UNIVERSITAT POLITÈCNICA DE VALÈNCIA

etsinf

Escola Tècnica Superior d'Enginyeria Informàtica

Escola Tècnica Superior d'Enginyeria Informàtica
Universitat Politècnica de València

Análisis de frecuencia de hashtags en Twitter

TRABAJO FINAL DE MÁSTER

Máster Universitario en Gestión de la Información

Autor: José Alberto Pérez Melián

Tutores: Cèsar Ferri Ramírez
José Alberto Conejero Casares

Curso 2015-2016

Agradecimientos

A D. Cèsar Ferri Ramírez y D. José Alberto Conejero Casares por su labor en la Dirección de este Proyecto.

A D. Francisco Rangel Pardo por cederme los datasets `Hispatweets` y `Bárcenas` para la realización de este Proyecto.

A Francisco Almenar Pedrós, a José Francisco García Cantos y Mirella Oreto Martínez Murillo por permitirme trabajar con el dataset `Elecciones`.

A los profesores del Máster Universitario en Gestión de la Información de la Universitat Politècnica de València.

A mis padres Eudaldo y Andrea y a mi hermana Ana, por el apoyo y consejo que siempre me han brindado.

A mis abuelos, por darles a sus hijos la oportunidad de estudiar que ellos no tuvieron.

A D. José Luis Cubero Somed por su comprensión y apoyo.

A mis compañeros del Máster Universitario en Gestión de la Información y a todas aquellas personas que, de alguna manera, me han ayudado y apoyado en la realización de este Proyecto.

Canarias son siete islas
arrulladas por el mar,
siete corazones guanches,
siete notas de un cantar.

Una guitarra en la mano
y en el aire una folía;
no hay canto como mi canto
ni tierra como la mía.

Popular

Resumen

Las redes sociales han transformado la comunicación de manera drástica en los últimos años, a través del auge de nuevas plataformas y el desarrollo de un lenguaje propio de comunicación. Este nuevo panorama digital requiere de nuevas formas de estudio para describir y predecir el comportamiento de los usuarios en la red.

En este trabajo se realiza un análisis del comportamiento de los hashtags en una conversación de Twitter, estudiando cómo se distribuye su frecuencia de acuerdo a su popularidad. Se ha observado que se sigue una distribución potencial, según lo esperado por la Ley de Zipf, que indica que existen pocos hashtags con mucha repercusión y muchos que no han tenido mucho éxito. Debido al comportamiento humano en estas redes, muchos de los hashtags no han alcanzado gran repercusión debido a que contienen alguna falta de ortografía o están mal escritos. Para corregir esto se han utilizado distancias de edición de cadenas que han permitido agrupar los hashtags similares entre sí para disminuir el efecto de los hashtags escritos de manera errónea en los estudios realizados.

También se muestra la aplicación que tiene la Ley de Benford aplicada al estudio del comportamiento de los usuarios en las redes sociales, donde las distribuciones del primer y segundo dígito más significativo de las frecuencias de los hashtags siguen la distribución esperada por Benford. Esto permite que se pueda utilizar para validar datos provenientes de Twitter y analizarlos en busca de comportamientos sospechosos.

Palabras clave: Redes sociales, Twitter, hashtags, Ley de Zipf, Ley de Benford

Abstract

Social networks have transformed communication dramatically in recent years through the rise of new platforms and the development of a new language of communication. This landscape requires new forms to describe and predict the behaviour of users in networks.

This paper presents an analysis of hashtag's behaviour in a conversation in Twitter by studying the frecuency distribution according to their popularity. It has been observed that follows a power law, as espected by Zipf's Law, which states that there are few hastags with huge impact and many others who have not been very successful. Due to the human's behaviour in social networks, many hashtags have not achieved a great impact because they contain mispelling errors. String distances has been used to correct the errors to group similar hashtags together to decrease the possible impact of these hashtags for future studies.

It is also shown that Benford's Law could be applied to study the user's behaviour in social networks, where the first and second significant digit distribution of hashtag's frequencies follow the expected Benford's distribution and allow it's use to validate data from Twitter and analyze it for suspicious behaviours.

Key words: Social networks, Twitter, hashtags, Zipf's Law, Benford's Law

Índice general

Índice de figuras

Índice de tablas

CAPÍTULO 1
Introducción

Las redes sociales como Facebook, Twitter o Instagram se pueden definir como servicios que proveen a los usuarios la posibilidad de establecer conexiones con sus amigos a través de una aplicación y compartir información con ellos. Son, con diferencia, las aplicaciones más populares de lo que se conoce como Web 2.0, término que comprende «aquellos sitios web que facilitan el compartir información, la interoperabilidad, el diseño centrado en el usuario y la colaboración en la World Wide Web»[1]. Se cuentan por millones los usuarios que las utilizan para estar en contacto con amigos, para conocer gente nueva o para hacer vínculos profesionales.

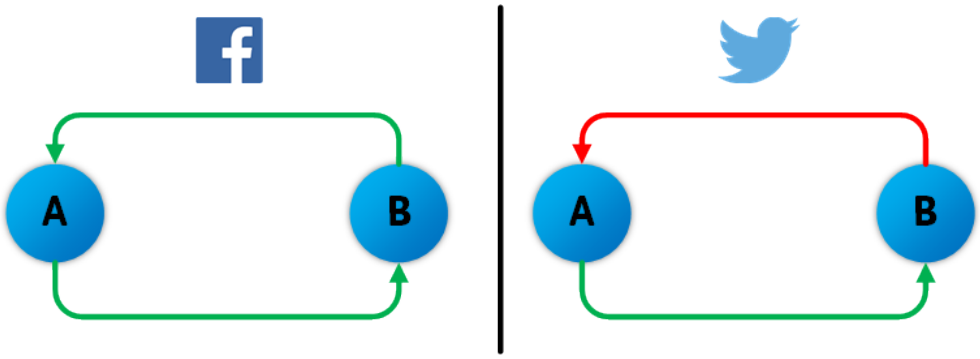


Figura 1.1: Tipos de relaciones entre usuarios en Facebook y en Twitter

Estas redes pueden se pueden representar mediante grafos, donde cada usuario es un nodo que se encuentra conectado con otros nodos mediante aristas si hay una relación entre ellos. Si hacemos una comparativa entre Facebook y Twitter, los nodos de ambos grafos representan a los usuarios y las aristas a la relación que puede haber entre ellos. En Facebook las aristas serían no dirigidas, ya que la relación de amistad es recíproca (un usuario A es amigo de B y B es amigo de A), mientras que en Twitter las aristas serían dirigidas, ya que un usuario A puede seguir a otro usuario B sin la necesidad de que el usuario B siga al usuario A (ver Figura 1.1).

Este tipo de relaciones entre los usuarios ha suscitado el interés por estudiar las características de este tipo de redes, analizando cómo se propaga la información en ellas [1], especialmente frente a fenómenos de actualidad como las inun-

[1] Wikipedia. Accedido el 25 de junio de 2016. [Enlace]

daciones de varios estados de Australia en 2010 [2], las protestas que tuvieron lugar en el Reino Unido en el año 2011 [3] o el análisis de las cuentas de Twitter de Mariano Rajoy y Alfredo Pérez Rubalcaba en las Elecciones a Cortes Generales de España de 2011 [4].

En este trabajo se cuenta con tres conjuntos de datos con conversaciones de Twitter en los que se estudia el comportamiento de la frecuencia de los hashtags en base a su popularidad. Con objeto de entender el comportamiento y manejar el lenguaje propio de Twitter, a continuación se hace una pequeña introducción donde se ponen de manifiesto sus aspectos más relevantes.

1.1 Twitter

Twitter es una red social de *microblogging* lanzada en el año 2006 cuya misión es «ofrecer a todo el mundo la posibilidad de generar y compartir ideas e información al instante y sin obstáculos»[2]. Cuenta con alrededor de 310 millones de usuarios activos mensuales y en el año 2012[3] se generaban en ella un total de 340 millones de *tweets* diarios[4].

Los mensajes o *tweets* no son más que pequeños mensajes de texto limitados a 140 caracteres de longitud en los que los usuarios pueden insertar fotos, vídeos, hipervínculos y emoticonos. Las siglas `RT` hacen referencia a los retweets, que son mensajes escritos por un usuario que otro comparte y propaga a su red de seguidores, dando más difusión al tweet original. También se ofrece la posibilidad de mencionar en el tweet a otros usuarios, añadiendo el símbolo @ seguido del nombre de la cuenta del usuario, y de poner etiquetas o *hashtags*, precedidas por el símbolo #, permitiendo así clasificar el tweet en una o varias temáticas.

Los hashtags sirven para clasificar mensajes, difundir ideas y promocionar temas específicos o personas destacadas. Son creados mayoritariamente por los propios usuarios y pueden llegar a tener éxito si se propagan adecuadamente por la red. En caso contrario, el hashtag tiene poca repercusión y muere. La popularidad de un hashtag se cuantifica en base al número de veces que ha sido mencionado en el conjunto de tweets. A mayor popularidad de un hashtag, más visible será y más probabilidad tendrá de ser usado por una cantidad de usuarios mayor.

Asimismo, Twitter muestra en su pantalla de inicio los hashtags más utilizados, que reciben el nombre de *Trending Topics*, y que pueden agruparse según diferentes ámbitos geográficos. De esta manera, los usuarios pueden ver los temas con más relevancia y repercusión del momento. En la actualidad, Twitter no ha implementado ningún sistema recomendador para la creación de hashtags que sugiera a un usuario qué hashtag debe usar, o qué características debe tener para que llegue a tener éxito. Esto permite que cada usuario genere tantos como quiera, dando lugar a la aparición de hashtags muy heterogéneos que suelen dificultar en algunos casos la indexación o búsqueda de algunos tweets.

[2] About Twitter. Accedido el 25 de junio de 2016. [Enlace]
[3] No se han encontrado datos oficiales más recientes
[4] Twitter Blog, 12 de marzo de 2012. Accedido el 25 de junio de 2016. [Enlace]

Los hashtags pueden ser creados de manera espontánea por los usuarios o pueden ser patrocinados por alguna entidad, como empresas o programas de entretenimiento, que los utilizan para promocionarse y alcanzar popularidad en la red. En este último caso, es posible que los hashtags sean *trending topic* no porque hayan sido difundidos de manera natural por los usuarios, sino debido a estas campañas de márketing, que han generado una popularidad artificial.

1.2 Objetivo

El objetivo de este trabajo es realizar un análisis de los hashtags de las conversaciones de Twitter para conocer su comportamiento, estudiando cómo se distribuye la frecuencia de acuerdo a su popularidad.

Para ello es necesario:

- Extraer la información necesaria para la realización del estudio de los datos que se tienen en bruto
- Plantear un método de almacenamiento que facilite el posterior tratamiento de la información
- Calcular la frecuencia de los hashtags para todos los datasets
- Estudiar la distribución de la frecuencia de los hashtags y determinar si existen anomalías en ella

1.3 Estructura de la memoria

Esta memoria se encuentra dividida tal como sigue:

- **Capítulo 2**
  Análisis exploratorio de los datasets que son objeto de estudio en este trabajo, presentando el modelo de datos diseñado para el almacenamiento de la información junto con los datos y estadísticas más significativos de todos los datasets.
- **Capítulo 3**
  Presentación de las leyes que son objeto de estudio, los estadísticos para contrastar su idoneidad y las distancias de edición entre cadenas usadas en este trabajo.
- **Capítulo 4**
  Estudio de Ley de Zipf y aplicación de la Ley de Benford para la distribución de los hahstags en Twitter.
- **Capítulo 5**
  Conclusiones y líneas de trabajo futuras de este trabajo.

CAPÍTULO 2
Análisis exploratorio de los datasets

En este capítulo se plantea el modelo de datos mediante el cual se almacenará la información de los tweets de los distintos datasets junto a un análisis con las estadísticas más importantes de todos ellos.

2.1 Modelado de datos

Los tweets almacenados en los datasets se encuentran en formato `JSON` (ver Figura 2.1) y contienen 26 campos que pueden ser consultados en la página de la documentación oficial de Twitter[1]. En la Tabla 2.1 se listan algunos de ellos junto a su significado.

