Disciplina: Data Mining **Data:** 29/10/2016

Professor: Geraldo Zimbrão da Silva

Aluno: Leniel Macaferi **DRE:** 116005784

Mestrado em Engenharia de Dados e Conhecimento @ COPPE\UFRJ - Universidade Federal do Rio de Janeiro

Relatório
KMeans with Movie Lens 1 MB Dataset

http://grouplens.org/datasets/movielens/

The test was conducted using the small dataset which contains: 100,000 ratings and 1,300 tag applications applied to 9,066 movies by 671 users. Last updated 10/2016.

Data was massaged in Microsoft Excel before serving as input to the KMeans algorithm. The timestamp column was discarded. The columns used were: userId, movieId and rating.

The data format was transformed as follows:

movieId 1	user1 rating	user2 rating	user3 rating	.....
movieId 2	user1 rating	user2 rating	user3 rating	.....
movieId 3	user1 rating	user2 rating	user3 rating	.....

Missing user ratings were filled by the average movie rating.

All the development was done using Visual Studio Code as the IDE and Python programming language with scikit-learn Machine Learning library and NumPy

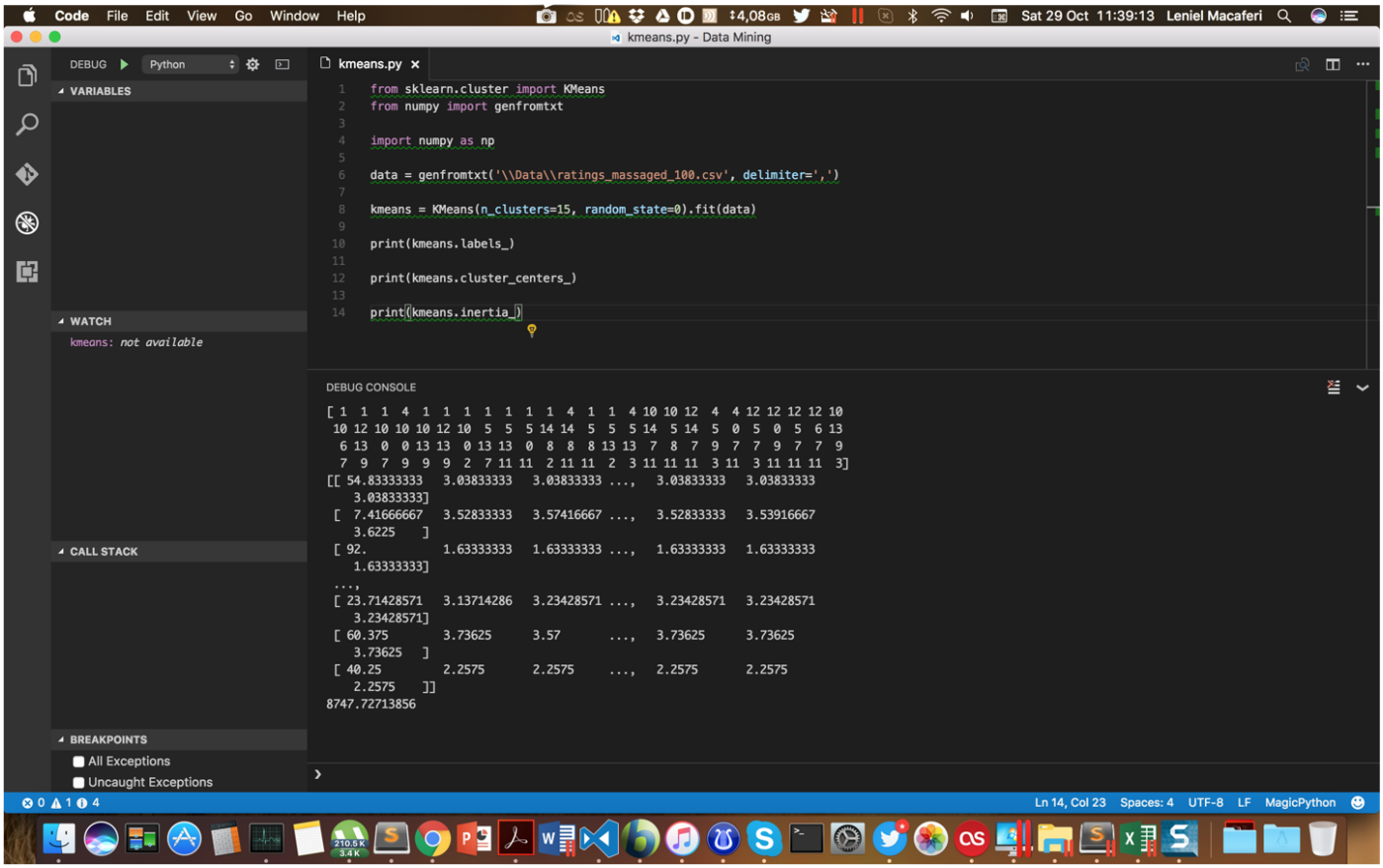
Source code is available @ https://github.com/leniel/DataMining

