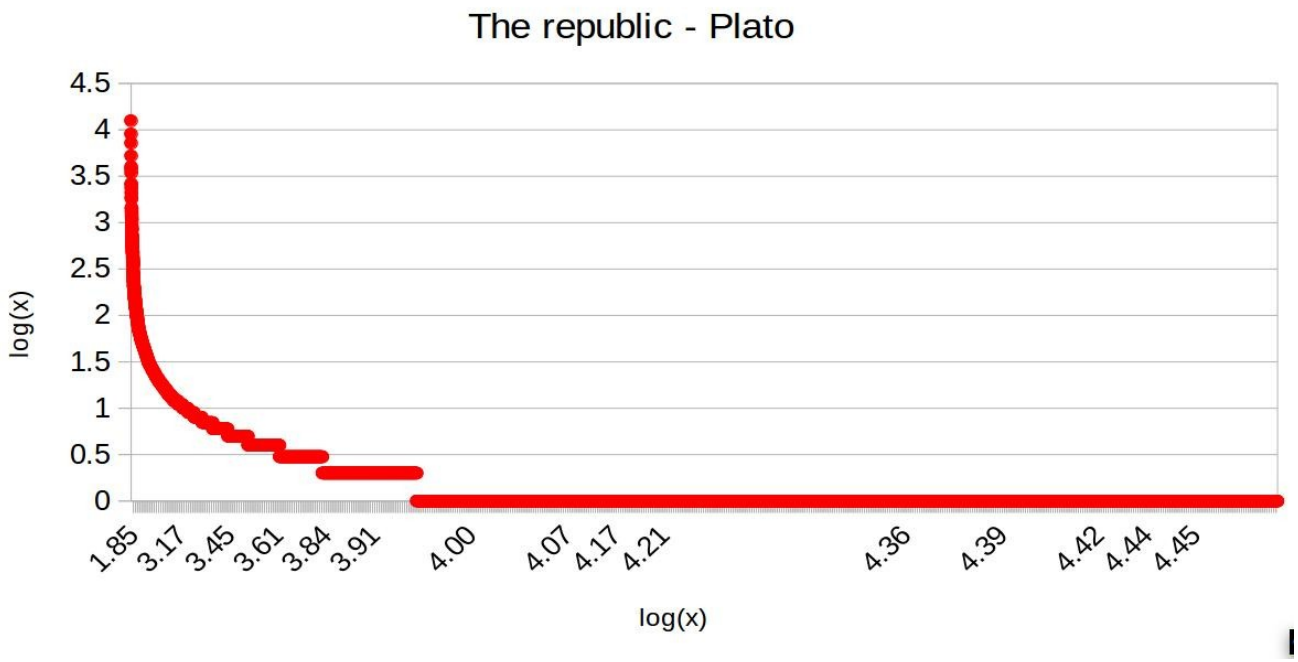
Práctica 1
Preprocesado de Documentos

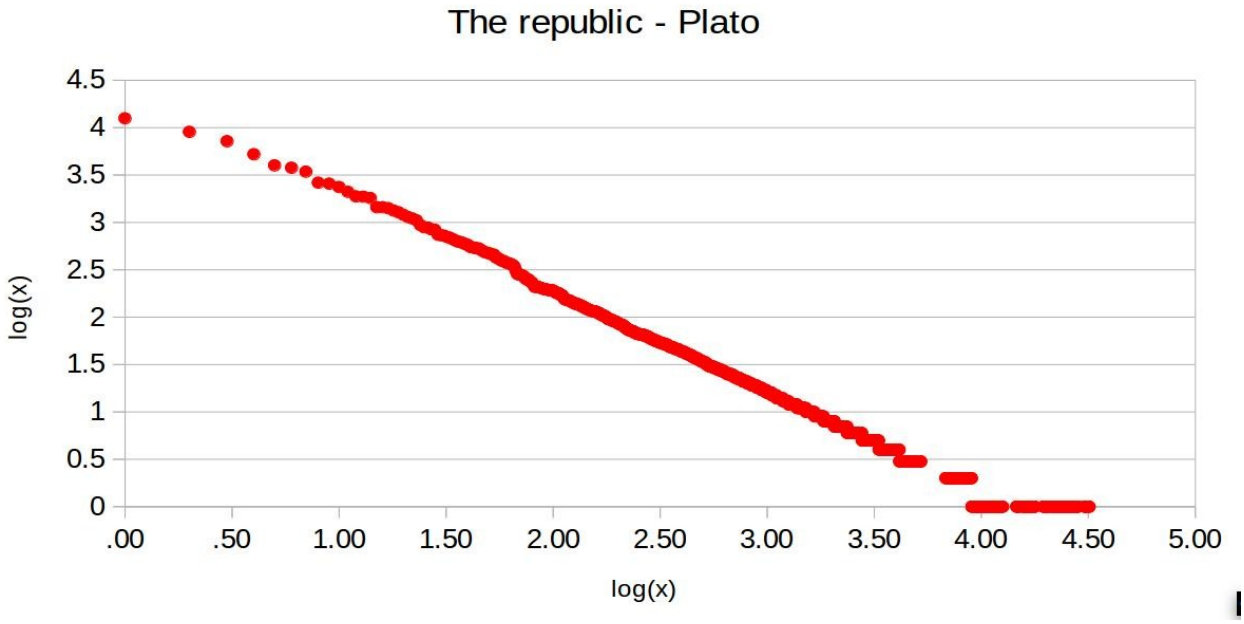
Jose Manuel Rodríguez Calvo
***Juan Miguel* Vilchez**

Resultados

El primer libro sobre el que hemos aplicado nuestro codigo está escrito en Ingles, se trata de 'the republic by Plato'. La grafia que hemos obtenido es la siguiente:



En la que podemos observar que se cumple la ley de Zipf puesto que la inmesa mayoría de las palabras se usan muy pocas veces, mientras que un pequeño numero son las que frecuentemente se repiten, y entre estas dos partes