Assignment #6: Mass Collaboration

Sanittawan Nikki Tan

11/15/2018

Note: files submitted include:

1) Sanittawan_three_digit_palindrome.py

Part 1

(a) The submissions to the Netflix Prize open call competition were judged based on how much better each team's movie recommendation model performs compared with Netflix's existing Cinematch model. The measurement criterion is root mean squared error (RMSE). The winning team must be able to improve RMSE at least by 10 percent compared with Cinematch. The formula is included below:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{1}{n}(\hat{y}_i - y_i)^2}$$

According to the Netflix Prize contest's rule page, it appears that every team's submission would be judged. There does not appear to be cutoffs for models that performed poorly. However, the 10 percent threshold is a minimum for any submissions to be considered for a prize.

(b) The most commonly used method at the outset of the competition was nearest neighbors method, which is an algorithm that makes a recommendation based on items with similar features (or neighbors) of an item that users selected (Bell et al. 2010, 25). However, as the competition progressed, various most successful models were latent factor models using matrix factorization (Bell et al. 2010, 26). Eventually, each team took the "ensemble" approach by weighing and combining results from a variety of models including nearest neighbors and latent factor (Bell et al. 2010, 28).

(c) The characteristic of one model relative to other models is that a model must have a low correlation with other models in the ensembles. According to Bell et al. (2010, 28-29), including several models which produce different levels of RMSEs into the ensembles will boost the prediction performance as long as each model is not highly correlated with others. This approach is called "ensembles of heterogeneous methods." For example, if two models can produce similar RMSEs and even if one model produces inferior RMSEs, including both models will improve the overall performance as long as one component is not highly correlated with the others (Bell et al. 2010, 28).

Part 2

(a) My Project Euler username is sanittawan. My friend key is: 1408688_MkG8yIFjfAW72hMZjp2KKiCli7VvWNbA

(b) I solved Problem 4: Largest palindrome product in the problem archives using Python. The answer is 906609. The code is included below and submitted as a separate `.py` file.

```
def pal_three_digit():
    '''
    Find the product of three digit numbers that is a palindrome

    Input: None

    Returns: (int) a product of three digit number which
            is a palindrome
    '''
    max_product = 0
```

```
    for i in range(100, 1000):
        for j in range(100, 1000):
            product = i * j
            if is_palindrome(str(product)):
                if product > max_product:
                    max_product = product
    return max_product

def is_palindrome(string):
    '''
    Test if a string is a palindrome

    Input: (str) a string

    Returns: (bool) True if a string is a palindrome, False otherwise

    '''
    return string == string[::-1]

if __name__=='__main__':

    print(pal_three_digit())
```

(c) Three awards that I most aspire to achieving are:

1) "Baby Steps" - I like this award because it serves as a realistic goal in the short run. Solving three problems will build my confidence to proceed with more challenging problems.

2) "C for Commitment" - I like this award because it serves as a longer term goal.

3) "Fibonacci Fever" - I like this award because I like Fibonacci sequence. I also suspect that some of the problem may involve recursive programming which I want to get more practice.

Part 3

(a) The MTurk human intelligence task that I chose is an advertisement tagging task. The task requires workers to identify brands in banner advertisement and record brand names.

(b) The full payment structure information is, unfortunately, quite limited. However, the provided rate is USD 0.6 and the time allotted is 15 minutes. Therefore, I assume that the rate is USD 0.6 per 15 minutes.

(c) The only qualification required is that workers must be granted "Premuim Ad Tagger" status. Workers without such qualification must submit a qualification request to Ad Tagger.

(d) The allotted time is 15 minutes. It is unclear from the task description how many ad banners that workers have to work on per task. Ad Tagger posted several similar MTurk requests as shown in the screenshot below, so it is possible that workers can complete four tasks in an hour, earning $0.6 * 4 = 2.4$ USD. Assuming that the task can be completed in 15 minutes, the hourly rate is $0.6 * 60/15 = 2.4$ USD per hour.

(e) This job is going to expire on Thursday, November 22, 2018.

