Project Planning Exercise: Solution

Seq.	Activity to perform	The expected way to perform	Activity Resource	Estimated Duration
1	Data Selection	Skipping few columns	Deleting from worksheet	3 hour
2	Data grouping	Ticket based grouping of conversation	Data group-by facility in python	3 hour
3	Deal with multiple languages	Translate them all into English	May be using some NLP resources using Python's libraries	2 days
4	Deal with noices in data	E.g., removing frequent less important words (thank you for your email)	Regular expressions in python	1 days
5	Dealing with multi-level data	Multi-level modelling or multi-class	Brainstrorming activity: Language do have support for both	1 day
6	Textual data representation change	Numeric representation of text	Libraries in python	1 day
7	Dealing imbalanced data	E.g., similar number of minimum records for each class	Libraries in python	Half day
8	Decide on we want supervised or un-supervised learning: Brain storming activity			
9	Data preparation for modelling	Separate training testing data	Libraries in python e.g., train_test split in sklearn	2 hour
10	Model selection for email classification	Use an appropriate (SOTA) model that can be use textual data for classification purpose	E.g., Random forest implementation in Python	4 hour
11	Model training, and testing for email classification	Input training data, train model, then use trained model for testing on test data	E.g., Random forest implementation in Python	3 days