encontrariamos las palabras importantes que nos hacen deducir de que trata el documento.



Comoo podemos observar en la figura anterior podemos usar la ecuación de Booth y Federowicz debido a que es una linea bien definida con la pendiente, con la que gracias a los calculos podemos obtener los valores caracteristicos para definirla matematicamente:

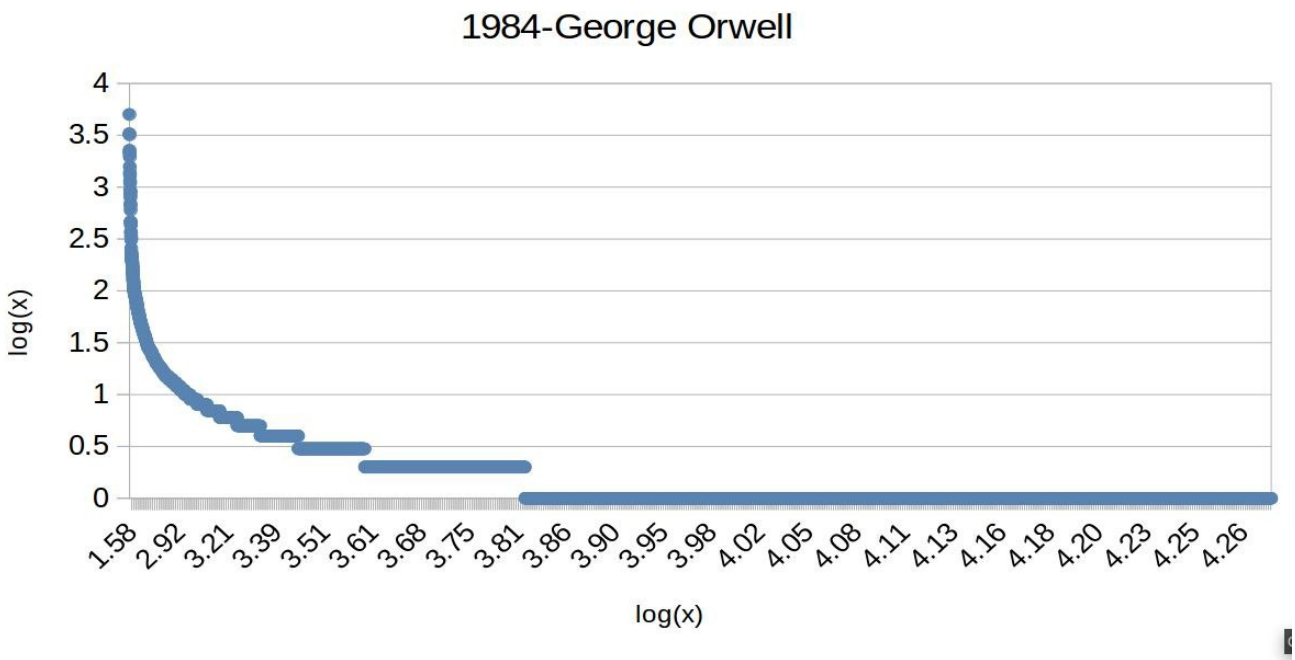
Con el preambulo de: $\log(F)=\log(k)-m\cdot\log(R)$ y utilizando la primera palabra en el Ranking $R=1\Rightarrow\log(R)=0$, facilmente obtenemos: $4.09=\log(k)\Rightarrow k=10^{4.09}$.