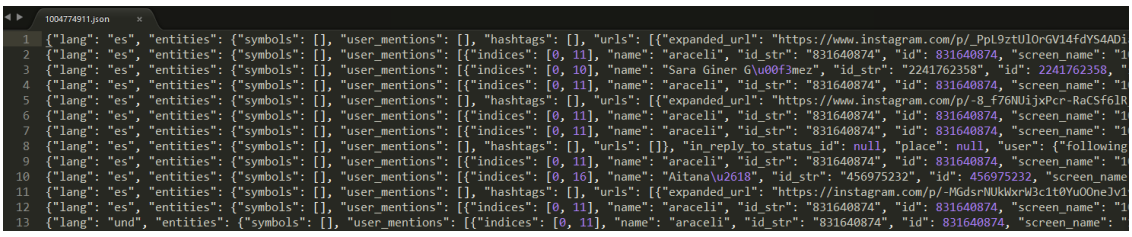


Figura 2.1: Muestra de los tweets almacenados en formato `JSON`

El tamaño de los datasets con los datos almacenados en bruto puede verse en la Tabla 2.2. Debido a su gran tamaño resulta inviable trabajar con toda la información almacenada, por lo que es necesario determinar qué datos es necesario extraer para la consecución del objetivo del trabajo.

Para estudiar la distribución de la frecuencia de los hashtags es necesario almacenar únicamente el usuario que ha mencionado el hashtag y qué hashtag ha sido mencionado por ese usuario. Esta información se almacenará en un fichero de texto `CSV` con dos columnas: `USUARIO` y `HASHTAG`, y para obtenerla es necesario recorrer todos los tweets y extraerla de los campos donde se encuentra almacenada.

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

Campo	Descripción
`created_at`	Fecha y hora de cuando se creó el tweet
`id`	ID del tweet
`text`	Texto del tweet
`user`	Información relativa al usuario que emitió el tweet
`favorite_count`	Número aproximado de veces que el tweet ha recibido *likes*
`retweet_count`	Número de veces que el tweet ha sido retuiteado

Tabla 2.1: Muestra de algunos campos del tweet en formato `JSON`

Dataset	Tamaño original
`Hispatweets`	11.9 GB
`Elecciones`	1.35 GB
`Barcenas`	52.6 MB

Tabla 2.2: Tamaño original de los datasets

Para la extracción de la información de los tweets se ha hecho uso de scripts escritos en `Python` debido a la facilidad de uso del lenguaje y a las librerías específicas para el tratamiento de datos que se encuentran en formato `JSON`.

2.1.1. Extracción y almacenamiento de la información

No todos los tweets contienen la misma información, y eso puede deberse a que haya habido errores en el proceso de descarga de los mismos. Si ha habido algún error, el tweet no se descarga completamente. En caso de que el tweet se haya descargado sin errores, se presupone que todos los campos existen y se procede a extraer la información.

La información variará según el tipo de tweet almacenado, por lo que se hace preciso mencionar qué tipos de mensajes se pueden encontrar en los datasets:

- **Tweets con interacción:** aquellos que contienen alguna mención a usuarios o a hashtags. Éstos a su vez se pueden subdividir en:
  - **Tweets de mención o respuesta:** aquellos que son escritos originalmente por un usuario o son fruto de una respuesta a un tweet emitido por un usuario
  - **Retweets:** aquellos que son compartidos por otros usuarios
- **Tweets sin interacción:** aquellos que no contienen mención alguna a usuarios o a hashtags

En este trabajo se usarán los tweets con interacción, ya sean de mención o respuesta o retweets, ya que son los que contienen algún hashtag en su interior.

Tweets con interacción de tipo mención o de respuesta

Los tweets con interacción de tipo de mención o de respuesta contienen, en el JSON descargado, un solo tweet. Para acceder al usuario que ha emitido el mensaje basta con acceder al campo `["user"]["screen_name"]` y obtendremos el nombre del usuario emisor. Los hashtags mencionados en el mensaje se pueden obtener de dos formas:

- Accediendo al campo `["text"]` y mediante el uso de expresiones regulares obtener los hashtags
- Accediendo al campo `["entities"]["hashtags"]` y obteniendo una lista con los hashtags mencinados

Mediante la primera opción es necesario utilizar expresiones regulares que extraigan del texto del tweet los hashtags, mientras que la segunda opción proporciona ya los hashtags mencionados. Por ello se ha escogido trabajar con esta última opción, debido a la facilidad y fiabilidad respecto a la primera.

Una vez se tiene el usuario y los hashtags que ha mencionado, se pasa a almacenar esta información en el fichero de texto siguiendo la notación arriba indicada. Por ejemplo: el usuario `@policia` escribe un mensaje donde hace mención a los hashtags `#estafa` y `#RedesSociales`. En el fichero de datos se almacenarán 2 líneas, tal como puede verse en la Tabla 2.3. Esto quiere decir que el usuario `@policia` ha mencionado en un tweet el hashtag `#estafa` y que el mismo usuario también ha mencionado el hashtag `#RedesSociales`.

USUARIO	HASHTAG
@policia	#estafa
@policia	#RedesSociales

Tabla 2.3: Fichero de ejemplo con los hashtags mencionados por `@policia`

Tweets con interacción de tipo retweet

Los tweets con interacción de tipo retweet contienen, además, dos tweets en el JSON descargado:

- El tweet original
- El tweet original retuiteado

La extracción de la información es similar al caso anterior pero con alguna variación. Por ejemplo: el usuario `@japmelian` retuitea el tweet visto anteriormente del usuario `@policia`. En el JSON, el mensaje de `@policia` se encuentra en el campo `["retweeted_status"]`, y la lista de hashtags mencionados en el campo `["retweeted_status"]["entities"]["hashtags"]`. Hasta aquí se construiría

una lista como la vista en la Tabla 2.3. Sin embargo, hay que añadir los hashtags mencionados en el tweet original de `@japmelian`, que se encuentran en el campo `["entities"]["hashtags"]`. En este ejemplo, al tratarse de un retweet, los hashtags coincidirán con los del mensaje de `@policia`. El fichero final quedaría como en la Tabla 2.4.

USUARIO	HASHTAG
@policia	#estafa
@policia	#RedesSociales
@japmelian	#estafa
@japmelian	#RedesSociales

Tabla 2.4: Fichero de texto con los hashtags mencionados por `@policia` y por `@japmelian`

2.1.2. Estandarización de hashtags

Una vez se han obtenido los hashtags que han sido utilizados y mencionados en el dataset, es necesario realizar una operación de normalización o estandarización sobre ellos. Para Twitter, el hashtag `#España` y el hashtag `#espana` hacen referencia a lo mismo, por lo que se debería sumar el número de menciones de cada uno para obtener su número de menciones total. De igual forma pasa con los hashtags `#elpaís` y el hashtag `#elpais` por la tilde y con los hashtags `#ferrerasARV` y el hashtag `#ferrerasarv` por las mayúsculas y las minúsculas, por ejemplo.

Para ello se han realizado las siguientes operaciones sobre los hashtags:

- Paso de mayúsculas a minúsculas
- Sustitución de tildes. Por ejemplo: `autónoma` por `autonoma`
- Sustitución de caracteres especiales del alfabeto: ñ, acento circunflejo, diéresis...

2.1.3. Cálculo de la frecuencia de los hashtags

Tras la operación anterior se procede a calcular el número de menciones que ha tenido cada hashtag, eliminando la columna `USUARIO` y añadiendo una nueva columna `CONTADOR` con el número de veces que cada hashtag ha sido mencionado. A continuación se detallan los pasos seguidos (ver Tabla 2.5) para la constricción del fichero final cogiendo como ejemplo los datos de la Tabla 2.4.

1. **Ordenar alfabéticamente por la columna** `HASHTAG`

   Se ordena alfabéticamente de la `a` a la `z`.

2. **Agrupar por hashtag y calcular el número de impresiones de cada uno**

   Una vez se tienen ordenados los datos por la columna HASHTAG, se realiza una agrupación por dicha columna y se contabilizan los elementos que forman parte de cada agrupación.

3. **Añadir al fichero la columna CONTADOR con el número de impresiones de cada hashtag**

   Calculada ya la frecuencia de cada hashtag, solo falta almacenarla añadiendo una nueva columna CONTADOR.

PASO 1 - Ordenar alfabéticamente por la columna HASHTAG

USUARIO	HASHTAG
@policia	#estafa
@japmelian	#estafa
@policia	#redessociales
@japmelian	#redessociales

PASO 2 - Agrupar por hashtag y calcular el número de impresiones de cada uno

HASHTAG	Nº DE IMPRESIONES
#estafa	2
#redessociales	2

PASO 3 - Añadir al fichero la columna CONTADOR con el número de impresiones de cada hashtag

HASHTAG	CONTADOR
#estafa	2
#redessociales	2

Tabla 2.5: Paso a paso de la construcción del fichero final

Tras estos pasos, y como se observa en la Tabla 2.6, el tamaño de los datasets desciende considerablemente al disminuir la información almacenada, permitiendo así realizar análisis posteriores con mayor rapidez.

Dataset	Tamaño final
Hispatweets	6.33 MB
Elecciones	130 KB
Barcenas	60 KB

Tabla 2.6: Tamaño final de los datasets tras aplicar el modelo de datos propuesto

A continuación se presentan los datos y estadísticas más significativas para cada uno de los tres datasets con los que se trabaja.

2.2 Estadísticas más sigfnificativas

2.2.1. Dataset `Hispatweets`

El dataset `Hispatweets` contiene tweets etiquetados en 7 países de habla española: Argentina, Chile, Colombia, España, México, Perú y Venezuela. Dichos tweets se encuentran almacenados en carpetas distintas (una por cada país). Dentro de cada una de ellas se encuentran varios ficheros de texto plano, cuyo nombre hace referencia al identificador o `ID` de un usuario. Cada fichero, en su interior, contiene un número de tweets almacenados en formato `JSON`.

	Nº de personas	Nº de tweets	Nº tweets con hashtags
Argentina	650	635765	89643
Chile	650	625739	63387
Colombia	650	616046	144352
España	650	623670	176167
México	650	624161	138631
Perú	650	621325	144561
Venezuela	650	610692	173906
Total países	4550	4357398	930647

Tabla 2.7: Estadísticas más significativas del dataset `Hispatweets`

En la Tabla 2.7 pueden verse las estadísticas más significativas del dataset. Contiene 4357398 tweets, los cuales se encuentran repartidos equitativamente entre los 7 países. Si atendemos a aquellos tweets que contienen interacción con hashtags, éste número decrece, quedando un total de 930647 (un 21.36 % del total). El país con mayor número de tweets con hashtag es España con 176167, mientras que el país con menor número es Chile con 63387.

	Mujer	Hombre	Indefinido	Total
Argentina	84	128	178	390
Chile	78	171	141	390
Colombia	83	152	155	390
España	88	156	146	390
México	58	159	173	390
Perú	101	149	140	390
Venezuela	54	160	176	390
Total países	546	1075	1109	2730

Tabla 2.8: Estadísticas por sexos del dataset `Hispatweets`

El número de usuarios es el mismo en cada país, 650, siendo el total de usuarios del dataset 4550. Sin embargo, no todos ellos se encuentran clasificados por

género. Para cada país se tienen el mismo número de personas clasificadas. Como puede verse en la Tabla 2.8, hay más personas etiquetadas como «indefinido» (1109) u «hombre» (1075) que como «mujeres» (546). En total se tienen unas 2730 personas etiquetadas, un 60 % del total.

País	Nº de hashtags únicos
España	76762
México	66955
Perú	65156
Chile	60262
Venezuela	59839
Colombia	52248
Argentina	44235

Tabla 2.9: Número de hashtags únicos por país del dataset `Hispatweets`

Tras aplicar las técnicas de limpieza de datos vistas en el apartado 2.1.2 para los hashtags, se ha procedido a contar el número de hashtags únicos para cada país (ver Tabla 2.9). El país con mayor número de hashtags únicos es España con 76762, mientras que el menor es Argentina con 44325.

En la Figura 2.2 se pueden ver los mapas de los 50 hashtags más populares por cada país del dataset.

2.2.2. Dataset `Elecciones`

El dataset `Elecciones` contiene tweets referentes a las Elecciones a Cortes Generales de España del mes de diciembre de 2015. Se trata de un conjunto de tweets fechados entre el día 1 y el día 22 de diciembre de 2015 que se encuentran almacenados en formato `JSON` en 33 ficheros de texto plano.

Obtención de datos

Los datos fueron extraídos entre el día 1 y el día 22 de diciembre de 2015 mediante un script en `Python` ejecutado durante 3 veces cada día (mañana, tarde y noche) usando distintas palabras de búsqueda (ver Tabla 2.10).

Palabras
`#PartidoPopular - Mariano Rajoy - @marianorajoy - Soraya Saenz - @Sorayapp`
`#Ciudadanos - Albert Rivera - @Albert_Rivera`
`#PSOE - Pedro Sánchez - @sanchezcastejon`
`#Podemos - Pablo Iglesias - @Pablo_Iglesias_`
`#IzquierdaUnida - Alberto Garzón - @agarzon`

Tabla 2.10: Palabras de búsqueda del dataset `Elecciones`

En la Tabla 2.11 se muestran las estadísticas más significativas, y en la Figura 2.3 se ve un mapa con los 50 hashtags más populares del dataset.