(f) Ad Tagger, the owner of this request, posted 15 similar tasks. Assume that each worker completes one task in 15 minutes and receives USD 0.6 in remuneration, the cost of one million people completing a single task would be $0.6 * 1,000,000 = 600,000$. If everybody completes 15 tasks in 15 minutes per task receiving USD 0.6 per task, it will cost the company $600,000 * 15 = 9,000,000$

Part 4

(a) I have registered a Kaggle account. My username is sanittawan.

(b)

The open competition that I am interested in is "Quora Insincere Questions Classification," which is sponsored by the online forum Quora. Quora, Inc. is an internet company that maintains an online question-and-answer forum of the same name. The website centers on questions which can be posted by anyone and are answered by members of the Quora communities. Because some of the questions are of high quality and some are not, the website wants to develop an algorithm to detect low-quality questions. Low quality, in this context, means toxic, not constructive or is considered "trolls." They also want to weed out those that aim to make statements or are based on false claims to improve the quality of conversation. Participants in this competition must categorize if a question is insincere (coded as 1) or not insincere (coded as 0). The submissions will be evaluated on the F1 score. Quora provides training datasets to participants, but their models will be run against a test set which is privately-held and is not provided. The F1 score from the test set will determine winners. Participants must submit their work directly from Kaggle Kernel. Kaggle Kernels submissions can be in a form of script or a notebook.

Even though there is no specific honor code page, the Quora competition page includes some rules. For example, private code sharing outside of teams is not allowed. A team can consist of a maximum of 8 members and team mergers can be done by the team leader. There can be a maximum of 5 submissions per day and participants can select up to two final submissions to be evaluated. In addition, no outside data sources can be used.

The competition was launched on November 6, 2018, and the closing date is February 5, 2019. Key dates are as follows. January 29, 2019, is the entry and team merger deadline. Teams or individuals must enter the competition before this date. It is also the last day that teams can merge or individuals can join teams. February 5, 2019, is the final submission deadline. All deadlines are at 11:59 PM UTC. Regarding the prizes, the winner will receive USD 12,000. The first runner-up and the second runner-up will receive USD 8,000 and 5,000 respectively.

(c) I believe that Quara, Inc. will implement the winners' models to improve their insincere question detection algorithms. In fact, the Rules page of the competition specifies winners obligations that winners have to provide the final model's software used to generate the winning submission to Quora.

Link to the competition: https://www.kaggle.com/c/quora-insincere-questions-classification

