**Explain in detail about the contribution of zipf’s law & luhn's law for IR Text operation?**

Zipf’s Law is a statistical distribution in certain data sets, such as words in a linguistic corpus, in which the frequencies of certain words are inversely proportional to their ranks. Named for linguist George Kingsley Zipf, who around 1935 was the first to draw attention to this phenomenon, the law examines the frequency of words in natural language and how the most common word occurs twice as often as the second most frequent word, three times as often as the subsequent word and so on until the least frequent word. The word in the position n appears 1/n times as often as the most frequent one.

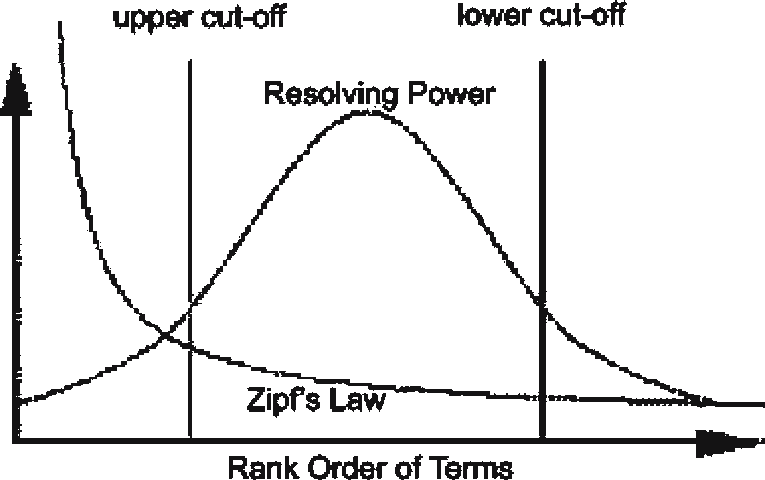
Zipf's Law is a statistical distribution in certain data sets, such as words in a linguistic corpus, in which the frequencies of certain words are inversely proportional to their ranks

The Luhn Algorithm—also known as the “Modulus 10 Algorithm”—is a formula that is used to determine whether the identification number provided by a user is accurate. The formula is widely used in validating credit card numbers, as well as other number sequences such as government Social Security Numbers (SSNs).

This section presents background material on the vector space model. Term weighting schemes are introduced and Luhn’s (1958) and Zipf’s (1949) contributions to IR are summarised. The vector space model is one of the most widely known and studied IR models. The classic vector space model represents each document in the collection as a vector of terms with weights associated with each term. The weight of each term is based on the frequency of the term in the document and collection. The query (user need) is also modelled as a vector. A matching function is used to compare each document vector to the query vector. Once the documents are compared, they are sorted into a ranked list and returned to the user. The  
tf-idf family of weighting schemes (Salton and Buckley, 1988) is one of the most widely used weighting schemes for the vector space model. The term-frequency (tf) is a document specific local measure and is calculated as follows:

|  |
| --- |
| *t f* =rt f  *max f req* |

Zipf (1949) recognized that the frequency of terms in a collection when placed in rank order approximately follows a log curve. The Zipfian distribution for terms in a collection states that the product of the rank order of terms in a collection and their frequencies is approximately constant. Luhn (1958) further proposed that terms that occur too frequently in a collection have little power to distinguish between documents and that terms that appear infrequently are also of little use in distinguishing between documents. Luhn used the Zipfian characteristics to devise 2 cut-off points for determining terms with a high resolving power. Terms that appear outside these points are considered as having a low distinguishing (resolving) power. Thus in Figure 1 (Schultz, 1968), the bell-shaped curve in the graph relates the frequency of terms to their resolving power. Salton and Yang (1973) validate much of Luhn’s and Zipf’s work with empirical analysis. The standard tf-idf weighting scheme stems from these ideas. As previously discussed the tfpart of the scheme identifies terms that appear more frequently as more important within a document (i.e. on a local level). This curve would follow the Zipfian curve in a local context. The idf part of this scheme weights the high frequency terms lower on a global scale. Thus, the curve of the idf measure follows an inverse of the Zipfian curve in Figure 1. Yu and Salton (1976) suggest that the best distinguishing terms are terms that occur with a high frequency in certain documents but whose overall frequency across a collection is low (low document frequency). They conclude from this that term weighting should vary directly with term frequency and inversely with document frequency. However, these weighting schemes

**Fig. 1** Zipf’s law and Luhn’s  
 

A Zipfian distribution of words is universal in natural language: It can be found in the speech of children less than 32 months old as well as in the specialized vocabulary of university textbooks. Studies show that this phenomenon also applies in nearly every language.

Individually, neither syntax nor semantics is sufficient to induce a Zipfian distribution on its own. However, syntax and semantics work together for a Zipfian distribution.

Only recently has Zipf’s Law been tested rigorously on databases large enough to ensure statistical validity. Researchers at the Centre de Recerca Matematica, part of the Government of Catalonia's CERCA network, who are attached to the Universitat Autonoma de Barcelona Department of Mathematics, analyzed the full collection of English-language texts in the Project Gutenberg, a free database with more than 30,000 works. When the rarest words were left out, Zipf’s Law applied to more than half of the words.

The law can be applied to fields other than literature. Zipfian distributions have been found in the population ranks of cities in various countries, corporation sizes, income rankings and ranks of the number of people watching the same TV channel.**luhn's law**