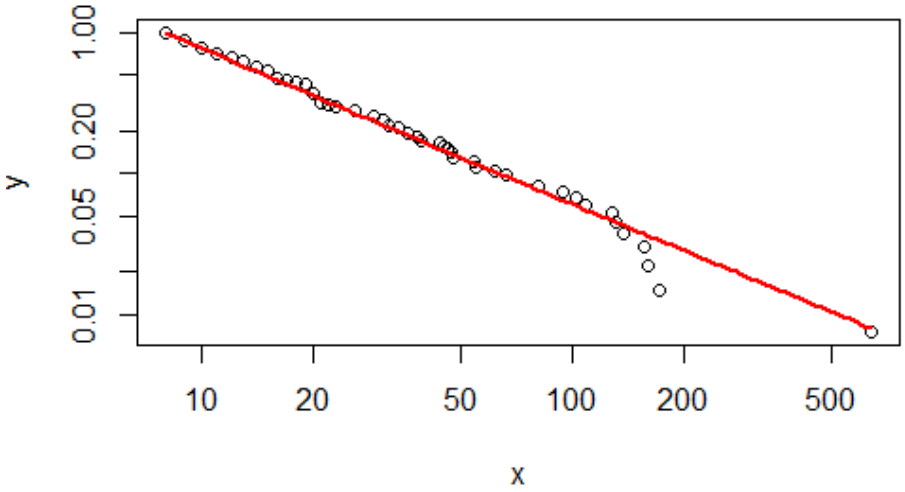
```
xmin = 8 +/- 2.149134
alpha = 1.926005 +/- 0.1060304
p_value = 0.9
p_value>0.1 indica que es plausible la hipotesis power-law
```

(c) Parámetros de la *Power Law*

Figura 5.27: Resultados *Power Law* para la ciudad de Phoenix



(a) Histograma de frecuencias de los *hashtags* en San Francisco

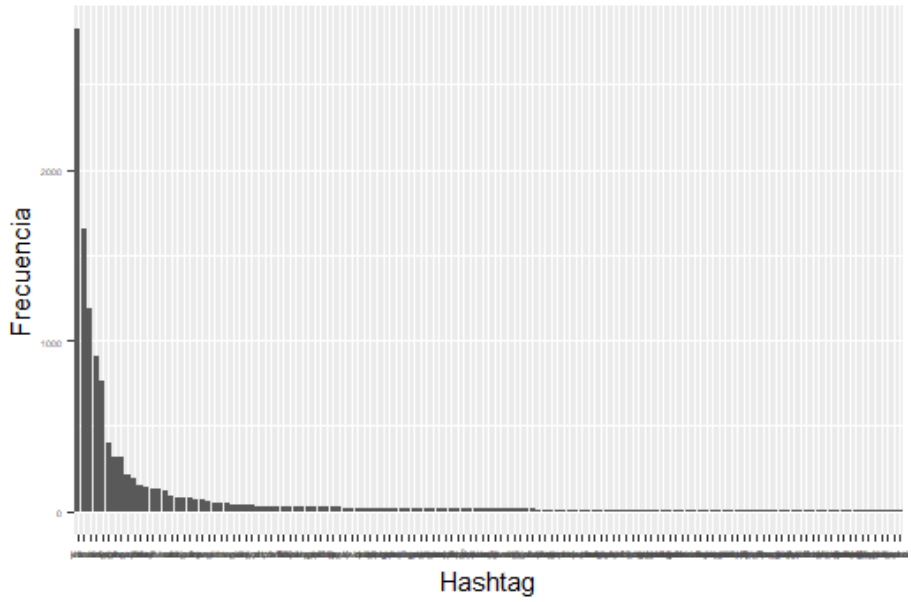


(b) Frecuencias en escala logarítmica

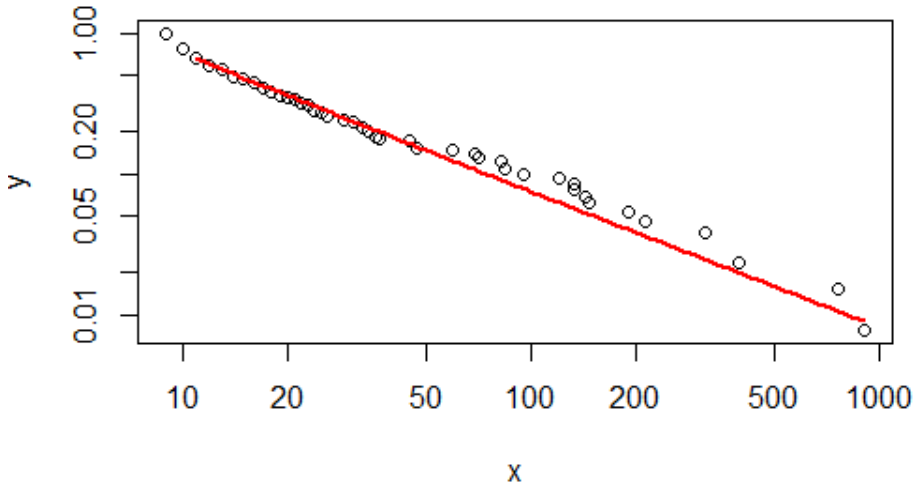
```
xmin = 8 +/- 4.095991
alpha = 2.085928 +/- 0.1390088
p_value = 0.14
p_value>0.1 indica que es plausible la hipotesis power-law
```

(c) Parámetros de la *Power Law*

Figura 5.28: Resultados *Power Law* para la ciudad de San Francisco



(a) Histograma de frecuencias de los *hashtags* en Washington

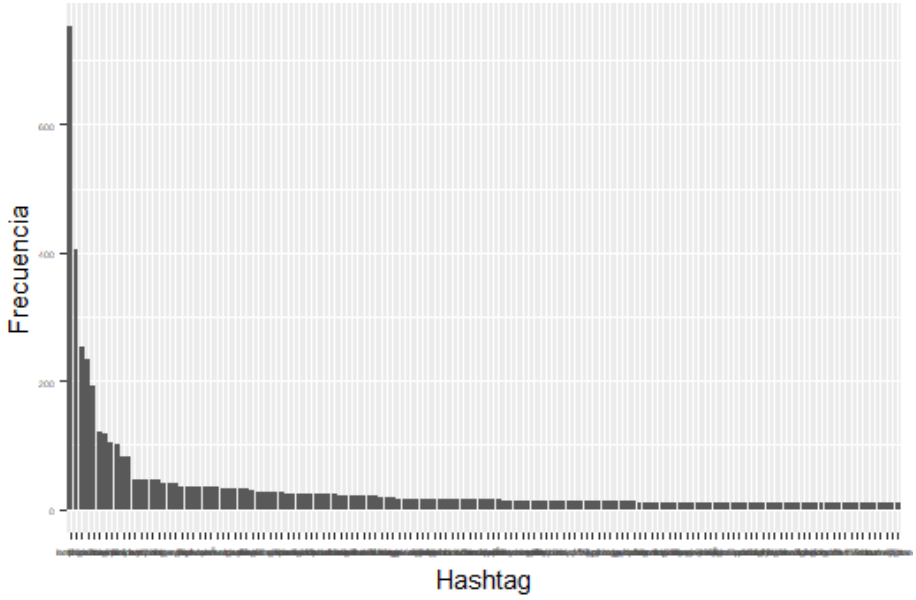


(b) Frecuencias en escala logarítmica

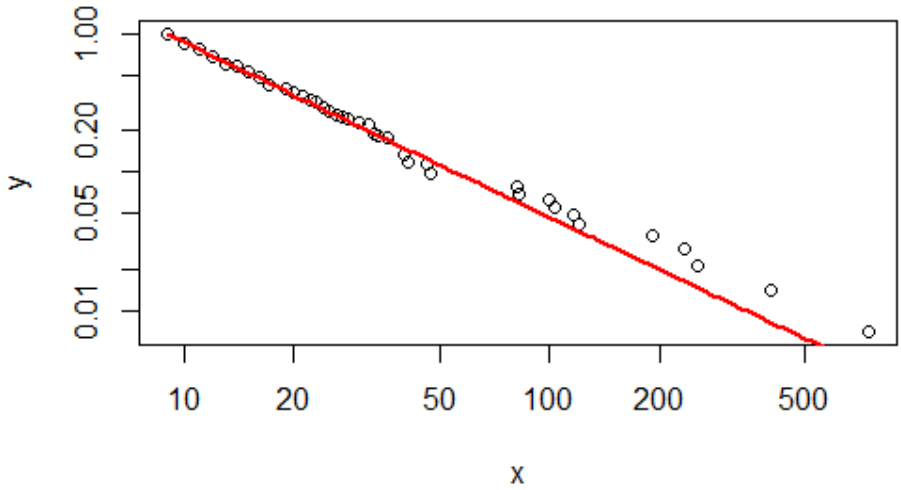
```
xmin = 11 +/- 1.628321
alpha = 1.962455 +/- 0.1388752
p_value = 0.72
p_value>0.1 indica que es plausible la hipotesis power-law
```

(c) Parámetros de la *Power Law*

Figura 5.29: Resultados *Power Law* para la ciudad de Washington



(a) Histograma de frecuencias de los *hashtags* en Londres

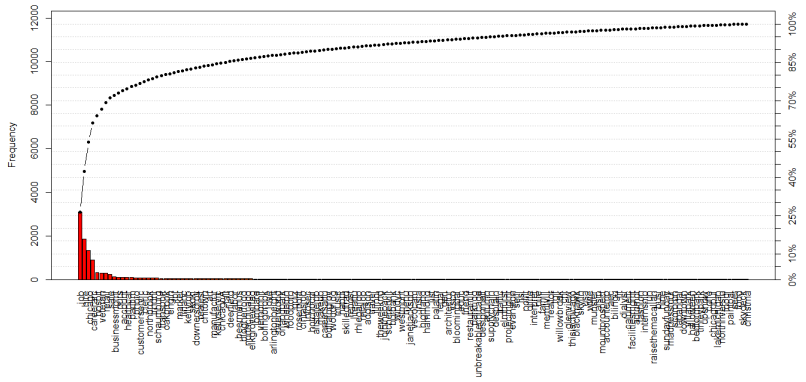


(b) Frecuencias en escala logarítmica

```
xmin = 9 +/- 0.4218521
alpha = 2.241616 +/- 0.08510036
p_value = 0.62
p_value>0.1 indica que es plausible la hipotesis power-law
```

(c) Parámetros de la *Power Law*

Figura 5.30: Resultados *Power Law* para la ciudad de Londres

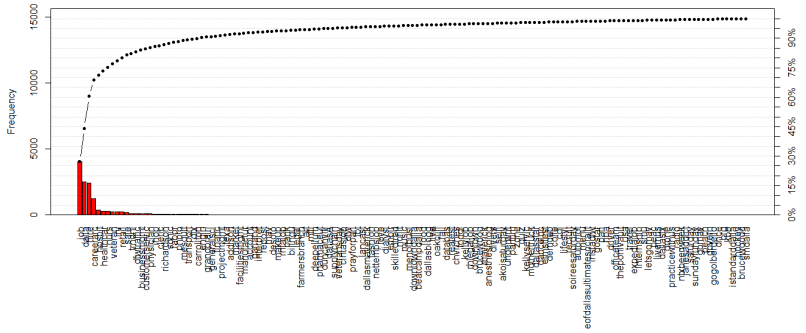


(a) Diagrama de Pareto

```
Pareto chart analysis for defect
           Frequency Cum.Freq.   Percentage Cum.Percent.
job             4046      4046 27.69146533     27.69147
dalla           2524      6570 17.27465608     44.96612
hire            2413      8983 16.51495449     61.48108
careerarc       1264     10247  8.65101636     70.13209
hospit           364     10611  2.49127370     72.62337
healthcar        313     10924  2.14222161     74.76559
nurs             272     11196  1.86161111     76.62720
veteran          250     11446  1.71103963     78.33824
irv              239     11685  1.63575388     79.97399
retail           226     11911  1.54677982     81.52077
```

