CAIM Lab, Session 1: ElasticSearch and Zipf's and Heaps' laws

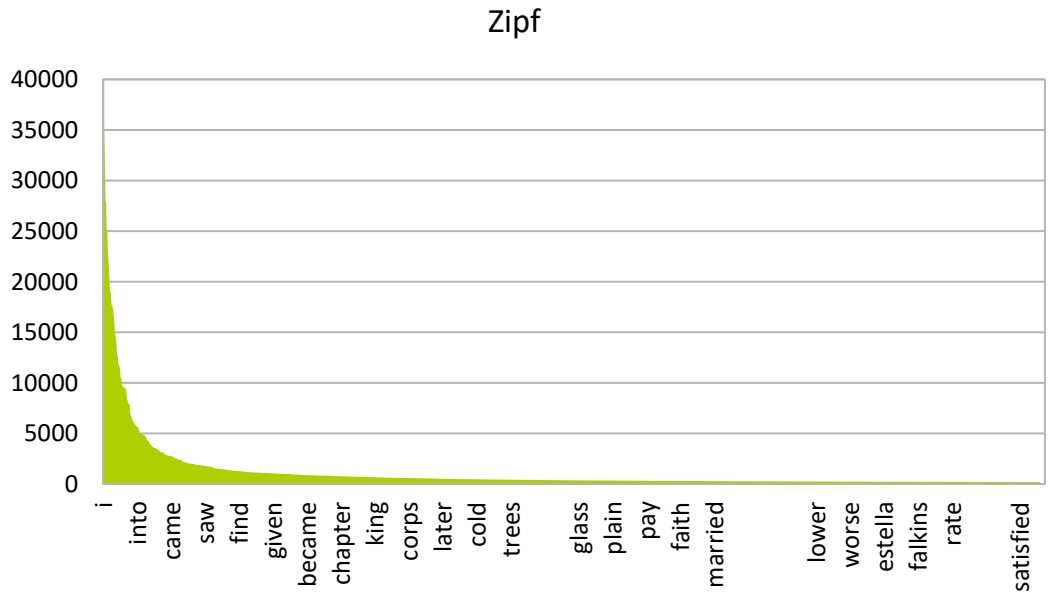
Fet per: Enrique Reyes Illescas i Ferran Noguera Vall.
Data: 06/10/2017

Zipf's Test

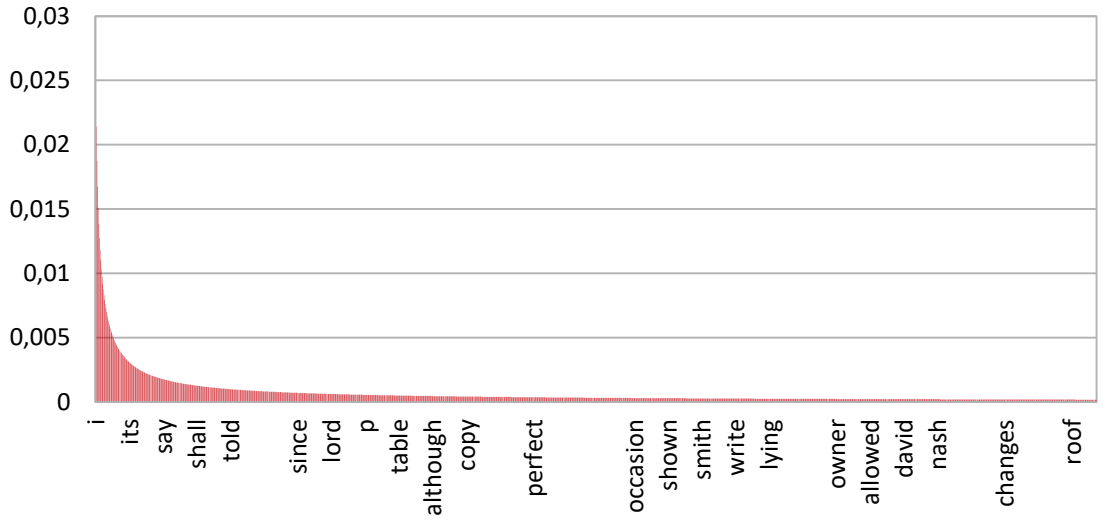
Tots el anàlisis han sigut realitzat al "novels corups", tal com ens recomana l'enunciat, degut a que les dades són molt més clares.

Ens ha sortit un total de 61823 paraules, però la gran majoria no era rellevant per l'anàlisi per tant hem aplicat un filtre per netejar el conjunt de paraules residuals. Aquest filtre ha consistit en eliminar totes les paraules que no tinguessin un mínim de 200 aparicions i les que en tinguessin un nombre desmesurat respecte el general, després d'aplicar-lo ens ha quedat un conjunt de 1553 paraules rellevants.

Seguidament hem col·locat aquestes 1553 paraules en un full d'excel i n'hem fet una gràfica del nombre d'ocurrències respecte la paraula en qüestió, el gràfic resultant ha sigut aquest:



Després hem calculat la freqüència de la paraula, com el número d'ocurrencies d'aquesta entre el número d'ocurrències totals, i la Power Law (amb una a, b i c random) amb la fórmula proporcionada en l'enunciat. Per tal d'obtenir una a, b i c òptimes havíem d'aconseguir que la diferència entre la freqüència i la Power Law al quadrat fos el més pròxima a zero possible, per tant hem anat aproximant-los per aconseguir-ho (aproximant primer c, llavors b i finalment a, ja que c era la que tenia més impacte en el resultat final i a la que menys). El resultat final ha sigut que la a, b i c més òptimes (amb una diferència de 0.00013 total) són: 0.86, 4 i 0.1, respectivament. Hem fet un gràfic de la Power Law respecte les paraules:



Com es pot comprovar és casi idèntic al d'ocurrències respecte paraules, que ja és el que buscàvem. Tots els càlculs es podran trobar en l'excel zipf.ods adjuntat juntament amb la documentació.