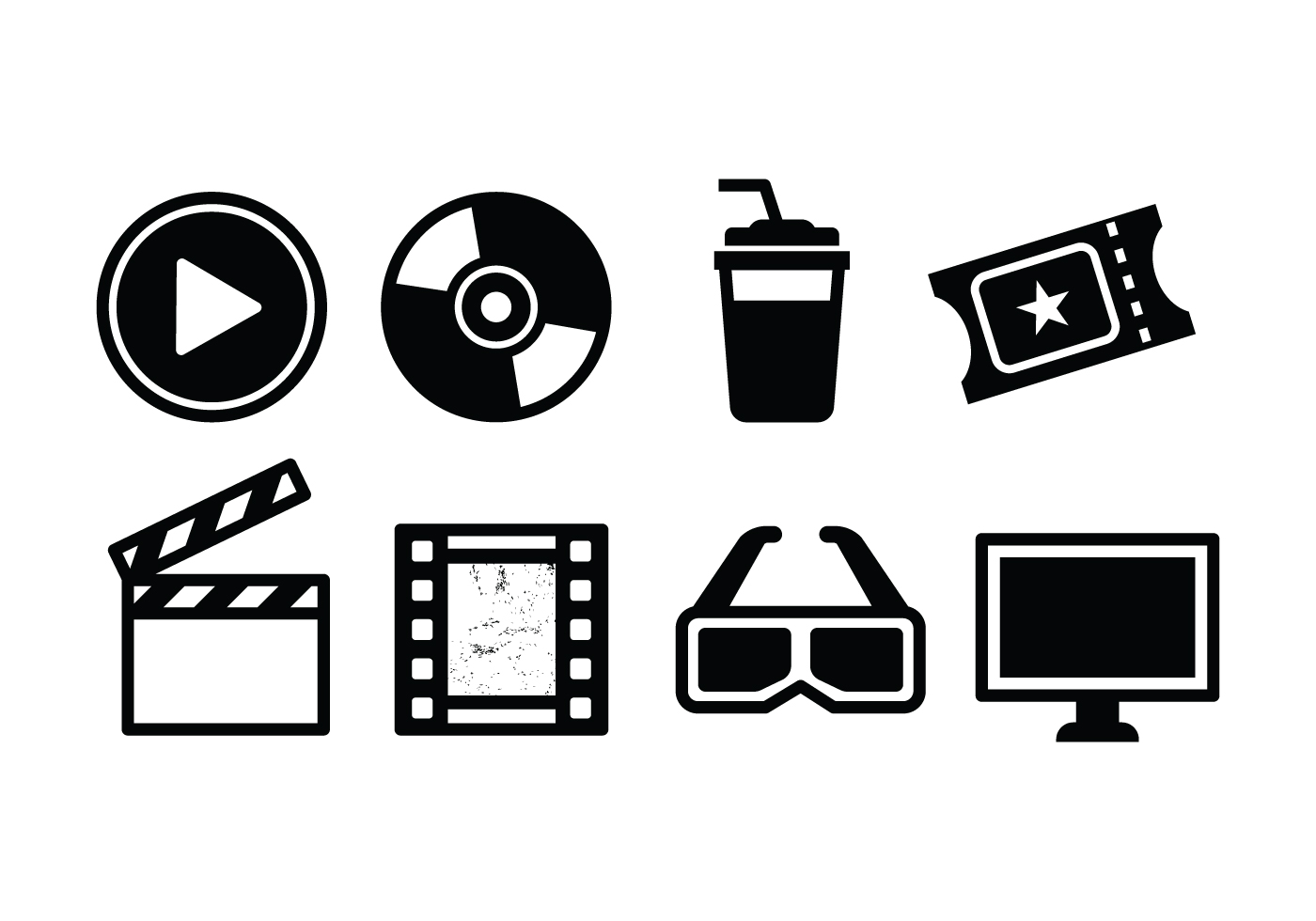
**MOVIE RECOMMENDATION SYSTEM**



**Team 14 :**

1. Surya Ushasri
2. Saurabh Sankhe
3. Viraj Sapre
4. Shravankumar Hiregoudar
5. Madhur Nagaraj

**GUIDELINES:**

The final project folder includes the following:

1. Project Report (Stats & ML Report.docx)
2. ALS Files folder (.csv and .ipynb files)
3. Cosine Similarity Files folder(.scala, .jar & .dat files)
4. KNN files folder(.csv and .ipynb files)

**Use the csv files to load the data in .ipynb files.**

**The instructions to run the scala and jar files are commented in the beginning of the scala file**

**PROBLEM STATEMENT :**

Build a good movie recommendation engine for users based on existing movie ratings using collaborative filtering approach.

**ABOUT THE DATASET:**

The data can be found at  <http://grouplens.org/datasets/movielens/100k/>

The dataset consists of:

* 100,000 ratings (1-5) from 943 users on 1682 movies.
* Each user has rated at least 20 movies.
* Simple demographic info for the users (age, gender, occupation, zip)

There is one more data set consisting of 1 million ratings. We will train our models on this data set also to see if accuracy of the model increases.

The data was collected through the MovieLens web site (movielens.umn.edu) during the seven-month period from September 19th, 1997 through April 22nd, 1998. This data has

been cleaned up - users who had less than 20 ratings or did not have complete demographic information were removed from this data set.

**MODELS :**

1. Alternative Least Squares Method (ALS)

ALS is the most common and effective method in matrix factorization.

1. K- Nearest Neighbor Algorithm

We used k-NN with cosine distance for classification. We can also use k-NN with Euclidean distance for classification.

