# INTRODUCTION:

In this project we will study the basics of sentiment analysis with Natural Language Processing techniques using both spark and gensim. We will try to build a predictive model for a real-world dataset and provide results and recommendations. We begin by looking at the basics of

## Sentiment Analysis:

Sentiment Analysis aims to determine the attitude of the speaker or a writer with respect to some topic using NLP. A basic task in this analysis is classifying the polarity of a given text at the document, sentence or feature/aspect level - whether the expressed opinion in a document, sentence or an entity feature/aspect is positive or negative.

The rise of social media such a blogs and social networks have fueled interest in sentiment analysis. With the proliferation of reviews, ratings, recommendations etc, there is a great opportunity for businesses to identify new opportunities and manage their reputations.

NLP comprises of

# EXECUTIVE SUMMARY:

## 2.1 PROBLEM STATEMENT:

The given dataset provides the 25000 reviews of movies on IMDB. The sentiment is associated with each review as 0(thumbs down) or 1(thumbs up).

Our approach aims to apply simple TF-IDF as well as Word2Vec models to try and build a system which can predict the polarity of the review as thumbs down(0) or thumbs up(1).

## 2.2 APPROACH:

EDA:

Clean the data for reviews, remove special characters

Plot a chart of words against count

case 1: a very naive word count:

<https://plot.ly/~mubeen_spark/26/>

The words like : the, of, is, I etc show up, which are not really useful

When we start using nltk:

<https://plot.ly/~mubeen_spark/47/>

words: movie, like , even, really, much, people, great, make, think, watch

Plotted the same count for negative reviews:

<https://plot.ly/~mubeen_spark/53/>

words like : bad, never,old, horror

Plotted the same count for positive reviews:

love, funny, right