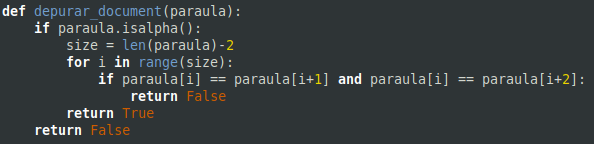
# Pràctica 1: ElasticSearch and Zipf’s and Heaps’ laws

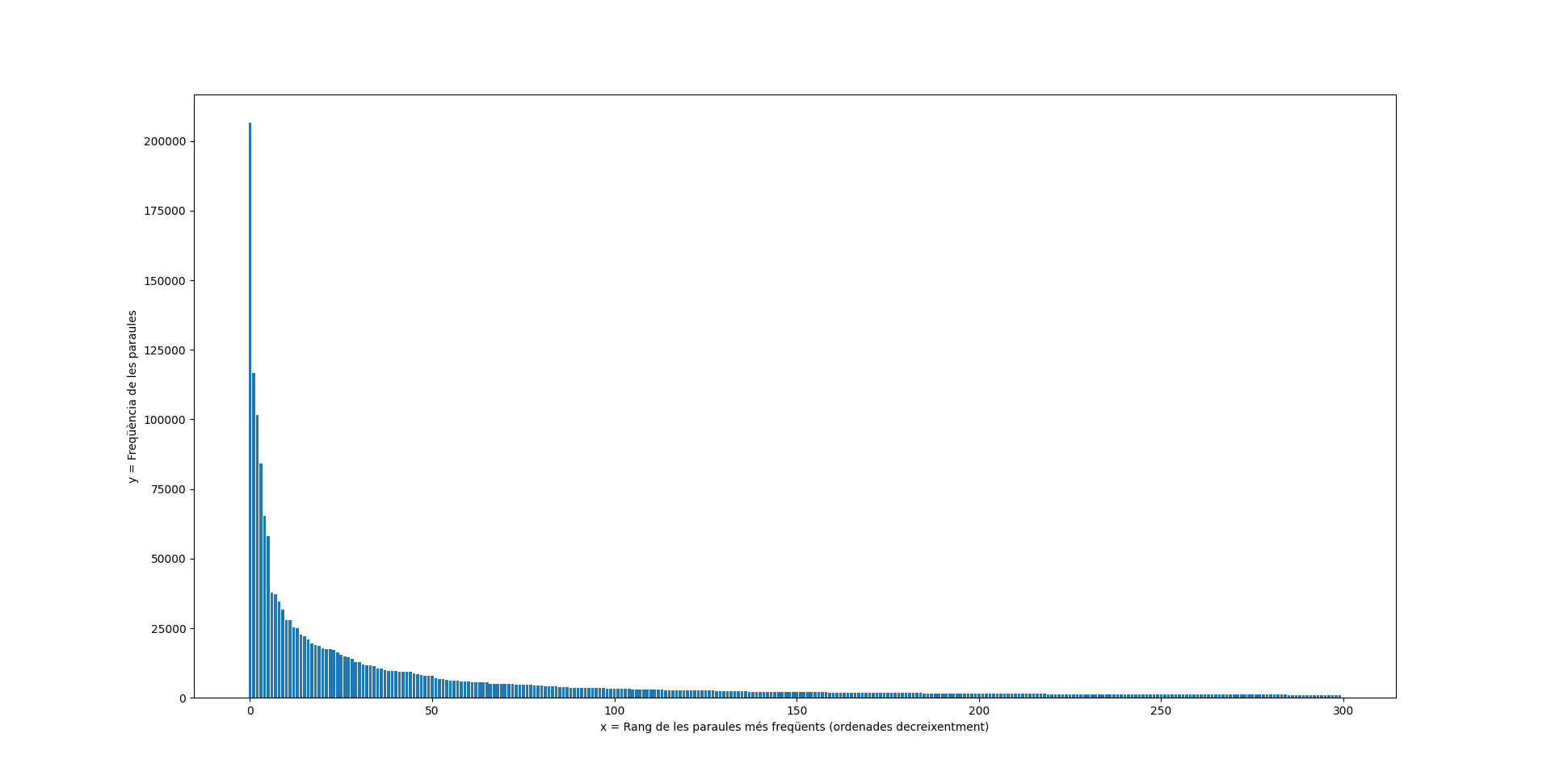
L’objectiu d’aquesta primera pràctica era familiaritzar-nos amb l’entorn ElasticSearch i aprendre a indexar documents per al posterior tractament d’aquests. Per això, hem treballat amb diferents documents de diferents mides i hem observat aspectes com el nombre de paraules que contenen. En concret, hem focalitzat el nostre análisis sobre la col·lecció de *novels,* al ser el grup de documents més petit i net.

