**Project Title**

– Recommendation hotel system

**Data set**

* Expedia hotel recommendation data set – kaggle.com

**Project Idea**

* We aim to use k-means clustering, logistic regression and random forest to predict hotel to any new user
* Implement any possible custom algorithm using pyspark to make recommendations
* The data consists of ~37 million rows (user and hotel information) for 150 hotel clusters. There are 24 attributes for training.csv (hotel and user data) and 150 attributes for destinations.csv.
* We want to do Principle component analysis to determine most significant components explaining the accuracy.
* We will do data cleaning and exploratory data analysis to find trends/ patterns in the data.

**Algorithm** – K-means clustering, random forest and logistic regression

**Team mates** – Sathya Anurag Siruguppa, Sree Pradyumna Davuloori, LakshmiNarasimhan GN

Sathya Anurag Siruguppa – Random forest in R/UI

Sree Pradyumna Davuloori – Data cleaning, Exploratory Data Analysis and PCA with logistic regression

LakshmiNarasimhan GN – Custom algorithm/Pyspark implementation

**Milestones** –

* 1st milestone – March 31St – Setup the AWS environment and do EDA
* 2nd milestone – April 15th – Machine learning algorithms
* 3rd milestone – May 10th – UI