Utilizando el preambulo $\log(F)=\log(k)-m\cdot\log(R)$ y en vez de aplicarlo a la primera palabra en el ranking, usamos la segunda, que nos da como solucion:

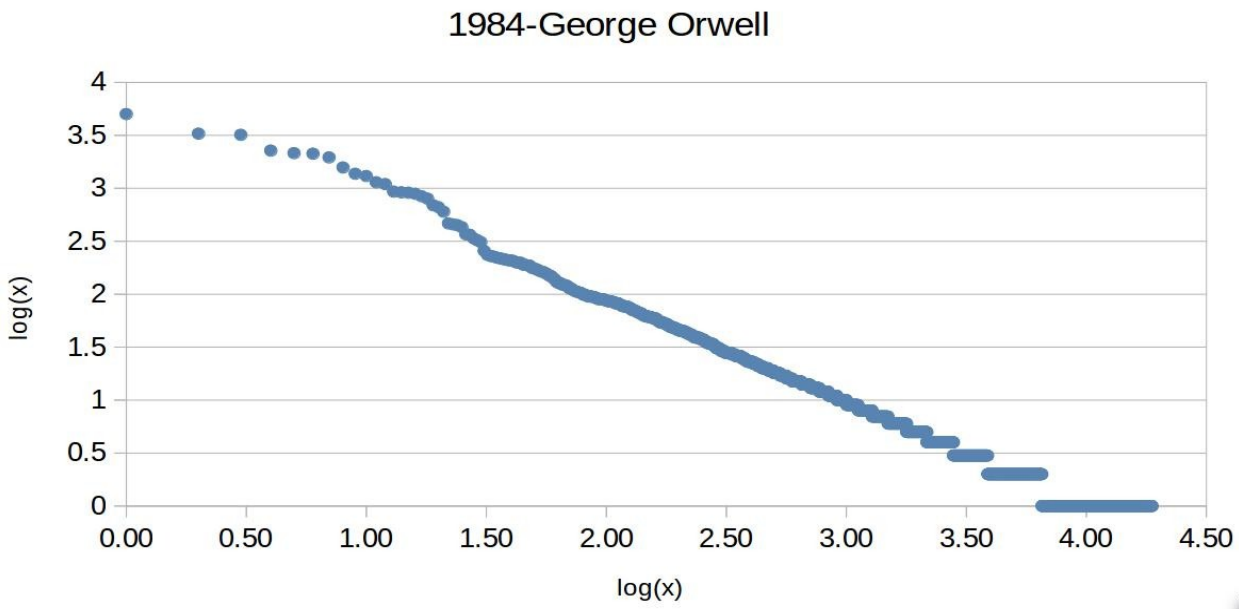
$3.92=\log(10^{4.09})-m\cdot 0.3\approx m=0.57$

Por tanto como solucion obtenemos **k = $10^{4.09}$** y **m = 0.57.**

El segundo libro sobre el que hemos aplicado nuestro codigo también está escrito en Ingles, se trata de '1984 by George Orwell'. La grafia que hemos obtenido es la siguiente:



En la que podemos observar que se cumple la ley de Zipf puesto que la inmesa mayoría de las palabras se usan muy pocas veces, mientras que un pequeño numero son las que frecuentemente se repiten, y entre estas dos partes encontrariamos las palabras importantes que nos hacen deducir de que trata el documento.