En primer lloc, hem hagut de depurar les paraules dels documents, ja que el programa *CountWords.py* mostrava termes que no són considerats paraules com a tals. Per a això hem decidit treure mots que continguessin números, guions, barres baixes, signes de puntuació… amb l’ús de la funció isalpha(), la qual indicava si es tractava d’una paraula (mot amb només lletres de l’abecedari) amb un boolean. Un cop feta la primera depuració, hem vist que encara continuaven apareixent molts termes que no són considerats paraules, i per aconseguir un output més net hem inclós al condicional que les paraules amb repeticions de 3 o més cops d’una mateixa lletra tampoc s’imprimissin.



*Figura 1: Codi de depuració de textos*

Després d’aquest tractament previ de l’índex, hem començat amb l’estudi de les lleis de Zipf i de Heaps, per observar com es comporten els documents i si satisfan aquests principis.



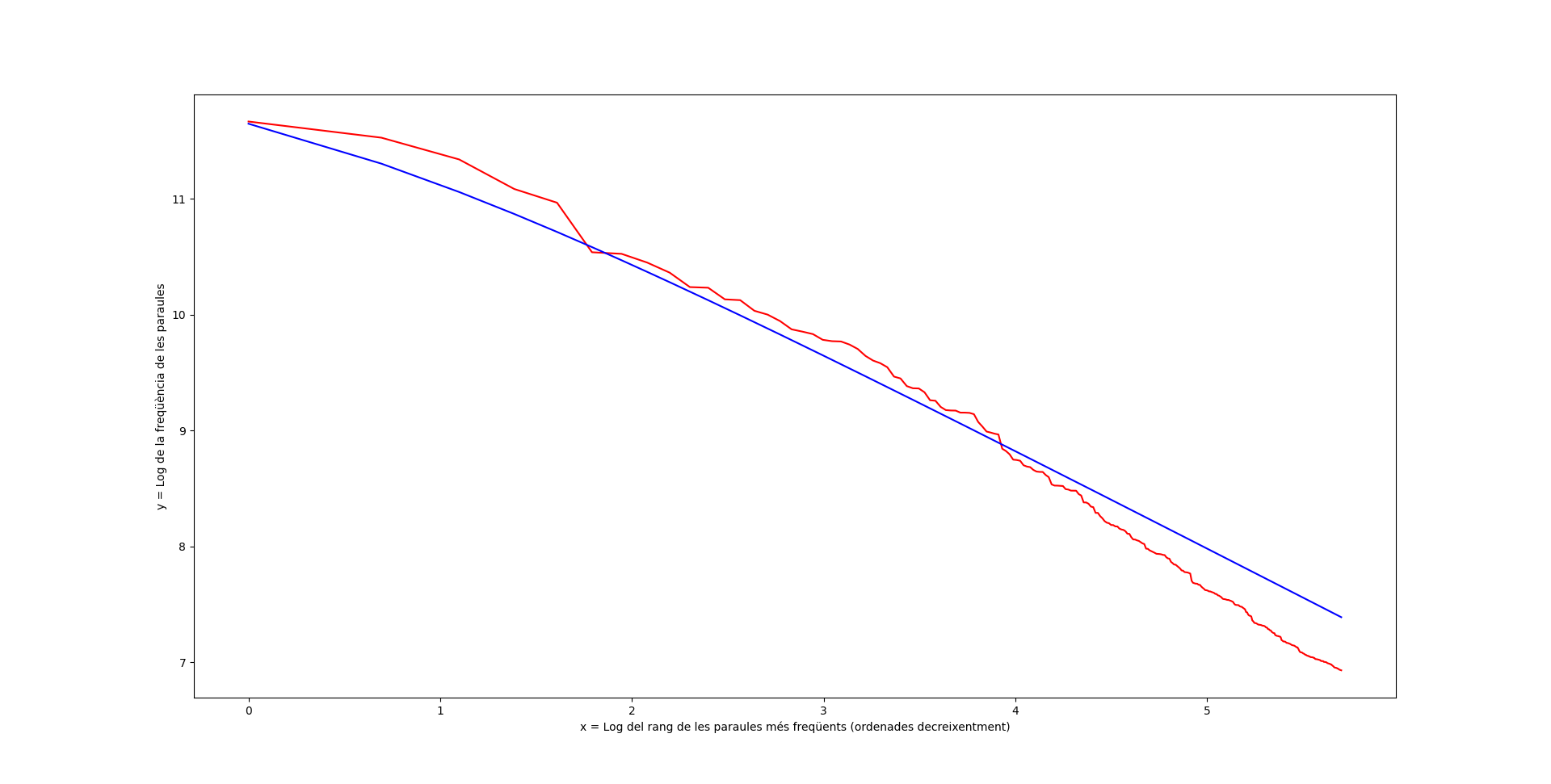
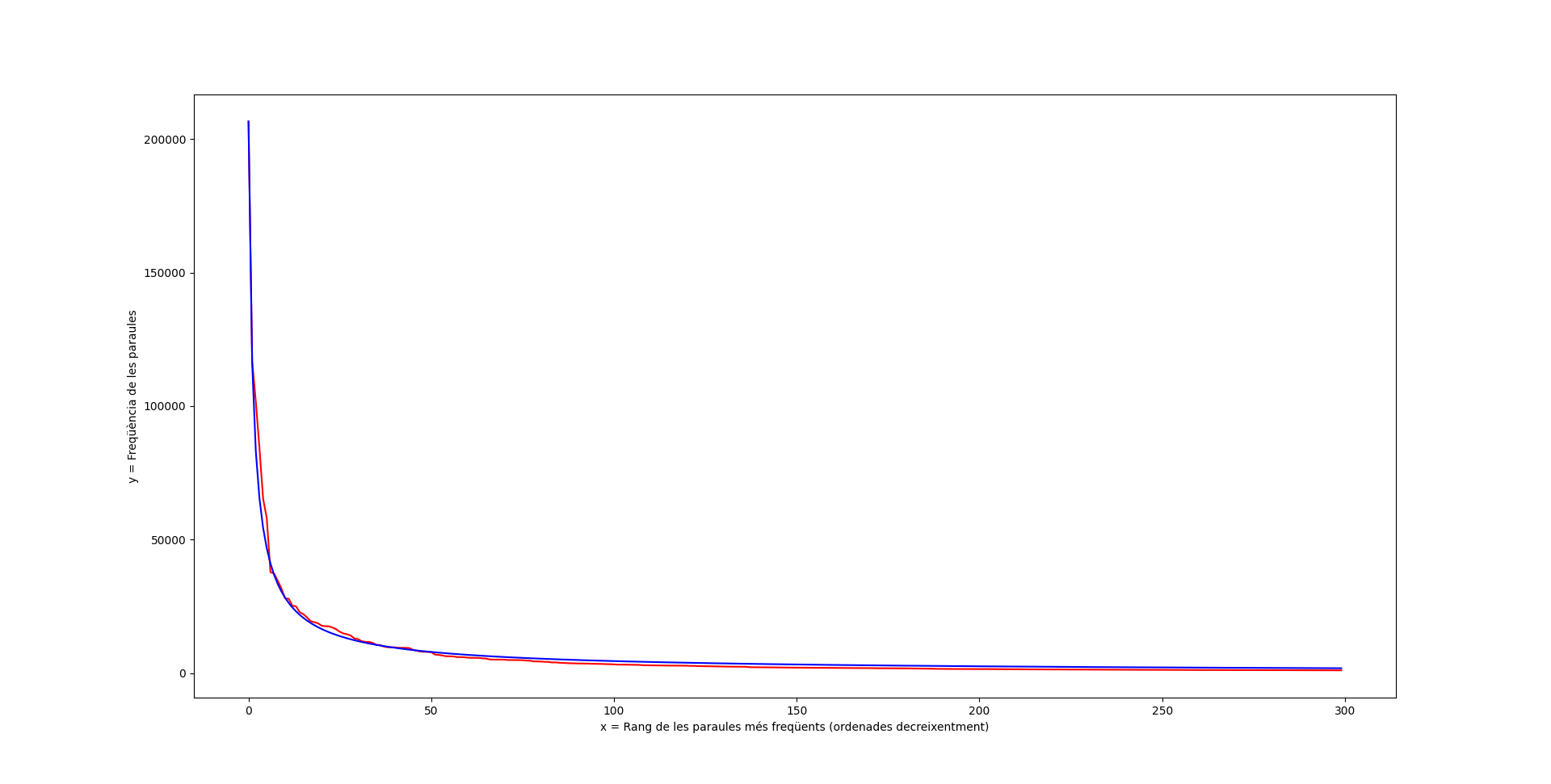
*Figura 2: Relació del rang amb la freqüència de les 300 paraules més freqüents*

