

PROBLEM DESCRIPTION

The meta kaggle dataset is used in our project to build a recommender system for users. The system is aimed at recommending kernels according to the current kernel they users are on. Data scientists, customers or just data enthusiasts who are looking for new kernels or datasets to experiment with or work on need new ideas for data analysis. When they search for a kernel, a system that recommends or suggests certain related kernels would be a very useful system to have. This is what we aim to achieve with this project.

Kaggle is the most popular website to find datasets and associated kernels. The dataset we use here is also from kaggle - about kaggle. Hence, it is the metadata that we are working on. The dataset contains various tables with useful related information. However, what we use for our project is primarily the kernel ID and the tag ID. The project uses these two values to build a clustering system that aims at finding related kernels.

Using the current kernel, we work on tags associated with that kernel and find other kernels which are most likely to be similar to the current kernel.

PROBLEM STATEMENT

“To build a recommender system for kernels based on the meta kaggle dataset. The system is primarily aimed at suggesting new kernel for kaggle users that are related to the current kernel by one or more factors. The kernel being built primarily uses tags for the current kernel and finding the correlation to other kernels with similar tags”

OBJECTIVES

- To build a system that recommends kernel based on current kernel the user is working on
- To find related kernels using tags of the current kernel
- To find a correlation between said kernels with common tags based on multiple factors using Pearson coefficient
- Using KNN to find a threshold value for the recommended kernel. KNN algorithm will find and rank the degree to which the kernels are related.