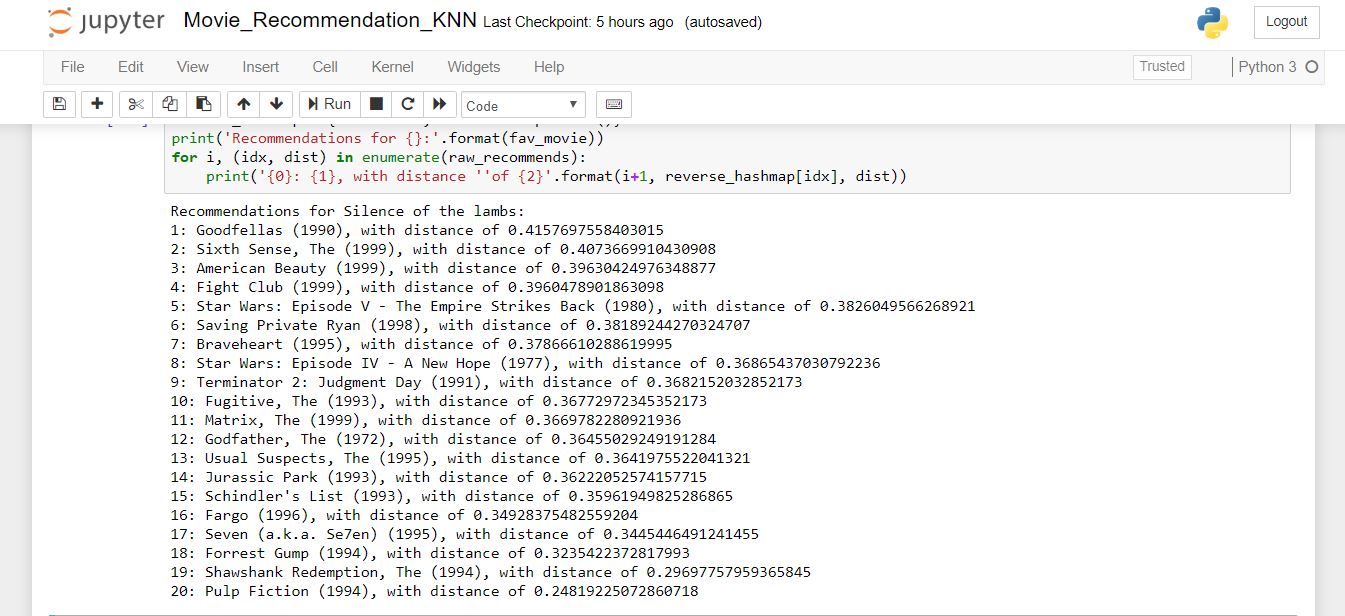
1. Cosine Similarity:

In cosine similarity we manually use the cosine distance formula for classification

This is the parameter by which we calculated the similarity score and number of users that have rated.

**ANALYSIS & RESULTS:**

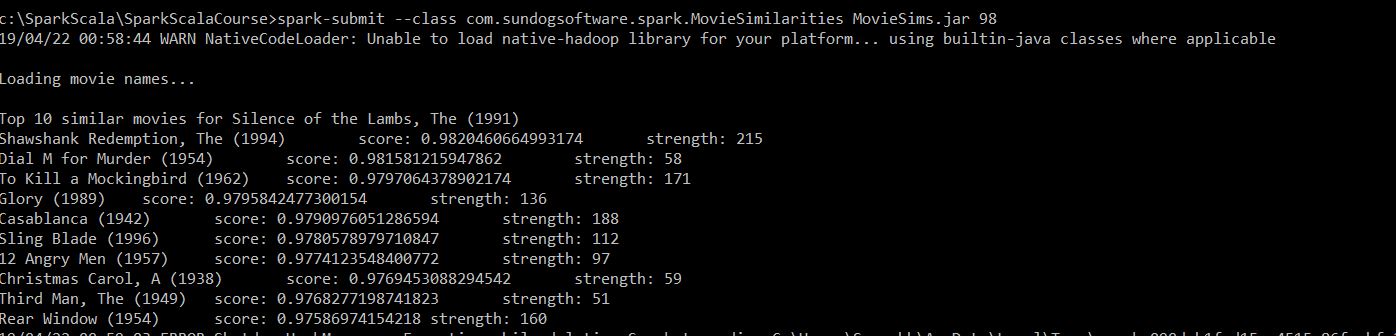
* We saw that the user 6 liked the movies : Sense & Sensibility, Father of the Bride – part 2, Eye for eye and ALS recommended the movies : Police Academy, Lost Tango in Paris, Zero Dark Thirty.
* Thus, ALS recommended similar movies to the user 6 by learning that the user liked Thriller/Drama movies.
* We compared the recommendations from the K-NN with the IMDB recommendations and found 12 recommendations to be same out of the 20 for a movie “The Silence of the Lambs”.







* We used the similar approach for verifying the cosine similarity model with IMDB recommendations and found out that 3 of the 10 recommendations for the movie “The Silence of the Lambs” matched.



**CONCLUSION**

* ALS Model accuracy improved with larger training data

Dataset with 100K observations showed RMSE value 1.14

whereas dataset with 1M observations showed RMSE value 0.904

* **Cold Start Strategy** - In production, for new users or items that have no rating history and on which the model has not been trained. During cross-validation, it is actually very common to encounter users and/or items in the evaluation set that are not in the training set. We solve this by setting cold start strategy as off making sure that those users not included in test set.
* Although ALS performs better than other matrix factorization methods and it’s RMSE is less, better recommendations can be given when we do a hybrid system consisting of content filtering and collaborative filtering both.