(a) Argentina

(b) Chile

(c) Colombia

(d) España

(e) México

(f) Perú

(g) Venezuela

Figura 2.2: Mapas con los 50 hashtags más populares de cada país

Item	Cantidad
Tweets	256293
Tweets con alguna interacción (@ o #)	245207
Tweets con hashtags	171650
Hashtags únicos	90920
Usuarios en la conversación	7950

Tabla 2.11: Estadísticas más significativas del dataset `Elecciones`

Figura 2.3: Mapas con los 50 hashtags más populares del dataset `Elecciones`

2.2.3. Dataset Bárcenas

El dataset Bárcenas contiene tweets emitidos por personas de habla española almacenados cada uno en un fichero en formato JSON, donde el nombre del fichero se corresponde con el ID del tweet descargado.

Item	Cantidad
Tweets	12620
Tweets con hashtags	12620
Hashtags únicos	3606
Usuarios en la conversación	7144

Tabla 2.12: Estadísticas más significativas del dataset Bárcenas

En la Tabla 2.12 se pueden ver las estadísticas más significativas del dataset. En la Figura 2.4 se ve un mapa con los 50 hashtags más populares.

Figura 2.4: Mapas con los 50 hashtags más populares del dataset Bárcenas

Obtención de datos

Los datos fueron extraídos entre el día 1 y el día 31 de agosto de 2013 mediante un script en Python descargando aquellos mensajes que contuvieran la palabra Bárcenas. Se trata de 12620 tweets únicos (no son retweets) y todos ellos contienen, al menos, un hashtag.

CAPÍTULO 3
Metodología

En este capítulo se presentan las dos leyes que son objeto de estudio: la Ley de Zipf y la Ley de Benford aplicadas a la distribución de los hashtags junto a los estadísticos más usados.

Se explican, además, las medidas de similitud de palabras de Levenshtein, Jaro y Jaro-Winkler, usadas para agrupar hashtags similares entre sí.

3.1 Ley de Zipf

Zipf [5] formuló una ley en la que se establece que, dado un corpus de frecuencias de palabras de un lenguaje, la frecuencia de cualquier palabra es inversamente proporcional a la posición que ocupa en el ranking de la tabla de frecuencias. Esto es, la palabra más frecuente en el corpus tendrá una frecuencia el doble de mayor que la frecuencia de la segunda palabra que más aparece en el corpus, y el triple que la tercera palabra y así sucesivamente.

La distribución del ranking y de la frecuencia sigue, por tanto, una relación inversa que puede aproximarse mediante:

$$P_n \sim \frac{1}{n^a} \tag{3.1}$$

donde P_n representa la frecuencia de una palabra que se encuentra en la tabla de frecuencias ordenada en la n-ésima posición y el exponente a es próximo a 1.

Para este trabajo, el corpus de palabras son los hashtags, y su frecuencia el número de veces que cada hashtag ha sido mencionado. Así, para aplicar la Ley de Zipf al dataset, es necesario ordenar de mayor a menor el número de ocurrencias de cada hashtag por país y graficar el resultado en una gráfica de escala logarítmica, donde el eje x representa el ranking del hashtag y el eje y representa el número de ocurrencias del mismo.

3.2 Ley de Benford

La Ley de Benford [6] establece que, en los fenómenos que ocurren de forma natural, la cantidad de sucesos cuyo primer dígito de la frecuencia es 1 tienen lugar un mayor número de veces que los fenómenos cuyo primer dígito de la frecuencia es 2, 3... y así hasta 9.

Si tomamos como ejemplo el número `81291`, el primer dígito más significativo es el `8`; el segundo dígito más significativo es el `1`, y así sucesivamente.

Simon Newcomb, un astrónomo y matemático estadounidense, y Frank Benford, un físico de General Electric, plantearon que la frecuencia exacta P de un dígito d es la siguiente:

$$P(d) = \log_{10}(1 + \frac{1}{d}) \tag{3.2}$$

Mediante la fórmula anterior, la probabilidad de que el primer dígito de un número sea 1 es de alrededor del 30 % mientras que la probabilidad de que sea 9 es de solo el 4.6 %. En la Tabla 3.1 se muestran las frecuencias para todos los dígitos del 0 al 9 en cada una de las primeras cuatro posiciones de cualquier número. En la Tabla B.1 se pueden ver también las frecuencias calculadas para los dos primeros dígitos más significativos.

Dígito	**1ª posición**	**2ª posición**	**3ª posición**	**4ª posición**
0	-	0.11968	0.10178	0.10018
1	0.30103	0.11389	0.10138	0.10014
2	0.17609	0.10988	0.10097	0.10010
3	0.12494	0.10433	0.10057	0.10006
4	0.09691	0.10031	0.10018	0.10002
5	0.07918	0.09668	0.09979	0.09998
6	0.06695	0.09337	0.09940	0.09994
7	0.05799	0.09035	0.09902	0.09990
8	0.05115	0.08757	0.09864	0.09986
9	0.04576	0.08500	0.09827	0.09982

Tabla 3.1: Frecuencias esperadas calculadas mediante la Ley de Benford

Para calcular los datos de las Tablas 3.1 y B.1 se han usado las siguientes fórmulas:

$$P(D_1 = d_1) = \log_{10}(1 + \frac{1}{d_1}) \tag{3.3}$$

$$P(D_2 = d_2) = \sum_{d_1=1}^{9} \log_{10}(1 + \frac{1}{d_1 d_2}) \tag{3.4}$$

$$P(D_1 D_2 = d_1 d_2) = \log_{10}(1 + \frac{1}{d_1 d_2}) \tag{3.5}$$

En la Figura 3.1 se muestra la distribución de Benford de acuerdo al primer dígito más significativo.

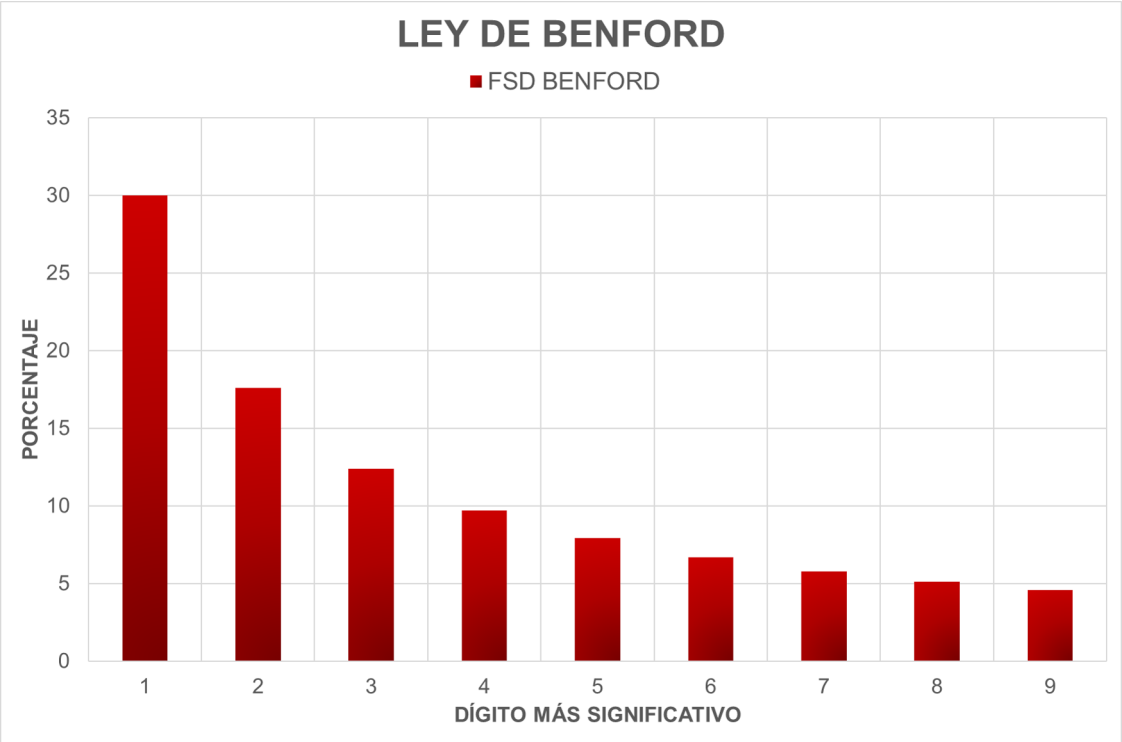


Figura 3.1: Distribución de frecuencias del primer dígito más significativo según la Ley de Benford

La Ley de Benford se usa a menudo en el análisis forense para la banca [7], donde una distribución anómala en los primeros dígitos puede ser indicio de fraude. Algunos investigadores la han aplicado también para datos genéticos y para resultados electorales [8], entre otros.

Para las redes sociales se ha aplicado la Ley de Benford sobre conjuntos de datos de varias redes sociales (Facebook, Twitter, Google Plus, Pinterest y LiveJournal), estudiando la distribución del primer dígito más significativo en el número de amigos y seguidores de los usuarios de dichas redes, y para todas ellas obteniendo resultados positivos que indicaban que se seguía la Ley de Benford [9].

Este Trabajo de Fin de Máster se centra en el estudio de la frecuencia de los hashtags de una conversación de Twitter y, para poder aplicar la Ley de Benford, es necesario calcular la frecuencia de todos los hashtags de las conversaciones y estudiar las distribuciones de los dígitos más significativos del número de la frecuencia de cada hashtag.

3.3 Estadísticos

En esta sección se presentan los tres principales estadísticos utilizados para calcular el ajuste a la Ley de Benford: el test de la χ^2 de Pearson, la Desviación Absoluta Media (MAD) y el coeficiente de correlación de Pearson.

3.3.1. Test de la χ^2 de Pearson

El estadístico es el siguiente:

$$\chi^2 = \sum_{d=m}^{9} \frac{(P_{obs}(d) - P_t(d))^2}{P_t(d)} \tag{3.6}$$

donde:

- $P_t(d)$ y $P_{obs}(d)$ son las frecuencias teóricas y observadas de cada dígito
- $m = 1$ para la distribución del primer dígito más significativo según la Ley de Benford

En la Tabla 3.2 se observan los grados de libertad y los valores críticos que hacen aceptar o rechazar la hipótesis nula H_0 de que los datos cumplen la Ley de Benford. Si el valor observado del estadístico χ^2 es mayor al valor teórico $\chi^2_{n,\alpha}$ se rechaza la hipótesis nula con un nivel de confianza α.

	g.l.	**95 %**	**99 %**
Primer dígito más significativo	8	15.507	20.090
Segundo dígito más significativo	9	16.919	21.666

Tabla 3.2: Valores del estadístico χ^2 de Pearson

3.3.2. Desviación Absoluta Media

La fórmula es la siguiente:

$$MAD = \frac{1}{9} \sum_{d=1}^{9} |P_{obs}(d) - P_t(d)| \tag{3.7}$$

Cuando se utiliza este estadístico para la Ley de Benford se sigue la tabla de valores determinada por Nigrini [10]. Dependiendo de los valores obtenidos por el estadístico se puede determinar el nivel de conformidad de una distribución con la Ley de Benford (ver Tabla 3.3).

Dígitos	Rango	Nivel de conformidad
Primer dígito	0.000 a 0.006	Mucho
	0.006 a 0.012	Aceptable
	0.012 a 0.015	Medio
	más de 0.015	Nada
Segundo dígito	0.000 a 0.008	Mucho
	0.008 a 0.010	Aceptable
	0.010 a 0.012	Medio
	más de 0.012	Nada

Tabla 3.3: Rango de valores críticos y niveles de conformidad para el ajuste a la Ley de Benford

3.3.3. Coeficiente de correlación de Pearson

El coeficiente de correlación de Pearson es una medida de la relación lineal entre dos variables aleatorias cuantitativas X e Y. Suele utilizarse para cuantificar el ajuste de una distribución de datos a la Ley de Benford [11]. Su fórmula es la siguiente:

$$\rho_{X,Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y} \tag{3.8}$$

donde:

- E es la esperanza matemática
- μ_X es la media de la variable X
- μ_Y es la media de la variable Y
- σ_{XY} es la covarianza de (X, Y)
- σ_X es la desviación típica de la variable X
- σ_Y es la desviación típica de la variable Y

De manera análoga podemos calcular este coeficiente sobre un estadístico muestral, denotado como r_{xy} a:

$$r_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x s_y} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}. \tag{3.9}$$

El valor del índice de correlación varía en el intervalo $[-1, 1]$:

- Si $r = 1$ existe una correlación positiva perfecta
- Si $0 < r < 1$ existe una correlación positiva

- Si $r = 0$ no existe relación lineal
- Si $-1 < r < 0$ existe una correlación negativa
- Si $r = -1$ existe una correlación negativa perfecta

3.4 Distancias de edición entre cadenas

En los tres datasets que son objeto de estudio se ha observado que pocos hashtags son los que han obtenido gran repercusión y, a su alrededor, se encuentran hashtags similares con pocas menciones. Estos hashtags con bajo número de menciones suelen contener faltas de ortografía o estar mal escritos y ser muy similares a los hashtags con gran repercusión. Para posteriores análisis resulta útil «corregir» estos fallos agrupando los hashtags que son muy similares entre sí, y para ello se hace uso de las distancias de edición entre cadenas. En este apartado se presentan tres de ellas: la distancia de Levenshtein, la de Jaro y la de Jaro-Winkler.

