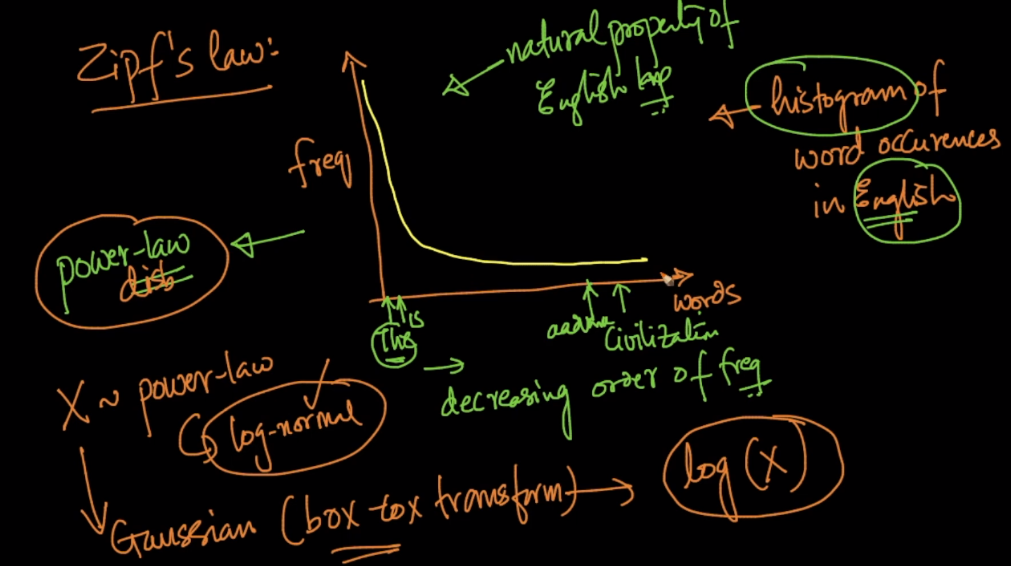


We can think of why log is used using zipf’s law which states that, there are few words in English which occurs a lot in any sentence (like: This, is), and a large population of words which occurs rarely like civilization.

And if we plot this we will get log-normal distribution or power law, and we can convert this in Gaussian distribution using box-cox transform.



Let’s understand using example.

For ‘The’, ni would be larger (approximately equal to N), therefore it’s division would be 1, and therefor log of division would be 0. Similarly for ‘is’.

Now for civilization ni occurs very rarely (hardly 1 in 1000) therefore its division would be 1000 and it’s log will be 6.9

As we can see division is much larger than log, and if use directly division while calculating TF-IDF, then IDF will be more dominant as it’s value is 1000 in case of civilization. And if we take log the IDF will be less dominant as it’s value is 6.9

