Q1:

The highest ranking model took 36 hours to train. It involved many steps and models in their pipeline, starting with Distance Matrix Completion (with utilization of gradient descent, the spherical law of cosine forumla and nesterov momentum), Field-aware Factorization Machines, then finalized with an xgboost using pairwise rank. This final step involved expanding each row into 100 (ie. =size of the hotel clusters), causing a huge enlargement in the data. The analyst chose to report the top 5 cluster scores for the competition, which amounted to 60.44%. The next 3 submissions all got a score of around ~53%.

The most important variables used is the set of variables:

* user\_location\_country
* user\_location\_region
* user\_location\_city
* hotel\_country
* hotel\_market
* srch\_destination\_id

These variables reflect a strong geographical affinity. It seems that the geography of the "inbound" and "outbound" destinations summarize critical important about this categorization task. For example, certain countries with higher purchasing power are more likely to choose certain hotels, or that certain countries tend to have favorite destinations that can be modeled into a predictable behavior.