(b) *Hashtags* más importantes

Figura 5.31: Diagrama de Pareto para la ciudad de Chicago

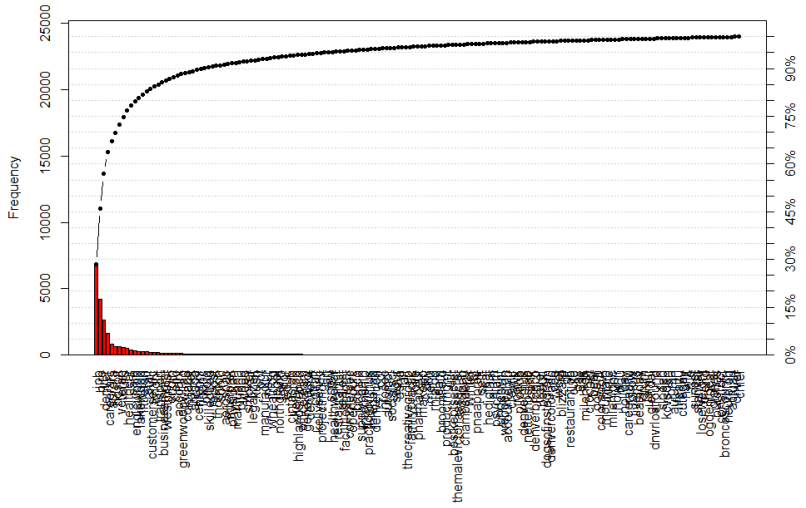


(a) Diagrama de Pareto

```
Pareto chart analysis for defect
           Frequency Cum.Freq.   Percentage Cum.Percent.
job             4046      4046 27.20182869     27.20183
dalla           2524      6570 16.96920801     44.17104
hire            2413      8983 16.22293936     60.39398
careerarc       1264     10247  8.49805029     68.89203
hospit           364     10611  2.44722334     71.33925
healthcar        313     10924  2.10434315     73.44359
nurs             272     11196  1.82869437     75.27229
veteran          250     11446  1.68078526     76.95307
irv              239     11685  1.60683071     78.55990
retail           226     11911  1.51942988     80.07933
```

(b) *Hashtags* más importantes

Figura 5.32: Diagrama de Pareto para la ciudad de Dallas

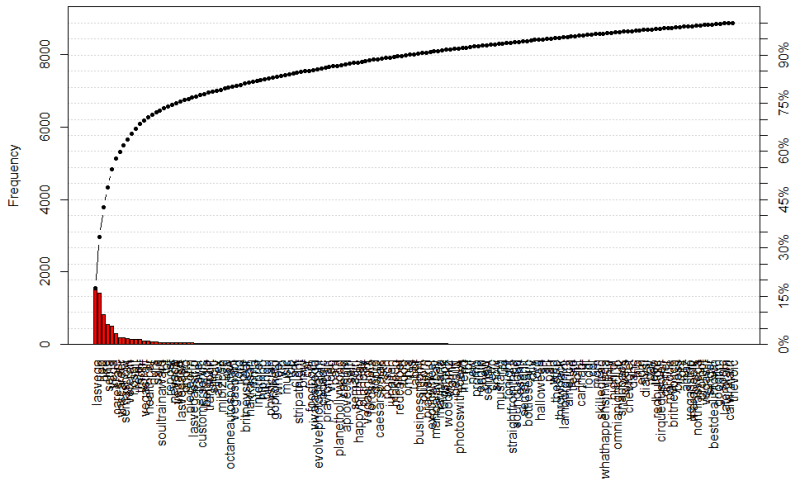


(a) Diagrama de Pareto

```
Pareto chart analysis for defect
           Frequency Cum.Freq.  Percentage Cum.Percent.
  job           6821      6821 28.42320193     28.42320
  hire          4230     11051 17.62646887     46.04967
  denver        2621     13672 10.92174348     56.97141
  careerarc     1638     15310  6.82556880     63.79698
  aurora         827     16137  3.44612051     67.24310
  hospit         615     16752  2.56271356     69.80582
  veteran        596     17348  2.48354030     72.28936
  nurs           588     17936  2.45020418     74.73956
  healthcar      473     18409  1.97099758     76.71056
  sale           401     18810  1.67097258     78.38153
  englewood      333     19143  1.38761563     79.76915
  lakewood       245     19388  1.02091841     80.79007
```

(b) *Hashtags* más importantes

Figura 5.33: Diagrama de Pareto para la ciudad de Denver



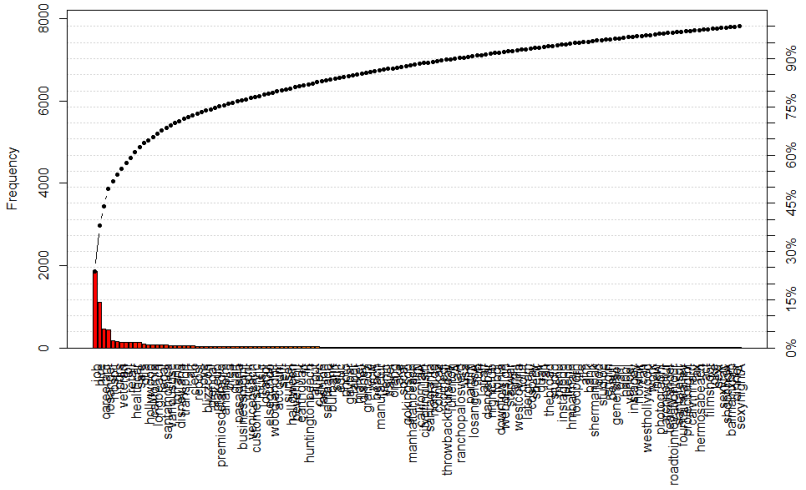
(a) Diagrama de Pareto

Pareto chart analysis for defect

	Frequency	Cum.Freq.	Percentage	Cum.Percent.
lasvega	1551	1551	17.4740874	17.47409
job	1413	2964	15.9193330	33.39342
hire	816	3780	9.1933303	42.58675
sema	555	4335	6.2528166	48.83957
vega	500	4835	5.6331681	54.47274
careerarc	291	5126	3.2785038	57.75124
vhscareer	190	5316	2.1406039	59.89184
semashow	189	5505	2.1293375	62.02118
veteran	158	5663	1.7800811	63.80126
hospit	149	5812	1.6786841	65.47995
retail	135	5947	1.5209554	67.00090
trndnl	130	6077	1.4646237	68.46553
vegastraff	100	6177	1.1266336	69.59216
healthcar	88	6265	0.9914376	70.58360
nurs	74	6339	0.8337089	71.41731
sale	61	6400	0.6872465	72.10455
soultrainaward	58	6458	0.6534475	72.75800
tbt	56	6514	0.6309148	73.38891
repost	54	6568	0.6083822	73.99730
nevada	48	6616	0.5407841	74.53808
vegasÃ	45	6661	0.5069851	75.04507
lasvegasÃ	39	6700	0.4393871	75.48445
photo	38	6738	0.4281208	75.91257
pieceofm	37	6775	0.4168544	76.32943
lasvegasstrip	37	6812	0.4168544	76.74628
travel	34	6846	0.3830554	77.12934
customerservic	34	6880	0.3830554	77.51239
bellagio	33	6913	0.3717891	77.88418
transport	32	6945	0.3605228	78.24470
sinciti	31	6976	0.3492564	78.59396
aapex	29	7005	0.3267237	78.92068
mjbizcon	29	7034	0.3267237	79.24741
usa	28	7062	0.3154574	79.56287
octaneautosportsÃ	27	7089	0.3041911	79.86706
vegasbabi	27	7116	0.3041911	80.17125

(b) *Hashtags* más importantes

Figura 5.34: Diagrama de Pareto para la ciudad de Las Vegas



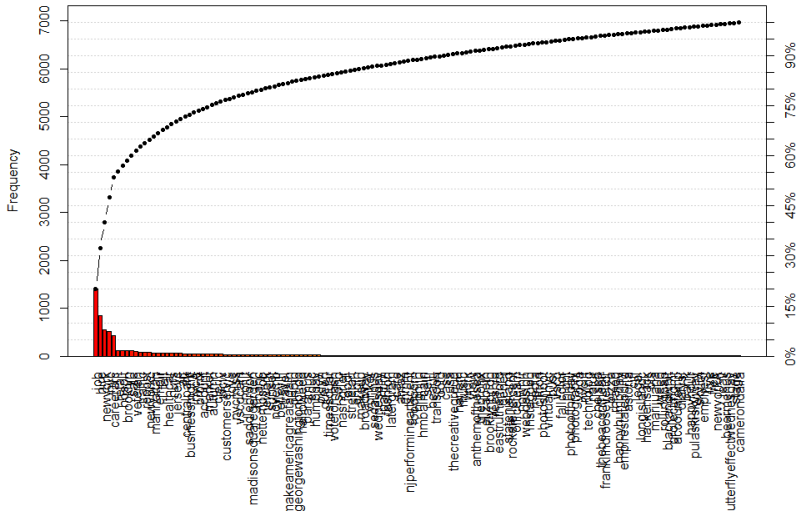
(a) Diagrama de Pareto

Pareto chart analysis for defect

	Frequency	Cum.Freq.	Percentage	Cum.Percent.
job	1864	1864	23.8546199	23.85462
hire	1110	2974	14.2052726	38.05989
careerarc	465	3439	5.9508574	44.01075
losangel	432	3871	5.5285385	49.53929
hospit	178	4049	2.2779626	51.81725
nurs	158	4207	2.0220118	53.83926
veteran	141	4348	1.8044535	55.64372
tbt	139	4487	1.7788585	57.42257
retail	135	4622	1.7276683	59.15024
healthcar	133	4755	1.7020732	60.85232
sale	130	4885	1.6636806	62.51600
dtla	89	4974	1.1389813	63.65498
hollywood	79	5053	1.0110059	64.66598
california	79	5132	1.0110059	65.67699
longbeach	78	5210	0.9982083	66.67520
legal	72	5282	0.9214231	67.59662
santamonica	63	5345	0.8062452	68.40287
varioususc	62	5407	0.7934477	69.19631
torranc	57	5464	0.7294599	69.92577
disneyland	52	5516	0.6654722	70.59125
transport	48	5564	0.6142821	71.20553
art	47	5611	0.6014845	71.80701
cleric	44	5655	0.5630919	72.37010
repost	40	5695	0.5119017	72.88201
love	36	5731	0.4607115	73.34272
blizzcon	35	5766	0.4479140	73.79063
glendal	34	5800	0.4351165	74.22575
makeup	34	5834	0.4351165	74.66087
premiosdelaradio	34	5868	0.4351165	75.09598
anaheim	33	5901	0.4223189	75.51830
brea	31	5932	0.3967238	75.91502
duart	31	5963	0.3967238	76.31175
pasadena	31	5994	0.3967238	76.70847
businessmgmt	27	6021	0.3455337	77.05401
trndnl	27	6048	0.3455337	77.39954
venicebeach	27	6075	0.3455337	77.74507
customerservic	27	6102	0.3455337	78.09061
music	27	6129	0.3455337	78.43614
losangl	27	6156	0.3455337	78.78167
elsegundo	26	6182	0.3327361	79.11441
woodlandhil	26	6208	0.3327361	79.44715
account	26	6234	0.3327361	79.77988
selfi	26	6260	0.3327361	80.11262