# Llei de Zipf

L’objectiu d’aquest apartat és descobrir si la relació entre el rang i la freqüència de les paraules dels textos estudiats segueixen la funció de Zipf, definida per:

On rank = Rang de les paraules estudiades. I a, b i c són paràmetres lliures que hem d’assignar amb el fi de que s’assembli al màxim als valors reals.

Primerament hem calculat la relació de rang i freqüència reals entre les 300 paraules més freqüents, ja que si agafavem les del mig, la freqüència es mostrava amb molt poca variació (representada com una línia horitzontal). La gràfica resultant:



*Figura 3: Relació de rang-freq i llei de Zipf* *Figura 4: Relació de log(rang-freq) i log(llei de Zipf)*

Un cop obtinguda la gràfica amb els paràmetres reals hem configurat les variables a, c i b de la funció de Zipf per a que aquesta sigui el màxim semblant possible a la relació real. Per fer-ho de la millor manera es fa servir la funció curve\_fit de la llibreria scipy, però ens ha donat molts problemes i finalment ho hem fet de forma manual. Els valors obtinguts són:

**a = 0.85, b = 0, c = 206546 (freq. màx)**

I com es pot veure a la Figura 3 més amunt, els resultats son qüasi idèntics. Per tant es pot concloure que la relació de rang i freqüència segueix la llei de Zipf.

Tot i així per a acabar de confirmar la conclusió hem provat si en el cas que estudiem el logaritme d’ambdues aquesta similitud es segueix donant. Com es pot comprovar a la gràfica de la Figura 4 (més amunt), donem per bona la conclusió anterior.