Useful reading
Clustering http://scikit-learn.org/stable/modules/clustering.html
KMean http://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html
PCA Decomposition
http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

First test - [small sample of 100 movies' ratings] - NO PCA

This 1st test was conducted to better understand how the KMeans clustering algorithm works.

K = 15

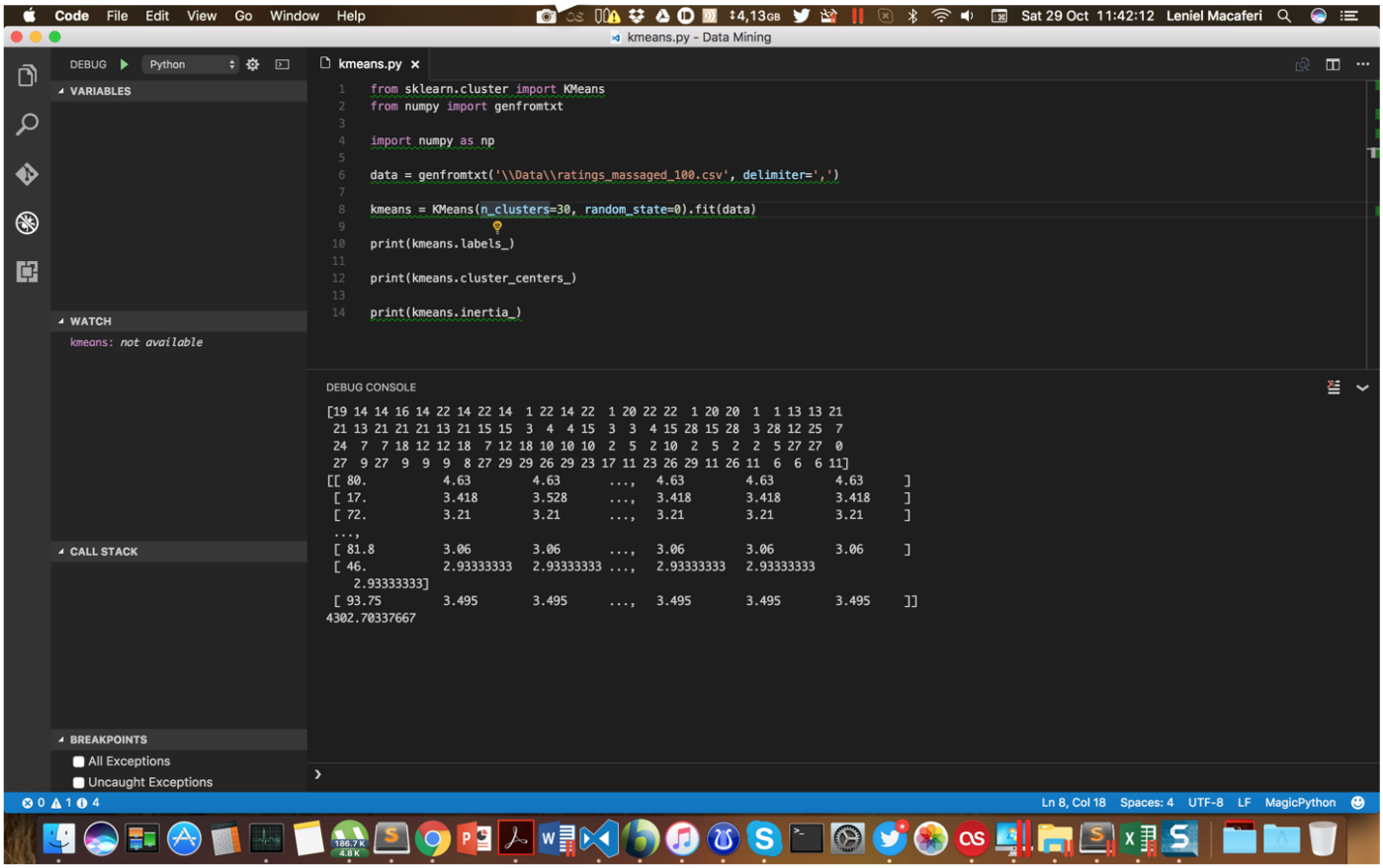


The labels property identifies the clusters, that is, each of the 100 movies is labeled with a given cluster number\identifier.

For K = 15, Inertia (inter-group distance) = 8747.72.

