ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE INGENIERÍA DE SISTEMAS

RECUPERACIÓN DE LA INFORMACIÓN

Integrantes:

- KEVIN MALDONADO
- PAOLA AUCAPIÑA
- RAQUEL ZUMBA

DOCENTE: IVAN CARRERA

FECHA DE ENTREGA: 30-05-2024

LEY DE ZIPF

1. **Recopilación de Datos**

1.1 Crear una lista con todas las palabras y sus respectivas frecuencias.

En este primer paso, creamos una lista con todas las palabras del corpus con sus respectivas frecuencias.

	A	B
1	**Palabra**	**Frecuencia**
2	impresión	1
3	domingraf	1
4	impressors	1
5	pol	1
6	ind	1
7	can	1
8	magarola	1
9	pasaje	1
10	autopista	1
11	nave	1
12	mollet	1
13	vallés	1
14	jessica	1
15	gustan	1
16	historias	1
17	anne	1
18	gustaban	1
19	di	1
20	oyó	1
21	ésta	1
22	primero	2
23	niño	6
24	vivió	1
25	señor	24
26	señora	13

26	señora	13
27	dursley	36
28	vivían	1
29	número	3
30	privet	3
31	drive	3
32	orgullosos	1
33	decir	1
34	normales	1
35	afortunada	1
36	últimas	1
37	personas	4
38	esperaría	1
39	encontrar	1
40	relacionad	1
41	extraño	1
42	misterioso	2
43	tales	1
44	tonterías	1
45	director	1
46	empresa	1
47	llamada	1
48	grunnings	2
49	fabricaba	1
50	taladros	6
51	hombre	3

2. Ordenación de las Palabras

2.1 Ordenar las palabras por frecuencia, de mayor a menor

Como se puede observar en las imágenes tenemos las palabras ordenadas de mayor a menor frecuencia.

Palabra	Frecuencia
dursley	36
señor	24
señora	13
mientras	11
potter	9
gato	8
aquel	7
lechuzas	7
niño	6
taladros	6
dudley	6
hermana	6
día	6
gente	6
noche	6
si	5
capa	5
personas	4
casi	4
tiempo	4
hijo	4
visto	4
vio	4
casa	4
cuenta	4
ser	4

privet	3	30
drive	3	31
hombre	3	32
aunque	3	33
habitual	3	34
mayor	3	35
llamado	3	36
así	3	37
nunca	3	38
después	3	39
gran	3	40
trató	3	41
coche	3	42
alejó	3	43
esquina	3	44
volvió	3	45
vez	3	46
debía	3	47
mirada	3	48
calle	3	49
llevaba	3	50
mañana	3	51
normal	3	52
harry	3	53
quedó	3	54
seguro	3	55
hoy	3	56

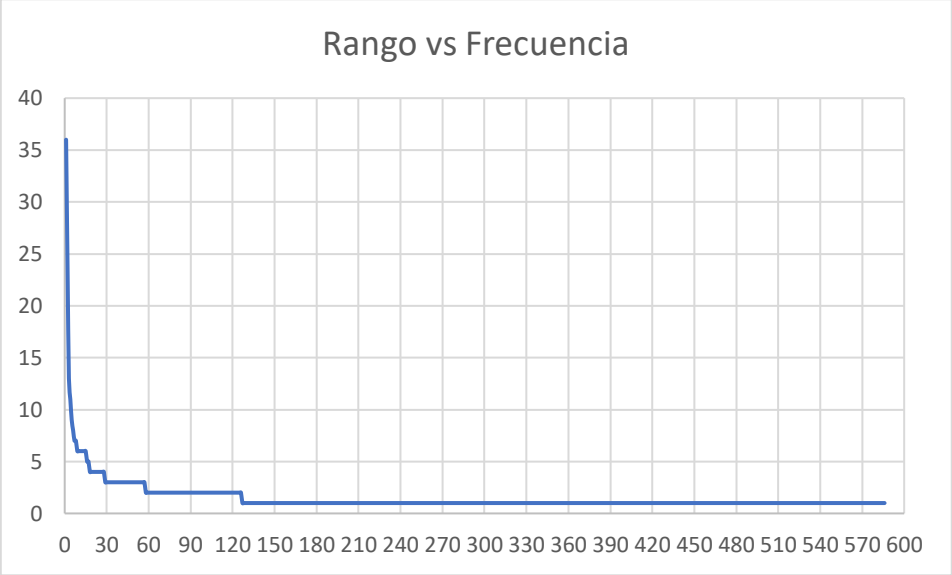
2.2 Asignar un rango a cada palabra basado en su posición en la lista ordenada (1 para la palabra más frecuente, 2 para la siguiente, etc.).