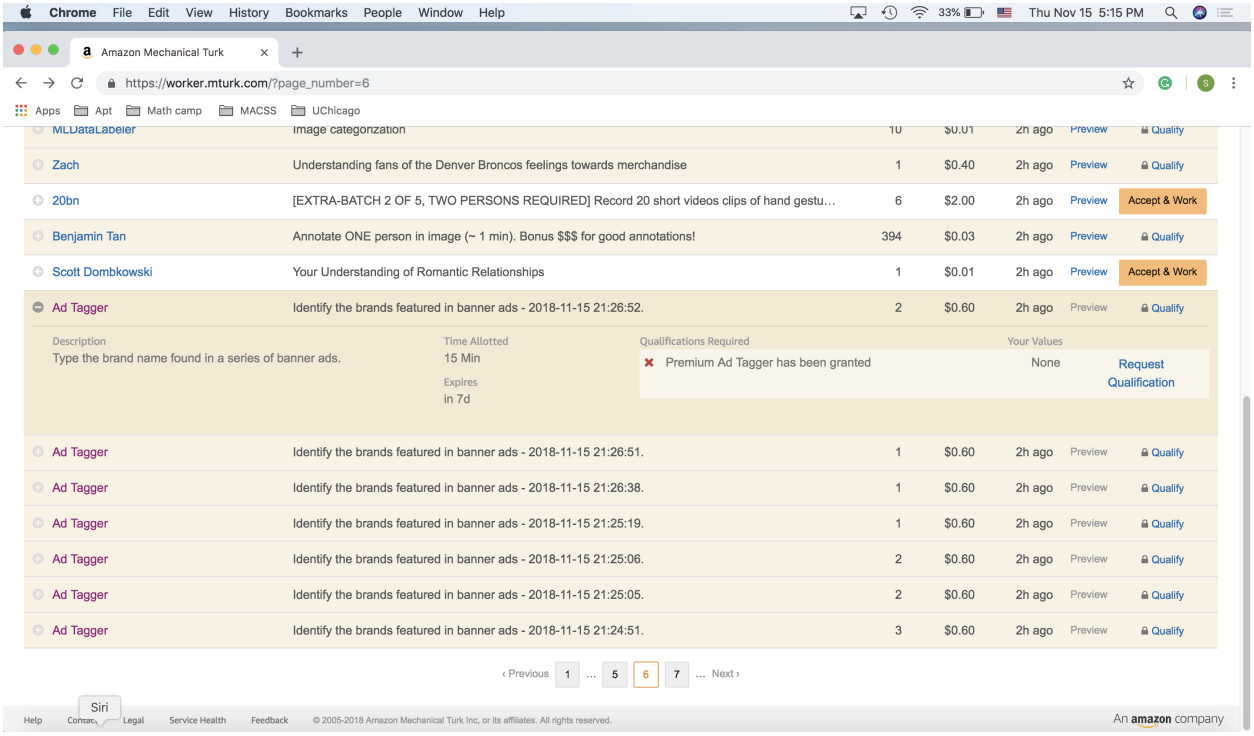

Figure 1: MTurk Screenshot

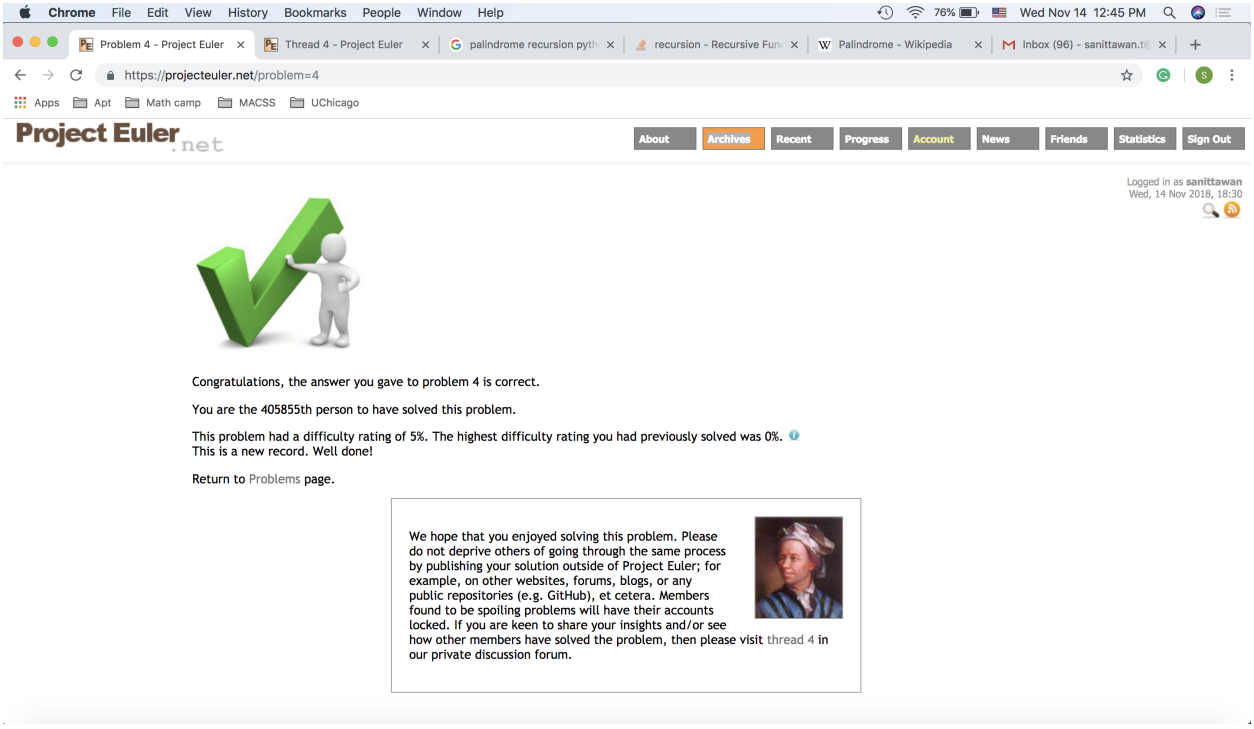

Figure 2: Project Euler's Problem 4