Design Document on SVD and CUR Decomposition

Submitted on: - Submitted to: -Dr. Aruna Malapati

21th November, 2016

Submitted by: -

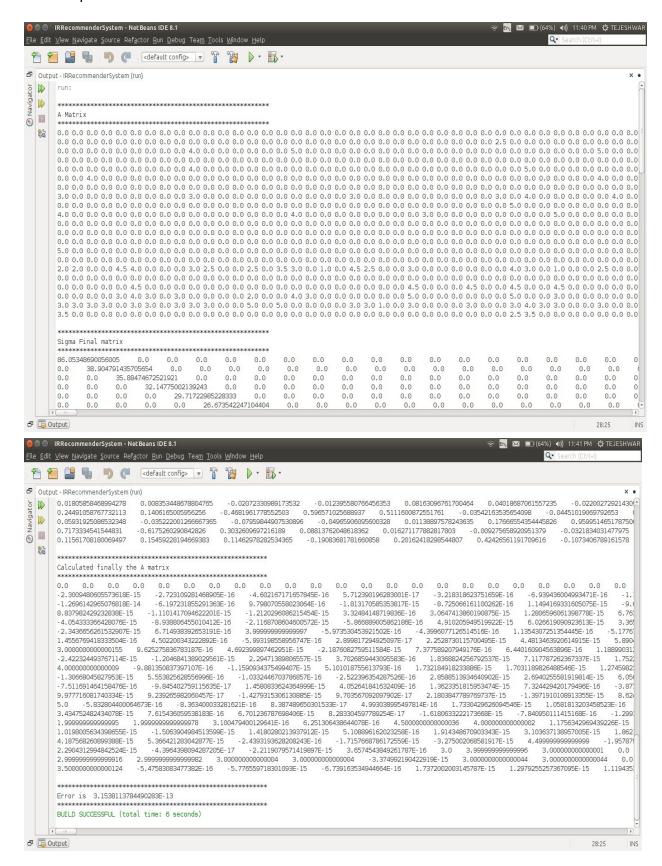
Soamya Agrawal 2014A7PS185H

Ayushi Behl 2014A7PS145H

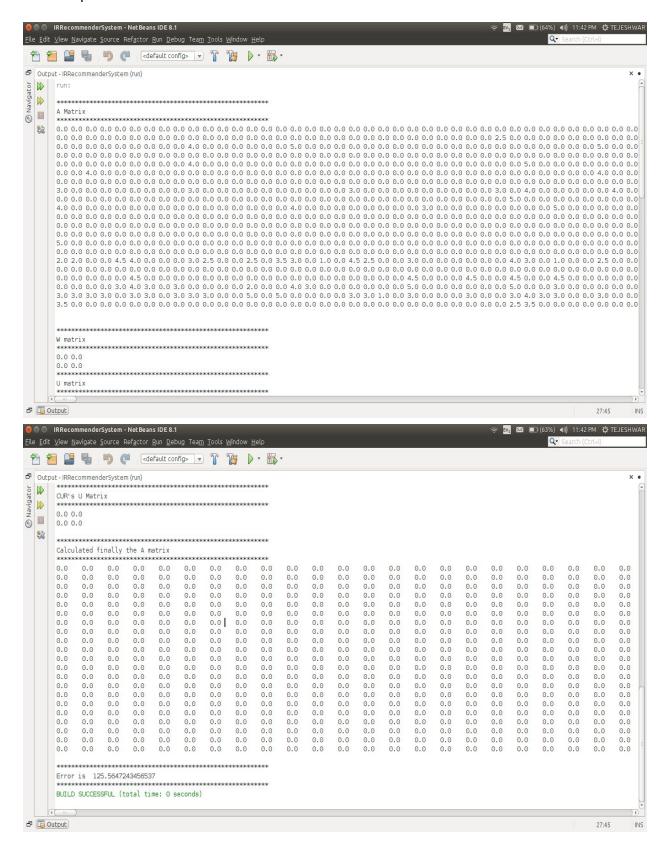
Tejeshwar Reddy 2014A8PS492H

Ruthvi Reddy 2014A7PS040H

SVD Output:-



CUR Output:



About our Assignment:

- Corpus: http://grouplens.org/datasets/movielens/ which includes user movie ratings and the movies are grouped by genre.
- Data Structures: We used HashMap for mapping Movie ID with the column's index of A matrix where A represents user-movie matrix. and Arrays for storing and for computational purposes
- Formulae used of SVD and CUR:
 - SVD->

SVD - Definition



 $\mathbf{A}_{[m \times n]} = \mathbf{U}_{[m \times r]} \, \mathbf{\Sigma}_{[r \times r]} \, (\mathbf{V}_{[n \times r]})^{\mathsf{T}}$

A: Input data matrix

- m x n matrix (e.g., m users, n movies)

U: Left singular vectors

- m x r matrix (m users, r concepts)

Σ: Singular values

- r x r diagonal matrix (strength of each 'concept')

(r: rank of the matrix A)

V: Right singular vectors

n x r matrix (n movies, r concepts)

CUR: How it Works



Sampling columns (similarly for rows):

Total length of all the columns

Input: matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, sample size cOutput: $\mathbf{C}_d \in \mathbb{R}^{m \times c}$

- 1. for x = 1 : n [column distribution] 2. $P(x) = \sum_{i} \mathbf{A}(i, x)^{2} / \sum_{i,j} \mathbf{A}(i, j)^{2}$ 3. for i = 1 : c [sample columns]
- 4. Pick $j \in 1$: n based on distribution P(j)
- 5. Compute $\mathbf{C}_d(:,i) = \mathbf{A}(:,j)/\sqrt{cP(j)}$

Note this is a randomized algorithm, same column can be sampled more than once

• Libraries :

 Jama package- We used Jama Package for all the matrix computations like finding rank, eigen values, eigen vectors etc.

Error Analysis + Time Analysis

For SVD -> The number of columns of A matrix are fixed as the number of distinct movies in the corpus and number of rows to be changing based on the users.

Using functions ,time and error were calculated.

For **CUR** -> The number of rows and number of columns for C and R matrix have been kept same. And for calculating error and time taken different values were passed.

SVD and CUR Error and Time Analysis ->

