Simple is good

‘Simple is good’ is a term that one will get to hear a lot while learning data analysis. Each and every tutorial videos or blogs does mention it at-least once. As a newcomer it might seem weird, as you would like to try that shiny analysis method that all the top Kagglers uses. You don’t wanna be lagging behind your competition. But starting out with simple methods does help as they are very intuitive and many of the complex methods can be seen as an extension to these simple methods. In this post I will use one of the most basic analysis tool to make some rather interesting observations about the movies on IMDB to showcase the power of simple

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.