# Llei de Heaps

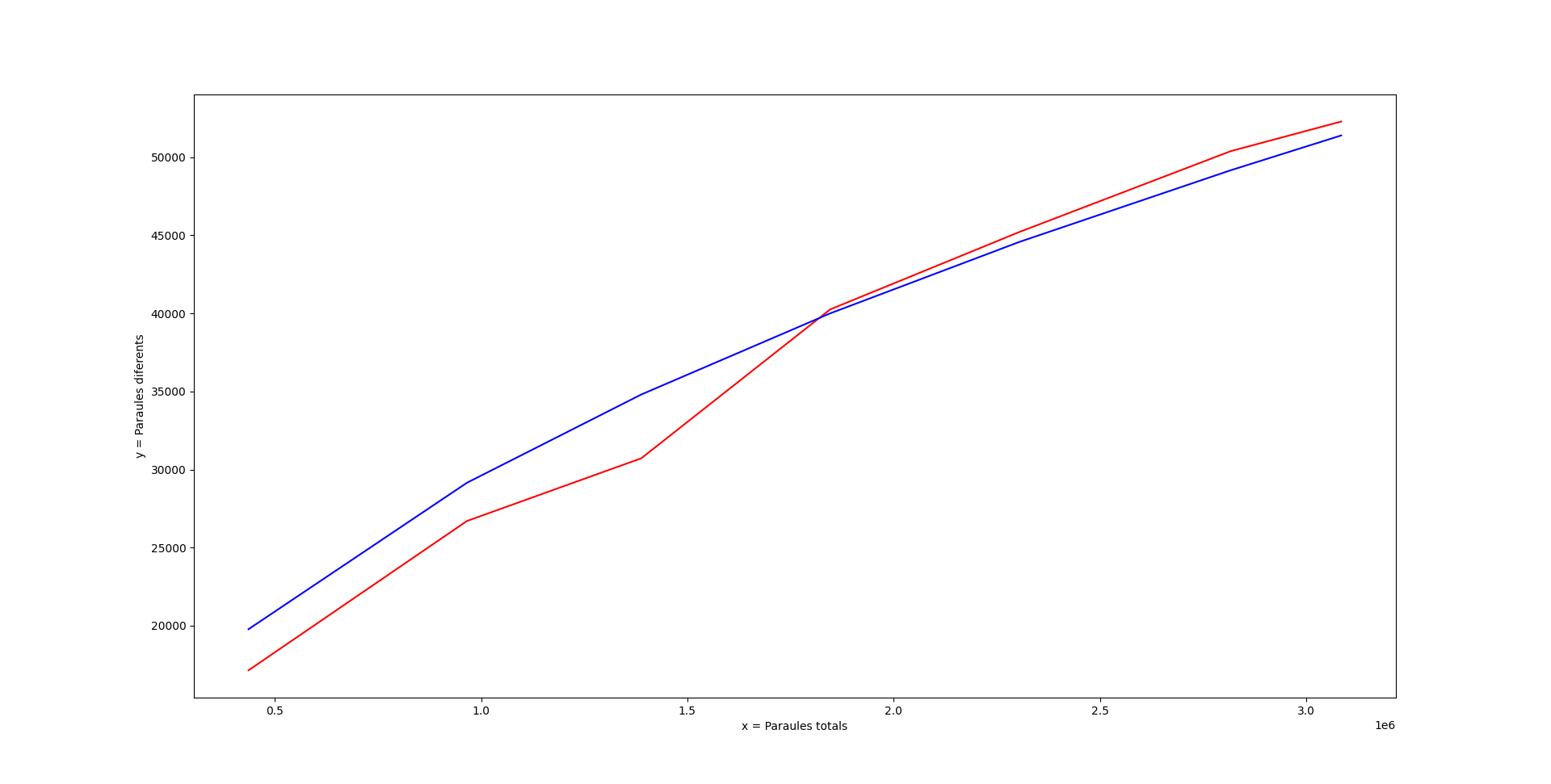
La llei de Heaps descriu el nombre diferent de paraules en un document. La seva fórmula és la següent:

on d representa el nombre de paraules diferents, N representa el nombre total de paraules del document i, K i β són paràmetres lliures que hem d’escollir per aproximar al gràfic real, i demostrar que les nostres dades segueixen la llei de Heaps.

En primer lloc, vam decidir reagrupar l’índex de *novels* en diferents grups de documents amb mides diferents. Així, vam crear 7 nous indexos (n0..n6), començant amb una mida aproximada de 2,5 MB al n0, i augmentant progressivament fins arribar al n6, que conté tots els documents de la col·lecció. A més, vam utilitzar el mateix mètode de depuració de paraules descrit al principi del document, per quedar-nos amb un output més net i lliure de termes estranys.