Comoo podemos observar en la figura anterior podemos usar la ecuación de Booth y Federowicz debido a que es una linea bien definida con la pendiente, con la que gracias a los calculos podemos obtener los valores caracteristicos para definirla matematicamente:

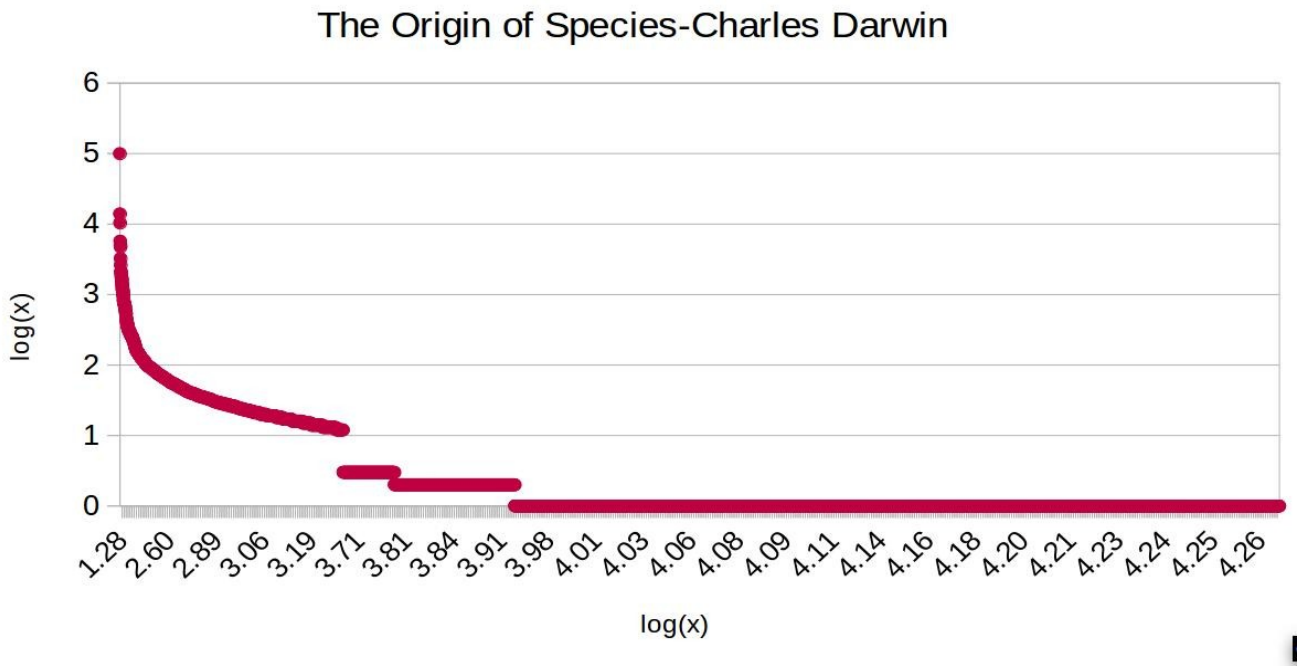
Con el preambulo de: $\log(F)=\log(k)-m\cdot\log(R)$ y utilizando la primera palabra en el Ranking $R=1\Rightarrow\log(R)=0$, facilmente obtenemos: $3.7=\log(k)\Rightarrow k=10^{3.7}$.

Utilizando el preambulo $\log(F)=\log(k)-m\cdot\log(R)$ y en vez de aplicarlo a la primera palabra en el ranking, usamos la segunda, que nos da como solucion:

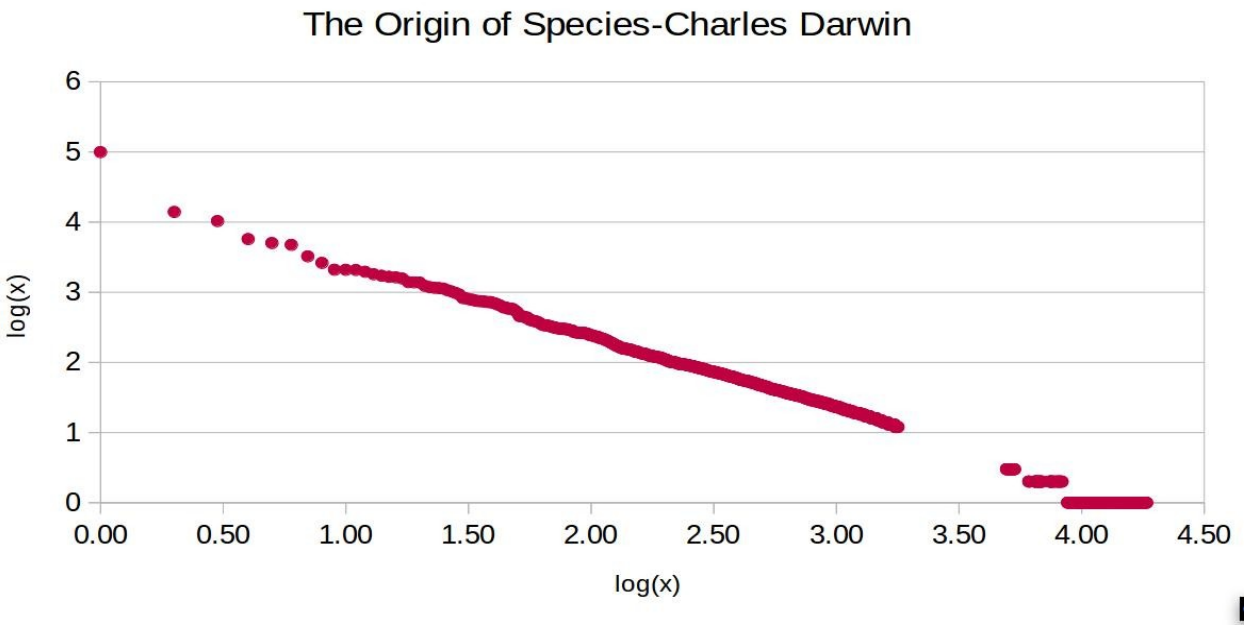
$3.51=\log(10^{3.7})-m\cdot 0.3\approx m=0.63$

Por tanto como solucion obtenemos **k = $10^{3.7}$** y **m = 0.63.**

El tercer libro sobre el que hemos aplicado nuestro codigo también está escrito en Ingles, se trata de 'The Origin of Species by Charles Darwin'. La grafia que hemos obtenido es la siguiente:



En la que podemos observar que se cumple la ley de Zipf puesto que la inmesa mayoría de las palabras se usan muy pocas veces, mientras que un pequeño numero son las que frecuentemente se repiten, y entre estas dos partes encontrariamos las palabras importantes que nos hacen deducir de que trata el documento.



Comoo podemos observar en la figura anterior podemos usar la ecuación de Booth y Federowicz debido a que es una linea bien definida con la pendiente, con la que gracias a los calculos podemos obtener los valores caracteristicos para definirla matematicamente:

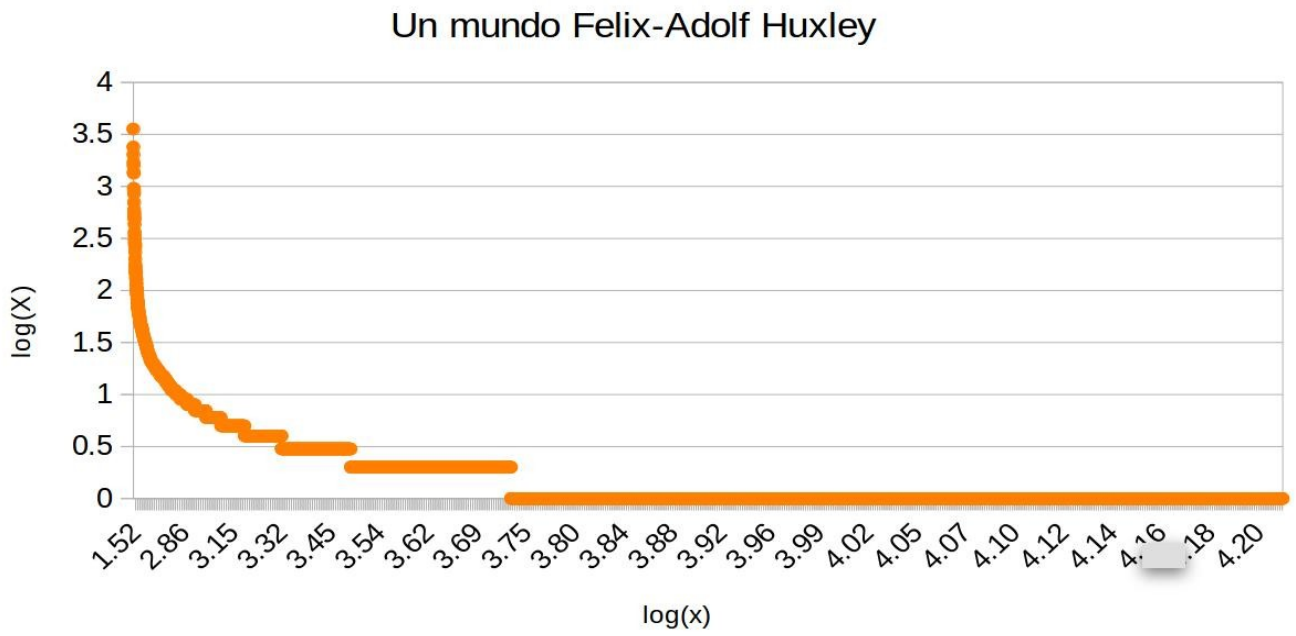
Con el preambulo de: $\log(F)=\log(k)-m\cdot\log(R)$ y utilizando la primera palabra en el Ranking $R=1\Rightarrow\log(R)=0$, facilmente obtenemos: $4.99=\log(k)\Rightarrow k=10^{4.99}$.

Utilizando el preambulo $\log(F)=\log(k)-m\cdot\log(R)$ y en vez de aplicarlo a la primera palabra en el ranking, usamos la segunda, que nos da como solucion:

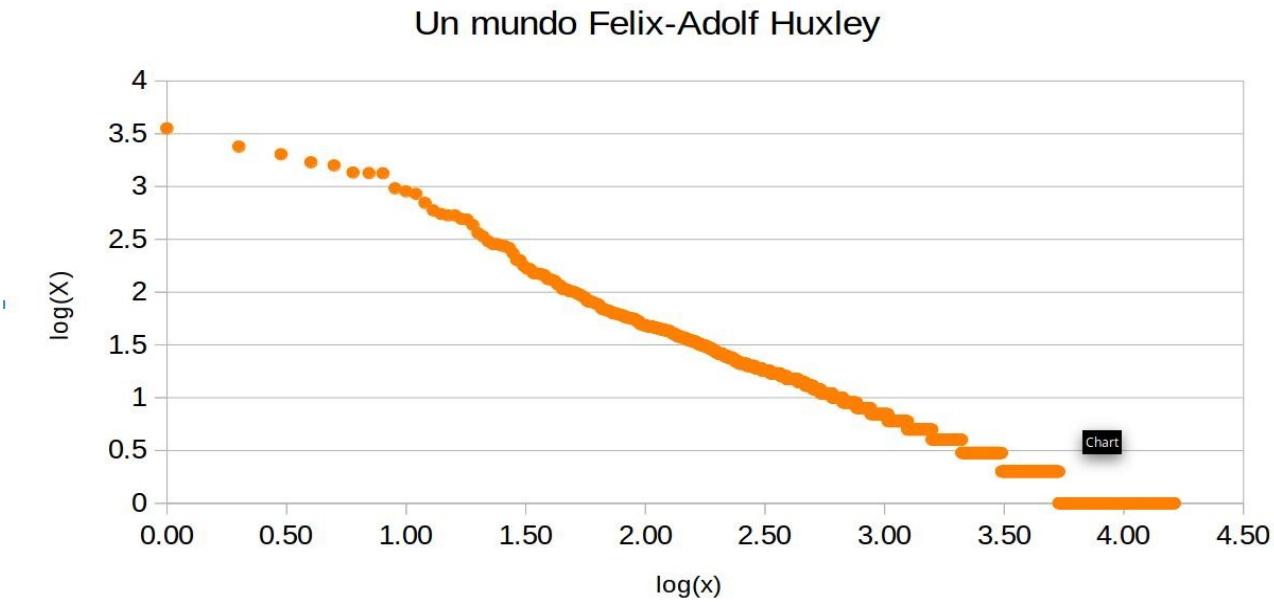
$4.14=\log(10^{4.99})-m\cdot 0.3\approx m=2.83$