3.4.1. Distancia de Levenshtein

La distancia de Levenshtein $dist_{lev}$, también conocida como distancia de edición, se define como el menor número de operaciones de edición (inserciones, borrados y sustituciones) que se requieren para transformar una palabra en otra.

De forma general, la distancia de Levenshtein $dist_{lev}(s_1, s_2)$ entre dos cadenas s_1 y s_2 (de tamaño $|s_1|$ y $|s_2|$ respectivamente) viene dada por $dist_{lev}(|s_1|, |s_2|)$ y se cumple que $0 \leq dist_{lev}(s_1, s_2) \leq max(|s_1|, |s_2|)$

Para su cálculo se requiere el uso de una matriz de tamaño $[|s_1| + 1, |s_2| + 1]$ y el procedimiento para rellenarla puede consultarse en el Apéndice A.

La distancia de Levenshtein puede convertirse en una función de similitud sim_{lev} con rango de valores $[0, 1]$ mediante la ecuación 3.10, donde s_1 y s_2 son cadenas de caracteres. Un valor de la función de similitud cercano a 0 significa que las cadenas s_1 y s_2 son distintas; un valor próximo a 1 significa que ambas cadenas son muy parecidas.

$$sim_{lev}(s_1, s_2) = 1 - \frac{dist_{lev}(s_1, s_2)}{max(|s_1|, |s_2|)} \tag{3.10}$$

3.4.2. Distancia de Jaro

La distancia de Jaro se usa normalmente para el mapeado de datos en colecciones de datos enlazados (*data linkage systems*), como pueden ser datos relacionados con el censo [12].

El algoritmo calcula el número c de caracteres en común entre dos cadenas s_1 y s_2, tomando como referencia aquellos caracteres que se encuentran en la primera mitad de la cadena más larga, y el número de trasposiciones t.

La distancia de Jaro $dist_{jar}(s_1, s_2)$ entre dos cadenas s_1 y s_2 (de tamaño $|s_1|$ y $|s_2|$) es:

$$dist_{jar}(s_1, s_2) = \begin{cases} 0 & \text{si } m = 0 \\ \frac{1}{3}\left(\frac{c}{|s_1|} + \frac{c}{|s_2|} + \frac{c-t}{c}\right) & \text{en otro caso} \end{cases} \tag{3.11}$$

Dos caracteres de cada cadena s_1 y s_2 se considera que casan entre ellos (*matching*) solamente si son iguales y no están separados entre ellos más de

$$\left\lfloor \frac{\text{máx}(|s_1|, |s_2|)}{2} \right\rfloor - 1$$

posiciones.

Cada carácter de s_1 es comparado con los caracteres iguales a él de s_2. El número de trasposiciones t es el número de caracteres coincidentes (en orden distinto) dividido por 2.

La distancia de Jaro puede convertirse en una función de similitud sim_{jar} con rango de valores $[0, 1]$ mediante la ecuación 3.12.

$$sim_{jar}(s_1, s_2) = \frac{1}{3}\left(\frac{c}{|s_1|} + \frac{c}{|s_2|} + \frac{c-t}{c}\right) \tag{3.12}$$

3.4.3. Distancia de Jaro-Winkler

La distancia de Jaro-Winkler mejora la distancia de Jaro al aplicar ideas basadas en estudios empíricos que evidencian que los errores de escritura se cometen menos en el comienzo de las palabras que en el final. El algoritmo usa un prefijo p que puntúa más favorablemente las cadenas cuyos l primeros caracteres coinciden.

Dadas dos cadenas s_1 y s_2, la distancia de Jaro-Winkler entre ellas es de:

$$dist_{jar-wink}(s_1, s_2) = dist_{jar}(s_1, s_2) + \left(lp\left(1 - dist_{jar}(s_1, s_2)\right)\right) \tag{3.13}$$

donde:

- $dist_{jar}(s_1, s_2)$ es la distancia de Jaro de las cadenas s_1 y s_2
- l es la longitud del prefijo común de ambas cadenas (hasta un máximo de $l = 4$)
- p es una constante cuyo valor estándar es $p = 0.1$ para el ajuste de los prefijos coincidentes de las cadenas

La distancia de Jaro-Winkler puede convertirse en una función de similitud $sim_{jar-wink}$ con rango de valores $[0, 1]$ mediante la ecuación 3.14.

$$sim_{jar-wink}(s_1, s_2) = dist_{jar}(s_1, s_2) + \left(lp\left(1 - dist_{jar}(s_1, s_2)\right)\right) \tag{3.14}$$

3.4.4. Implementación

Para el cálculo de las distancias y de las funciones de similitud se ha hecho uso de la librería `python-Levenshtein`[1].

La citada librería dispone de métodos para calcular la función de similitud de Jaro y de Jaro-Winkler, no así la de Levenshtein, que ha tenido que ser implementada (ver Código 3.1).

```
import Levenshtein as lv

# Funcion de similitud de Levenshtein
def sim_lev(s1, s2):
    return 1 - (float(lv.distance(s1,s2)) / float(max(len(s1), len(
        s2))))

cadena1 = "hola"
cadena2 = "ola"

# Levenshtein
sim_lev(cadena1, cadena2)

# Jaro
lv.jaro(cadena1, cadena2)

# Jaro-Winkler
lv.jaro_winkler(cadena1, cadena2)
```

Listing 3.1: Implementación en Python de las distancias de edición y sus funciones de similitud

A modo de ejemplo puede verse en la Tabla 3.4 la similitud entre varios hashtags calculada mediante las funciones de similitud vistas. Como se observa, la similitud entre los hashtags varía dependiendo de la función de similitud utilizada. Esto se debe a que, por ejemplo, Levenshtein contempla el número de «cambios» realizados entre dos hashtags independientemente de dónde se realicen (al principio o al final de cualquier hashtag), mientras que Jaro y Jaro-Winkler distinguen y ponderan de distinta forma aquellos fallos producidos bien en la primera o bien en la segunda mitad de los hashtags.

Hashtag 1	Hashtag 2	Levenshtein	Jaro	Jaro-Winkler
#20delecciones	#20democracia	0.3846	0.6773	0.8064
#20delecciones	#20dediciembre	0.3846	0.7019	0.8211
#7deldebatedecisivo	#7deldebateadecisivo	0.9473	0.9824	1.0000
#7deldebatedecisivo	#7deldebatedecsivo	0.9445	0.9618	1.0000
#canarias	#valencia	0.2500	0.5834	0.5834
#marianorajoy	#pedrosanchez	0.0834	0.3889	0.3889

Tabla 3.4: Similitud entre varios hashtags usando las tres distancias de edición

Para agrupar los hashtags similares entre sí es necesario conocer la similitud entre ellos y determinar un cierto nivel de significación α para crear los grupos.

[1] https://pypi.python.org/pypi/python-Levenshtein/0.12.0

En primer lugar se optó por calcular la similitud entre todos los hashtags, opción que resultó inviable debido al alto coste computacional que requería. Si se pone como ejemplo el dataset `Elecciones`, que contiene 7950 hashtags únicos, se deben calcular un total de 31,597,275 combinaciones distintas, generando una matriz de similitud dispersa de tamaño 7950 filas por 7950 columnas.

Debido a su alta complejidad, se optó por ordenar alfabéticamente los hashtags para luego calcular la similitud entre un hashtag i y un hashtag $i+1$ de la lista ordenada, calculando así la similitud entre un hashtag y su hashtag más cercano. Una vez calculadas las similitudes entre cada hashtag y su siguiente en la lista, se pasa a agrupar aquellos que son similares entre sí. Para ello se establecen varios niveles de α y, recorriendo la lista ordenada alfabéticamente, se agrupan aquellos que son similares a un nivel α determinado y se suman las frecuencias de los hashtags incluídos en el grupo generado.

En este trabajo se han establecido los siguientes valores:

$$\alpha = [0.95, 0.90, 0.85, 0.80]$$

No se ha tomado en cuenta valores inferiores a 0.80 debido a que agruparían hashtags significativamente diferentes.

3.4.5. Ejemplo

Para ilustrar lo anterior se escoge una muestra de 10 hashtags del dataset `Elecciones` (ver Tabla 3.5) sobre los que aplicar las distancias de edición para agrupar luego aquellos similares entre sí. El fichero de ejemplo

Hashtag	Número de menciones
caraacara2015	23283
caraacara2015ra	1
caraacara2015raj	1
caraacara2015rajoy	732
caraacara2016	11
caraacara2025	7
caraacara2105	10
caraacara3n	6
caraacara6	1
caraacara6f	2

Tabla 3.5: Muestra de hashtags del dataset `Elecciones`

Para poder agruparlos es necesario calcular la similitud entre un hashtag i y el hashtag $i+1$. En la Tabla 3.6 se encuentra calculada la similitud entre ellos con todas las distancias de edición.

Para proceder a agruparlos se recorre la lista de arriba hacia abajo, agrupando aquellos hashtags que presentan una similitud igual o superior al nivel de significación α escogido. Si un hastag y su siguiente en la lista presentan un nivel

Hashtag i	Hashtag $i+1$	Levenshtein	Jaro	Jaro-Winkler
caraacara2015	caraacara2015ra	0.87	0.96	1.00
caraacara2015ra	caraacara2015raj	0.93	0.98	1.00
caraacara2015raj	caraacara2015rajoy	0.89	0.96	1.00
caraacara2015rajoy	caraacara2016	0.67	0.86	1.00
caraacara2016	caraacara2025	0.84	0.89	1.00
caraacara2025	caraacara2105	0.84	0.94	1.00
caraacara2105	caraacara3n	0.69	0.83	0.98
caraacara3n	caraacara6	0.81	0.91	0.99
caraacara6	caraacara6f	0.90	0.97	1.00

Tabla 3.6: Similitud entre la muestra de hashtags

de similitud inferior se creará un nuevo grupo. En la Figura 3.2 se ven los hashtags agrupados junto al número de impresiones total de cada grupo, observando que las agrupaciones varían en número dependiendo de la distancia de similitud escogida.

Figura 3.2: Grupos formados en la muestra de hashtags escogida

CAPÍTULO 4
Experimentación y resultados

En este capítulo se realiza el estudio de las distribuciones de los hashtags de los distintos datasets en base a su popularidad según la Ley de Zipf. A continuación, mediante la Ley de Benford y sus estadísticos, se comprueba si hay o no indicios de anomalías en dichas distribuciones.

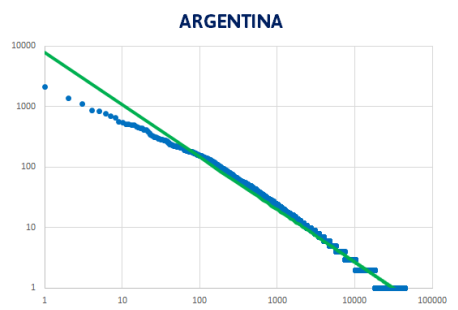
4.1 Ley de Zipf

4.1.1. Dataset `Hispatweets`

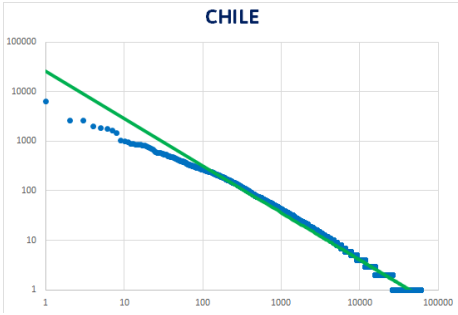
Para cada país del dataset `Hispatweets` se cumple la Ley de Zipf. La distribución de los hashtags de acuerdo a su popularidad se puede aproximar a una recta de regresión con un coeficiente de regresión R^2 muy próximo a 1, lo que indica que el ajuste es muy bueno (ver Figura 4.1). En la Tabla 4.1 se puede ver más en detalle la función aproximada para cada país junto a su coeficiente de regresión R^2.

