***Machine Learning Nanodegree Capstone project:***

**Problem statement:** Building a collaborative filtering movie recommendation system using Word embeddings, Matrix factorization methods by defining and implementing a Neural network with Keras.

**Datasets:**

The dataset is obtained from grouplens.org: https://grouplens.org/datasets/movielens/20m/ MovieLens itself is a research site run by GroupLens Research group at the University of Minnesota. The first automated recommender system was developed there in 1993.

The datasets describe ratings and free-text tagging activities from MovieLens, a movie recommendation service. It contains 20000263 ratings and 465564 tag applications across 27278 movies. These data were created by 138493 users between January 09, 1995 and March 31, 2015. This dataset was generated on October 17, 2016.

Users were selected at random for inclusion. All selected users had rated at least 20 movies.

**Content:**

No demographic information is included. Each user is represented by an id, and no other information is provided.

The data are contained in six files.

**1. tag.csv** that contains tags applied to movies by users:

 userId

 movieId

 tag

 timestamp

**2. rating.csv** that contains ratings of movies by users:

 userId

 movieId

 rating

 timestamp

**3. movie.csv** that contains movie information:

 movieId

 title

 genres

**4. link.csv** that contains identifiers that can be used to link to other sources:

 movieId

 imdbId

 tmbdId

**5. genome\_scores.csv** that contains movie-tag relevance data:

 movieId

 tagId

 relevance

**6. genome\_tags.csv** that contains tag descriptions:

 tagId

 tag

**Acknowledgements:** To acknowledge use of the dataset in publications, the following paper:

F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. ACM Transactions on Interactive Intelligent Systems (TiiS) 5, 4, Article 19 (December 2015), 19 pages. DOI=http://dx.doi.org/10.1145/2827872

We make use of only **Ratings.csv and movie.csv** files to build a recommendation system.

Please download and place the “rating.csv” file in the same folder the notebook files are present. <https://www.kaggle.com/grouplens/movielens-20m-dataset#rating.csv>,

We want to build a model that takes a user, ui and a movie, mj, and outputs a number from 0.5-5, representing how many stars we think this user would give that movie.

**Software Requirements:**

* Anaconda
* Python 3.5 and above
  + Keras
  + TensorFlow
  + Scikit-learn
  + Genism
  + NumPy
  + Pandas
  + Matplotlib

**Hardware Requirements:**

* Operating System: Windows 10
* GPU: Nvidia GTX 1050(minimum)
* 8 GB RAM, 1TB Hard Drive.

**Instructions to check Notebook files:**

As the dataset is very large please **run** the notebook files in the order mentioned below:

1. **Data Preprocessing.ipynb**: Notebook file consists of all data preprocessing process.
2. **Embeddings.ipynb:** Notebook file consists of “Basic DNN model” definition and implementation. It includes most of the “embeddings” study, its definition, uses etc.,
3. **Matrix\_Factorization.ipynb:** Notebook file consists of Matrix factorization architecture implementation and best model finding and implementing.
4. **Gensim.ipynb:** Notebook file consists of exploration of embeddings using **“Gensim tools”.**
5. **t-SNE.ipynb:** Notebook file consists of all visualizations made using t-SNE