En las imágenes podemos ver que la palabra más frecuente esta asignada el rango de 1 y para la siguiente le asignamos el número 2 y así sucesivamente.

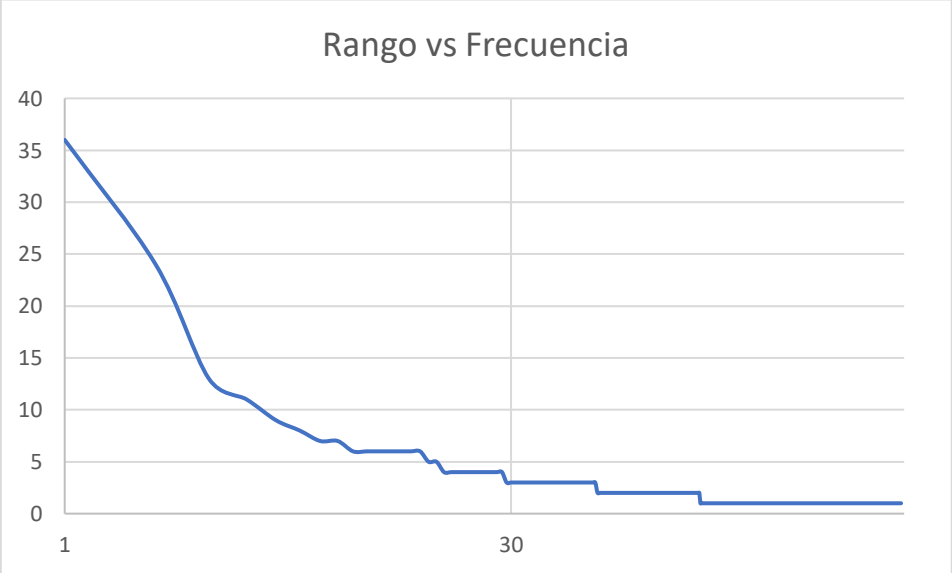
Palabra	Frecuencia	Rango
dursley	36	1
señor	24	2
señora	13	3
mientras	11	4
potter	9	5
gato	8	6
aquel	7	7
lechuzas	7	8
niño	6	9
taladros	6	10
dudley	6	11
hermana	6	12
día	6	13
gente	6	14
noche	6	15
si	5	16
capa	5	17
personas	4	18
casi	4	19
tiempo	4	20
hijo	4	21
visto	4	22
vio	4	23
casa	4	24
cuenta	4	25
ser	4	26
aunque	3	33
habitual	3	34
mayor	3	35
llamado	3	36
así	3	37
nunca	3	38
después	3	39
gran	3	40
trató	3	41
coche	3	42
alejó	3	43
esquina	3	44
volvió	3	45
vez	3	46
debía	3	47
mirada	3	48
calle	3	49
llevaba	3	50
mañana	3	51
normal	3	52
harry	3	53
quedó	3	54
seguro	3	55
hoy	3	56
entró	3	57
primero	2	58
misterioso	2	59

