Movie Search

Objective

Objective of this project is to use text data of movies to calculate inverted term frequency and use that to find similarity between user's search query and movies.

Text Extraction

In the TMDB Movies dataset, there are many text fields such as movie title, tagline, overview, keywords, genres, production companies, cast, crew, languages, etc.

Title, tagline, overview are in string format and can be used directly, but fields like keywords, genres, production companies are in list of dict format so first I have converted string to literal structure then I have extracted only necessary fields such as name.

Other source of information are casts and characters of movies, I can use information about actors and role they played. Similarly from crew information I have only extracted name of director, producer, writer which is good source of information.

In the end I have constructed document for each movies that contains all the text data mentioned above.

For better results I have applied snowball stemmer.

Search Result

To find similarity between search query and movies first I have generated word vector for each movie and using word vectors inverted term frequency was calculated.

Now for new query we need to generate word vector of query and then calculate inverted term frequency. After that we need to calculate cosine similarity between inverted term frequency of query and of each movie. Then we will find top results whose cosine similarity is maximum.