País	Recta de regresión	R^2
Argentina	$-1{,}1011x + 4{,}4794$	0.9549
Chile	$-0{,}9538x + 4{,}4206$	0.9617
Colombia	$-0{,}9550x + 4{,}3778$	0.9641
España	$-0{,}9496x + 4{,}5036$	0.9628
México	$-0{,}8612x + 4{,}0208$	0.9527
Perú	$-0{,}8953x + 4{,}1562$	0.9549
Venezuela	$-1{,}0394x + 4{,}8159$	0.9617

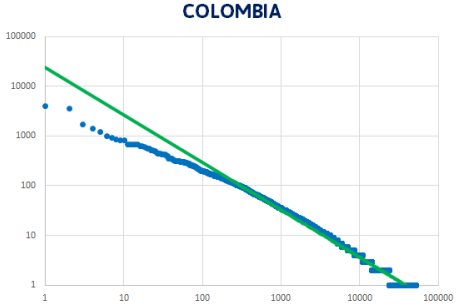
Tabla 4.1: Rectas de regresión para cada país del dataset `Hispatweets`



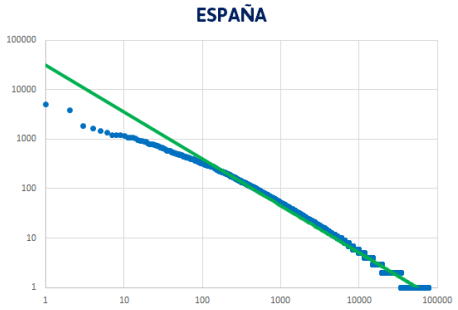
(a) Argentina



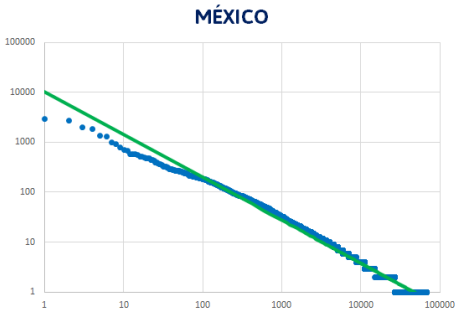
(b) Chile



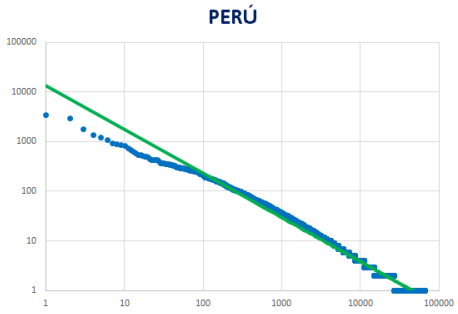
(c) Colombia



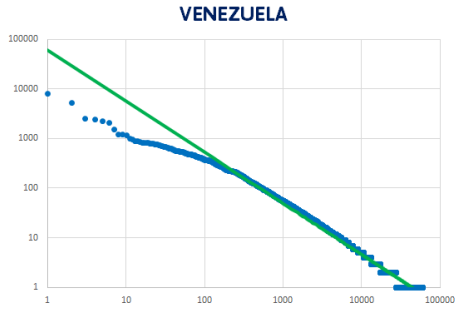
(d) España



(e) México



(f) Perú



(g) Venezuela

Figura 4.1: Gráficas log-log con la distribución de hashtags por país del dataset `Hispatweets`

4.1.2. Dataset `Elecciones`

La distribución de la popularidad de los hashtags del dataset `Elecciones` cumple la Ley de Zipf. La distribución se puede aproximar a una recta de regresión $-1{,}4909x + 5{,}7644$ con un coeficiente de regresión $R^2 = 0.9879$, muy próximo a 1, signo de un buen ajuste (ver Figura 4.2).

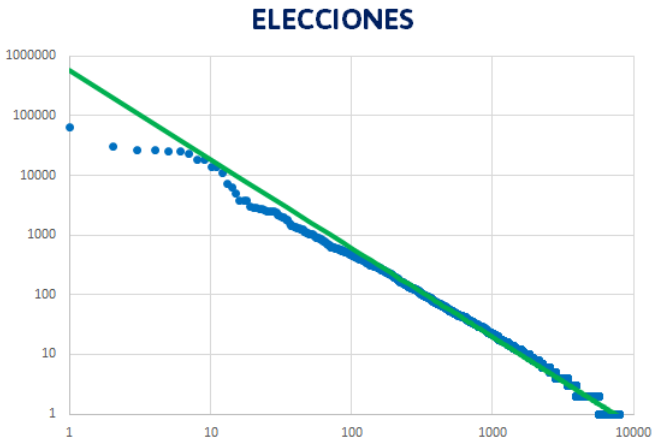


Figura 4.2: Gráfica log-log con la distribución de hashtags del dataset `Elecciones`

4.1.3. Dataset `Bárcenas`

Para el dataset `Bárcenas` la distribución de la popularidad de los hashtags se puede aproximar a una función de regresión potencial $-0{,}7508x + 2{,}5057$ con un coeficiente de regresión $R^2 = 0.8785$ (ver Figura 4.3), un ajuste bajo comparado con el observado en el resto de datasets. Esto puede deberse a la cantidad de hashtags con poca popularidad (el 73 % de los hashtags han sido mencionados tan solo 1 vez).

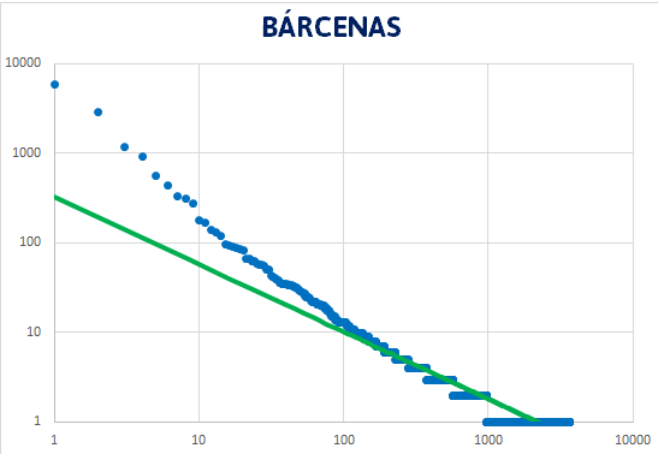


Figura 4.3: Gráfica log-log con la distribución de hashtags del dataset `Bárcenas`

4.2 Ley de Benford

4.2.1. Dataset `Hispatweets`

En la Tabla 4.2 se muestran los datos en porcentaje obtenidos por cada país y el porcentaje teórico de la distribución del dígito más significativo o FSD, por sus siglas en Inglés (*First Significant Digit*).

	1	2	3	4	5	6	7	8	9	Total
FSD Benford	30.01 %	17.60 %	12.40 %	9.69 %	7.91 %	6.69 %	5.79 %	5.11 %	4.57 %	100 %
FSD Argentina	62.54 %	13.37 %	6.69 %	4.14 %	2.41 %	1.80 %	1.18 %	1.03 %	0.83 %	100 %
FSD Chile	60.88 %	19.19 %	6.95 %	4.51 %	2.86 %	2.11 %	1.47 %	1.16 %	0.87 %	100 %
FSD Colombia	59.45 %	19.41 %	7.36 %	4.71 %	3.05 %	2.30 %	1.49 %	1.28 %	0.95 %	100 %
FSD España	60.17 %	19.89 %	7.00 %	4.56 %	2.74 %	2.03 %	1.52 %	1.16 %	0.92 %	100 %
FSD México	63.53 %	18.61 %	6.41 %	3.98 %	2.52 %	1.84 %	1.33 %	0.96 %	0.82 %	100 %
FSD Perú	62.60 %	19.15 %	6.78 %	4.08 %	2.47 %	1.84 %	1.23 %	1.05 %	0.79 %	100 %
FSD Venezuela	58.71 %	19.55 %	7.48 %	4.89 %	3.01 %	2.01 %	1.61 %	1.32 %	1.12 %	100 %

Tabla 4.2: Distribución del primer dígito más significativo para cada país

De los resultados obtenidos se desprende que la distribución del FSD para todos los países es similar, y dista bastante en porcentaje de la frecuencia del FSD esperada. Por ello, y para simplificar el desarrollo, se escoge España como país de referencia para realizar el estudio.

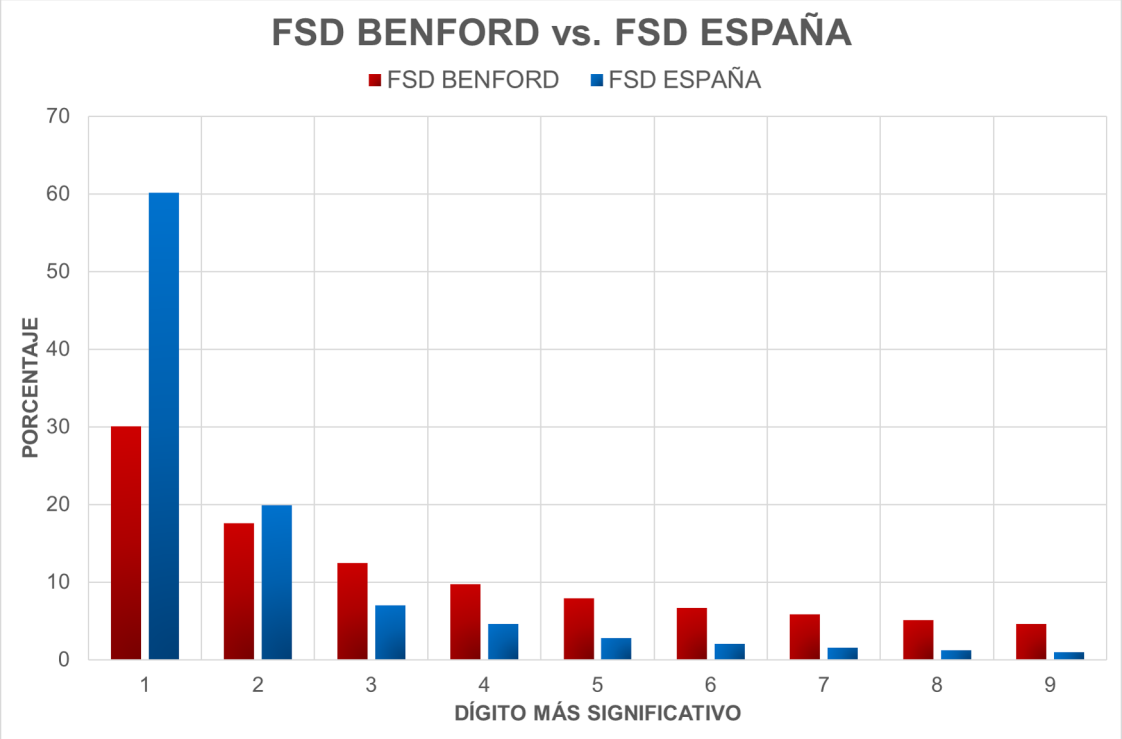


Figura 4.4: Gráfica comparativa entre la distribución del FSD esperada y la del FSD obtenida para España

En la Figura 4.4 puede verse una gráfica comparativa entre la distribución FSD esperada y la distribución FSD obtenida para España, y se observa que para

el dígito 1 la proporción excede ostensiblemente lo esperado, ya que se espera un 30 % mientras que se observa un 60 %.

Para mitigar este suceso se aplican las tres distancias de edición de palabras estudiadas, agrupando los hashtags similares entre sí para los valores predeterminados de α.

Tras agrupar los hashtags similares se calcula nuevamente la distribución del dígito más significativo. Los resultados, que pueden verse en las Tablas B.2, B.3 y B.4, varían muy poco de los datos obtenidos para la distribución FSD de España, por lo que no se ha conseguido el objetivo de reducir considerablemente los casos «anómalos».

Otro motivo que podría ser la causa de esa diferencia para el dígito 1 sería la existencia de hashtags que solo han sido mencionados una sola vez. Al eliminarlos se reduce considerablemente el porcentaje, pero aumentan el del resto de dígitos (distribución FSD España (# $>$ 1)). Debido a la descompensación generada por eliminar únicamente aquellos mencionados una única vez, se considera eliminar aquellos que tienen menos de 10 menciones (distribución FSD España (# $\geq$ 10)).

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Total**
FSD Benford	30.01 %	17.60 %	12.40 %	9.69 %	7.91 %	6.69 %	5.79 %	5.11 %	4.57 %	100 %
FSD España	60.17 %	19.89 %	7.00 %	4.56 %	2.74 %	2.03 %	1.52 %	1.16 %	0.92 %	100 %
FSD España (# $>$ 1)	9.68 %	45.07 %	15.89 %	10.36 %	6.19 %	4.61 %	3.46 %	2.65 %	2.09 %	100 %
FSD España (# $\geq$ 10)	56.12 %	18.74 %	9.24 %	5.37 %	3.65 %	2.33 %	1.94 %	1.27 %	1.18 %	100 %

Tabla 4.3: Distribución del primer dígito más significativo para España con variaciones

En la Tabla 4.3 se observan los resultados obtenidos para las distintas combinaciones antes mencionadas. Ninguna de ellas presenta porcentajes similares a la distribución esperada del FSD de Benford.

