

# Movie Reviews - Rotten Tomatoes

The dataset is comprised of tab-separated files with phrases from the **Rotten Tomatoes dataset**. The train/test split has been preserved for the purposes of benchmarking, but the sentences have been shuffled from their original order. Each Sentence has been parsed into many phrases by the Stanford parser. Each phrase has a Phraseld. Each sentence has a Sentenceld. Phrases that are repeated (such as short/common words) are only included once in the data.

train.tsv contains the phrases and their associated sentiment labels. We have additionally provided a Sentenceld so that you can track which phrases belong to a single sentence.

test.tsv contains just phrases. You must assign a sentiment label to each phrase.

The sentiment labels are:

0 - negative, 1 - somewhat negative, 2 - neutral, 3 - somewhat positive, 4 - positive

## Task and Approach:

- Like IMDb , Rotten Tomatoes is also one of the most popular website where people can find reviews and ratings for nearly any movie.
- Using TF - IDF we are going Transform phrase in to a normalized tf or tf-idf representation .

TF - How many times a particular word appears in a single doc

IDF - It is calculated by taking the log of {number of docs in your corpus divided by the number of docs in which this term appears}