28.8 Why do we use 'log' in the IDF?

The Inverse Document Frequency of a word 'w_i' across the document corpus 'D_c' is given as

$$IDF(w_i, D_c) = \log(N/n_i)$$

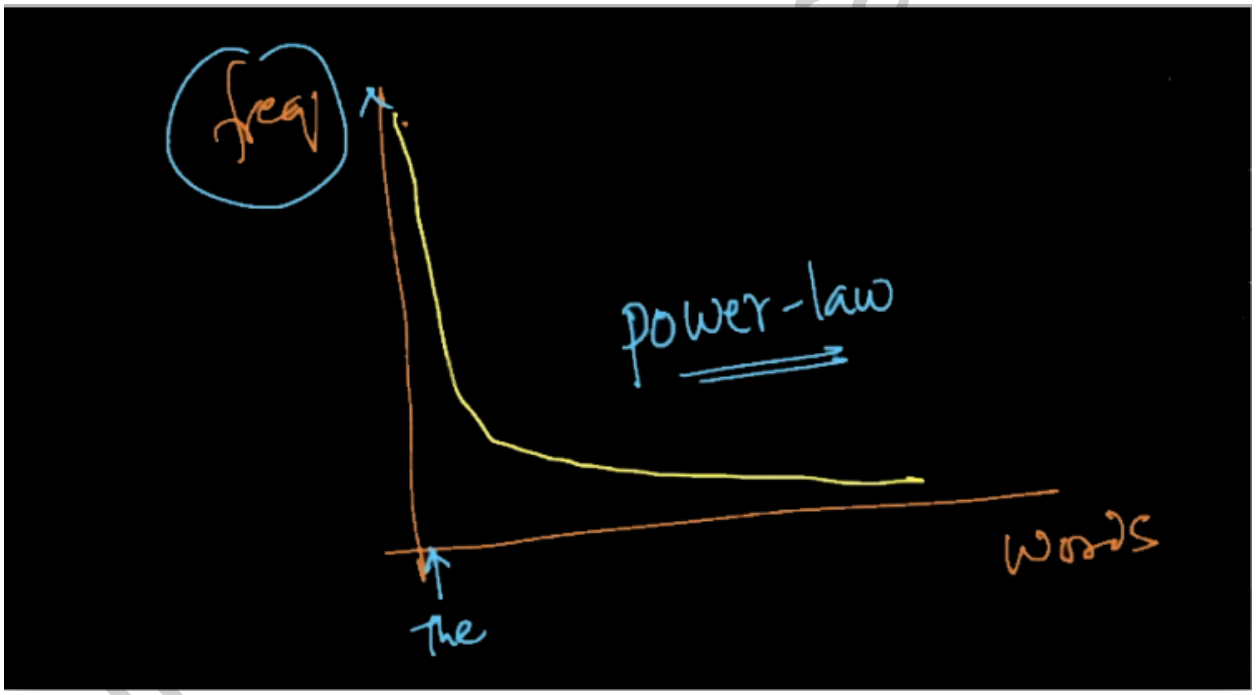
Where N → Total number of documents in the corpus

n_i → Total number of documents containing the word 'w_i'

In order to know why we use logarithm in the IDF, we shall go through Zipf's law first.

Zipf's Law

As mentioned in the video, starting from the timestamp 6:35, if we have all the words on the 'X' axis and the frequency of occurrence of each word on 'Y' axis, then if we plot a histogram, the it looks like below