Inertia is not a normalized metric: we just know that lower values are better and zero is optimal. But in very high-dimensional spaces, Euclidean distances tend to become inflated (this is an instance of the so-called “curse of dimensionality”). Running a dimensionality reduction algorithm such as PCA prior to k-means clustering can alleviate this problem and speed up the computations.

K = 30



For K = 30, Inertia (inter-group distance) = 4302.70, that is, half the inertia of K = 15.

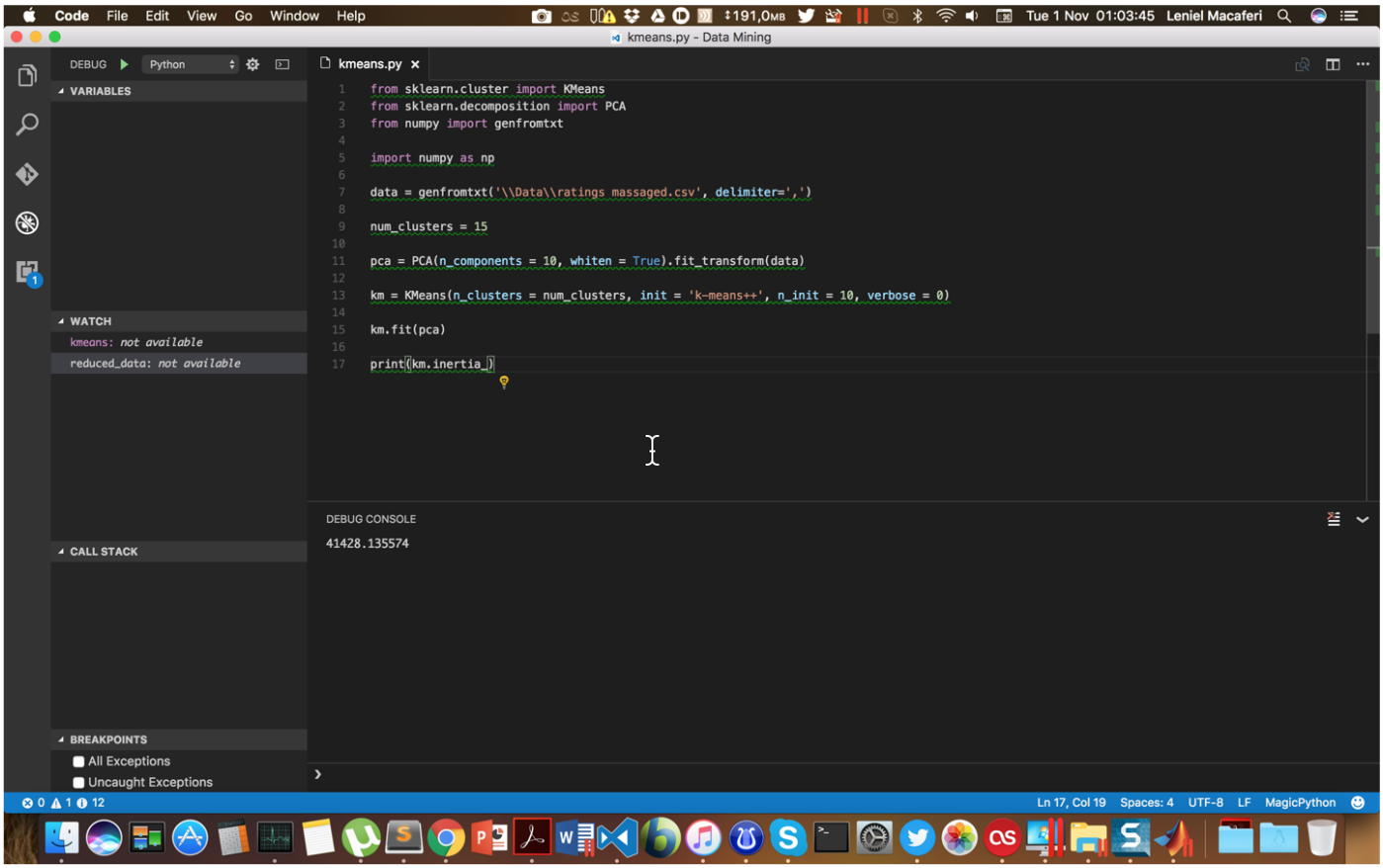
Second test - [full sample of 9066 movies' ratings] - NO PCA