Si se estudia la distribución del segundo dígito más significativo (*Second Significant Digit*), los resultados tampoco arrojan indicios de que se cumpla la distribución esperada salvo para el caso que contiene hashtags con un número de impresiones igual o superior a 10 (ver Tabla 4.4)

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Total**
SSD Benford	11.96 %	11.38 %	10.98 %	10.43 %	10.03 %	9.66 %	9.33 %	9.03 %	8.75 %	8.5 %	100 %
SSD España	93.68 %	1.08 %	0.92 %	0.81 %	0.76 %	0.66 %	0.54 %	0.56 %	0.51 %	0.56 %	100 %
SSD España (# $>$ 1)	85.67 %	2.45 %	2.09 %	1.84 %	1.73 %	1.51 %	1.23 %	1.27 %	1.17 %	1.05 %	100 %
SSD España (# $\geq$ 10)	16.76 %	14.25 %	12.12 %	10.70 %	10.04 %	8.74 %	7.14 %	7.37 %	6.78 %	6.10 %	100 %

Tabla 4.4: Distribución del segundo dígito más significativo para España con variaciones

4.2.2. Dataset `Elecciones`

En la Tabla 4.5 se observa, para el primer dígito más significativo, el porcentaje esperado y el observado para el dataset.

	1	2	3	4	5	6	7	8	9	Total
FSD Benford	30.01 %	17.60 %	12.40 %	9.69 %	7.91 %	6.69 %	5.79 %	5.11 %	4.57 %	100 %
FSD `Elecciones`	40.39 %	25.99 %	8.97 %	9.23 %	3.91 %	4.50 %	2.54 %	2.69 %	1.77 %	100 %

Tabla 4.5: Distribución del primer dígito más significativo del dataset `Elecciones`

En este caso vuelve a haber una diferencia entre el valor teórico y observado para el dígito 1 (ver Figura 4.5), por lo que se aplican las funciones de edición nuevamente con los mismos parámetros de configuración. En las Tablas B.5, B.6 y B.7 se pueden observar los resultados obtenidos.

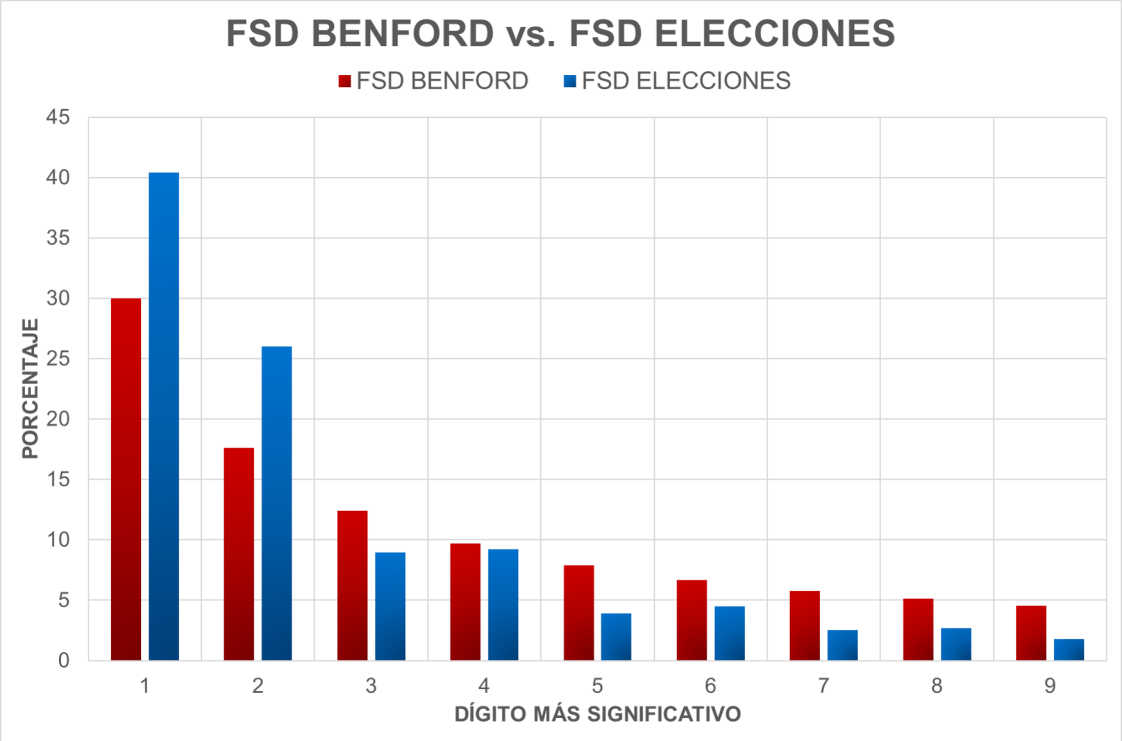


Figura 4.5: Gráfica comparativa entre la distribución del FSD esperada y la del FSD `Elecciones`

Al igual que se hizo para el datatset `Hispatweets`, se calculan las distribuciones FSD y SSD quitando aquellos hashtags con una única mención y con menos de 10 menciones (ver Tablas 4.6 y 4.7). En ningún caso se logra obtener unos porcentajes similares a los esperados por la Ley de Benford.

	1	2	3	4	5	6	7	8	9	Total
FSD Benford	30.01 %	17.60 %	12.40 %	9.69 %	7.91 %	6.69 %	5.79 %	5.11 %	4.57 %	100 %
FSD `Elecciones`	40.39 %	25.99 %	8.97 %	9.23 %	3.91 %	4.50 %	2.54 %	2.69 %	1.77 %	100 %
FSD `Elecciones` **(# $>$ 1)**	15.14 %	36.82 %	12.71 %	13.08 %	5.54 %	6.38 %	3.60 %	3.81 %	2.51 %	100 %
FSD `Elecciones` **(# $\geq$ 10)**	48.36 %	18.25 %	11.26 %	7.38 %	4.94 %	3.49 %	2.88 %	1.33 %	2.11 %	100 %

Tabla 4.6: Distribución del primer dígito más significativo del dataset `Elecciones` con variaciones

	0	1	2	3	4	5	6	7	8	9	Total
SSD Benford	11.96 %	11.38 %	10.98 %	10.43 %	10.03 %	9.66 %	9.33 %	9.03 %	8.75 %	8.5 %	100 %
SSD `Elecciones`	81.17 %	2.35 %	3.27 %	2.34 %	2.30 %	1.67 %	2.29 %	1.46 %	1.86 %	1.28 %	100 %
SSD `Elecciones` **(# $>$ 1)**	73.32 %	3.33 %	4.63 %	3.31 %	3.26 %	2.37 %	3.24 %	2.07 %	2.64 %	1.82 %	100 %
SSD `Elecciones` **(# $\geq$ 10)**	16.97 %	10.37 %	14.42 %	10.32 %	10.15 %	7.38 %	10.09 %	6.43 %	8.21 %	5.66 %	100 %

Tabla 4.7: Distribución del segundo dígito más significativo del dataset `Elecciones` con variaciones

4.2.3. Dataset `Bárcenas`

En la Tabla 4.8 se observan los porcentajes teóricos y observados para el dígito más significativo del dataset `Bárcenas`.

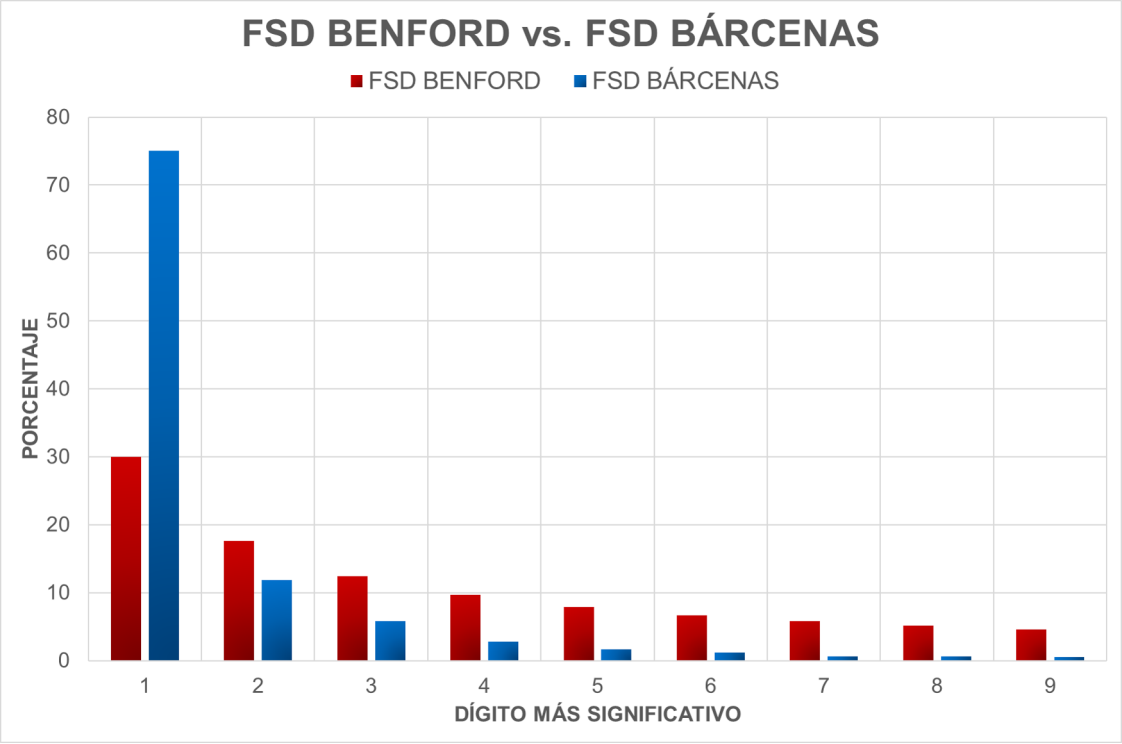


Figura 4.6: Gráfica comparativa entre la distribución del FSD esperada y la del FSD `Bárcenas`

Como se observa en la Figura 4.6, vuelve a haber una gran diferencia entre el valor esperado para el dígito 1, un 30.01 %, frente al valor observado, un 75.07 %. Al aplicar las distancias de edición quedan los resultados de las Tablas B.8, B.9 y B.10.

	1	2	3	4	5	6	7	8	9	Total
FSD Benford	30.01 %	17.60 %	12.40 %	9.69 %	7.91 %	6.69 %	5.79 %	5.11 %	4.57 %	100 %
FSD Bárcenas	75.07 %	11.84 %	5.85 %	2.75 %	1.61 %	1.16 %	0.64 %	0.58 %	0.50 %	100 %

Tabla 4.8: Distribución del primer dígito más significativo del dataset Bárcenas

Al igual que para los dos anteriores datasets, se calculan los valores de las distribuciones de hashtags para el primer dígito más significativo y para el segundo más significativo, eliminando los hashtags mencionados en una única ocasión y mencionados más de 10 veces (ver Tablas 4.9 y 4.10). En ningún caso se logra obtener unos porcentajes similares a los esperados por la Ley de Benford.

	1	2	3	4	5	6	7	8	9	Total
FSD Benford	30.01 %	17.60 %	12.40 %	9.69 %	7.91 %	6.69 %	5.79 %	5.11 %	4.57 %	100 %
FSD Bárcenas	75.07 %	11.84 %	5.85 %	2.75 %	1.61 %	1.16 %	0.64 %	0.58 %	0.50 %	100 %
FSD Bárcenas **(# $>$ 1)**	7.03 %	44.16 %	21.82 %	10.24 %	6.00 %	4.34 %	2.38 %	2.17 %	1.86 %	100 %
FSD Bárcenas **(# $\geq$ 10)**	50.75 %	18.66 %	12.69 %	3.73 %	5.22 %	3.73 %	0.00 %	2.24 %	2.99 %	100 %

Tabla 4.9: Distribución del primer dígito más significativo el dataset Bárcenas con variaciones

	0	1	2	3	4	5	6	7	8	9	Total
SSD Benford	11.96 %	11.38 %	10.98 %	10.43 %	10.03 %	9.66 %	9.33 %	9.03 %	8.75 %	8.5 %	100 %
SSD Bárcenas	96.92 %	0.61 %	0.39 %	0.58 %	0.31 %	0.39 %	0.19 %	0.22 %	0.25 %	0.14 %	100 %
SSD Bárcenas **(# $>$ 1)**	88.52 %	2.28 %	1.45 %	2.17 %	1.14 %	1.45 %	0.72 %	0.83 %	0.93 %	0.52 %	100 %
SSD Bárcenas **(# $\geq$ 10)**	17.16 %	16.42 %	10.45 %	15.67 %	8.21 %	10.45 %	5.22 %	5.97 %	6.72 %	3.73 %	100 %

Tabla 4.10: Distribución del segundo dígito más significativo el dataset Bárcenas con variaciones

4.2.4. Resultados

Los resultados obtenidos para las distribuciones de los dígitos más significativos en las frecuencias de los hashtags de todos los datasets permiten afirmar que son similares a la distribución esperada según la Ley de Benford para el primer y para el segundo dígito más significativo, por lo que se confirma la aplicación de la Ley de Benford a redes sociales como Twitter.