(b) *Hashtags* más importantes

Figura 5.35: Diagrama de Pareto para la ciudad de Los Ángeles



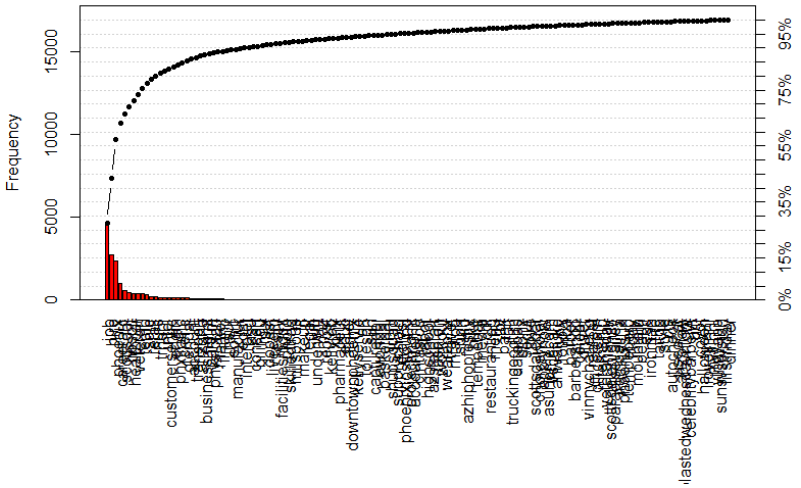
(a) Diagrama de Pareto

Pareto chart analysis for defect

	Frequency	Cum.Freq.	Percentage	Cum.Percent.
job	1404	1404	20.1637225	20.16372
hire	853	2257	12.2504668	32.41419
newyork	542	2799	7.7840011	40.19819
nyc	514	3313	7.3818756	47.58007
careerarc	423	3736	6.0749677	53.65503
retail	119	3855	1.7090335	55.36407
hospit	117	3972	1.6803102	57.04438
brooklyn	108	4080	1.5510556	58.59543
sale	106	4186	1.5223323	60.11777
veteran	99	4285	1.4218009	61.53957
traffic	86	4371	1.2350998	62.77467
repost	78	4449	1.1202068	63.89487
newyorkc	73	4522	1.0483987	64.94327
manhattan	68	4590	0.9765906	65.91986
empir	64	4654	0.9191440	66.83901
fall	63	4717	0.9047824	67.74379
healthcar	63	4780	0.9047824	68.64857
nurs	58	4838	0.8329743	69.48155
jerseyc	54	4892	0.7755278	70.25707
art	54	4946	0.7755278	71.03260
centralpark	49	4995	0.7037197	71.73632
businessmgmt	45	5040	0.6462732	72.38259
nycÃ	44	5084	0.6319115	73.01451
bronx	40	5124	0.5744650	73.58897
account	40	5164	0.5744650	74.16344
market	37	5201	0.5313802	74.69482
autumn	36	5237	0.5170185	75.21183
cleric	36	5273	0.5170185	75.72885
wcw	36	5309	0.5170185	76.24587
customerservic	32	5341	0.4595720	76.70544
love	32	5373	0.4595720	77.16502
nycmiss	29	5402	0.4164871	77.58150
vscocam	29	5431	0.4164871	77.99799
nofilt	28	5459	0.4021255	78.40011
saddlebrook	27	5486	0.3877639	78.78788
madisonsquaregarden	26	5512	0.3734023	79.16128
financ	25	5537	0.3590406	79.52032
nettempsjob	25	5562	0.3590406	79.87936
newark	25	5587	0.3590406	80.23840

(b) *Hashtags* más importantes

Figura 5.36: Diagrama de Pareto para la ciudad de Nueva York



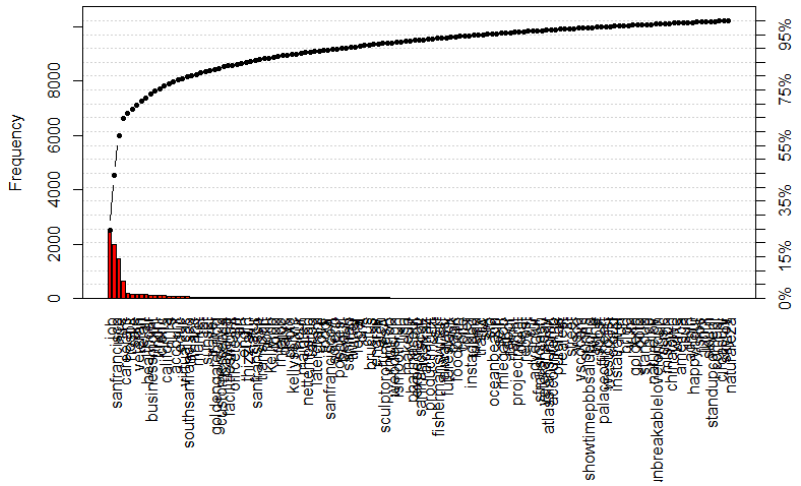
(a) Diagrama de Pareto

Pareto chart analysis for defect

	Frequency	Cum.Freq.	Percentage	Cum.Percent.
job	4614	4614	27.28885735	27.28886
hire	2731	7345	16.15211734	43.44097
phoenix	2347	9692	13.88100308	57.32198
careerarc	991	10683	5.86113083	63.18311
scottsdal	540	11223	3.19375444	66.37686
hospit	433	11656	2.56091791	68.93778
healthcar	371	12027	2.19422758	71.13201
veteran	362	12389	2.14099834	73.27301
temp	361	12750	2.13508398	75.40809
sale	347	13097	2.05228294	77.46037
retail	229	13326	1.35438846	78.81476
nurs	211	13537	1.24792997	80.06269

(b) *Hashtags* más importantes

Figura 5.37: Diagrama de Pareto para la ciudad de Phoenix

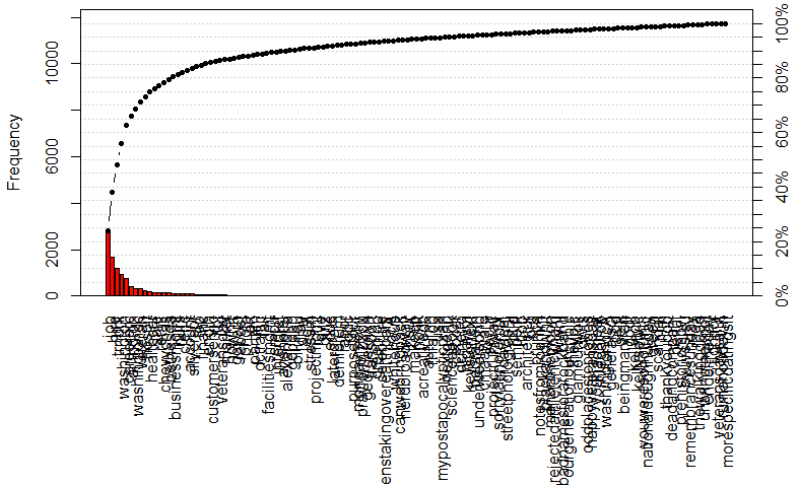


(a) Diagrama de Pareto

```
Pareto chart analysis for defect
                    Frequency Cum.Freq.  Percentage Cum.Percent.
job                      2520      2520 24.64306669     24.64307
sanfrancisco             1996      4516 19.51887346     44.16194
hire                     1466      5982 14.33600626     58.49795
careerarc                 644      6626  6.29767260     64.79562
hospit                    173      6799  1.69176609     66.48739
sale                      160      6959  1.56463916     68.05202
veteran                   157      7116  1.53530217     69.58733
retail                    138      7254  1.34950127     70.93683
healthcar                 131      7385  1.28104831     72.21788
businessmgmt              128      7513  1.25171132     73.46959
trndnl                    109      7622  1.06591042     74.53550
cleric                    102      7724  0.99745746     75.53296
california                 94      7818  0.91922550     76.45218
nurs                       81      7899  0.79209857     77.24428
account                    66      7965  0.64541365     77.88969
educ                       66      8031  0.64541365     78.53511
virtualcac                 62      8093  0.60629767     79.14140
southsanfrancisco          55      8148  0.53784471     79.67925
financ                     54      8202  0.52806571     80.20731
```

(b) *Hashtags* más importantes

Figura 5.38: Diagrama de Pareto para la ciudad de San Francisco

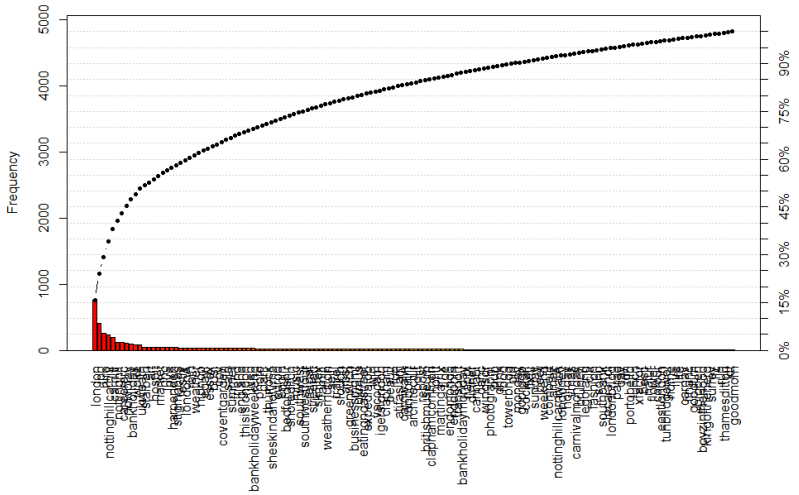


(a) Diagrama de Pareto

```
Pareto chart analysis for defect
           Frequency Cum.Freq.   Percentage Cum.Percent.
  job           2820      2820 24.02453570     24.02454
  hire          1653      4473 14.08246720     38.10700
  trndnl        1190      5663 10.13801329     48.24502
  washington     912      6575  7.76963708     56.01465
  careerarc      764      7339  6.50877492     62.52343
  arlington      396      7735  3.37365820     65.89709
  washingtondc   319      8054  2.71766911     68.61476
  veteran        319      8373  2.71766911     71.33242
  hospit         213      8586  1.81461919     73.14704
  healthcar      190      8776  1.61867439     74.76572
  sale           148      8924  1.26086216     76.02658
  retail         144      9068  1.22678480     77.25337
  chevychas      134      9202  1.14159141     78.39496
  bethesda       133      9335  1.13307207     79.52803
  businessmgmt   121      9456  1.03084001     80.55887
```