Por tanto como solucion obtenemos **k = $10^{4.99}$** y **m = 2.83.**

El cuarto libro sobre el que hemos aplicado nuestro codigo está escrito en castellano, se trata de 'Un Mundo Feliz de Adolf Huxley'. La grafia que hemos obtenido es la siguiente:



En la que podemos observar que se cumple la ley de Zipf puesto que la inmesa mayoría de las palabras se usan muy pocas veces, mientras que un pequeño numero son las que frecuentemente se repiten, y entre estas dos partes encontrariamos las palabras importantes que nos hacen deducir de que trata el documento.



Comoo podemos observar en la figura anterior podemos usar la ecuación de Booth y Federowicz debido a que es una linea bien definida con la pendiente, con la que gracias a los calculos podemos obtener los valores caracteristicos para definirla matematicamente:

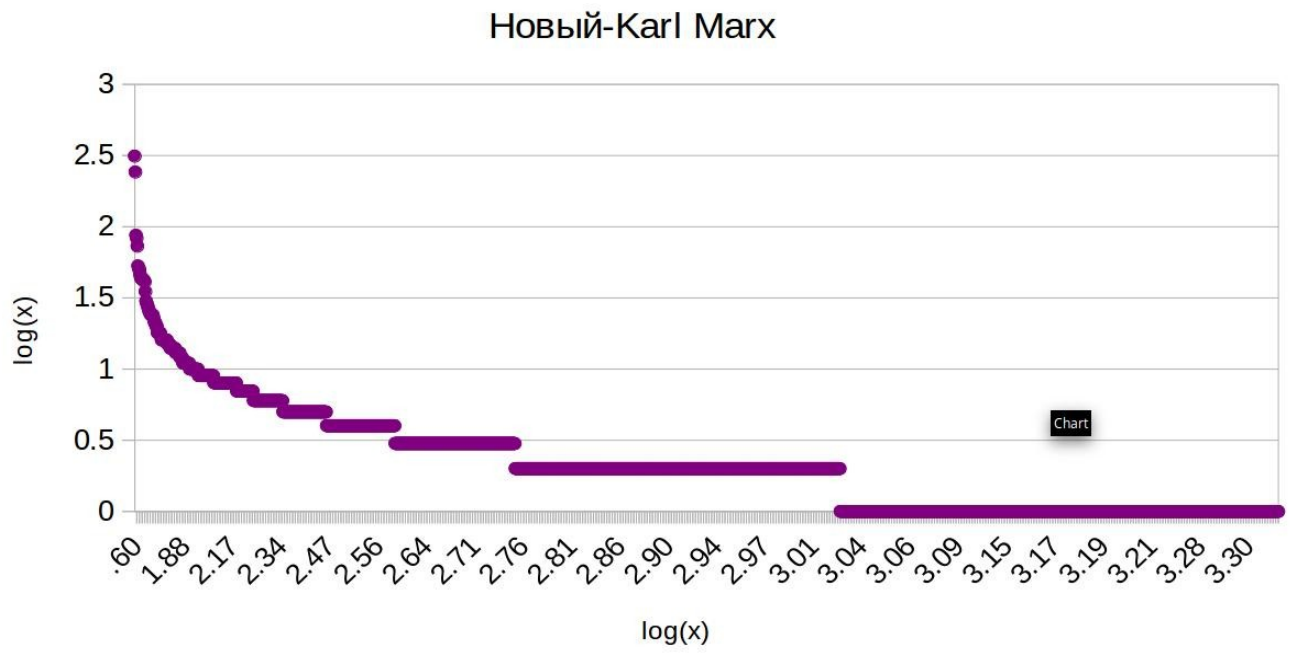
Con el preambulo de: $\log(F)=\log(k)-m\cdot\log(R)$ y utilizando la primera palabra en el Ranking $R=1\Rightarrow\log(R)=0$, facilmente obtenemos: $3.55=\log(k)\Rightarrow k=10^{3.55}$.