Si se observan los resultados del Coeficiente de Correlación de Pearson, la correlación entre la distribución del primer dígito más significativo y la distribución esperada de Benford es muy fuerte, ya que todos los valores son superiores a 0.9, como se observa en la Tabla 4.11. Se ve también que correlación de las distribuciones que contienen hashtags con un número de menciones mayor a 1 es débil, debido al sesgo producido al eliminar dichos hashtags. Sin embargo, cuando esto se corrige, las distribuciones de los hashtags con menciones igual o mayores a 10 muestran una correlación también fuerte.

Distribución	Correlación	χ^2	MAD
FSD España	0.9699	51.41	0.071
FSD España (# $>$ 1)	0.4460	62.16	0.070
FSD España (# $\geq$ 10)	0.9760	38.61	0.060
FSD `Elecciones`	0.9835	23.78	0.038
FSD `Elecciones` **(# $>$ 1)**	0.5971	123.44	0.068
FSD `Elecciones` **(# $\geq$ 10)**	0.9865	34.22	0.046
FSD `Bárcenas`	0.9288	99.78	0.099
FSD `Bárcenas` **(# $>$ 1)**	0.3942	71.45	0.081
FSD `Bárcenas` **(# $\geq$ 10)**	0.9786	28.24	0.048

Tabla 4.11: Valores del índice de Correlación de Pearson, del test de la χ^2 y de la Desviación Absoluta Media entre las distribuciones del primer dígito más significativo de los datasets y los valores esperados por la Ley de Benford

Si estudiamos ahora la distribución del segundo dígito más significativo (ver Tabla 4.12), se observa que las distribuciones observadas presentan unos valores de cercanos al 0.6, es decir, bajos. Para la distribución que toma en cuenta los hashtags con un número de menciones superior a 1, los valores vuelven a ser bajos, mientras que la distribución de los hashtags con un número de menciones igual o superior a 10 presenta valores muy altos para España mientras que algo más bajos para el resto de datasets.

Distribución	Correlación	χ^2	MAD
SSD España	0.5946	634.14	0.163
SSD España (# $>$ 1)	0.6029	516.06	0.147
SSD España (# $\geq$ 10)	0.9850	4.80	0.018
SSD `Elecciones`	0.6039	445.14	0.138
SSD `Elecciones` **(# $>$ 1)**	0.6120	357.95	0.122
SSD `Elecciones` **(# $\geq$ 10)**	0.8644	5.59	0.018
SSD `Bárcenas`	0.5933	685.58	0.169
SSD `Bárcenas` **(# $>$ 1)**	0.6043	556.91	0.153
SSD `Bárcenas` **(# $\geq$ 10)**	0.8963	13.54	0.032

Tabla 4.12: Valores del índice de Correlación de Pearson, del test de la χ^2 y de la Desviación Absoluta Media entre las distribuciones del segundo dígito más significativo de los datasets y los valores esperados por la Ley de Benford

Los resultados obtenidos por el test de la χ^2 no permiten afirmar que las distribuciones del primer dígito más significativo sigan la Ley de Benford. Tampoco las distribuciones del segundo dígito más significativo permiten realizar la afirmación, a excepción de las distribuciones que contienen los hashtags con un número de menciones igual o superior a 10 que sí que presentan resultados que permiten afirmar que se asemejan a lo esperado por la Ley de Benford.

El test de la Desviación Absoluta media presenta resultados que hacen rechazar, para todos los casos, la hipótesis de que los datos se asemejan a la Ley de Benford.

Los resultados de los test de la χ^2 y la Desviación Absoluta Media presentan resultados desfavorables debido a que para datasets de gran tamaño no se recomienda su uso [9] para comprobar si se cumple la Ley de Benford, ya que una pequeña variación en los datos hará que automáticamente se rechace la hipótesis nula formulada.

Tras aplicar en los tres datasets las medidas de similitud propuestas los resultados siguen siendo desfavorables, como se observa en las Tablas 4.13, 4.14 y 4.15. Puede verse también que el índice de Correlación de Pearson es muy alto en todos los casos.

	Levenshtein			**Jaro**			**Jaro-Winkler**		
	Correlación	χ^2	**MAD**	**Correlación**	χ^2	**MAD**	**Correlación**	χ^2	**MAD**
$\alpha = 0.95$	0.9699	51.28	0.071	0.9709	49.42	0.070	0.9798	35.91	0.060
$\alpha = 0.90$	0.9689	50.04	0.070	0.9639	46.29	0.068	0.9869	24.59	0.050
$\alpha = 0.85$	0.9695	48.76	0.070	0.9773	39.70	0.063	0.9926	14.52	0.038
$\alpha = 0.80$	0.9725	47.15	0.069	0.9844	28.34	0.054	0.9966	7.48	0.028

Tabla 4.13: Valores del índice de Correlación de Pearson, del test de la χ^2 y de la Desviación Absoluta Media entre las distribuciones del primer dígito más significativo para España aplicando las medidas de similitud y los valores esperados por la Ley de Benford

	Levenshtein			**Jaro**			**Jaro-Winkler**		
	Correlación	χ^2	**MAD**	**Correlación**	χ^2	**MAD**	**Correlación**	χ^2	**MAD**
$\alpha = 0.95$	0.9836	16.00	0.041	0.9845	15.61	0.040	0.9880	12.58	0.036
$\alpha = 0.90$	0.9845	15.70	0.041	0.9852	14.75	0.039	0.9926	9.15	0.030
$\alpha = 0.85$	0.9855	15.14	0.041	0.9889	13.29	0.037	0.9958	5.79	0.023
$\alpha = 0.80$	0.9857	15.23	0.069	0.9908	10.76	0.033	0.9979	2.72	0.014

Tabla 4.14: Valores del índice de Correlación de Pearson, del test de la χ^2 y de la Desviación Absoluta Media entre las distribuciones del primer dígito más significativo para el dataset `Elecciones` aplicando las medidas de similitud y los valores esperados por la Ley de Benford

	Levenshtein			Jaro			Jaro-Winkler		
	Correlación	χ^2	**MAD**	**Correlación**	χ^2	**MAD**	**Correlación**	χ^2	**MAD**
$\alpha = 0.95$	0.9290	99.62	0.099	0.9296	98.33	0.099	0.9410	83.42	0.090
$\alpha = 0.90$	0.9290	98.66	0.099	0.9352	94.41	0.096	0.9524	69.75	0.081
$\alpha = 0.85$	0.9304	96.58	0.097	0.9390	85.97	0.091	0.9614	56.71	0.072
$\alpha = 0.80$	0.9314	94.02	0.096	0.9494	71.83	0.083	0.9781	36.38	0.056

Tabla 4.15: Valores del índice de Correlación de Pearson, del test de la χ^2 y de la Desviación Absoluta Media entre las distribuciones del primer dígito más significativo para el dataset `Bárceas` aplicando las medidas de similitud y los valores esperados por la Ley de Benford

CAPÍTULO 5
Conclusiones y líneas de trabajo futuras

En esta memoria se muestra que la distribución de los hahstags de acuerdo a su popularidad sigue una distribución potencial, según la Ley de Zipf, donde pocos hashtags han conseguido un gran número de menciones y muchos de ellos no han tenido éxito y han obtenido pocas menciones. Esto se explica porque Twitter no pone ningún obstáculo para la creación de hashtags ni ha desarrollado un sistema recomendador que dé indicaciones al usuario sobre las características que debe tener un hashtag para que tenga éxito.

Se puede aseverar, además, que la distribución de la frecuencia de los hashtags, en concreto la distribución del primer y segundo dígito más significativo de la frecuencia, siguen la distribución esperada por la Ley de Benford atendiendo a los valores del Coeficiente de Correlación de Pearson. Sin embargo, en las distribuciones del primer dígito más significativo se observan diferencias notorias entre los valores obtenidos y esperados para el dígito 1. Este fenómeno se debe a la cantidad de hashtags que tienen una sola mención, debido entre otros factores a contener faltas de ortografía, alcanzando así bajos niveles de popularidad. Para subsanar esta diferencia en las distribuciones se han aplicado las distancias de edición entre los hashtags y se han agrupado los que son similares entre sí, no obteniendo resultados que la redujeran.

Para calcular el nivel de conformidad de las distribuciones obtenidas con las distribuciones esperadas por la Ley de Benford se hace uso de tres estadísticos. Aquellos que realizan un contraste de hipótesis, los estadísticos de la χ^2 y la Desviación Absoluta Media, arrojan valores dispares que hacen aceptar o rechazar la hipótesis nula de conformidad con la Ley de Benford dependiendo de la distribución del dígito que se analice. Sin embargo, el Coeficiente de Correlación de Pearson sí que permite aseverar que los datos se asemejan a la distribución esperada por Benford. Esta disparidad se produce debido a que desde que se haya una pequeña variación entre los datos obtenidos y los esperados la hipótesis nula será rechazada, unido a que no se dispone de un procedimiento estandarizado y válido para el estudio de la conformidad con la Ley de Benford, ni siquiera en el trabajo original publicado [10]. En el estudio realizado por Golbeck [9] se presenta este caso y se propone el Coeficiente de Correlación de Pearson como nuevo método para calcular la semejanza de una distribución a la Ley de Benford.

Este trabajo presenta la utilidad de la Ley de Benford, que aplicada al estudio del comportamiento de los usuarios en las redes sociales, permite validar los datos que se generan en ellas para analizarlos y validarlos.

Como trabajo futuro se considerarán y añadirán más variables al estudio, como la fecha y hora de publicación de los tweets, para analizar la distribución a lo largo del tiempo de la popularidad de los hashtags.

Bibliografía

[1] W. Chen, L. V. Lakshmanan, and C. Castillo, "Information and influence propagation in social networks," *Synthesis Lectures on Data Management*, vol. 5, no. 4, pp. 1–177, 2013.

[2] D. Bird, M. Ling, K. Haynes, *et al.*, "Flooding Facebook-the use of social media during the Queensland and Victorian floods," *Australian Journal of Emergency Management, The*, vol. 27, no. 1, p. 27, 2012.

[3] M. Beguerisse-Díaz, G. Garduno-Hernández, B. Vangelov, S. N. Yaliraki, and M. Barahona, "Interest communities and flow roles in directed networks: the Twitter network of the UK riots," *Journal of The Royal Society Interface*, vol. 11, no. 101, p. 20140940, 2014.

[4] C. G. Ortega and R. Z. Azagra, "La campaña virtual en Twitter: análisis de las cuentas de Rajoy y de Rubalcaba en las elecciones generales de 2011," *Historia y Comunicación Social*, vol. 19, pp. 299–311, 2014.

[5] G. K. Zipf, *Human behavior and the principle of least effort: An introduction to human ecology*. Ravenio Books, 2016.

[6] F. Benford, "The law of anomalous numbers," *Proceedings of the American Philosophical Society*, pp. 551–572, 1938.

[7] C. Durtschi, W. Hillison, and C. Pacini, "The effective use of Benford's Law to assist in detecting fraud in accounting data," *Journal of forensic accounting*, vol. 5, no. 1, pp. 17–34, 2004.

[8] B. F. Roukema, "Benford's Law anomalies in the 2009 Iranian presidential election," *Unpublished manuscript*, 2009.

[9] J. Golbeck, "Benford's Law Applies to Online Social Networks," *PLOS ONE*, vol. 10, no. 8, p. e0135169, 2015.

[10] M. Nigrini, *Benford's Law: Applications for forensic accounting, auditing, and fraud detection*, vol. 586. John Wiley & Sons, 2012.

[11] G. Judge and L. Schechter, "Detecting problems in survey data using Benford's Law," *Journal of Human Resources*, vol. 44, no. 1, pp. 1–24, 2009.

[12] E. H. Porter, W. E. Winkler, *et al.*, "Approximate string comparison and its effect on an advanced record linkage system," in *Advanced record linkage system. US Bureau of the Census, Research Report*, Citeseer, 1997.