(b) *Hashtags* más importantes

Figura 5.39: Diagrama de Pareto para la ciudad de Washington



(a) Diagrama de Pareto

Pareto chart analysis for defect

	Frequency	Cum.Freq.	Percentage	Cum.Percent.
london	752	752	15.6243507	15.62435
job	406	1158	8.4354872	24.05984
hire	254	1412	5.2773738	29.33721
nottinghillcarniv	234	1646	4.8618325	34.19904
carniv	192	1838	3.9891959	38.18824
nottinghil	120	1958	2.4932475	40.68149
careerarc	117	2075	2.4309163	43.11240
nowplay	104	2179	2.1608145	45.27322
bankholiday	100	2279	2.0777062	47.35092
woke	83	2362	1.7244962	49.07542
ukweath	82	2444	1.7037191	50.77914
stalban	47	2491	0.9765219	51.75566
art	47	2538	0.9765219	52.73218
hospit	47	2585	0.9765219	53.70871
market	46	2631	0.9557449	54.66445
nhc	46	2677	0.9557449	55.62020
love	41	2718	0.8518595	56.47205
harpenden	40	2758	0.8310825	57.30314
stigmabas	40	2798	0.8310825	58.13422
festiv	36	2834	0.7479742	58.88219
londonÃ	36	2870	0.7479742	59.63017
rain	36	2906	0.7479742	60.37814
weather	36	2942	0.7479742	61.12612
hiphop	36	2978	0.7479742	61.87409
egayl	36	3014	0.7479742	62.62207
essex	34	3048	0.7064201	63.32849
busi	33	3081	0.6856431	64.01413
coy	32	3113	0.6648660	64.67899
coventgarden	32	3145	0.6648660	65.34386
sale	32	3177	0.6648660	66.00873
summer	32	3209	0.6648660	66.67359
friend	30	3239	0.6233119	67.29690
england	28	3267	0.5817577	67.87866
thisislondon	28	3295	0.5817577	68.46042
wed	27	3322	0.5609807	69.02140
bankholidayweekend	27	3349	0.5609807	69.58238
photo	26	3375	0.5402036	70.12258
parti	25	3400	0.5194266	70.64201
thurrock	25	3425	0.5194266	71.16144
sheskindahotvma	24	3449	0.4986495	71.66009
europ	24	3473	0.4986495	72.15874
engin	24	3497	0.4986495	72.65739
bedfordshir	23	3520	0.4778724	73.13526
shoreditch	23	3543	0.4778724	73.61313
music	23	3566	0.4778724	74.09100
southwest	23	3589	0.4778724	74.56888
southwestfour	22	3611	0.4570954	75.02597
reigat	21	3632	0.4363183	75.46229
streetart	21	3653	0.4363183	75.89861
skrillex	21	3674	0.4363183	76.33493
famili	20	3694	0.4155412	76.75047
weatherhutton	20	3714	0.4155412	77.16601
selfi	20	3734	0.4155412	77.58155
travel	19	3753	0.3947642	77.97631
slough	19	3772	0.3947642	78.37108
artist	19	3791	0.3947642	78.76584
greenwich	17	3808	0.3532101	79.11905
businessmgmt	17	3825	0.3532101	79.47226
ukmus	17	3842	0.3532101	79.82547
eatingndastreet	17	3859	0.3532101	80.17868

(b) *Hashtags* más importantes

Figura 5.40: Diagrama de Pareto para la ciudad de Londres

Se han observado ciertas discrepancias en cuanto al número de temas que se han obtenido con cada técnica, tal como se observa en la tabla 5.2. Con la técnica de la media armónica existe menor diferencia entre la cantidad de temas por ciudades, mientras que con las frecuencias de los *hashtags* sí existen bastantes diferencias en algunas ciudades. Esto puede deberse a que los *hashtags* se ven muy influenciados por eventos que estén ocurriendo en el momento, lo que provoca que *hashtags* alrededor de un mismo suceso tengan frecuencias muy parecidas y por tanto aumente el número de ellos con más relevancia.

Ciudad	**Cantidad de Temas**	
	Media armónica	**Hashtags**
Chicago	15	10
Dallas	16	10
Denver	14	12
Las Vegas	15	35
Los Ángeles	13	43
Nueva York	19	39
Phoenix	17	12
San Francisco	13	19
Washington	18	15
Londres	18	60

Tabla 5.2: Comparación número de temas con ambas técnicas

Con el propósito de determinar si en un mismo intervalo de tiempo el número de temas en las ciudades se comportaría de forma similar, se decide aplicar las técnicas anteriores en un intervalo de diecinueve días, con excepción de Londres porque como se observó en la tabla 4.3, en esta ciudad el número de tuits generados supera en gran cantidad el de las demás, por lo que solo se tomaron catorce días. Cabe recordar que a pesar de la gran cantidad de datos que se generan en Twitter, menos del uno por ciento de los tuits están geoposicionados (Jurgens et al., 2015), lo cual supone una limitación de información para este trabajo.

Después de realizado todo el proceso necesario, se extrae el número de temas tal como se observa en la tabla 5.3. En ésta se comprueba lo dicho anteriormente en cuanto a la cantidad de temas que surgen con ambas técnicas, es decir, que con la frecuencia de los *hashtags* existe mayor discrepancia entre las cantidades.

Ciudad	**Cantidad de Temas**	
	Media armónica	**Hashtags**
Chicago	17	8
Nueva York	16	23
Washington	13	10
Londres	18	57

Tabla 5.3: Comparación número de temas con ambas técnicas en un mismo período de tiempo

5.2 Extracción de los Temas con el Algoritmo LDA

En esta sección se muestran los temas obtenidos con el algoritmo LDA. Al final se realiza una comparación entre los temas que éste genera y los que realmente se observan estudiando cada una de las agrupaciones.

Las dos técnicas implementadas para determinar el número de temas dan una aproximación de la cantidad que podrían encontrarse en los tuits de cada una de las ciudades. En este caso se utiliza el número obtenido con la media armónica, porque con ésta las diferencias entre una ciudad y otra es menor, a parte de que este método ya se ha probado en varios estudios de este tipo.

Los temas obtenidos en las ciudades se clasifican en catorce categorías, que son: trabajo, entretenimiento, salud, viaje, música, deporte, clima, transporte, turismo, tecnología, eventos, educación, *trending topics* y finalmente una categoría denominada "indefinido" para aquellos temas en los que exista una mezcla de términos y que puedan clasificarse en varias categorías a la vez.

Dentro de la categoría trabajo se realizaron dos subdivisiones, una de trabajos relacionados con el área de salud y otra de trabajos en general. Se realiza de esta manera debido a que algunos temas relacionados con trabajo también contienen las palabras enfermera, cuidado y seguro de salud, lo que lleva a pensar que se trata de algún trabajo relacionado con esa área.

La categoría de entretenimiento se subdividió en tres categorías, las cuales son diurno, nocturno y general. La primera se debe a que algunos temas incluyen términos relacionados con el día, tales como naturaleza y parque, lo mismo pasa con el segundo, en el que se encuentra el término *"night"* que en español significa noche. La última es para los temas relacionados con entretenimiento que directamente no hacen referencia al momento del día en el que son publicados.

La última categoría subdividida es turismo, en la que se encuentran: infraestructura de la ciudad, cultura y general. En la primera pueden encontrarse términos como construcción, avenida y calle, en la segunda algunos como galería, museo y arte. La subcategoría general es aquella en la que aparecen términos que hacen referencia al turismo en general, como foto, hotel, el nombre de la ciudad y algún lugar en específico de la ciudad.

De las categorías, una que está presente en todas las ciudades es la de trabajo, tal como se aprecia en las tablas desde la 5.4 hasta la 5.13. El tema del trabajo es común para todas la ciudades, tanto de Estados Unidos como Londres, en ellas siempre están presente los términos contratación y trabajo. Dentro de las investigaciones realizadas se encontró que existen muchos anuncios dentro de los tuits publicados, los cuales están relacionados con ofertas de trabajo, por lo que es compresible la popularidad del tema.

Otro tema que está presente en el 80 por ciento de las ciudades es el del clima, se debe a que en Twitter existen usuarios que se denominan "bots", estos son usuarios no humanos, detrás de ellos existe un algoritmo que genera información automática ya sea del tiempo, tráfico o cualquier tema de interés tal como lo define Chu et al. (2012) en su artículo relacionado con la detección de ese tipo de usuarios. También el tema de los *trending topic* es un "bot", este genera los te-

mas de tendencia en Twitter en cada momento del día. Por este motivo siempre aparecen temas relacionados con la temperatura, humedad y temas de tendencia.

El primer resultado corresponde a la ciudad de Dallas (ver tabla 5.4), en esta se observan 16 temas, de los que el más repetido es entretenimiento, esto denota que las personas tienen a publicar mensajes en Twitter cuando se encuentran en momentos de diversión o relajación. En esta ciudad al igual que en Phoenix (ver tabla 5.11) son las únicas en las que aparece un tema relacionado con deporte, esto puede deberse a que en las fechas en las que se tomaron los datos en las demás ciudades no había ningún partido muy relevante.

En la ciudad de Londres (ver tabla 5.5) también se habla mucho de temas relacionados con entretenimiento, pero en esta coincidió la fecha del análisis con su carnaval, *Notting Hill Carnival*, por este motivo también surgió la categoría evento.

En Chicago (ver tabla 5.6) también está en primer lugar la categoría entretenimiento a la que le sigue la de trabajo. Lo mismo sucede en Denver (ver tabla 5.7), Los Ángeles (ver tabla 5.9), Nueva York (ver tabla 5.10), Phoenix (ver tabla 5.11), San Francisco (ver tabla 5.13) y Washington (ver tabla 5.12). Al igual que en todas las ciudades, los usuarios suelen publicar tuits relacionados con los lugares en donde se encuentran o las cosas que están haciendo en ese instante.

La ciudad de Las Vegas (ver tabla 5.8) es la que más contiene respecto a las categorías de turismo y entretenimiento. Esto era de esperar, porque Las Vegas es uno de los principales destinos turísticos de los Estados Unidos gracias a sus zonas comerciales y vacacionales, pero sobre todo por sus casinos.

Algunos temas muy específicos que denotan eventos del momento fueron encontrados en algunas ciudades como, Washington (ver tabla 5.12), en donde surgió el evento del día de los Veteranos, término encontrado en las demás ciudades de Estados Unidos pero que no se definía como un tema por sí solo. También en Los Ángeles, un temblor de tierra cercano al Barrio de Pacoima el cual ocurrió en noviembre 2015, fecha en la que fueron tomados los datos. Estos eventos corroboran el hecho de que Twitter es un medio informativo en el que muchos sucesos importantes son difundidos de forma instantánea.

