

Hi Mentoring Committee,

The BAP_2017_Ranking_Application_Transformed Google Sheets in this folder contains the survey data given to mentors and candidates, reshaped in a way that will best help your decision-making abilities.

The transformed dataset has 7 fields:

1. Mentor_name:
 - a. The name of the mentor.
2. Candidate_name:
 - a. The name of the candidate.
3. Mentor_ranking_of_candidate:
 - a. The rank the mentor assigned to the candidate. If the mentor did not assign the candidate any rank, this field will be marked as NA - not applicable.
4. Candidate_ranking_of_mentor:
 - a. The rank the candidate assigned to the mentor. As above.
5. Profession_score:
 - a. A score ranging between 0 and 1 giving a measure of the similarity of the mentor and candidates professional interests. More explanation below.
6. Hobby_score:
 - a. Same as above.
7. Valid_match:
 - a. If the candidate is either a junior or senior and the mentor is a junior, the pairing will be classified as FALSE; all other pairings are classified as TRUE. Invalid pairings have been filtered out of the transformed dataset for your convenience.

Each row of the dataset contains the above information for a mentor-candidate pairing. All possible pairings are represented in the dataset. However, the dataset has been sorted so that the relevant pairings go to the top of the dataset.

Rows are arranged in this order:

Mentor_ranking_of_candidate, candidate_ranking_of_mentor, NOT(both_juniors), profession_score, hobby_score.

That is to say, a higher ranking (i.e. smaller number) assigned to a candidate by a mentor will push the mentor-candidate pairing up to the top. This choice was made because we want to prioritize mentor's choices.

Given the same mentor_ranking_of_candidate scores, a mentor-candidate pairing which assigns a higher ranking by a candidate to a mentor will take precedence over one which assigns a lower ranking to a mentor..

Etc.

If you would like to change the way the data is arranged, you can do so in Google Sheets by following this guide:

<https://support.google.com/docs/answer/3540681?co=GENIE.Platform%3DDesktop&hl=en>

However, please do it in your local copy on Excel, since we want everyone to have a clean Google Sheets table. Sorting by multiple columns in Excel can be done by following this guide:

<http://www.contextures.com/xlSort01.html>

If you'd like to sort using R (yay you):

1. `read.csv()`
2. `dplyr::arrange()`

If you'd like to sort using Python:

1. `pandas.read_csv()`
2. `pandas.sort_values()`

Profession and hobby scores are calculated as follows:

How many of the candidate's interests, out of the total, overlap with the mentor's interests?

How many of the mentor's interests, out of the total, overlap with the candidate's interests?

We take the average of these two percentages.

E.g.:

Candidate A is interested in Investment Banking, Investment Banking, and Consulting.

Mentor Z is interested in Investment Banking, Equity/Credit Research, and NA.

2 of Candidate A's 3 interests are shared with Mentor Z, since they are both interested in investment banking and candidate A put investment banking twice. Thus, the candidate similarity score with the mentor is $\frac{2}{3}$.

1 of Mentor Z's 2 interests are shared with candidate A, since they are both interested in investment banking. Because NA is not an interest, mentor Z is only considered to have 2 interests. Thus, the mentor similarity score with the candidate is $\frac{1}{2}$.

Then the profession score for Candidate A and Mentor Z is the average of $\frac{2}{3}$ and $\frac{1}{2}$: 0.58333...

The same formula is used for hobby similarity scorings.

Note:

- Repeating the same interest results in higher similarity scores.
- Both putting Other counts as a match.
- Putting NA does not count as an interest; as such, remaining interests are weighted higher. Thus, selecting investment banking 3 times is the same as selecting investment banking one time and NA two times.

Best,
Eric