# Automatic Test Generation via ML: Draft Documentation

Arjun Jauhari

10 December 2015

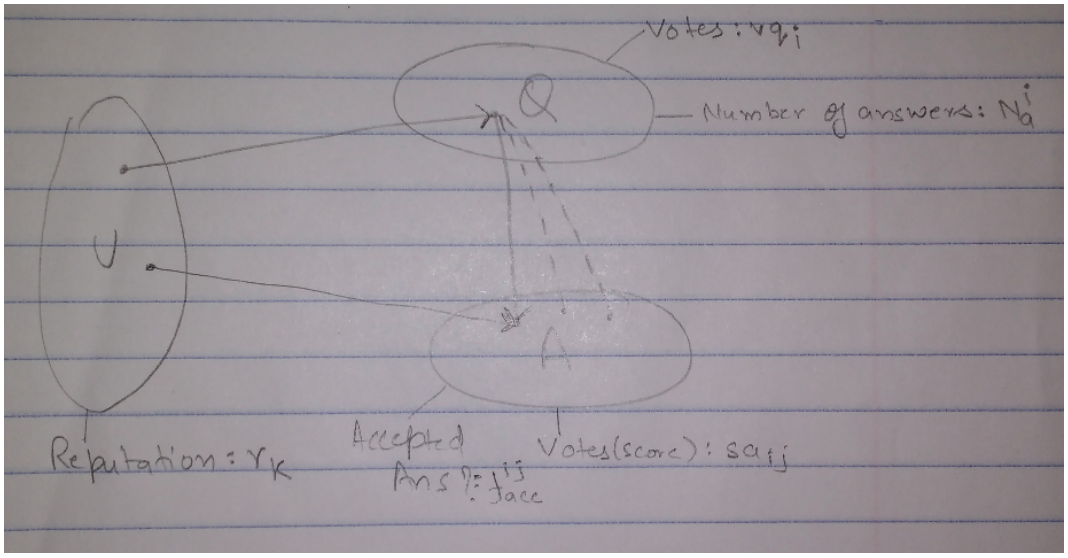


Figure 1: Overview

## Subscripts

- $i$ is subscript for question.
- $j$ is subscript for answer.
- $k$ is subscript for user.

## Notations

1. $u_k$: Quality measure of the $k$th user.
2. $q_i$: Quality measure of the $i$th question.
3. $va_{ij}$: Normalized votes corresponding to $j$th answer of $i$th question.
   Calculated as: $va_{ij} = \frac{|sa_{ij}|}{\sum_j |sa_{ij}|}$
   where $sa_{ij}$ is the actual votes(score) read from data dump.
4. $a_{ijk}$: Quality measure of $a_{ij}$th answer given by the $k$th user.
5. $f_{acc}^{ij}$: Boolean flag telling if this answer was Accepted, read from data dump.
6. $r_k$: Reputation of the $k$th user, read from data dump.

7. $N_a^i$: Number of answer to $i$th question, read from data dump.
8. $vq_i$: Number of votes to $i$th question, read from data dump.

## Equations

Below equations model the relation/dependence between the above defined parameters.

1. $a_{ijk} = f_a(u_k, q_i, va_{ij}, f_{acc}^{ij})$
2. $u_k = f_u(\{a_{ijk}\}_{ij}, q_i, r_k)$ , where $\{a_{ijk}\}_{ij}$ is set of all answers by user k
3. $q_i = f_q(u_k, N_a^i, vq_i, \{a_{ijk}\}_{jk})$, where $\{a_{ijk}\}_{jk}$ is set of all answers to $i$th question

## Status

Currently, the model is very simple with $va_{ij} = \frac{|sa_{ij}|}{\sum_j |sa_{ij}|}$, where $sa_{ij}$ is the actual votes(score) read from data dump.
and $a_{ijk} = w_1 * u_k + w_2 * q_i + w_3 * va_{ij} + w_4 * f_{acc}^{ij}$

As of now weights $w_1, w_2, w_4$ are 0 and $w_3$ is 1, so $a_{ijk} = va_{ij}$

User quality is being modeled as: $u_k \sim \mathcal{N}(mean(\{a_{ijk}\}_{ij}), Var(\{a_{ijk}\}_{ij}))$

Question quality is still not modeled.