PART 1

1. You are given a textfile that contains the age (Integer) of each people in Jakarta, let's assume Jakarta 10 Million people. The format of the file is very simple, each line is an integer representing age of one people.

Example age.txt:
==========
23
51
35
==========

Write a script/program in **Python or Java or Scala** that takes this file name as an input, and produce an output file that contains the sorted age in the same format. You are not allowed to use any external library besides the built-in library for the language that you choose.

Example sorted_age.txt:
==========
23
35
51
==========

2. Now, the age.txt contains the age of not only people in Jakarta, but the whole world (7 Billion People). Also, you only have a shitty laptop with 1GB of RAM. Can your script in part 1 handle this file? If not, how would you modify it so it can handle this kind of big file?

3. Let's go back to Jakarta again. Now, you are the owner of an online shop. You have been running your shop for a long time, and have a list of 1 Million blacklisted name and phone number. Each line is one word(name), followed by space, then the phone number.

Example blacklist.txt:
==========
Andi 1341441
Melisa 8565467
Aslam 2908345
==========

You want to build an API server that receive the name and phone number as an input, then output boolean whether this name and phone number is in the blacklist. **How would you write these two functions to optimize the latency for each API call (no need to write an API server)**:

- initialize(blacklist)
This function takes string input, which is the file name of the blacklist you have, and called when the API server is starting.

-check_blacklist(name, phone_number)
This function takes 2 arguments, name(string) and phone number(int). This function is called whenever the API is called, and return boolean the input name and phone number is in the blacklist.

You can use **Python or Java or Scala**, and again, you are not allowed to use any external library.

PART 2

1. Traditional SVM can do classification between 2 classes. What if we have more than 2 classes and we want to use SVM to classify?

2. Suppose we have so many features available, but we cant use all of them directly because of computational cost. How do you do deal with this kind of situation?

3. What is your favourite package for machine learning? And why?

4. If you are about to choose which tree ensemble method to use, between Random Forest and Gradient Boosted Trees, what is your consideration for choosing one over the other?

5. What is the expected value of a dice roll? Now, you are given a chance to re-roll the dice once if you want, you want to get high value for the final roll, what is the expected value of the final roll? Explain your steps.

6. What is the expected number of coin flips required until you get 2 consecutive heads? Explain your steps.