**Sentiment Analysis on IMDB Movie reviews**

**Challenges:**

1. Data extracted from internet usually contains a lot of html entities like <br>, <p>, <h1> etc. which gets embedded in the original data. We have to remove them from the data by using python modules like Beautiful soup.
2. In every sentence in data. there are consists of stop words like of, is, the etc. we have to get rid of that stop words from data. Python module Nltk is used for text pre-processing. One of the major forms of pre-processing is to filter out useless data. Nltk module has predefined stop words. by using them we can remove stop words from our data.
3. In every sentence in data. there are consists of punctuations. we have to get rid of that punctuations from data. Python module re (Regular Expressions) is used for removing those punctuations by giving some patterns that matches respective punctuation.
4. Sometimes extracted data consists of some symbols, we have to use regular expressions techniques to get rid of that symbols. But those removing that symbols, we may loss information. Honestly, we can do anything.
5. Sometimes we may get ambiguous and sarcastic review that may not going to help.
6. After converting this huge amount of data to DTM (document term matrix) by using vectorization. There will increase of columns that makes support vector machine algorithm computationally cost and time consuming.