Why did I choose this project?

I looked for inspiration on the website Kaggle which was used several times to provide examples of things that would make for acceptable projects. I decided upon one of their contest entries: Data Science for Good: Careervillage.org. I was interested in this project in part because it seemed like a mission which could have a positive impact, and the fact that there is potential prize money involved doesn't hurt.

What is it?

Careervillage.org is a website where people with career related questions can go to get advice and answers to those questions from professionals in the field. The challenge that they're trying to address here is to increase response rates from the professionals who are volunteers. The prompt of this competition is to try to match questions to volunteers in the related profession, but also to try to drive questions to the volunteers most likely to answer that question.

What is the source of data?

Careervillage.org has provided several large files that provide the source of the data. Included in the data are questions that were asked, answers to questions, anonymized lists of users including students and professionals, hash tagged topics, and users who follow specific tags. There are several other files as well. The provided data seems to be pretty well organized and there is quite a bit of it. It looks like there is some data that could be used as a training set if I wanted to use a learning algorithm, but it looks like a collaborative filtering algorithm might also work for this. The availability of (what looks like) good data also factored into my choice of this topic.

How do I plan to approach this task and solve it?

At first glance, it appears that an item-item collaborative filtering method would be a reasonable approach to take here. One could compare questions with other questions to find the most similar questions that were answered and then send those questions to the people who answered the original ones. However, potential concerns here could be 1. Not enough diversity in questions for volunteers to keep them interested and 2. Not enough variation of opinion (if the same type of question ended up always being answered by the same person or group of persons).

A riskier but potentially interesting approach would be to approach this as a classification problem for an algorithm such as SVM where questions are predicted to either be answered or go unanswered by each volunteer. The principal difficulty here might be the sheer number of possible labels in this comparison, since there are literally tens of thousands of volunteers. Narrowing this number of categorizations down to a handful by previously categorizing questions by topic could help, but further inspection should be done to try and bring down the number of possible labels for each asked question down to (ideally) a handful.

What be the outcome of this project?

Hopefully I will meet the goal stated by the contest prompt that will match (recommend) questions asked to volunteers who will answer those questions. Unfortunately there is no guarantee that this algorithm will get live tested, so it's not guaranteed that this will ever actually be used, but there is enough data here that I could potentially split some of it off to use as a training set and leave more of it

as a testing/verification set to give some idea of how well this works. The best possible outcome is that I not only complete this on time for class, but also before the contest entry deadline and win the contest, which would likely mean my algorithm was adopted for use and I could win $15,000. That is the ideal (but probably unlikely) outcome. In practice, I think the outcome will be personal growth and expanded knowledge of recommendation/learning algorithms.