Assignment #6

MACS 30000, Dr. Evans
Ying Sun

1. Netflix Prize and Bell, Koren, and Volinsky (2010)

(a) The submissions to the Netflix Prize open call contest would be judged based on "the improvements in root mean squared error (RMSE)" (Bell et al., 2010, p.24) compared with "Netflix's internal algorithm, Cinematch" (Bell et al., 2010, p.24). The criterion function is:

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

where $\hat{y}_i$ is the predicted rating and y_i is the actual rating. According to Netflix's requirement, the winner has to improve the movie recommendation system at least 10% (Bell et al., 2010, p.24). So 10% improvement can be regard as a threshold.

(b) At the beginning of the Netflix Prize contest, "nearest neighbors" was the most commonly used method for predicting ratings on movies (Bell et al., 2010, p.25).

(c) The winning model of the first year was "a linear combination of 107 prediction sets" (Bell et al., 2010, p.29). According to the article, it is almost certain that the more models and algorithms combined, the better the prediction result. A blend could be improved if "it was not highly correlated with the other components" (Bell et al., 2010, p.28).

Reference:

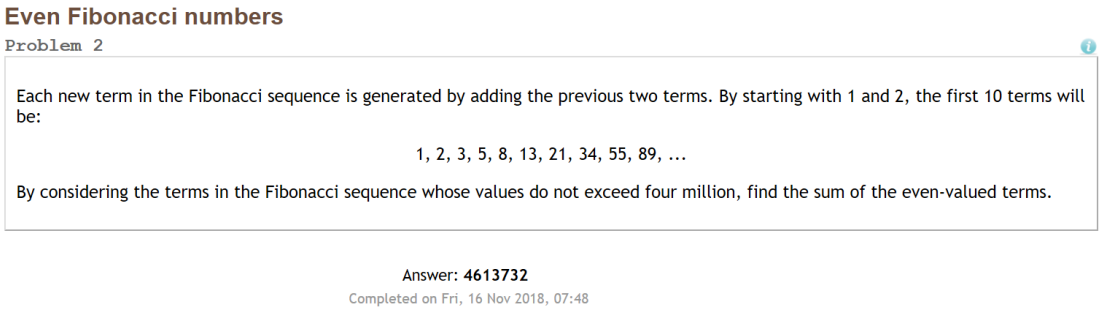
Bell, Robert M., Yehuda Koren, and Chris Volinsky, "All Together Now: A Perspective on the Netflix Prize," Chance, 2010, 23 (1), 24–29.

2. Collaborative problem solving: Project Euler

(a) Username: sunying2018

Friend Key: 1408719_lE0RWLDsGK5dm0WQrqawk0WRGDzgcyWj

(b) Problem:



Even Fibonacci numbers

Problem 2

Each new term in the Fibonacci sequence is generated by adding the previous two terms. By starting with 1 and 2, the first 10 terms will be:

1, 2, 3, 5, 8, 13, 21, 34, 55, 89, ...

By considering the terms in the Fibonacci sequence whose values do not exceed four million, find the sum of the even-valued terms.

Answer: **4613732**

Completed on Fri, 16 Nov 2018, 07:48

My answer is 4613732

Code (Python):

```
lst = []
lst.append(1)
lst.append(2)
while lst[-1] < 4000000:
    lst.append(lst[-1] + lst[-2])
total = 0
for val in lst:
    if val % 2 == 0:
        total += val
print(total)

4613732
[Finished in 0.1s]
```

(c) Three awards: Gold Medal, Ten Out of Ten, Valued Contributor

The Gold Medal is awarded for the first to solve a problem. I like it because it can give me a great sense of achievement and I really hope to do something creative. For Ten Out of Ten, it is awarded for solving the ten most recent problems. I think most recent problems are more challenging because you may not find any idea from previous solutions. I hope to challenge myself in innovative thinking. I like Valued Contributor because I hope to come up with more effective solutions and contribute more on this sharing platform.

Reference:

https://projecteuler.net/

3. Human computation projects on Amazon Mechanical Turk

(a) The human computation project I selected is “Transcribe up to 35 Seconds of Media to Text”. It requires participants to transcribe up to 35 seconds of Media into text.

(b) Basically, the reward is $ 0.05 for completing the task. In addition, reward amount is based on media length. Additional reward amounts will be paid as a bonus. But the total amount doesn’t exceed $0.17 per HIT.

(c) It has some required qualifications. First, HIT approval rate (%) is not less than 95. In addition, qualified to work on Transcription tasks is not less than 900.

(d) The allotted time for this task is 15 minutes. I think I could complete 6 items in an hour. The implied hourly rate is $0.30 per hour.

(e) The job expires on 11/18/2019.

(f) The project would cost at most $170,000 if one million people participated in the task.

Reference:

https://worker.mturk.com/

4. Kaggle open calls

(a) Profile page: https://www.kaggle.com/sunying2018

(b) The link of the competition: https://www.kaggle.com/c/ga-customer-revenue-prediction

The competition is "Google Analytics Customer Revenue Prediction", which is sponsored by RStudio and Google Cloud. RStudio is a famous company engaged in developing free and open tools for R and professional products for enterprise teams. Google Cloud is a company which is passionate about providing a suite of cloud computing services to facilitate data analysis and improve the security and reliability of data sharing. In this competition, participants are required to predict revenue per customer by analyzing a dataset of a GStore.

The submissions will be evaluated by the "root mean squared error (RMSE)", which is the standard deviation of the residuals. As a measure of how spread out residuals are, it can reflect how concentrated the data is around the line of best fit.

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

The prizes can be categorized into two types. The first type is "Leaderboard Prizes", which is based on final private leaderboard ranking. The top 3 winners will get $12,000, $8000 and $5000 respectively. The second type is "Special R Usage Prizes" which is exclusively designed for top ranking selected solution using R. The top 3 performers will receive $10,000, $7,000 and $3,000 respectively. The winners are able to win these two types of prizes at the same time.

There are several important honor code issues. First of all, participants are not allowed to submit from multiple accounts. Second, participants are not permitted to share code or data outside of teams. What's more, the open source code used by participants must be licensed under "an Open Source Initiative-approved license" (see www.opensource.org).

This competition started at September 13, 2018. Participants must accept the competition rules before November 23, 2018. Also, participants have to make decisions on team merge before November 23, 2018. As for the final submission, participants must submit their solutions before November 30, 2018. Besides, all the solutions will be tested using the real transaction data from December 1st, 2018 to

January 31st 2019.

For the submission, participants must submit their final solution before November 30, 2018. Each team can submit up to 5 entries per day. Every team is not allowed to submit more than 2 final submissions for judging.

(c) For Google Merchandise Store, the winning answer could be used for predicting future revenues. Based on this prediction, the company could make better market strategies in price, product, promotion and place (4P strategies). For the RStudio company, it could gain more popularity and promote the use of R in the area of data analysis.

Reference:

https://www.kaggle.com/c/ga-customer-revenue-prediction