Simple is good

This is just a motivational post, if you want to skip all my gibberish and head straight to material click here.

When I first started learning data science and machine learning, I had a few goals in my mind:

1. Learn all these fancy machine learning techniques
2. Win big bucks in Kaggle competitions by using the above learned techniques

Naturally, with such goals in mind I sought out the fastest way to accomplish all of this. The hacker inside me knew that there was a shortcut to it all and reading up on basic techniques was not worth as they were never used in a real world scenario, or so I thought. After reading up a lot of blogs on the cutting edge models used by Kaggle competition winner, I decided to enter a competition and try out my new learned skill :D. So after creating a lot of parametric and non-parametric (I will talk about them in a later blog) models and ensembling them (yup first attempt at my first competition I was ensembling models), I decided to upload my solution. Before uploading I saw that scores of quite a few people were less than even the sample submission provided by the competition master and I thought to myself how bad can you suck at this that the sample is better than your solution. So, laughing at these people I uploaded my solution and after seeing the results my laughter changed into bewilderment, I was at the bottom of the leaderboard, lower than any of the people I was just laughing at. My brain went into complete denial, this can’t be true I applied at methods used by competition winners, how can I be at the bottom? There must have been a mistake, I must have uploaded the wrong file. So, after re-uploading the files and trying to change the parameters in the models used, it slowly dawned on me that my solution sucked. So I started looking at the discussion forum and kernels that other people had created to try an understand what went wrong. I found a kernel in which a guy discussed his approach and his approach was fairly simple, he used only one of the techniques used by me and yet his solution was a hundred times better than mine. So what was the difference, he applied a small transformation to his data to handle highly skewed variables and that was it. This one simple thing made his solution perform far better than mine.

It was then that I realized that fancy shiny methods are nice but they really cannot work if the dataset does not suit them. And the best to understand a dataset as to which methods can be applied is to try out the simple things first

tldr: Tried fancy stuff without understanding the basics.

Now let’s see how one of the easiest method Term frequency can reveal interesting information a dataset

Term Frequency

Term frequency as the name suggest represents the frequency of a term in a corpus of documents. This term is related to the information retrieval domain where it was kind of used to rank documents based on relevance to the search query by search engines. It is such a simple concept that even when it was used by search engines it was not used alone, it had a partner IDF – inverse document frequency. More on these two together later.