Temas	Palabras	Categoría
Tema 1	job, dalla, click, see, hire, appli, work, manag, latest, want	Trabajo
Tema 2	new, team, park, hous, citi, will, join, season, art, food	Entretenimiento diurno
Tema 3	worth, trend, dallasft, trndnl, know, last, hour, night, place, friend	Viaje
Tema 4	great, job, fit, might, sale, hire, engin, richardson, group, transport	Trabajo
Tema 5	job, dalla, hire, careerarc, nurs, healthcar, utsw, care, addison, alert	Trabajo/salud
Tema 6	amp, come, live, make, one, week, girl, tri, amaz, celebr	Música / estación de radio
Tema 7	job, hire, dalla, open, can, careerarc, hospit, veteran, latest, anyon	Trabajo/salud
Tema 8	drink, good, got, morn, photo, blue, way, ball, watch, meet	Entretenimiento diurno
Tema 9	center, american, game, airlin, start, star, mav, let, play, first	Deporte (baloncesto)
Tema 10	love, like, man, beauti, fun, home, say, peopl, yall, much	Entretenimiento
Tema 11	get, now, mph, year, wind, take, fall, cant, factori, humid	Clima
Tema 12	dalla, texa, happi, birthday, bank, dal, field, shop, map, theatr	Entretenimiento
Tema 13	just, tonight, post, photo, thank, today, downtown, big, studio, readi	Entretenimiento/fotografia
Tema 14	time, dont, look, music, bar, even, well, life, event, parti	Entretenimiento nocturno / fiesta
Tema 15	irv, back, dalla, stop, traffic, area, accid, right, lane, min	Transporte / accidente de tráfico en Irving
Tema 16	drink, new, photo, smu, ball, leav, parti, fashion, part, sweet	Entretenimiento

Tabla 5.4: Temas en la ciudad de Dallas

Temas	Palabras	Categoría
Tema 1	west, south, clapham, four, common, beauti, festiv, old, studio, head	Evento
Tema 2	london, greater, station, bridg, tower, hounslow, eye, lhr, underground, railway	Viaje/ turismo
Tema 3	one, bong, will, end, world, made, anoth, stalban, harpenden, boy	Viaje
Tema 4	year, bankholiday, home, girl, ive, ever, place, tomorrow, let, selfi	Evento (día festivo)
Tema 5	love, take, peopl, yesterday, airport, heathrow, termin, two, club, town	Turismo
Tema 6	amp, new, look, weather, right, fun, week, wed, guy, famili	Entretenimiento en familia
Tema 7	now, today, bar, make, nowplay, first, watch, need, download, wait	Entretenimiento
Tema 8	time, garden, big, way, citi, covent, even, hotel, green, squar	Turismo
Tema 9	best, summer, realli, england, your, check, man, hiphop, greenwich, free	Turismo
Tema 10	job, hire, see, great, work, want, careerarc, open, latest, team	Trabajo
Tema 11	just, photo, post, thank, can, drink, life, bit, game, give	Entretenimiento
Tema 12	day, good, night, like, last, morn, dont, cant, well, think	Indefinido
Tema 13	get, rain, happi, lol, today, friend, birthday, essex, food, show	Entretenimiento
Tema 14	carniv, hill, not, nottinghillcarniv, nottinghil, parti, windsor, nhc, legoland, noth	Evento
Tema 15	holiday, bank, back, unit, much, weekend, kingdom, know, monday, pleas	Evento (día festivo)
Tema 16	come, road, art, still, feel, shoreditch, sunday, hope, restaur, soho	Turismo
Tema 17	street, got, hous, miss, play, that, littl, video, run, music	Entretenimiento
Tema 18	park, follow, live, next, final, amaz, better, win, nice, queen	Entretenimiento diurno

Tabla 5.5: Temas en la ciudad de Londres

Temas	Palabras	Categoría
Tema 1	just, photo, post, today, happi, morn, game, music, video, boy	Entretenimiento
Tema 2	love, one, like, miss, much, even, lake, school, avail, girl	Indefinido
Tema 3	drink, get, come, littl, way, right, favorit, weekend, still, famili	Entretenimiento en familia
Tema 4	park, fall, mph, let, cloudi, humid, peopl, year, wind, start	Clima
Tema 5	job, hire, latest, open, click, see, work, appli, careerarc, want	Trabajo
Tema 6	chicago, hous, take, follow, line, downtown, theatr, world, blue, chitown	Entretenimiento
Tema 7	citi, beauti, back, will, tower, fun, street, hotel, made, thing	Turismo
Tema 8	great, fit, center, might, unit, near, interest, market, custom, repres	Compras
Tema 9	job, hire, careerarc, sale, retail, can, hospit, veteran, anyon, recommend	Trabajo
Tema 10	amp, dont, bar, show, week, tri, feel, big, grill, tonight	Entretenimiento nocturno
Tema 11	time, thank, account, part, make, season, nurs, care, maci, man	Salud
Tema 12	chicago, illinoi, good, intern, airport, home, hot, ohar, chocol, morn	Viaje
Tema 13	night, friend, trend, place, look, parti, friday, play, readi, trndnl	Entretenimiento nocturno
Tema 14	day, got, last, need, best, first, night, birthday, amaz, food	Entretenimiento nocturno
Tema 15	new, art, check, club, life, your, know, univers, bean, event	Evento

Tabla 5.6: Temas en la ciudad de Chicago

Temas	Palabras	Categoría
Tema 1	job, hire, denver, appli, click, sale, retail, manag, careerarc, littleton	Trabajo
Tema 2	denver, citi, look, interpret, beauti, littl, season, bilingu, bar, creek	Entretenimiento
Tema 3	latest, open, see, aurora, read, team, join, view, music, servic	Entretenimiento
Tema 4	time, good, tonight, get, come, bronco, news, full, new, man	Entretenimiento
Tema 5	today, thank, one, will, best, home, tomorrow, morn, repost, live	Indefinido
Tema 6	denver, trend, trndnl, place, hour, know, top, topic, hashtag, took	Trending topic
Tema 7	job, veteran, hire, great, denver, fit, might, engin, account, careerarc	Trabajo
Tema 8	day, got, art, theÃ, final, denverÃ, press, anoth, fun, amaz	Entretenimiento
Tema 9	denver, colorado, amp, center, transport, driver, downtown, life, hotel, school	Viaje
Tema 10	just, night, high, mile, theatr, take, food, say, sport, marijuana	Entretenimiento
Tema 11	job, hire, hospit, can, anyon, recommend, careerarc, aurora, denver, healthcar	Trabajo
Tema 12	drink, love, photo, make, like, tri, brew, feel, lol, girl	Entretenimiento
Tema 13	job, hire, nurs, work, want, denver, health, lakewood, click, detail	Trabajo
Tema 14	mph, wind, humid, pressur, just, fall, post, temperatur, now, photo	Clima

Tabla 5.7: Temas en la ciudad de Denver

Temas	Palabras	Categoría
Tema 1	vegastraff, accid, club, fun, downtown, fremont, blvd, girl, say, thing	Evento (accidente)
Tema 2	got, trend, drink, trndnl, make, hour, know, place, week, top	Entretenimiento nocturno
Tema 3	thank, mph, now, beauti, wind, part, humid, pressur, right, spring	Clima
Tema 4	vega, las, nevada, cosmopolitan, wynn, hakkasan, via, north, hello, disturb	Turismo-hoteles
Tema 5	night, get, day, last, parti, first, can, tonight, readi, saturday	Entretenimiento nocturno
Tema 6	lasvega, show, valley, vhscareer, need, health, system, season, booth, support	Salud
Tema 7	sema, new, show, semashow, york, car, start, repost, wheel, ford	Evento Sema Show
Tema 8	today, airport, mccarran, intern, hollywood, morn, planet, life, britney, still	Transporte- aeropuerto
Tema 9	nightclub, birthday, happi, dont, amaz, lol, rock, much, que, guy	Entretenimiento nocturno
Tema 10	just, photo, post, good, like, look, grand, bar, back, mgm	Turismo
Tema 11	job, lasvega, hire, great, see, careerarc, work, click, veteran, hospit	Trabajo
Tema 12	amp, casino, hotel, one, resort, meet, dinner, aria, luxor, stratospher	Turismo / hoteles
Tema 13	time, strip, tonight, center, pari, will, take, let, year, next	Entretemiento nocturno
Tema 14	love, bellagio, venetian, high, travel, citi, game, stop, roller, man	Turismo / hoteles
Tema 15	come, best, friend, palac, caesar, miss, big, soultrainaward, room, restaur	Turismo / hoteles

Tabla 5.8: Temas en la ciudad de Las Vegas

Temas	Palabras	Categoría
Tema 1	hire, california, angel, will, amp, day, job, los, one, sunset	Trabajo
Tema 2	hollywood, santa, hire, just, monica, year, new, losangel, job, dtla	Trabajo
Tema 3	amp, nurs, hire, sunset, california, citi, los, studio, great, sunni	Indefinido
Tema 4	angel, los, california, job, hire, beach, thank, great, happi, new	Trabajo
Tema 5	job, work, want, latest, hire, los, click, angel, see, view	Trabajo
Tema 6	humid, mph, wind, pressur, weather, rise, fair, visibl, current, los	Clima
Tema 7	angel, day, los, get, like, now, earthquak, pacoima, hollywood, amaz	Evento (catástrofe natural)
Tema 8	hire, job, hospit, click, california, appli, angel, beach, starbuck, latest	Trabajo
Tema 9	california, angel, amp, get, got, back, center, night, best, time	Entretenimiento Nocturno
Tema 10	hollywood, good, losangel, photo, love, night, los, hire, open, fame	Turismo
Tema 11	los, angel, california, just, losangel, hire, time, post, see, get	Turismo
Tema 12	"job, careerarc, can, hire, retail, anyon, recommend, healthcar, sale, veteran"	Trabajo/Salud
Tema 13	california, hollywood, today, los, amp, night, just, job, temperatur, get	Clima

Tabla 5.9: Temas en la ciudad de Los Ángeles

Temas	Palabras	Categoría
Tema 1	nyc, littl, newyorkc, top, west, shop, harlem, rock, broadway, updat	Entretenimiento
Tema 2	citi, time, squar, east, friend, live, studio, club, madison, theatr	Entretenimiento
Tema 3	today, come, like, friday, restaur, fun, center, shot, tri, light	Entretenimiento / almuerzo
Tema 4	job, hire, great, careerarc, can, jersey, retail, fit, hospit, anyon	Trabajo
Tema 5	take, alway, long, call, big, never, class, lol, run, red	Indefinido
Tema 6	empir, bar, life, need, anoth, lunch, astoria, shoot, grand, termin	Entretenimiento/almuerzo
Tema 7	just, photo, post, hall, school, video, book, webster, walk, williamsburg	Educación
Tema 8	job, newyork, hire, see, work, latest, open, careerarc, want, view	Trabajo
Tema 9	art, now, museum, check, your, event, high, wait, follow, center	Turismo
Tema 10	island, hous, part, sunset, stage, busi, natur, nice, plaza, season	Entretenimiento diurno
Tema 11	new, york, sight, librari, public, sport, penn, annual, nypl, sidelow	Educacion
Tema 12	love, best, morn, airport, intern, john, kennedi, fashion, liberti, weather	Viaje
Tema 13	tonight, happi, show, repost, music, dont, girl, miss, favorit, meet	Entretenimiento
Tema 14	brooklyn, back, manhattan, traffic, stop, ave, bronx, bridg, delay, state	Transporte / tráfico
Tema 15	get, look, good, make, way, thing, amaz, know, let, say	Indefinido
Tema 16	street, avenu, beauti, construct, place, build, even, que, con, special	Infraestructura de la ciudad
Tema 17	amp, got, week, still, next, dinner, wine, nov, man, media	Entretenimiento / almuerzo
Tema 18	night, one, thank, last, novemb, wednesday, birthday, drink, world, parti	Entretenimiento nocturno
Tema 19	park, day, central, fall, home, first, queen, station, autumn, soho	Entretenimiento diurno