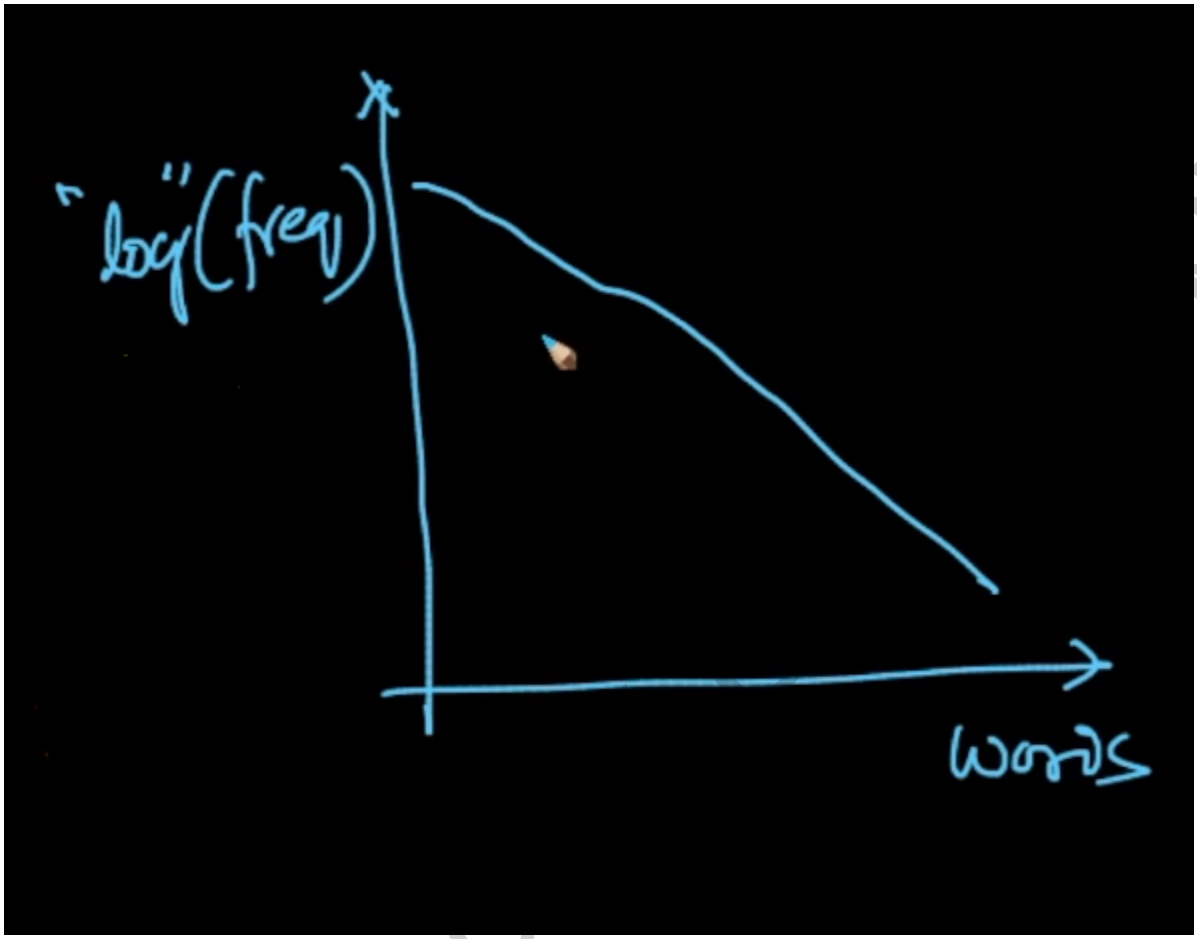


In the above plot, the most frequently occurring words are present towards the origin and the rare words are present away from the origin on the 'X' axis. The curve is in the decreasing order of the frequency. This is an example of Power Law.

If we have a random variable 'X' which follows Power Law, then we can convert it into a gaussian distribution by applying a box-cox transform. Also from the definition of Power Law, we also know that if random variables 'X' and 'Y' follow the power law, then the plot of log(X) vs log(Y) will be a straight line. Similarly, even if one feature(ie.,

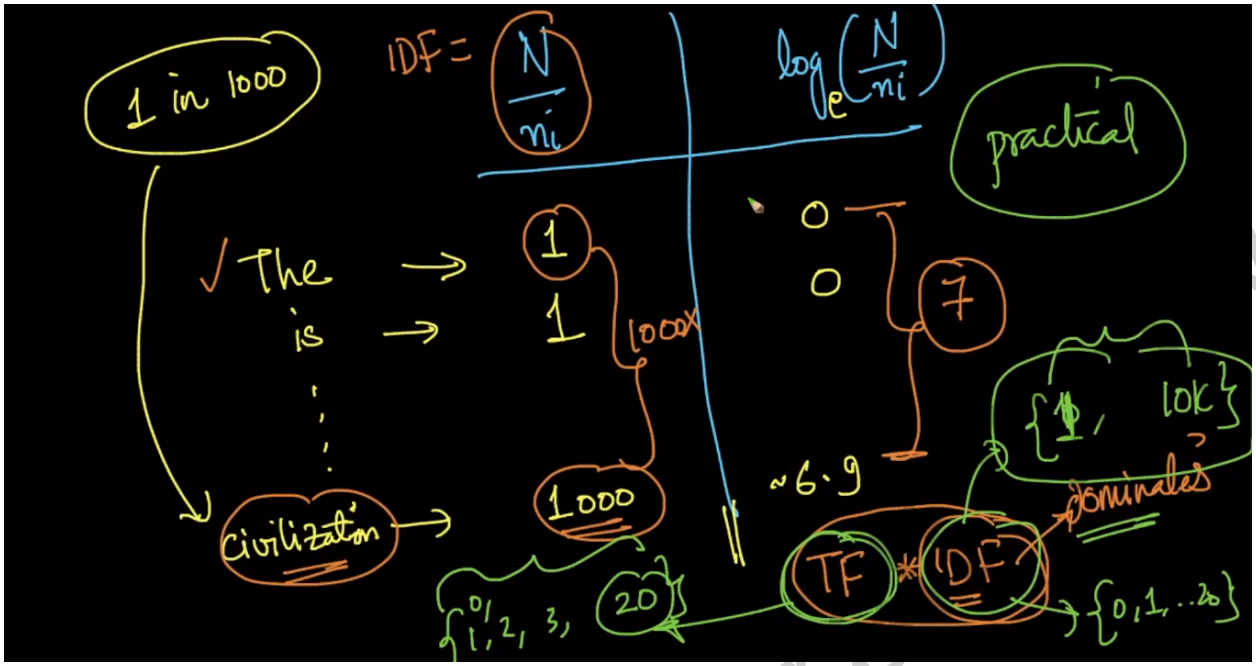
Appl... Draft Copy)

frequency) follows power law while the other feature(words) is discrete, then if we apply logarithm to the frequency, the curve gets transformed into a line as shown below.



Let us now apply logarithm to the values of (N/n_i) and check how the transformed values look like, as discussed in the video starting from the timestamp 8:00

AppliedRo



The range of (N/n_i) is 1 to 10000, whereas the range of $\log(N/n_i)$ is 1 to 7. So if we take (N/n_i), the range is very high, the IDF will dominate TF and result in a huge value. Hence we choose $\log(N/n_i)$ to get reasonable values. We apply logarithm to reduce the scale.

AppliedRoots (Dr