APÉNDICE A
Cálculo de la Distancia de Levenshtein

```
Sea m la longitud de la palabra s1
Sea n la longitud de la palabra s2
Sea D la matriz de tamano [m,n]

# PASO 1 - Comprobacion
Si (m) == 0:
  Devolver n y salir

Si (n) == 0:
  Devolver m y salir


# PASO 2 - Inicializacion
Desde i = 1 hasta m hacer:
  D[i,0] = i

Desde j = 1 hasta n hacer:
  D[0,j] = j


# PASO 3 - Calculo de la matriz
Desde i = 1 hasta m hacer:
    Desde j = 1 hasta n hacer:
        D[i,j] = minimo( D[i-1,j] + 1,
                         D[i,j-1] + 1,
                         D[i-1,j-1] + 1 si s1[i] != s2[j]
                                    + 0 si s1[i] == s2[j]
                       )

# PASO 4 - Resultado
D[m,n] es la distancia de Levenshtein
```

Listing A.1: Pseudocódigo para el cálculo de la Distancia de Levenshtein

APÉNDICE B
Tablas de resultados

Dígito	F&S-SD	Benford (%)	Dígito	F&S-SD	Benford (%)
10	4.14		55	0.78	
11	3.78		56	0.77	
12	3.48		57	0.76	
13	3.22		58	0.74	
14	3.00		59	0.73	
15	2.80		60	0.72	
16	2.63		61	0.71	
17	2.48		62	0.69	
18	2.35		63	0.68	
19	2.23		64	0.67	
20	2.12		65	0.66	
21	2.02		66	0.65	
22	1.93		67	0.64	
23	1.85		68	0.63	
24	1.77		69	0.62	
25	1.70		70	0.62	
26	1.64		71	0.61	
27	1.58		72	0.60	
28	1.52		73	0.59	
29	1.47		74	0.58	
30	1.42		75	0.58	
31	1.38		76	0.57	
32	1.34		77	0.56	
33	1.30		78	0.55	
34	1.26		79	0.55	
35	1.22		80	0.54	
36	1.19		81	0.53	
37	1.16		82	0.53	
38	1.13		83	0.52	
39	1.10		84	0.51	
40	1.07		85	0.51	
41	1.05		86	0.50	
42	1.02		87	0.50	
43	1.00		88	0.49	
44	0.98		89	0.49	
45	0.95		90	0.48	
46	0.93		91	0.47	
47	0.91		92	0.47	
48	0.90		93	0.46	
49	0.88		94	0.46	
50	0.86		95	0.45	
51	0.84		96	0.45	
52	0.83		97	0.45	
53	0.81		98	0.44	
54	0.80		99	0.44	

Tabla B.1: Distribución del primer y segundo dígito más significativo según la Ley de Benford

Distancia de Levenshtein									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
FSD Benford	30.01 %	17.60 %	12.40 %	9.69 %	7.91 %	6.69 %	5.79 %	5.11 %	4.57 %
FSD España	60.17 %	19.89 %	7.00 %	4.56 %	2.74 %	2.03 %	1.52 %	1.16 %	0.92 %
$\alpha = 0.95$	60.14 %	19.88 %	7.01 %	4.57 %	2.73 %	2.04 %	1.52 %	1.17 %	0.92 %
$\alpha = 0.90$	59.73 %	19.22 %	7.12 %	4.67 %	2.77 %	2.10 %	1.55 %	1.21 %	0.94 %
$\alpha = 0.85$	59.30 %	19.16 %	7.25 %	4.70 %	2.82 %	2.19 %	1.57 %	1.23 %	0.96 %
$\alpha = 0.80$	58.71 %	20.11 %	7.36 %	4.80 %	2.93 %	2.23 %	1.61 %	1.22 %	1.03 %

Tabla B.2: Porcentajes de la distribución FSD de España al aplicar la función de similitud de Levenshtein para varios niveles de α

Distancia de Jaro									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
FSD Benford	30.01 %	17.60 %	12.40 %	9.69 %	7.91 %	6.69 %	5.79 %	5.11 %	4.57 %
FSD España	60.17 %	19.89 %	7.00 %	4.56 %	2.74 %	2.03 %	1.52 %	1.16 %	0.92 %
$\alpha = 0.95$	59.53 %	19.93 %	7.19 %	4.67 %	2.79 %	2.13 %	1.56 %	1.24 %	0.96 %
$\alpha = 0.90$	58.49 %	16.77 %	7.42 %	4.89 %	3.14 %	2.14 %	1.68 %	1.23 %	1.03 %
$\alpha = 0.85$	56.00 %	20.44 %	8.07 %	5.24 %	3.35 %	2.46 %	1.85 %	1.43 %	1.17 %
$\alpha = 0.80$	51.50 %	20.58 %	9.17 %	6.17 %	4.10 %	2.97 %	2.29 %	1.81 %	1.42 %

Tabla B.3: Porcentajes de la distribución FSD de España al aplicar la función de similitud de Jaro para varios niveles de α

Distancia de Jaro-Winkler									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
FSD Benford	30.01 %	17.60 %	12.40 %	9.69 %	7.91 %	6.69 %	5.79 %	5.11 %	4.57 %
FSD España	60.17 %	19.89 %	7.00 %	4.56 %	2.74 %	2.03 %	1.52 %	1.16 %	0.92 %
$\alpha = 0.95$	54.57 %	20.48 %	8.46 %	5.58 %	3.60 %	2.61 %	1.96 %	1.47 %	1.27 %
$\alpha = 0.90$	49.78 %	20.70 %	9.48 %	6.55 %	4.40 %	3.13 %	2.41 %	1.92 %	1.62 %
$\alpha = 0.85$	44.74 %	20.51 %	10.62 %	7.18 %	5.25 %	4.02 %	2.95 %	2.67 %	2.06 %
$\alpha = 0.80$	39.89 %	20.56 %	11.12 %	8.33 %	5.92 %	4.72 %	3.45 %	3.49 %	2.55 %

Tabla B.4: Porcentajes de la distribución FSD de España al aplicar la función de similitud de Jaro-Winkler para varios niveles de α

Distancia de Levenshtein									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
FSD Benford	30.01 %	17.60 %	12.40 %	9.69 %	7.91 %	6.69 %	5.79 %	5.11 %	4.57 %
FSD `Elecciones`	40.39 %	25.99 %	8.97 %	9.23 %	3.91 %	4.50 %	2.54 %	2.69 %	1.77 %
$\alpha = 0.95$	40.43 %	25.98 %	8.96 %	9.21 %	3.92 %	4.50 %	2.55 %	2.70 %	1.75 %
$\alpha = 0.90$	40.47 %	25.77 %	9.02 %	9.22 %	3.98 %	4.50 %	2.58 %	2.68 %	1.77 %
$\alpha = 0.85$	40.31 %	25.53 %	9.21 %	9.28 %	4.04 %	4.63 %	2.57 %	2.67 %	1.77 %
$\alpha = 0.80$	40.36 %	25.43 %	9.16 %	9.51 %	4.08 %	4.53 %	2.53 %	2.70 %	1.70 %

Tabla B.5: Porcentajes de la distribución FSD del dataset `Elecciones` al aplicar la función de similitud de Levenshtein para varios niveles de α

Distancia de Jaro									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
FSD Benford	30.01 %	17.60 %	12.40 %	9.69 %	7.91 %	6.69 %	5.79 %	5.11 %	4.57 %
FSD `Elecciones`	40.39 %	25.99 %	8.97 %	9.23 %	3.91 %	4.50 %	2.54 %	2.69 %	1.77 %
$\alpha = 0.95$	40.36 %	25.77 %	9.06 %	9.28 %	4.00 %	4.55 %	2.56 %	2.66 %	1.75 %
$\alpha = 0.90$	39.87 %	25.56 %	9.26 %	9.52 %	4.23 %	4.57 %	2.53 %	2.77 %	1.68 %
$\alpha = 0.85$	39.67 %	24.64 %	9.73 %	9.82 %	4.37 %	4.69 %	2.64 %	2.73 %	1.71 %
$\alpha = 0.80$	38.61 %	24.02 %	10.36 %	9.56 %	4.73 %	4.96 %	2.85 %	2.90 %	2.00 %

Tabla B.6: Porcentajes de la distribución FSD del dataset `Elecciones` al aplicar la función de similitud de Jaro para varios niveles de α

Distancia de Jaro-Winkler									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
FSD Benford	30.01 %	17.60 %	12.40 %	9.69 %	7.91 %	6.69 %	5.79 %	5.11 %	4.57 %
FSD `Elecciones`	40.39 %	25.99 %	8.97 %	9.23 %	3.91 %	4.50 %	2.54 %	2.69 %	1.77 %
$\alpha = 0.95$	39.00 %	24.61 %	9.85 %	10.07 %	4.45 %	4.88 %	2.51 %	2.79 %	1.83 %
$\alpha = 0.90$	38.08 %	23.24 %	10.60 %	9.79 %	4.92 %	5.09 %	3.24 %	3.07 %	1.98 %
$\alpha = 0.85$	36.35 %	21.88 %	11.55 %	9.80 %	5.42 %	5.51 %	3.61 %	3.49 %	2.40 %
$\alpha = 0.80$	33.8 %	20.49 %	12.56 %	9.70 %	6.61 %	5.81 %	4.13 %	4.27 %	2.63 %

Tabla B.7: Porcentajes de la distribución FSD del dataset `Elecciones` al aplicar la función de similitud de Jaro-Winkler para varios niveles de α

Distancia de Levenshtein									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
FSD Benford	30.01 %	17.60 %	12.40 %	9.69 %	7.91 %	6.69 %	5.79 %	5.11 %	4.57 %
FSD `Bárcenas`	75.07 %	11.84 %	5.85 %	2.75 %	1.61 %	1.16 %	0.64 %	0.58 %	0.50 %
$\alpha = 0.95$	75.01 %	11.86 %	5.89 %	2.75 %	1.58 %	1.19 %	0.61 %	0.61 %	0.50 %
$\alpha = 0.90$	74.90 %	11.85 %	5.99 %	2.64 %	1.60 %	1.29 %	0.62 %	0.59 %	0.51 %
$\alpha = 0.85$	74.36 %	12.08 %	6.15 %	2.69 %	1.57 %	1.20 %	0.72 %	0.63 %	0.60 %
$\alpha = 0.80$	73.84 %	12.33 %	6.05 %	2.79 %	1.66 %	1.34 %	0.76 %	0.67 %	0.58 %

Tabla B.8: Porcentajes de la distribución FSD del dataset `Bárcenas` al aplicar la función de similitud de Levenshtein para varios niveles de α

Distancia de Jaro									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
FSD Benford	30.01 %	17.60 %	12.40 %	9.69 %	7.91 %	6.69 %	5.79 %	5.11 %	4.57 %
FSD `Bárcenas`	75.07 %	11.84 %	5.85 %	2.75 %	1.61 %	1.16 %	0.64 %	0.58 %	0.50 %
$\alpha = 0.95$	74.71 %	11.95 %	6.10 %	2.64 %	1.53 %	1.22 %	0.65 %	0.60 %	0.60 %
$\alpha = 0.90$	73.73 %	12.51 %	6.02 %	2.92 %	1.61 %	1.32 %	0.70 %	0.64 %	0.56 %
$\alpha = 0.85$	71.49 %	13.55 %	6.61 %	3.04 %	1.78 %	1.47 %	0.68 %	0.74 %	0.65 %
$\alpha = 0.80$	67.52 %	14.70 %	7.95 %	3.70 %	2.09 %	1.75 %	0.86 %	0.65 %	0.79 %

Tabla B.9: Porcentajes de la distribución FSD del dataset `Bárcenas` al aplicar la función de similitud de Jaro para varios niveles de α

	Distancia de Jaro-Winkler								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
FSD Benford	30.01 %	17.60 %	12.40 %	9.69 %	7.91 %	6.69 %	5.79 %	5.11 %	4.57 %
FSD `Bárcenas`	75.07 %	11.84 %	5.85 %	2.75 %	1.61 %	1.16 %	0.64 %	0.58 %	0.50 %
$\alpha = 0.95$	70.78 %	14.01 %	6.65 %	2.97 %	1.84 %	1.45 %	0.68 %	0.87 %	0.74 %
$\alpha = 0.90$	66.78 %	15.18 %	8.43 %	3.45 %	2.12 %	1.60 %	1.04 %	0.67 %	0.74 %
$\alpha = 0.85$	62.73 %	16.22 %	9.46 %	4.24 %	2.27 %	1.75 %	1.54 %	0.94 %	0.86 %
$\alpha = 0.80$	55.12 %	18.14 %	11.19 %	5.25 %	3.50 %	2.07 %	2.33 %	1.38 %	1.01 %

Tabla B.10: Porcentajes de la distribución FSD del dataset `Bárcenas` al aplicar la función de similitud de Jaro-Winkler para varios niveles de α