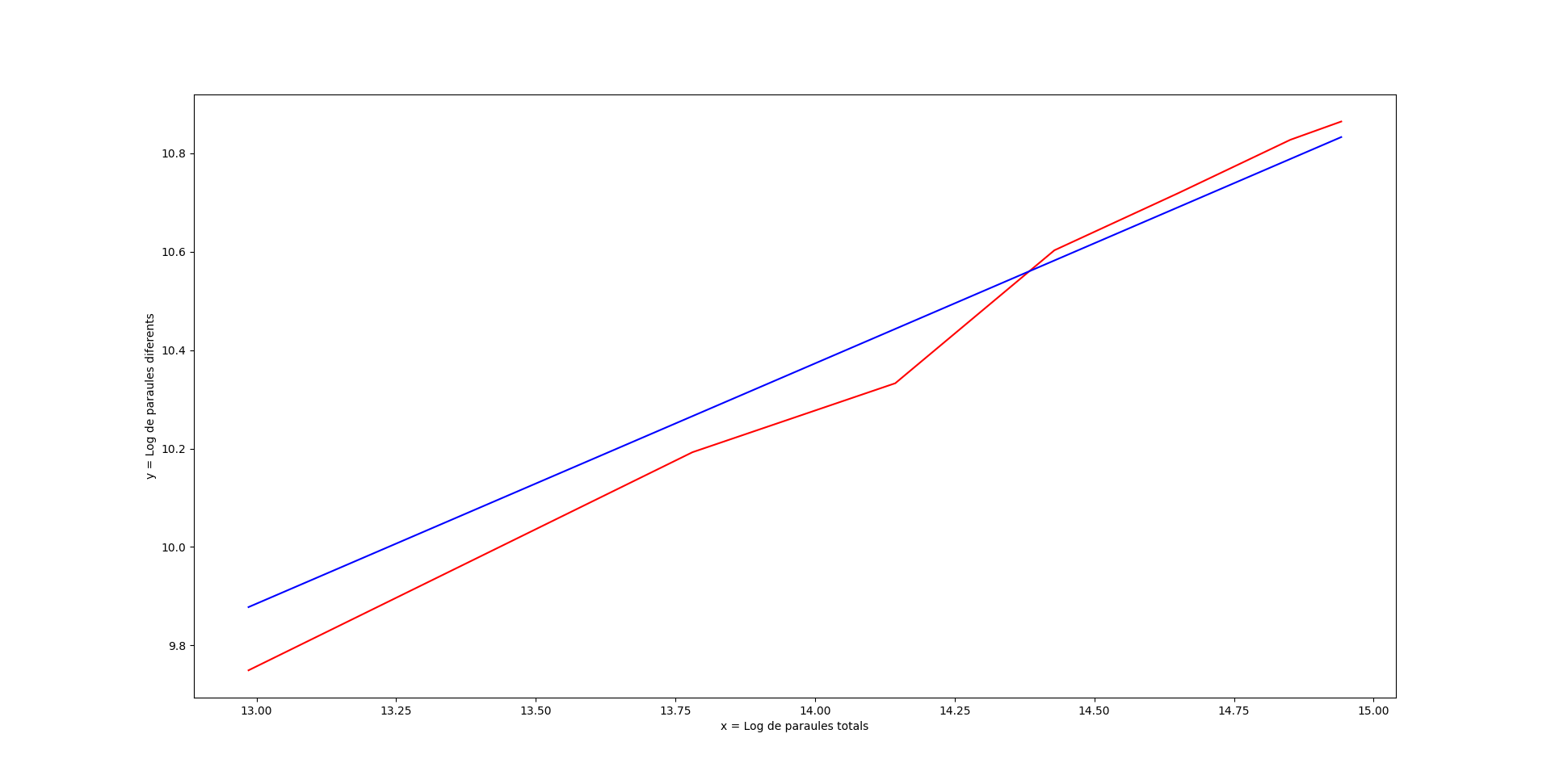
La sortida gràfica mostra un creixement lineal, amb una mica de deformació al mig, però en general podem veure que segueix aquesta tendència. Per aproximar els valors dels paràmetres lliures de la funció de Heaps vam provar valors entre 10-100 per K i entre 0,4 i 0,6 per β, que corresponen als valors típics d’un text en anglès. Fent diferentes proves, vam trobar que els valors pels que millor s’aproximava a la gràfica del output real eren **K = 35** i **β = 0,488:**

Per assegurar-nos que els documents satisfan la llei de Heaps vam decidir realitzar també el gràfic amb els logaritmes, el qual semba confirmar que les dades segueixen aquest principi, tot i que al principi de la recta les línies es troben una mica separades. Amb això concluïm que la col·lecció *novels* també satisfà la llei de Heaps.



*Figura 5: Comparació gràfica relació Paraules totals i paraules diferents i la sortida de la*

*llei de Heaps*

**

*Figura 6: Comparació gràfica relació Log(paraules totals) i Log(paraules diferents) i la sortida del Log(llei de Heaps)*