Tabla 5.10: Temas en la ciudad de Nueva York

Temas	Palabras	Categoría
Tema 1	arizona, phoenix, fun, scottsdal, airport, good, new, like, intern, name	Viaje
Tema 2	job, see, click, hospit, phoenix, appli, center, amp, engin, veteran	Trabajo
Tema 3	job, phoenix, hire, amp, will, intel, game, arizona, get, love	Trabajo
Tema 4	job, phoenix, temp, time, today, get, drink, love, first, one	Indefinido
Tema 5	phoenix, arizona, great, scottsdal, hire, see, insur, tonight,just, fit	Trabajo
Tema 6	phoenix, amp, arizona, hire, job, love, day, physician, grand,wine	Trabajo
Tema 7	mph, wind, humid, temperatur, near, pressur, just, fall, scan,weather	Clima
Tema 8	phoenix, arizona, sport, kid, scottsdal, athlet, tri, chicken,even, celebr	Deporte
Tema 9	phoenix, night, just, scottsdal, state, job, new, amp, latest,talk	Trabajo
Tema 10	phoenix, job, arizona, get, sky, phxtraffic, hire, temp, fit, latest	Indefinido
Tema 11	scottsdal, phoenix, arizona, time, hire, latest, retail, park, good, get	Trabajo
Tema 12	job, hire, phoenix, veteran, transport, hospit, driver, arizona, great, team	Trabajo / Salud
Tema 13	phoenix, amp, great, happi, job, day, arizona, place, top, thank	Trabajo
Tema 14	phoenix, trend, trndnl, hour, topic, thursday, monday, took, phx, tuesday	Trending topic
Tema 15	arizona, job, temp, open, market, latest, phoenix, state, read, happi	Indefinidod
Tema 16	job, phoenix, hire, scottsdal, careerarc, interpret, customerservic, latest, hospit, nurs	Trabajo
Tema 17	job, hire, temp, sale, careerarc, manag, account, phoenix, scottsdal, anyon	Trabajo

Tabla 5.11: Temas en la ciudad de Phoenix

Temas	Palabras	Categoría
Tema 1	just, see, open, photo, post, latest, nation, read, univers, georgetown	Turismo
Tema 2	new, one, tonight, club, event, big, real, market, much, dinner	Entretenimiento nocturno
Tema 3	can, anyon, look, recommend, know, station, need, metro, interpret, never	Viaje
Tema 4	twitter, memori, happi, monument, birthday, client, web, iphon, topapp, lincoln	Turismo
Tema 5	amp, today, drink, best, work, street, take, loung, right, tbt	Entretenimiento
Tema 6	night, thank, like, last, friday, friend, silver, spring, amaz, weekend	Entretenimiento
Tema 7	day, get, hous, white, park, capitol, way, veteransday, coffe, virginia	Evento
Tema 8	washington, live, nurs, healthcar, check, your, care, room, honor, cafe	Salud
Tema 9	job, arlington, hire, click, appli, want, engin, manag, retail, work	Trabajo
Tema 10	time, center, mph, will, fall, verizon, humid, wind, cloudi, show	Clima
Tema 11	trndnl, trend, unit, topic, state, now, mention, tweet, rts, hour	Trending topic
Tema 12	peopl, sinc, seen, becam, let, school, home, man, celebr, anoth	Indefinido
Tema 13	good, come, nation, lol, museum, realli, alway, life, galleri, bar	Cultura
Tema 14	district, love, columbia, dont, got, first, make, start, even, shit	Turismo
Tema 15	cst, trndnl, great, fit, might, daysuntilmitam, meet, babi, paul, grill	Entretenimiento diurno
Tema 16	washingtondc, morn, back, wait, cant, littl, fuck, old, two, ave	Indefinido
Tema 17	job, hire, careerarc, veteran, hospit, sale, bethesda, account, manag, starbuck	Trabajo
Tema 18	trend, alert, trndnl, place, top, saturday, tuesday, took, monday, wednesday	Trending topic

Tabla 5.12: Temas en la ciudad de Washington

Temas	Palabras	Categoría
Tema 1	open, amp, via, request, graffiti, iphon, android, park, street, sidewalk	Tecnología
Tema 2	just, citi, photo, post, golden, bridg, gate, dali, bay, happi	Entretenimiento
Tema 3	job, sanfrancisco, hire, careerarc, latest, click, appli, sale, manag, hospit	Trabajo
Tema 4	trend, know, trndnl, hour, place, top, room, light, friday, topic	Trending topic
Tema 5	AA, ÂÃ, â€šÃ, ÂÃ, ÂÃ, â€™â€, love, pier, take, amaz	Indefinido
Tema 6	san, francisco, can, anyon, recommend, hill, mile, full, heart, fran	Indefinido
Tema 7	california, sanfrancisco, center, art, nurs, home, need, medic, sunset, district	Salud
Tema 8	now, coffe, food, sky, right, loung, mph, new, weather, wind	Clima
Tema 9	year, like, start, hall, tonight, best, day, will, music, see	Entretenimiento
Tema 10	night, get, sanfrancisco, great, day, new, squar, back, today, bar	Entretenimiento
Tema 11	good, â€, drink, look, morn, beach, like, one, ËœÅ, ËœÃ	Entretenimiento
Tema 12	work, view, peopl, cafe, want, still, wait, fun, alway, travel	Viaje / trabajo
Tema 13	close, case, request, street, resolv, complet, graffiti, â€™Ã, â€™, south	Infraestructura de la ciudad

Tabla 5.13: Temas en la ciudad de San Francisco

Al analizar cada uno de los temas encontrados en las ciudades se observaron algunas categorías que se repiten aunque con diferentes palabras. Por tal motivo se realiza una tabla en la que se muestran todas las categorías que de acuerdo con la técnica de la media armónica se encuentran en las ciudades y al final de la misma una fila en la que se coloca el número de categorías que realmente representan los tuits de cada ciudad (ver figura 5.41).

Tal como muestran los resultados en la figura 5.41, el promedio de los temas encontrados por ciudad está alrededor de ocho, mientras que anteriormente era dieciséis, esto significa que después del análisis individual de cada tema estos se redujeron a la mitad.

Los temas más comunes en todas las ciudades son trabajo, entretenimiento, viaje, clima, turismo, eventos y salud. El resto de temas son poco comunes, esto no significa que de ellos no se hable, puesto que se debe tener en cuenta que la cantidad de tuits analizados son geoposicionados, lo que implica analizar una cantidad muy pequeña en comparación con la cantidad diaria que se genera en cada ciudad.

Categorías		Ciudades									
		Dallas	Londres	Chicago	Denver	Las Vegas	Los Ángeles	Nueva York	Phoenix	Washington	San Francisco
Trabajo	Salud	2					1		1		
	General	2	1	2	4	1	5	2	9	2	1
Entretenimiento	Diurno	2						2		1	
	Nocturno	1	1	3		4	1	1		1	
	General	4	5	3	6			6		2	4
Salud			2	1		1				1	1
Viaje		1	1	1	1			1	1	1	1
Música		1									
Deporte		1							1		
Clima		1		1	1	1	2		1	1	1
Transporte		1				1		1			
Turismo	Infraestructura de la ciudad							1			1
	Cultura							1		1	
	General		4	1		5	2			3	
Tecnología											1
Evento			4	1		2	1			1	
Compras				1							
Educación								2			
Trending Topic					1				1	2	1
Indefinido				1	1		1	2	3	2	2
Total (Temas)		10	7	10	6	7	7	10	7	12	9

Figura 5.41: Cantidad de temas por ciudad

Ahora se analizan los resultados de los tuits tomados durante un mismo lapso de tiempo con el objetivo de verificar si existe alguna variación en cuanto a las categorías determinadas o si siguen siendo las mismas a pesar del período de tiempo.

Observando las tablas desde la 5.14 hasta la 5.17, se obtiene que independientemente del tiempo en que hayan sido creados los tuits, las categorías que aparecen son las mismas. Lo que sí pueden variar son los términos utilizados y también los eventos que ocurren, porque como ya se ha mencionado, los usuarios suelen utilizar las redes sociales para tratar temas que estén ocurriendo en el momento.

En este caso se observan algunas palabras como *mother* y *day* que en ciudades como Nueva York y Washington hacen referencia al tema del día de las madres, el cual en Estados Unidos fue el domingo 14 de mayo de 2017, fecha contenida en el rango de tiempo en que fueron generados los tuits. En este caso, al ser tan pocos los días que fueron analizados, la cantidad real de temas está alrededor de cinco.

Temas	Palabras	Categoría
Tema 1	post, just, chicago, photo, hous, ball, lol, lisa, schmitz, check	Entretenimiento
Tema 2	engin, softwar, center, consult, summer, amp, hospit, ticket, drink, manag	Indefinido
Tema 3	just, fun, left, say, way, pint, night, tomorrow, one, last	Entretenimiento nocturno
Tema 4	job, hire, chicago, see, latest, open, join, team, careerarc, care	Trabajo
Tema 5	anyon, can, recommend, job, drink, photo, manag, hospit, hire, amp	Trabajo
Tema 6	morn, get, chicago, time, well, law, thing, launch, ootd, wow	Moda
Tema 7	now, taco, time, dont, one, love, know, ive, debonairsocialclub, especi	Entretenimiento
Tema 8	work, chicago, look, job, check, your, businessmgmt, side, south, top	Trabajo
Tema 9	job, chicago, hire, latest, open, read, sale, work, want, view	Trabajo
Tema 10	chicago, illinoi, citi, west, take, selfi, will, plane, latergram, open	Turismo
Tema 11	chanc, tstorm, may, forecast, tonight, mon, chicago, shower, today, stock	Clima
Tema 12	chicago, field, wrigley, park, cub, secur, beer, session, lincoln, work	Entretenimiento
Tema 13	lake, chicago, good, stage, sunset, will, park, day, stay, alway	Entretenimiento
Tema 14	fit, job, great, chicago, hire, might, interest, develop, careerarc, art	Trabajo
Tema 15	chicago, day, stomper, mother, humid, temperatur, mph, starbuck, nike, dare	Indefinido
Tema 16	job, click, hire, chicago, appli, detail, want, work, careerarc, businessmgmt	Trabajo
Tema 17	chicago, peopl, night, field, italian, guarante, heritag, rate, great, wrigley	Deporte