For K = 15, Inertia = 35,906,782,235.4

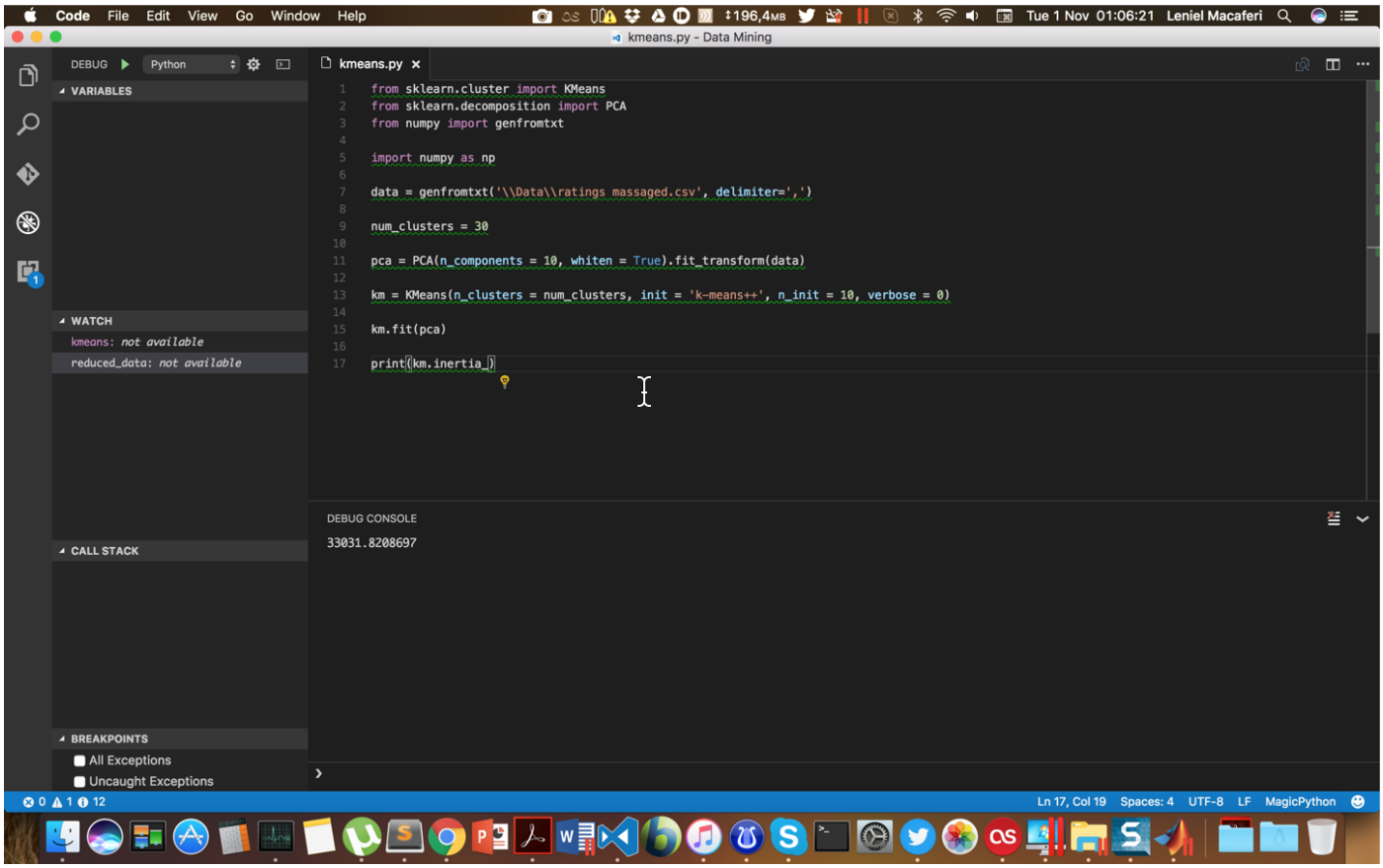
For K = 30, Inertia = 7,642,479,407.69

Inertia was reduced by 5 with a double K.

Third test - [full sample of 9066 movies' ratings] - WITH PCA

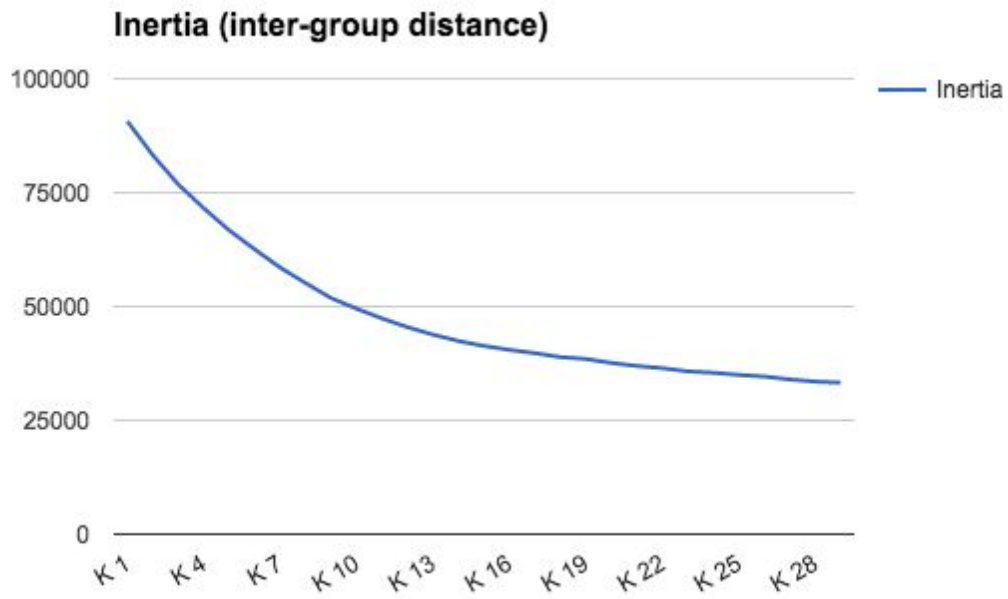


For PCA = 10 dimensions and K = 15, Inertia = 41428.13



For PCA = 10 dimensions and K = 30, Inertia = 33031.82

Graph showing how inertia decreases for greater K



Generated clusters are shown in the file clusters.txt.

Cluster 1 contains a good batch of animation movies.

Cluster 27 has 2 Star Wars movies: Episode V - The Empire Strikes Back (1980)' and Episode VI - Return of the Jedi (1983)