3. Análisis de la Ley de Zipf

3.1 Graficar las frecuencias (eje Y) contra los rangos (eje X) en una escala lineal.



3.2 Graficar las frecuencias (eje Y) contra los rangos (eje X) en una escala logarítmica.



3.3 Analizar si las gráficas siguen una línea recta, lo cual confirmaría la ley de Zipf.

En la gráfica lineal, la relación entre rango y frecuencia no forma una línea recta, sino que muestra una rápida caída inicial seguida de una cola larga y plana. Esto

indica una distribución de frecuencias en la que unas pocas palabras tienen frecuencias muy altas y muchas palabras tienen frecuencias bajas.

La gráfica lineal no muestra una evidencia clara de que la Ley de Zipf se cumpla en este conjunto de datos. Sin embargo, la tendencia lineal descendente de la gráfica es consistente con la Ley de Zipf.

La gráfica en escala logarítmica sugiere que los datos siguen una tendencia que se aproxima a una línea recta, lo cual es consistente con la ley de Zipf. Aunque no es una línea perfectamente recta, la forma de la gráfica log-log respalda la hipótesis de que la frecuencia de las palabras en el texto sigue la distribución de Zipf, especialmente en el rango medio.

4. Discusión

4.1 ¿Como afecta la alta frecuencia de palabras comunes (como stopwords) en la eficiencia de un sistema de búsqueda?

La alta frecuencia de palabras comunes en un sistema de búsqueda puede reducir su eficiencia al generar muchos resultados irrelevantes, por lo tanto, esto hace que el tiempo de respuesta sea lento. Esto ocurre porque las stopwords ocupan espacio y recursos sin aportar valor significativo, dificultando la identificación de contenido relevante.

4.2 ¿Qué técnicas se pueden utilizar para manejar palabras extremadamente frecuentes y palabras raras en sistemas de Recuperación de Información?

Existen varias técnicas para abordar estos desafíos, lo cual nos permite optimizar la eficacia del sistema de RI. Por ejemplo:

Para el Manejo de Palabras Extremadamente Frecuentes podemos utilizar las siguientes técnicas:

- **Eliminación de Stop Words:** Elimina palabras comunes (artículos, preposiciones) que no aportan valor al análisis.

- **Ponderación TF-IDF**: Reduce la importancia de palabras comunes ponderando su frecuencia inversa en el corpus.

Para el Manejo de Palabras Raras podemos utilizar la siguiente técnica:

- **Lematización y Stemming**: Reducen las palabras a sus formas base o raíces; la lematización usa la forma del diccionario, el stemming corta los sufijos.

4.3 ¿Cómo se pueden usar estos hallazgos para mejorar la precisión y el recall de un motor de búsqueda?

Para mejorar la precisión y el recall de un motor de búsqueda, se puede aplicar la eliminación de stop words para reducir el ruido y enfocar la búsqueda en términos relevantes, mientras que la ponderación TF-IDF ayuda a equilibrar la importancia de palabras comunes y raras, priorizando documentos más significativos. La lematización y el stemming normalizan las palabras a sus formas base, asegurando la recuperación de documentos que contienen diferentes variaciones de los términos de búsqueda.

Además de las técnicas mencionadas, también es importante considerar estrategias adicionales para mejorar la precisión y el recall del motor de búsqueda. Por ejemplo, la expansión de consultas mediante la inclusión de sinónimos y términos relacionados puede aumentar la cobertura de la búsqueda y mejorar el recall. Al implementar estas técnicas, el motor de búsqueda se vuelve más eficiente, proporcionando resultados más pertinentes y recuperando una mayor cantidad de documentos relevantes.

Anexo:

Hoja de Cálculo:

https://epnecuador-my.sharepoint.com/:x:/g/personal/paola_aucapina_epn_edu_ec/EVNL-AZD8RVFkd_592sgodsB8LFxfHcYkicUGLcvlWCx6Q?e=WKaMlf