Tabla 5.14: Temas durante un tiempo específico en la ciudad de Chicago

Temas	Palabras	Categoría
Tema 1	day, today, good, morn, may, hous, year, polit, lunch, sunset	Entretenimiento
Tema 2	beauti, life, sunday, check, weekend, way, anoth, thing, style, two	Entretenimiento Diurno
Tema 3	park, hyde, free, run, light, avail, full, finish, dinner, tomorrow	Entretenimiento Nocturno
Tema 4	amp, come, stigmabas, best, food, say, need, turn, bridg, help	Entretenimiento Nocturno
Tema 5	unit, london, kingdom, palac, buckingham, londr, around, buckinghampalac, chinatown, wednesday	Turismo
Tema 6	new, just, photo, post, back, video, fashion, yesterday, flower, pretti	Turismo
Tema 7	bst, trndnl, big, ben, david, west, break, fri, wall, star	Turismo
Tema 8	trend, trndnl, tweet, topic, alert, now, start, mention, rts, made	Temas Trending
Tema 9	twitter, happi, week, birthday, client, iphon, coffe, topapp, saturday, android	Entretenimiento
Tema 10	london, greater, white, kensington, royal, ferri, station, road, albert, hall	Transporte
Tema 11	london, job, great, see, work, hire, can, fit, england, open	Trabajo
Tema 12	just, nowplay, start, harrod, friday, feel, summer, mommi, walk, man	Entretenimiento Diurno
Tema 13	time, get, thank, will, know, peopl, first, even, take, tonight	Entretenimiento
Tema 14	night, last, make, travel, squar, fun, british, bigben, garden, cant	Entretenimiento Nocturno
Tema 15	love, amaz, littl, show, bar, music, much, play, think, blue	Entretenimiento Nocturno
Tema 16	one, art, street, got, still, design, repost, home, museum, westminst	Turismo / Cultura
Tema 17	look, chelsea, like, live, soho, well, book, que, meet, drink	Entretenimiento
Tema 18	friend, world, alway, final, let, everi, ive, tri, britain, enjoy	Turismo

Tabla 5.15: Temas durante un tiempo específico en la ciudad de Londres

Temas	Palabras	Categoría
Tema 1	job, washington, hire, nurs, white, hous, careerarc, great, fit, click	Trabajo
Tema 2	trend, trndnl, start, tweet, just, jone, demayo, priebus, laden, smith	Trending topic
Tema 3	washington, time, spring, amp, silver, joe, seafood, bekend, lijst, mensen	Entretenimiento
Tema 4	trend, trndnl, alert, just, celta, vigo, parti, may, advanc, bent	Trending topic
Tema 5	washington, amp, work, monument, best, drink, white, billiard, buffalo, coffeehous	Turismo
Tema 6	just, post, photo, better, say, somebodi, video, day, mother, famili	Evento (día de las madres)
Tema 7	morn, trndnl, power, saxwednesday, strike, cdt, horford, jesus, selfiesforariana, sprinkl	Indefinido
Tema 8	work, get, wait, glass, detail, want, hire, click, bro, perform	Trabajo
Tema 9	zoo, newbuild, washingtondc, like, tune, special, brookland, call, mans, nuageux	Turismo
Tema 10	nation, big, harbor, mgm, come, cheer, year, imperfect, maytribeslam, poetsofinstagram	Turismo
Tema 11	washington, state, unit, columbia, district, garden, join, park, csdc, topic	Entretenimiento
Tema 12	mph, will, wind, ano, atrÃ, minha, wed, hope, broken, cloud	Clima
Tema 13	twitter, iphon, topapp, client, web, dctraffic, washington, healthcar, wordstori, capitol	Tecnología

Tabla 5.16: Temas durante un tiempo específico en la ciudad de Washington

Temas	Palabras	Categoría
Tema 1	best, like, miss, crown, good, get, readi, come, height, just	Viaje
Tema 2	new, york, citi, squar, need, drink, time, georgewashingtonbridg, bridg, fish	Entretenimiento diurno
Tema 3	brooklyn, park, central, williamsburg, color, nyc, one, prospect, mind, beauti	Entretenimiento
Tema 4	day, center, state, unit, rockefel, rehears, summer, alic, life, top	Turismo
Tema 5	harlem, night, get, last, tonight, nyc, set, home, got, tattoo	Turismo
Tema 6	east, side, nyc, upper, come, thank, perform, dream, repost, let	Entretenimiento nocturno
Tema 7	manhattan, street, west, avenu, incid, updat, station, boulevard, metgala, construct	Infraestructura de la ciudad
Tema 8	nyc, today, favorit, look, may, book, one, ridgewood, day, tonight	Entretenimiento
Tema 9	just, photo, post, amp, yanke, avenu, american, one, histori, museum	Turismo / Cultura
Tema 10	time, music, dinner, ever, restaur, night, hot, right, bro, let	Entretenimiento diurno
Tema 11	work, nyc, let, hotel, view, astoria, week, journal, will, metropolitan	Turismo
Tema 12	job, newyork, hire, great, fit, appli, click, see, latest, open	Trabajo
Tema 13	nyc, mother, watch, spring, women, made, day, compani, met, date	Evento (día de las madres)
Tema 14	day, happi, art, bar, mother, museum, nyc, amp, video, best	Evento (día de las madres)
Tema 15	broadway, want, realli, see, team, model, share, chang, littl, bushwick	Entretenimiento
Tema 16	may, friday, yanke, made, stadium, els, dat, bar, mensfashion, tie	Deporte

Tabla 5.17: Temas durante un tiempo específico en la ciudad de Nueva York

Ahora se realiza una última prueba, pero esta vez se conoce a priori un evento que sucedió en los Estados Unidos en noviembre del 2016. Se trata de las elecciones presidenciales, las cuales tuvieron lugar el ocho de noviembre. En este caso se pretenden analizar los temas que surgen en el período desde el primero del noviembre hasta el dieciocho. El objetivo es descubrir si dentro de esos temas está contenida alguna información relevante acerca de ese evento tan importante mundialmente, es decir, verificar si realmente las redes sociales son utilizadas para compartir información relevante acerca de temas del momento. Para esto se analizan los tuits de las siguientes ciudades: Chicago, Nueva York, Washington y Londres.

Después de analizar los temas resulta una nueva categoría, denominada "elecciones presidenciales", la cual está presente en dos de las principales ciudades de los Estados Unidos (Nueva York y Washington). Cabe destacar que en Chicago esta categoría no está presente, pero sí un evento muy importante para la misma, el campeonato de la serie mundial de béisbol en el que ganaron los Chicago Cubs. Este equipo estuvo 108 años sin ganar la serie mundial de béisbol (Marca, 2017), por este motivo de los nueve temas que en ella surgen, cuatro están relacionados con deporte. Sin embargo, en Londres no aparece ninguna categoría referente a las elecciones de los Estados Unidos, lo cual resulta extraño dada la alta repercusión que tuvieron esas elecciones a nivel mundial.

Quizás si se hubiera analizado un mayor número de ciudades o un período de tiempo más prolongado se hubiera obtenido algún tipo de información más

relevante referente a la victoria de Donald Trump o bien a la derrota de Hillary Clinton. En este caso el nombre del presidente solo sale a relucir en la ciudad de Washington y el de la candidata Hillary no está presente en ninguna. Siempre hay que recordar las limitaciones que trae consigo el análisis de tuits geoposicionados, es probable que si se analizaran tuits independientemente de las coordenadas el resultado sería muy diferente, pero con la salvedad de que no se tendría conocimiento acerca de la ciudad en la que son generados los tuits.

Temas	Palabras	Categoría
Tema 1	chicago, flythew, wrigley, field, cub, just, photo, post, art, food	Deportes
Tema 2	job, chicago, hire, latest, careerarc, click, can, open, appli, anyon	Trabajo
Tema 3	chicago, amp, ipsento, center, yeah, model, feel, sushi, art, kill	Indefinido
Tema 4	chicago, cub, illinoi, morn, night, navi, worldserieschamp, good, pier, riverwalk	Deportes
Tema 5	chicago, cub, thank, low, someon, blue, new, got, like, qualiti	Deportes
Tema 6	park, grant, watch, get, gocubsgo, morn, new, unit, center, still	Deporte
Tema 7	chicago, citi, mph, illinoi, wind, los, room, church, clear, sky	Clima
Tema 8	one, world, night, photo, wrigleyvill, long, seri, brain, shrink, halloween	Indefinido
Tema 9	chicago, love, year, last, hope, night, illinoi, next, citi, get	Entretenimiento Nocturno

Tabla 5.18: Elecciones de los Estados Unidos en la ciudad de Chicago

Temas	Palabras	Categoría
Tema 1	day, vote, amp, home, hous, now, will, best, famili, like	Elecciones Presidenciales
Tema 2	just, photo, post, central, bar, amp, brooklyn, grand, nyc, park	Turismo
Tema 3	manhattan, nyc, museum, mta, midtown, subway, someth, made, place, alway	Turismo / cultura
Tema 4	brooklyn, love, williamsburg, elect, hotel, nyc, harlem, know, see, today	Elecciones Presidenciales
Tema 5	nyc, citi, happi, jobsearch, hire, tonight, drink, loung, lucki, inÃ	Trabajo
Tema 6	newyork, job, bronx, great, hire, work, nyc, fit, love, run	Trabajo
Tema 7	beauti, night, last, theatr, brooklyn, nov, bridg, friday, eye, grand	Entretenimiento nocturno
Tema 8	new, york, incid, brooklyn, citi, level, bridg, plaza, clear, georgewashingtonbridg	Turismo
Tema 9	time, squar, park, vote, amp, imwithh, good, nyc, parti, still	Elecciones Presidenciales
Tema 10	avenu, incid, station, brooklyn, street, nyc, birthday, thing, thank, line	Infraestructura de la ciudad
Tema 11	street, make, east, nyc, west, like, sure, construct, club, music	Entretenimiento nocturno

Tabla 5.19: Elecciones de los Estados Unidos en la ciudad de Nueva York

Temas	Palabras	Categoría
Tema 1	amp, drink, see, feder, bad, home, maryland, hous, floor, chevychas	Indefinido
Tema 2	washington, just, photo, post, old, club, need, station, washingtondc, memori	Turismo
Tema 3	ltd, happi, thank, one, babi, devic, monitor, system, grill, word	Indefinido
Tema 4	love, get, coffe, local, back, glad, brought, old, electionday, acr	Elecciones Presidenciales
Tema 5	washington, great, fit, job, hire, might, interest, columbia, district, careerarc	Trabajo
Tema 6	vote, mph, dont, life, takoma, sky, wind, clear, humid, park	Clima
Tema 7	trend, look, now, trndnl, check, your, topic, washington, unit, account	Trending topic
Tema 8	job, hire, washington, work, careerarc, can, recommend, anyon, want, veteran	Trabajo
Tema 9	night, friend, now, morn, wonder, special, artichok, amp, earlier, train	Entretenimiento diurno
Tema 10	hous, white, washington, day, just, beauti, rosslyn, trump, area, georgia	Elecciones Presidenciales
Tema 11	latest, job, open, hire, washington, see, appli, click, read, careerarc	Trabajo
Tema 12	nation, memori, elect, good, airport, flight, reagan, ronald, breakfast, live	Elecciones Presidenciales
Tema 13	washington, today, high, tonight, amaranth, night, ipa, nitro, food, grade	Entretenimiento nocturno

Tabla 5.20: Elecciones de los Estados Unidos en la ciudad de Washington

Temas	Palabras	Categoría
Tema 1	night, new, tonight, soho, brixton, last, theatr, even, polit, dont	Entretenimiento nocturno
Tema 2	art, mayfair, got, first, fashion, king, design, run, burger, ofÃ	Moda
Tema 3	hous, squar, can, see, day, work, make, now, tomorrow, alway	Indefinido
Tema 4	great, happi, amp, know, lunch, lane, theÃ, inspir, heaven, anoth	Entretenimiento diurno
Tema 5	twitter, get, like, back, week, best, client, peopl, topapp, iphon	Trending topic
Tema 6	street, park, nowplay, beauti, hyde, white, oxford, man, fun, green	Turismo
Tema 7	london, greater, big, england, palac, citi, bridg, westminst, buckingham, ben	Infraestructura de la ciudad
Tema 8	just, photo, post, good, come, live, market, music, british, parti	Turismo
Tema 9	love, old, friday, chelsea, will, galleri, realli, sound, borough, amp	Entretenimiento diurno
Tema 10	trend, trndnl, gmt, alert, start, tweet, just, let, novemb, hour	Trending topic
Tema 11	unit, kingdom, time, morn, view, friend, tate, red, modern, video	Indefinido
Tema 12	christma, look, one, drink, light, station, tower, circus, camden, carnabi	Entretenimiento diurno
Tema 13	thank, londonÃ, shoreditch, birthday, amaz, bar, garden, repost, job, play	Turismo

Tabla 5.21: Elecciones de los Estados Unidos en la ciudad de Londres

Como se ha visto en todas la pruebas realizadas, las categorías en las que se pueden dividir los temas suelen ser las mismas independientemente de la ciudad, con la salvedad de que se pueden ver influenciadas por eventos específicos tales como día festivo, carnaval, conciertos, entre otros que suelen causar mucha repercusión en las diferentes ciudades. También se ha visto cómo los temas pueden ir cambiando de forma impredecible dependiendo del tipo de suceso que esté ocurriendo en el momento, lo que hace que el realizar este tipo de investigaciones sea cada vez más importante.