Utilizando el preambulo $\log(F)=\log(k)-m\cdot\log(R)$ y en vez de aplicarlo a la primera palabra en el ranking, usamos la segunda, que nos da como solucion:

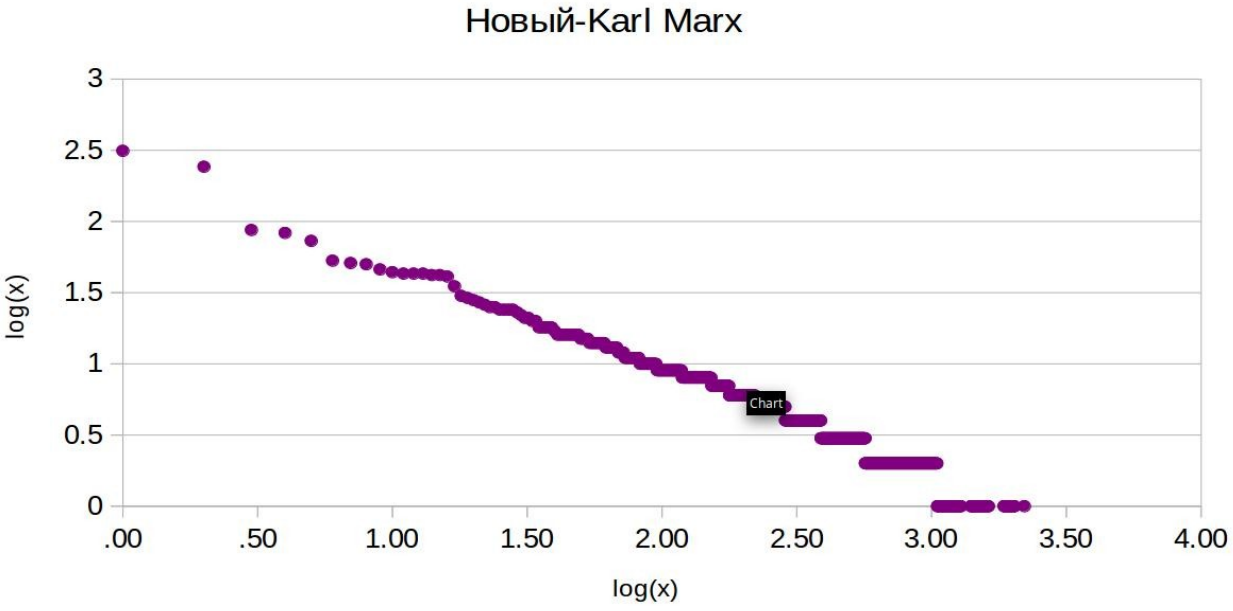
$3.38=\log(10^{3.55})-m\cdot 0.3\approx m=0.567$

Por tanto como solucion obtenemos **k = $10^{3.55}$** y **m = 0.567.**

El ultimo libro sobre el que hemos aplicado nuestro codigo está escrito en ruso, se trata de ‘El Manifiesto Comunista de Karl Marx’. La grafia que hemos obtenido es la siguiente:



En la que podemos observar que se cumple la ley de Zipf puesto que la inmesa mayoría de las palabras se usan muy pocas veces, mientras que un pequeño numero son las que frecuentemente se repiten, y entre estas dos partes encontrariamos las palabras importantes que nos hacen deducir de que trata el documento.



Comoo podemos observar en la figura anterior podemos usar la ecuación de Booth y Federowicz debido a que es una linea bien definida con la pendiente, con la que gracias a los calculos podemos obtener los valores caracteristicos para definirla matematicamente:

Con el preambulo de: $\log(F)=\log(k)-m\cdot\log(R)$ y utilizando la primera palabra en el Ranking $R=1\Rightarrow\log(R)=0$, facilmente obtenemos: $2.49=\log(k)\Rightarrow k=10^{2.49}$.

Utilizando el preambulo $\log(F)=\log(k)-m\cdot\log(R)$ y en vez de aplicarlo a la primera palabra en el ranking, usamos la segunda, que nos da como solucion:

$2.38=\log(10^{2.49})-m\cdot 0.3\approx m=0.367$

Por tanto como solucion obtenemos **k = $10^{2.49}$** y **m = 0.367.**