Este capítulo ha reflejado como las redes sociales, en este caso Twitter, se han convertido en la nueva forma de comunicarse para las personas. Por medio de Twitter se tiene acceso a noticias, publicidades, temas de entretenimiento, entre otros. De acuerdo con los casos estudiados, las personas realizan más publicaciones cuando se encuentran en momentos de diversión y entretenimiento. También se observó, a través del análisis de los *hashtags*, que los *hashtags* más utilizados por los usuarios son los que arrojan los temas más relevantes.

CAPÍTULO 6
Conclusiones y Trabajo Futuro

Las redes sociales son una herramienta que ha permitido una nueva forma de comunicarse entre las personas. Las acciones que se realizan en estas redes dejan una huella digital que puede ser recogida y procesada para un posterior análisis. Además, en algunos casos, esta huella digital viene etiquetada con coordenadas geo-espaciales. Este hecho nos permite conocer no solo el contenido del mensaje sino también el lugar en dónde se generó.

En este trabajo se ha desarrollado una metodología para el análisis y detección de manera automática de temas en mensajes geoposicionados extraídos de la red social Twitter. Para ello, se han implementado procedimientos que permiten la extracción y el almacenamiento de los mensajes. Se ha realizado un preprocesamiento de los mensajes con el propósito de eliminar el contenido no relevante para el objetivo de este proyecto. Para la detección automática del número de temas, se han analizado y evaluado dos métodos estadísticos: uno con la media armónica y el otro utilizando los *hashtags*. Estos métodos se han aplicado a la detección automática de temas en ciudades de los Estados Unidos y Europa.

En los resultados obtenidos al aplicar la metodología propuesta, en varias ciudades se ha observado que los temas son muy parecidos independientemente de la ciudad de la que se trate, es decir, en todas aparecen las mismas categorías sin importar el período de tiempo en el que hayan sido tomados los tuits. También se ha encontrado que al calcular la cantidad de temas mediante el número de *hashtags*, en todos los casos estudiados, esta cantidad superaba la obtenida con el método de la media armónica.

Por otro lado, se ha demostrado cómo los *hashtags* siguen una distribución *Power Law*, es decir, que la menor parte de ellos son nombrados por muchos usuarios mientras que el resto son encontrados en muy pocos tuits, tal y como se comporta la distribución de las riquezas estudiada por la Ley de Pareto. Debido a la limitación en la cantidad de caracteres que tiene permitido Twitter para cada mensaje, los *hashtags* pueden servir para tener una idea de los temas más sobresalientes en cada tuit, ya que suelen resumir el contenido de los mismos.

Durante el análisis de los resultados del trabajo, se observó que con el algoritmo LDA un 70 por ciento de las ciudades contienen temas indefinidos, pero estos a su vez forman una mínima parte del total de temas, es decir, que los términos que este algoritmo agrupa para un mismo tema están relacionados entre sí.

Durante el desarrollo del trabajo se tuvo la limitación sobre la cantidad de tuits que son geoposicionados, ya que muy pocos usuarios hacen pública su ubicación. A pesar de ello, se pudieron identificar diferentes categorías que pueden ser útiles para diferentes tipos de instituciones, por ejemplo empresas que quieran valorar los comentarios realizados por las personas acerca de sus productos o servicios. También valoraciones acerca de cualquier evento realizado en la ciudad en la que el ayuntamiento o una empresa privada sea responsable de su ejecución y quiera conocer la opinión pública acerca del desarrollo del mismo.

El resultado de este trabajo puede ser muy útil para saber lo que pasó en la ciudad en un determinado momento. Los usuarios de las redes sociales, en este caso Twitter, suelen publicar comentarios acerca de lo que están haciendo o lo que está ocurriendo en el momento. Es un buen medio de información instantánea en la que desde sucesos tales como accidentes hasta situaciones de festividad suelen ser tomados en cuenta por las personas en sus tuits.

Como trabajo futuro se plantea tener en cuenta, además de la geo-localización de los mensajes, información del perfil de los usuarios como por ejemplo su ciudad de procedencia. Con esta información se podrían diferenciar los temas de los que hablan los turistas o gente de fuera de la ciudad y los temas de los que hablan los residentes. Esta diferenciación podría permitir a las ciudades entender mejor las preocupaciones de distintos perfiles de personas y ayudar a la toma de decisiones.

Por otra parte, otro aspecto a considerar como trabajo futuro sería tener en cuenta no solo los tuits geoposicionados sino también aquellos tuits que tienen algún tipo de información relacionada con la posición (p.ej. zona horaria, lugares, localizaciones). De esta manera, el número de tuits sería mayor y se podrían obtener resultados más fiables.

Bibliografía

Adnan, M., Leak, A., and Longley, P. (2014). A geocomputational analysis of twitter activity around different world cities. *Geo-spatial Information Science*, 17(3):145–152.

Berrocal, J. L. A., Figuerola, C. G., and Rodríguez, Á. F. Z. (2016). Análisis de temas emergentes a través de twitter. *Scire: representación y organización del conocimiento*, 22(2):67–73.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Chaffey, D. (2017). Global social media research summary 2017. http://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/. [Online; accessed 15-July-2017].

Chu, Z., Gianvecchio, S., Wang, H., and Jajodia, S. (2012). Detecting automation of twitter accounts: Are you a human, bot, or cyborg? *IEEE Transactions on Dependable and Secure Computing*, 9(6):811–824.

Clauset, A., Shalizi, C. R., and Newman, M. E. (2009). Power-law distributions in empirical data. *SIAM review*, 51(4):661–703.

Cunha, E., Magno, G., Comarela, G., Almeida, V., Gonçalves, M. A., and Benevenuto, F. (2011). Analyzing the dynamic evolution of hashtags on twitter: a language-based approach. In *Proceedings of the Workshop on Languages in Social Media*, pages 58–65. Association for Computational Linguistics.

del Val, E., Martínez, C., and Botti, V. (2015a). A multi-agent framework for the analysis of users behavior over time in on-line social networks. In *10th International Conference on Soft Computing Models in Industrial and Environmental Applications*, pages 191–201. Springer.

del Val, E., Martínez, C., and Botti, V. (2016a). Analyzing users' activity in online social networks over time through a multi-agent framework. *Soft Computing*, 20(11):4331–4345.

del Val, E., Palanca, J., and Rebollo, M. (2016b). U-tool: A urban-toolkit for enhancing city maps through citizens' activity. In *Advances in Practical Applications of Scalable Multi-agent Systems. The PAAMS Collection*, pages 243–246. Springer.

del Val, E., Rebollo, M., and Botti, V. (2015b). Does the type of event influence how user interactions evolve on twitter? *PLOS one*, 10(5):e0124049.

Drăgulescu, A. and Yakovenko, V. M. (2001). Exponential and power-law probability distributions of wealth and income in the united kingdom and the united states. *Physica A: Statistical Mechanics and its Applications*, 299(1):213–221.

Förster, T., Lamerz, L., Mainka, A., and Peters, I. (2014). The tweet and the city: Comparing twitter activities in informational world cities. In *Proceedings of the 2014 Conference: Informationsqualität und Wissensgenerierung, Frankfurt am Main, Germany*, pages 8–9.

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6):721–741.

Gordon, K. (2017). Social Media Statistics & Facts. https://www.statista.com/topics/social-networks/. [Online; accessed 7-July-2017].

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1):5228–5235.

Griffiths, T. L., Steyvers, M., Blei, D. M., and Tenenbaum, J. B. (2005). Integrating topics and syntax. In *Advances in neural information processing systems*, pages 537–544.

Hawking, S. W. (1977). Zeta function regularization of path integrals in curved spacetime. *Communications in Mathematical Physics*, 55(2):133–148.

Hong, L. and Davison, B. D. (2010). Empirical study of topic modeling in twitter. In *Proceedings of the first workshop on social media analytics*, pages 80–88. ACM.

Jurgens, D., Finethy, T., McCorriston, J., Xu, Y. T., and Ruths, D. (2015). Geolocation prediction in twitter using social networks: A critical analysis and review of current practice. *ICWSM*, 15:188–197.

Ma, Z., Sun, A., Yuan, Q., and Cong, G. (2014). Tagging your tweets: A probabilistic modeling of hashtag annotation in twitter. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 999–1008. ACM.

Marca, D. (2017). Los Cubs ganan las Series Mundiales 108 años después. http://www.marca.com/otros-deportes/2016/11/03/581aeb66e2704e96608b4620.html. [Online; accessed 10-July-2017].

Newman, M. E. (2005). Power laws, pareto distributions and zipf's law. *Contemporary physics*, 46(5):323–351.

Pak, A. and Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10.

Pareto, V. (1964). *Cours d'économie politique*, volume 1. Librairie Droz.

Ponweiser, M. (2012). Latent dirichlet allocation in r.

Ramos, J. et al. (2003). Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 133–142.

Richardson, L.-J. (2015). Micro-blogging and online community. *Internet Archaeology*, (39).

Rios, M. and Lin, J. J. (2013). Visualizing the"pulse.ºf world cities on twitter. In *ICWSM*.

Santana Sepúlveda, J. S. and Mateos Farfán, E. (2014). El arte de programa en r: un lenguaje para la estadística.

Vivanco, E., Palanca, J., del Val, E., Rebollo, M., and Botti, V. (2017). Using geotagged sentiment to better understand social interactions. In *International Conference on Practical Applications of Agents and Multi-Agent Systems*, pages 369–372. Springer.

Wallach, H. M. (2006). Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd international conference on Machine learning*, pages 977–984. ACM.

Wang, Q., Bhandal, J., Huang, S., and Luo, B. (2017). Classification of private tweets using tweet content. In *Semantic Computing (ICSC), 2017 IEEE 11th International Conference on*, pages 65–68. IEEE.